



Visual Sentiment Feature Extraction and User Engagement Prediction for UGC in Community Operations: A Joint Learning Approach

Xueling Li[✉], Yan Hao^{*✉}

Business Administration College, Chengdu Jin Cheng University, Chengdu 610097, China

Corresponding Author Email: haoyan@cdjcc.edu.cn

Copyright: ©2026 The authors. This article is published by IETA and is licensed under the CC BY 4.0 license (<http://creativecommons.org/licenses/by/4.0/>).

<https://doi.org/10.18280/ts.430412>

ABSTRACT

Received: 10 March 2026

Revised: 28 July 2026

Accepted: 10 August 2026

Available online: 31 August 2026

Keywords:

user-generated content, visual sentiment analysis, multi-modal fusion, vision-language model, engagement prediction, causal inference

Visual sentiment analysis of user-generated content (UGC) serves as a core technical pillar for social media operations and content recommendation. However, in community scenarios, general-purpose vision-language models suffer a notable drop in sentiment recognition accuracy; the semantic inconsistency between image-text pairs remains poorly modeled; and sentiment feature extraction and engagement prediction are typically treated as disjoint tasks. To address these gaps, this paper proposes a Community-Adaptive Sentiment-Engagement Joint Learning Network (CASE-Net). Built upon a vision-language pre-trained model, we adopt a lightweight adapter-based fine-tuning strategy that introduces learnable bottleneck adapters while keeping the backbone frozen, enabling the model to adapt to the sentiment distribution of a target community with only ~0.5% trainable parameters. We further construct dynamic sentiment prototype representations to achieve multi-granularity sentiment distribution prediction. To tackle the pervasive phenomenon of image-text sentiment inconsistency in community content, we systematically design quantification metrics along three dimensions—distributional divergence, sentiment intensity, and sentiment polarity—encoding image-text sentiment tension as a computable feature for engagement prediction. At the prediction stage, we propose a sentiment-aware cross-attention mechanism that uses the global sentiment distribution as a guide to adaptively focus on emotion-discriminative regions in images, and a multi-task gated predictor that jointly optimizes sentiment classification and engagement prediction. On a self-constructed real-world community image dataset, our method achieves a sentiment classification accuracy of 74.6% and a Spearman correlation coefficient of 0.812 for engagement prediction. Ablation studies confirm the individual contribution of each module. Furthermore, we incorporate Double Machine Learning (DML) for causal effect estimation, revealing the differentiated driving mechanisms of distinct sentiment dimensions on likes versus comments, thereby providing causal-level decision support for optimizing community operation strategies.

1. INTRODUCTION

User-generated content (UGC) has become a core constituent element of the social media ecosystem. On visual-centric social platforms represented by Instagram, Rednote, and TikTok, image content dominates information dissemination [1, 2]. On these platforms, user engagement—including behaviors such as likes, comments, shares, and favorites—serves not only as a direct metric of content influence, but also as a key indicator for optimizing platform recommendation algorithms, formulating community operation strategies, and evaluating commercial value [3, 4]. Image sentiment is an important intrinsic factor driving user engagement behaviors [5, 6]. The sentiment conveyed by an image—whether it is an uplifting landscape, a resonant warm moment, or a contrasting scene that sparks discussion—can significantly affect the degree of users' emotional resonance and their subsequent willingness to interact. Therefore, accurately extracting sentiment features from UGC images and predicting user engagement on that basis has become a

frontier topic at the intersection of computer vision and multimedia analysis [7, 8].

There are essential differences between UGC image sentiment analysis in community scenarios and general-purpose image sentiment recognition [9, 10]. Community UGC images exhibit the following distinctive characteristics. First, semantic community specificity: the same visual element may carry entirely different sentiment connotations across different communities—minimalism represents high-end aesthetics in photography communities, yet may be perceived merely as an ordinary record in general lifestyle communities [11, 12]. Second, modality heterogeneity: UGC images are typically accompanied by multi-modal contexts such as captions, hashtags, and geolocation, where the sentiment relationship between image and text is complex and variable [13, 14]. Third, sentiment dynamics: users' sentiment interpretation of images in community contexts is subject to dynamic factors such as trending topics and community culture, which static sentiment lexicons struggle to cover [15, 16]. These specificities impose precision and adaptability

requirements that far exceed those of general-purpose image sentiment recognition, and also render the direct transfer of existing sentiment analysis techniques fundamentally difficult [17, 18].

Existing research on image sentiment analysis and engagement prediction suffers from three limitations. In terms of sentiment feature extraction, traditional methods mostly rely on pure-vision convolutional neural networks to extract low-level visual features and then map them onto predefined sentiment categories, making it difficult to capture fine-grained sentiment semantics [17, 18]. Although vision-language pre-trained models represented by CLIP have established a joint embedding space between images and natural language through cross-modal contrastive learning and demonstrated excellent performance in zero-shot sentiment classification, their direct application to specific community scenarios leads to a significant drop in sentiment recognition accuracy due to the distributional gap between training data and the target community [19, 20]. Although some studies have attempted to fine-tune CLIP for specific tasks, systematic adaptation methods oriented toward community UGC scenarios remain absent [21, 22]. In terms of cross-modal modeling, because of content creators' subjective expression habits or deliberate design, the sentiment information conveyed by the image and its accompanying text in UGC content frequently exhibits inconsistency or even conflict [23, 24]. Existing studies mostly adopt simple concatenation or average fusion of image-text sentiment features, failing to fully exploit the informational value embedded in this image-text sentiment tension—which, precisely, may be a key driver of comment-generation and controversial interactions [23, 24]. In terms of task synergy, sentiment recognition and engagement prediction are typically modeled as two independent tasks, overlooking their intrinsic association: image representations with strong sentiment discriminability often also carry predictive power for engagement [25, 26]. Existing methods also lack causal-level explanations of *how* and *which* sentiment dimensions drive user engagement, limiting their practical utility in community operation decision-making [5, 6].

To address the aforementioned gaps, this paper proposes EMO-Community Net, a joint learning network for community sentiment and engagement prediction. At the feature extraction level, building upon the CLIP dual-tower structure, we introduce a lightweight adapter-based fine-tuning strategy by inserting learnable bottleneck adapters after frozen Transformer layers, enabling the model to adapt to the target community's sentiment distribution while preserving cross-modal alignment capability, and outputting multi-level sentiment feature representations. At the cross-modal modeling level, we systematically measure image-text sentiment discrepancy through multi-dimensional metrics—including KL divergence, valence-arousal space distance, and sentiment polarity conflict—thereby elevating image-text sentiment tension from an empirical phenomenon to a computable modeling object. At the prediction framework level, we design a sentiment-aware cross-attention mechanism that uses the global sentiment distribution to guide visual feature focusing, construct a multi-task gated predictor for joint optimization of sentiment classification and engagement prediction, and incorporate Double Machine Learning (DML) for causal effect estimation to reveal the independent driving effects of each sentiment dimension on user engagement. This paper also constructs a community UGC benchmark dataset

containing images, captions, multi-dimensional sentiment labels, and engagement metrics. The remainder of this paper is organized as follows: Section 2 elaborates the methodological framework in detail, Section 3 introduces dataset construction and experimental result analysis, and Section 4 concludes the paper.

2. METHODOLOGY

This section systematically elaborates the complete technical framework of the proposed EMO-Community Net for community sentiment and engagement prediction. The design of this framework follows a progressive logic from semantic alignment and discrepancy quantification to association modeling and causal inference, and is composed of four core modules in cascade. The community-adaptive vision-language sentiment encoder takes the frozen CLIP dual-tower structure as its backbone; by inserting learnable bottleneck adapters in parallel after each Transformer layer, it adapts to the target community's sentiment semantic distribution with approximately 0.5% trainable parameters, and outputs multi-level sentiment distributions of image and text in the shared embedding space. Built on the premise that image and text sentiment distributions are aligned in the same probability simplex space, the cross-modal sentiment inconsistency quantification and encoding module systematically measures the semantic discrepancy between image and text from three dimensions—distributional morphology, sentiment intensity, and sentiment polarity—and encodes the inconsistency metrics into a compact feature vector via a small neural network. Taking the global sentiment distribution as a query, the sentiment-aware attention and multi-task gated predictor guides cross-attention computation over the visual Transformer to focus on emotion-discriminative regions, and enables the shared features to simultaneously possess sentiment discriminability and engagement predictability through multi-task joint optimization; the gating mechanism further achieves data-dependent adaptive fusion of inconsistency features. Serving as a post-training offline analysis component, the DML causal inference module strips away confounding bias from low-level visual descriptors through a cross-fitting strategy to estimate the average treatment effect of each sentiment dimension on engagement metrics. The first three modules constitute an end-to-end trainable joint framework, while the fourth module provides causal-level decision support at the back end. The entire network adopts a staged warm-up and cosine annealing learning rate scheduling strategy, and introduces gradient clipping to ensure training stability and convergence speed.

2.1 Problem formulation and overall framework

Let the community UGC dataset be $D = \{(I_i, T_i, y_i^{cls}, y_i^{reg})\}_{i=1}^N$, where I_i is the raw image, T_i is the corresponding caption and hashtag sequence, $y_i^{cls} \in \{0, 1\}^K$ is the one-hot vector of the ground-truth sentiment label of the image, taking $K = 15$ sentiment categories, and $y_i^{reg} \in \mathbb{R}^4$ is the observed value of four engagement metrics, corresponding in order to number of likes, comments, favorites, and shares. The goal of this paper is to learn a prediction function $F: (I, T) \mapsto (\hat{y}^{cls}, \hat{y}^{reg})$ such that sentiment classification is accurate and engagement prediction is precise. On this basis, causal

inference is performed to estimate the average treatment effect of each sentiment dimension on engagement, providing interpretable decision support for community operations and elevating predictive capability into strategic guidance capability.

The end-to-end forward computation flow of EMO-Community Net follows the progressive logic of semantic alignment, discrepancy quantification, and association modeling. First, the raw image and caption are respectively fed through the community-adaptive vision-language dual encoder to extract their respective sentiment distributions, which are aligned in the same probability simplex space. Subsequently, the inconsistency between image-text sentiment distributions is explicitly quantified into multi-dimensional metrics—including distributional morphology divergence, sentiment intensity distance, and sentiment polarity conflict—and mapped by an encoding network into a compact inconsistency feature vector. Thereafter, the image sentiment distribution serves as a query to guide cross-attention computation over the visual Transformer, focusing on discriminative visual regions that are highly relevant to the current sentiment semantics. Finally, after multi-task gated fusion, the sentiment classification result and engagement prediction value are output in parallel. All trainable parameters in this flow are globally optimized through the joint loss function L_{total} , enabling the adapter layers, inconsistency encoder, attention module, and gated predictor to achieve coordinated convergence under the guidance of a common

objective.

2.2 Community-adaptive vision-language sentiment encoder

Although general-purpose vision-language models have acquired rich semantic priors on large-scale image-text data, a significant distributional shift exists between their training data and specific community UGC scenarios, making it difficult for sentiment recognition accuracy to meet practical demands upon direct application. While full-parameter fine-tuning can adapt to the target domain, the rapidly emerging new aesthetic trends in community operation scenarios require models to possess rapid iteration capabilities, and the high tuning cost limits the feasibility of such approaches. This paper takes the CLIP dual-tower structure as the backbone network and introduces a lightweight adapter-based fine-tuning strategy. Figure 1 illustrates the architecture of the community-adaptive vision-language sentiment encoder. The image encoder Φ_{img} splits the input image $I \in \mathbb{R}^{3 \times 224 \times 224}$ into $N = 196$ visual patch tokens and concatenates a learnable CLS token; after 12 Transformer layers of encoding, it outputs the visual feature map $F_v^{raw} \in \mathbb{R}^{197 \times 768}$ and the global representation $f_{img}^{cls} \in \mathbb{R}^{768}$. The text encoder Φ_{txt} encodes the caption via BPE tokenization into a token sequence of maximum length $L = 77$, and outputs the text global representation $f_{txt} \in \mathbb{R}^{768}$.

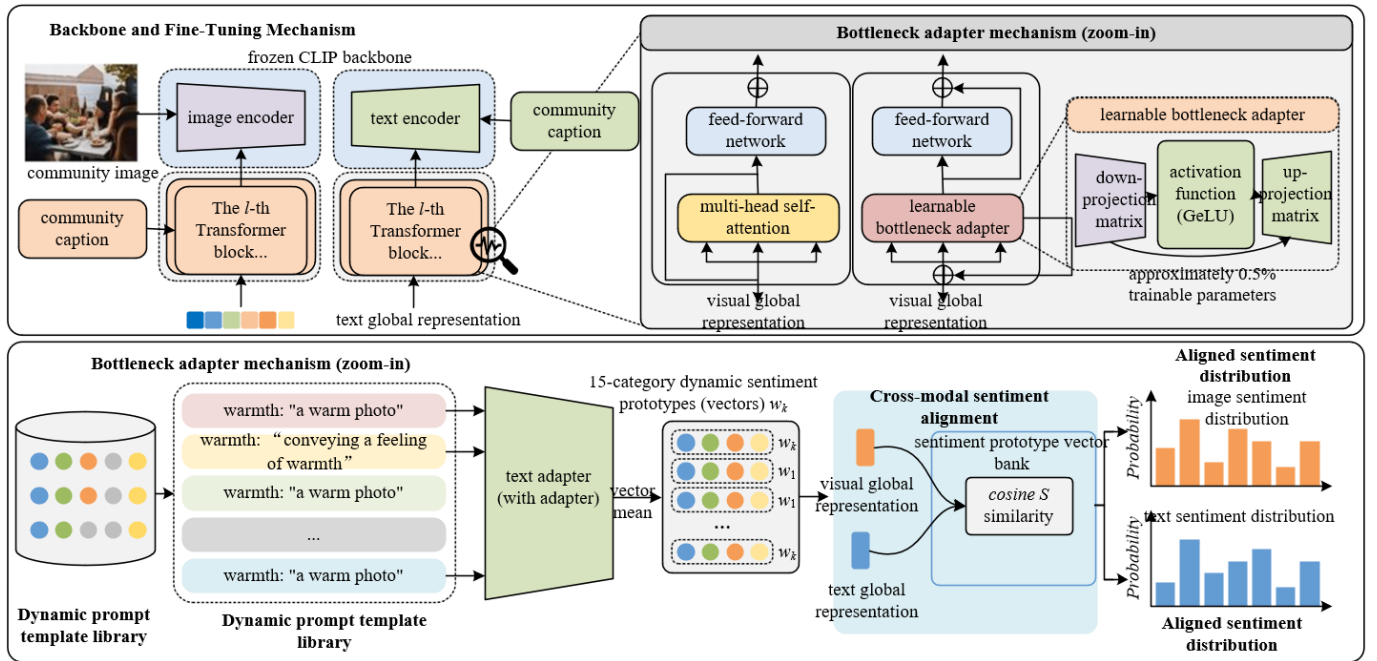


Figure 1. Architecture of the community-adaptive vision-language sentiment encoder

This paper freezes all backbone parameters of Φ_{img} and Φ_{txt} , and only inserts learnable bottleneck adapters in parallel after the multi-head self-attention module and feed-forward network of each Transformer layer. For the input hidden state $H^{(l)} \in \mathbb{R}^{197 \times 768}$ of the l -th Transformer layer, the adapter adopts a residual connection form to ensure that the network behavior approaches an identity mapping at the initial moment:

$$H_{adapt}^{(l)} = H^{(l)} + s \cdot \text{GeLU}(\text{LN}(H^{(l)}) \cdot W_{down} + b_{down}) W_{up} + b_{up} \quad (1)$$

where, the down-projection matrix $W_{down} \in \mathbb{R}^{768 \times 64}$ and the up-projection matrix $W_{up} \in \mathbb{R}^{64 \times 768}$ constitute the bottleneck structure, with the bottleneck dimension set to 64 to achieve parameter efficiency. $\text{LN}(\cdot)$ is layer normalization, $\text{GeLU}(\cdot)$ is the activation function, and the scaling factor s is initialized to 10^{-4} and gradually increased during training, so as to progressively inject target-domain knowledge and alleviate catastrophic forgetting during source-to-target domain shift. All adapter layers account for only 0.52% of the total network parameters, approximately 1.58 M trainable parameters,

reducing the training cost by nearly two orders of magnitude compared with the 302.5 M parameters of full-parameter fine-tuning.

To achieve fine-grained sentiment characterization in community scenarios, this paper constructs a dynamic prompt template library T containing $K = 15$ sentiment categories, covering Ekman's six basic emotions, five aesthetic sentiments, and four community-evolution-specific sentiments. To eliminate the semantic ambiguity of a single prompt, for each sentiment category k , $M = 5$ prompt templates with varying sentence structures are constructed; the M text descriptions are respectively fed into the text encoder with adapters, and their vector mean is taken as the stable prototype $p_k \in \mathbb{R}^{768}$ for that sentiment category:

$$p_k = \frac{1}{M} \sum_{m=1}^M \Phi_{\text{txt}}(t_k^{(m)}) \quad (2)$$

For the input image I , its global visual representation $f_{\text{img}} = \Phi_{\text{img}}(I)$ computes cosine similarity with all prototype vectors, and is mapped via a Softmax with a learnable temperature parameter γ to obtain the image sentiment distribution $E_{\text{img}} \in \mathbb{R}^K$:

$$E_{\text{img}}^{(k)} = \frac{\exp(\gamma \cdot \text{sim}(f_{\text{img}}, p_k))}{\sum_{j=1}^K \exp(\gamma \cdot \text{sim}(f_{\text{img}}, p_j))}, \text{sim}(u, v) = \frac{u^\top v}{\|u\| \|v\|} \quad (3)$$

The temperature parameter γ is initialized to $\log(1/0.07)$ and adaptively adjusted during training to control the

sharpness of the output distribution. Similarly, the caption sentiment distribution E_{txt} is obtained by contrasting the text representation f_{txt} with the same set of prototype vectors. Because image and text features are mapped into CLIP's same joint embedding space and compute similarity with the same prototypes, E_{img} and E_{txt} naturally reside in the same probability simplex space, laying a precise alignment foundation for the subsequent inconsistency quantification.

The optimization of the adapter layers is embedded in the end-to-end total loss. In the early training phase, an auxiliary loss is used for warm-up to ensure alignment between the prototype space and community semantics:

$$L_{\text{adapt}} = L_{\text{contrast}} + \lambda_{\text{cls}}^{\text{init}} \cdot L_{\text{cls}}^{\text{init}} \quad (4)$$

where, L_{contrast} is the cross-modal contrastive loss, which forces image features and samples with the same sentiment label within the same batch to move closer in the embedding space, and move apart from samples with different labels:

$$L_{\text{contrast}} = -\frac{1}{B} \sum_{i=1}^B \log \frac{\exp(\text{sim}(f_{\text{img}}^{(i)}, f_{\text{txt}}^{(i)})/\tau_0)}{\sum_{j=1}^B \exp(\text{sim}(f_{\text{img}}^{(i)}, f_{\text{txt}}^{(j)})/\tau_0)} \quad (5)$$

$\tau_0 = 0.07$ is the fixed contrastive temperature, and B is the batch size set to 128. $L_{\text{cls}}^{\text{init}} = \text{CE}(E_{\text{img}}^{(i)}, y_i)$ is the sentiment classification cross-entropy loss, with $L_{\text{cls}}^{\text{init}}$ set to 0.5. The warm-up phase lasts for the first 5 epochs, after which the adapter layers and all subsequent modules undergo end-to-end joint training, ensuring that the adaptation process consistently serves the ultimate engagement prediction objective.

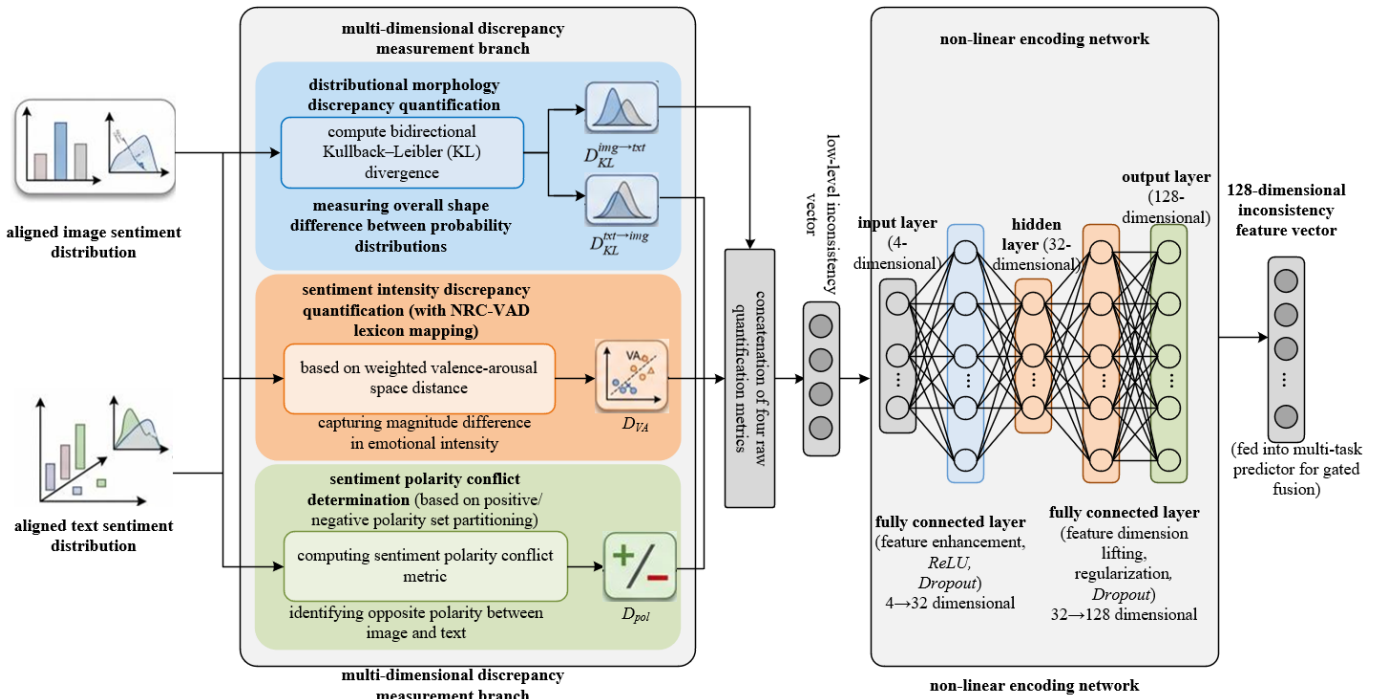


Figure 2. Flowchart of cross-modal sentiment inconsistency quantification and encoding

Note: NRC-VAD: National Research Council Canada-Valence-Arousal-Dominance.

2.3 Cross-modal sentiment inconsistency quantification and encoding

In community UGC scenarios, although image and caption jointly constitute a complete expressive unit, the sentiment

signals conveyed by the two often exhibit inconsistency or even mutual conflict. This image-text sentiment tension is not merely noise; rather, it may be a key factor in provoking users' curiosity, confusion, and desire to discuss. Traditional multi-modal fusion methods mostly adopt simple concatenation or

averaging of image-text features, implicitly assuming sentiment semantic consistency between the two, and failing to treat the inconsistency itself as a modelable information source. After obtaining E_{img} and E_{txt} in the same probability simplex space, this paper systematically constructs quantification metrics from three complementary dimensions—distributional morphology, sentiment intensity, and sentiment polarity—thereby transforming image-text sentiment tension from an empirical phenomenon into a computable modeling object. Figure 2 illustrates the flowchart of cross-modal sentiment inconsistency quantification and encoding.

Distributional morphology inconsistency is measured via bidirectional KL divergence to capture the overall shape difference between two probability distributions. To maintain numerical stability, a minimal offset $\epsilon=10^{-8}$ is applied to the probability values:

$$D_{KL}^{img \rightarrow txt} = \sum_{k=1}^K E_{img}^{(k)} \log \frac{E_{img}^{(k)} + \epsilon}{E_{txt}^{(k)} + \epsilon} \quad (6)$$

$$D_{KL}^{txt \rightarrow img} = \sum_{k=1}^K E_{txt}^{(k)} \log \frac{E_{txt}^{(k)} + \epsilon}{E_{img}^{(k)} + \epsilon} \quad (7)$$

The asymmetry of KL divergence enables the model to separately perceive the degree of surprise of the image relative to the text and that of the text relative to the image. Sentiment intensity inconsistency maps the 15 sentiment categories onto valence $v \in [-1, 1]$ and arousal $a \in [-1, 1]$ coordinates based on the NRC-VAD lexicon, computing the weighted expected coordinates of image and text in the VA space and their Euclidean distance:

$$v_{img} = \sum_{k=1}^K E_{img}^{(k)} v_k, a_{img} = \sum_{k=1}^K E_{img}^{(k)} a_k \quad (8)$$

$$D_{va} = \sqrt{(v_{img} - v_{txt})^2 + (a_{img} - a_{txt})^2} \quad (9)$$

This metric captures the magnitude difference in emotional intensity between image and text. Sentiment polarity conflict divides the labels into a positive polarity set P and a negative polarity set N , defines the polarity gap $\Delta = \sum_{k \in P} E^{(k)} - \sum_{k \in N} E^{(k)}$, and formulates the polarity conflict metric as:

$$D_{pol} = |\Delta_{img} - \Delta_{txt}| \quad (10)$$

When the sentiment polarities of image and text are opposite, this metric yields a significantly positive value. Thus far, the model has completed a comprehensive characterization of image-text sentiment inconsistency from three dimensions. The four raw metrics each have their own emphasis and complement one another, providing rich structural input for subsequent encoding.

After concatenating the above four metrics into $d_{raw} \in \mathbb{R}^4$, they are fed into an inconsistency-encoding Multi-Layer Perceptron (MLP) for nonlinear mapping and feature enhancement:

$$D_{fusion} = MLP_{\Delta}(d_{raw}) \in \mathbb{R}^{128} \quad (11)$$

The structure of MLP_{Δ} is: 4-dimensional input mapped to 32 dimensions, followed by ReLU activation and Dropout (rate 0.1) for regularization, then mapped to a 128-dimensional output. This encoder accomplishes the transformation from low-dimensional statistics to a high-dimensional feature space, enabling the inconsistency information to achieve dimensional compatibility with the 768-dimensional visual features in the subsequent gated fusion stage, and ensuring representational balance in the multi-source feature fusion process.

The encoding network is embedded in the overall end-to-end optimization framework; the gradient of D_{fusion} is back-propagated to the front-end adapter module via the total loss L_{total} . During training, the visual encoder adjusts the response strength of visual primitives based on feedback from the downstream engagement prediction task, driving the image-text sentiment distributions to evolve in a direction beneficial to task performance. This mechanism bridges the optimization link between sentiment representation learning and discrepancy measurement, avoiding the two-stage processing paradigm where feature extraction and discrepancy quantification are disconnected, and letting inconsistency measurement directly serve the modeling objective of community engagement.

2.4 Sentiment-aware attention and multi-task gated predictor

The preceding modules complete community-adaptive sentiment distribution extraction and image-text inconsistency quantification, yet the two types of information have not yet established effective interaction at the prediction stage. This module realizes the reverse guidance of sentiment semantics on visual features, while simultaneously accomplishing the joint optimization of sentiment classification and engagement prediction within the shared representation space. Traditional vision Transformers perform indiscriminate aggregation over all image patches, failing to reflect the differentiated contributions of different image regions to sentiment discrimination. This paper constructs a sentiment-aware cross-attention mechanism, taking the global sentiment distribution E_{img} as the query and the low-level visual feature map as the key and value, to adaptively enhance the representation output of regions possessing emotion-discriminative attributes. Figure 3 illustrates the schematic of the sentiment-aware attention and multi-task gated predictor. Let the visual feature map be $F_v \in \mathbb{R}^{N \times d}$, where $N = 196$ is the number of image patches and $d = 768$ is the feature dimension per patch. Linear projections generate the query, key, and value matrices:

$$Q = E_{img} W_Q \in \mathbb{R}^{K \times d_k} \quad (12)$$

$$K = F_v W_K \in \mathbb{R}^{N \times d_k} \quad (13)$$

$$V = F_v W_V \in \mathbb{R}^{N \times d_k} \quad (14)$$

where, $W_Q \in \mathbb{R}^{K \times d_k}$, $W_K \in \mathbb{R}^{d \times d_k}$, and $W_V \in \mathbb{R}^{d \times d_k}$ are learnable projection matrices, and the single-head attention dimension is set to $d_k = 64$. The model adopts a multi-head attention mechanism with the number of heads $H = 4$ to capture multi-pattern interaction relations between sentiment semantics and visual regions. The single-head attention computation takes the form:

$$Attn(Q, K, V) = \text{Softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \quad (15)$$

The query vectors of this attention module originate from the model's output sentiment distribution rather than randomly

initialized learnable vectors, thus the resulting attention weights possess interpretable semantics—the weight values indicate the degree of attention that each sentiment category pays to each image patch, providing spatial-level attribution for model decisions.

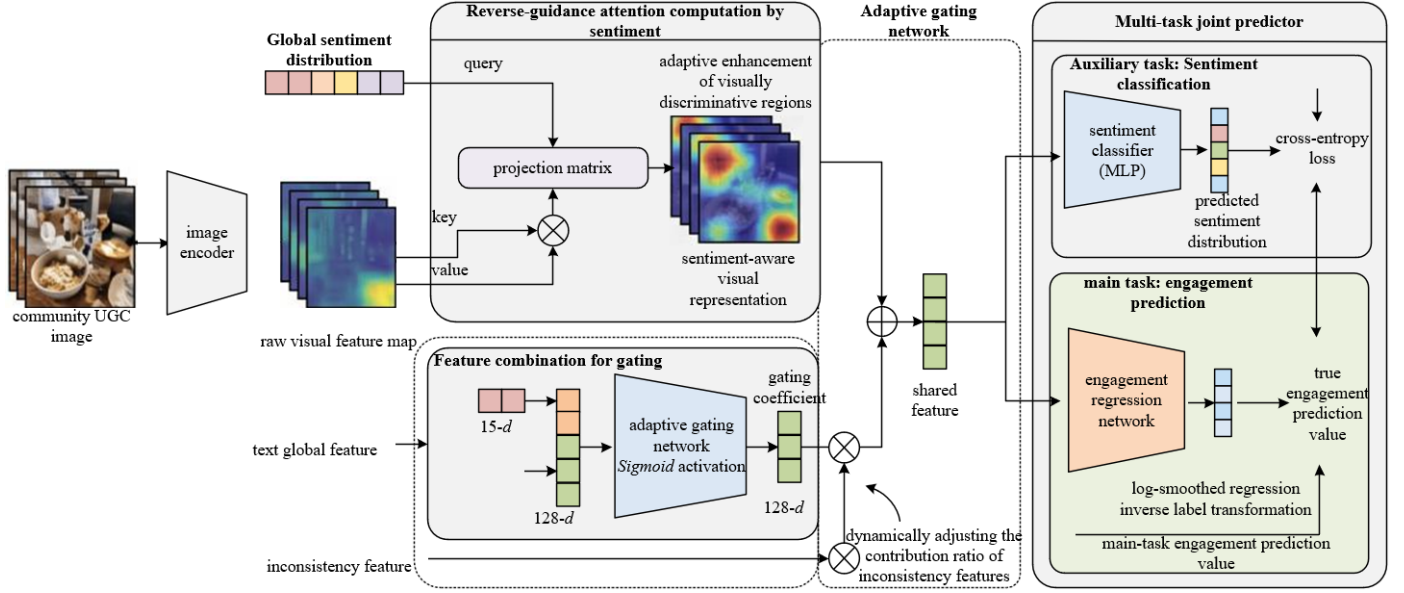


Figure 3. Schematic of the sentiment-aware attention and multi-task gated predictor

Note: UGC: User-Generated Content; MLP: Multi-Layer Perceptron.

Concatenating the multi-head attention outputs along the feature dimension and mapping back to the original feature dimension via the output projection matrix $W_O \in \mathbb{R}^{(H \cdot d_k) \times d}$ yields the sentiment-enhanced visual representation $F_v' \in \mathbb{R}^{K \times d}$. Weighted aggregation based on the sentiment distribution probabilities produces the final visual feature:

$$f_{visual} = \sum_{k=1}^K \alpha_k F_v'^{(k)}, \alpha_k = \text{Softmax}(E_{img})_k \quad (16)$$

where, α_k is the probability weight of the k -th sentiment category after Softmax normalization. This weighting operation reschedules the visual representation using the sentiment distribution: the attention output corresponding to high-probability sentiment categories occupies a larger feature weight. Sentiment semantics no longer serve merely as the model's output target; instead, they participate in reverse in the visual feature extraction process, constructing a computation closed loop in which semantics drive visual representation.

Engagement metrics of community UGC exhibit significant long-tail distribution: a small number of head samples possess extremely high interaction values. Directly adopting raw labels for mean squared error regression causes the model to bias toward fitting large-value samples, weakening the learning effect on tail samples. This paper applies a log-smoothing transformation $\tilde{y} = \log(1 + y)$ to the regression target, mapping labels into a numerically smoother interval; after the model outputs predictions, the inverse transformation $\hat{y} = \exp(\tilde{y}) - 1$ restores the true engagement prediction result. The sentiment-enhanced visual feature f_{visual} and the text global feature f_{txt} are concatenated to construct the basic shared feature $f_{shared} \in \mathbb{R}^{768}$. To accommodate the varying degrees of image-text conflict across different samples, a sample-adaptive gating coefficient is introduced to

dynamically adjust the contribution ratio of the inconsistency feature:

$$G = \sigma(\text{MLP}_{gate}([E_{img}; D_{fusion}])) \in \mathbb{R}^{128} \quad (17)$$

where, $\sigma()$ denotes the Sigmoid activation function, and MLP_{gate} receives the 15 + 128 concatenated feature of dimension and outputs a 128-dimensional gating vector. When image-text sentiments tend to be consistent, the gating output approaches zero, suppressing the effect of the inconsistency feature; when image-text sentiment conflict is significant, the gating output approaches one, amplifying the influence of image-text sentiment tension on prediction. The forward computation for engagement regression is expressed as:

$$\hat{y} = \text{MLP}_{reg}([f_{shared}; G \odot D_{fusion}]) \quad (18)$$

The symbol $\odot$ denotes element-wise multiplication. MLP_{reg} receives the combined feature of dimension 768 + 128, passes through hidden layers of 512 and 256 dimensions in sequence, and outputs the engagement prediction result in log space.

The network as a whole adopts a multi-task joint loss for end-to-end optimization. The loss expression is:

$$L_{total} = L_{reg} + \lambda_{cls} L_{cls}^{aux} + \lambda_{adapt} L_{adapt} \quad (19)$$

The main-task regression loss $L_{reg} = 1/B \sum_{i=1}^B (\hat{y}_i - \tilde{y}_i)^2$ takes the mean squared error form. The auxiliary sentiment classification loss $L_{cls}^{aux} = 1/B \sum_{i=1}^B CE(\hat{y}_i^{(cls)}, y_i)$ is the cross-entropy loss. L_{adapt} is the adapter warm-up loss defined in Section 2.2. Hyperparameters are set as $\lambda_{cls} = 0.8$ to prevent the auxiliary classification task from dominating gradient updates, and $\lambda_{adapt} = 1.0$. The optimizer is AdamW, with initial

learning rate set to 5×10^{-5} , weight decay of 1×10^{-4} , total training epochs of 50, and batch size $B = 128$. Sentiment classification as an auxiliary task introduces semantic regularization constraints into the shared feature space, improving the representation quality of engagement regression while learning sentiment decision boundaries, and alleviating the insufficient feature space structuring caused by relying solely on numerical supervision in a pure regression task.

2.5 Double machine learning causal inference module

The end-to-end network relies on multi-task joint optimization to complete the fused modeling of sentiment representation and engagement prediction; its output can only capture correlational relationships between variables and

cannot reflect causal changes brought by deliberate interventions. Community operation decisions need to answer counterfactual questions—that is, by how much user engagement will change after the expression intensity of a certain sentiment dimension within an image is actively increased. Relying solely on fitted results from observational samples is insufficient to support intervention strategy formulation. Therefore, this paper sets DML as an offline analysis module executed after network training is complete, leveraging non-parametric machine learning to flexibly control high-dimensional confounding variables and unbiasedly estimate the average treatment effect of the treatment variable T on the outcome variable Y at the linear solving stage. Figure 4 illustrates the six layers of cross-fitting and residual extraction.

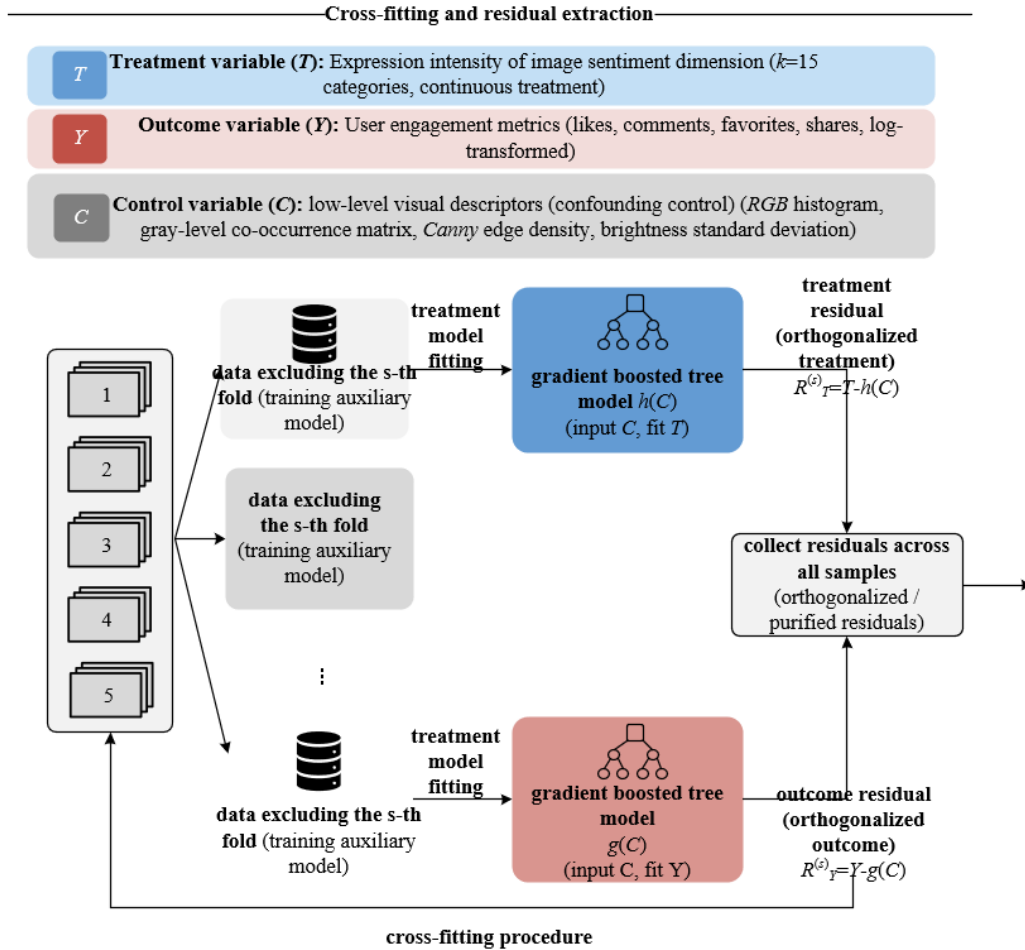


Figure 4. Framework of Double Machine Learning (DML) causal inference

For each sentiment dimension k , the image sentiment distribution output $E^{(k)}_{img}$ is selected as the continuous treatment variable T , and the engagement metrics—including likes, comments, favorites, and shares—are taken as the outcome variable Y in turn. The control variable $C \in \mathbb{R}^{101}$ adopts low-level visual descriptors of the image, used to eliminate confounding bias arising from objective image quality. This set of features contains the RGB three-channel color histogram, contrast, energy, and homogeneity texture obtained from the gray-level co-occurrence matrix (GLCM), while also incorporating Canny edge density and image brightness standard deviation. To suppress estimation bias caused by overfitting and obtain an asymptotically normal estimator, this module executes a cross-fitting procedure: the

dataset is randomly partitioned into $S = 5$ equal-sized, mutually non-overlapping subsets, where I_s denotes the sample index set corresponding to the s -th fold. For each fold s , two groups of auxiliary models based on gradient boosted trees are trained using all samples excluding the s -th fold. The outcome model $g(C)$ learns the mapping from confounding control variable C to the outcome variable Y , the treatment model $h(C)$ learns the mapping from the confounding control variable C to the treatment variable T . Orthogonalized residuals are solved on the s -th fold samples to strip away the interference brought by low-level visual features, yielding purified variables:

$$\tilde{Y}_i = Y_i - \hat{g}_{-s}(C_i), \tilde{T}_i = T_i - \hat{h}_{-s}(C_i), \forall i \in I_s \quad (20)$$

where, $\hat{g}_s(C_i)$ and $\hat{h}_s(C_i)$ denote the auxiliary models trained without using the s -th fold samples. Collecting the residual sample set $\{(\tilde{T}_i, \tilde{Y}_i)\}_{i=1}^N$ generated across all folds, the average treatment effect θ_k corresponding to the k -th sentiment category is solved via intercept-free linear regression:

$$\theta_k = \frac{\sum_{i=1}^N \tilde{T}_i \tilde{Y}_i}{\sum_{i=1}^N \tilde{T}_i^2} \quad (21)$$

This estimator satisfies the Neyman orthogonality condition; its estimation error converges to zero at a rate of $\sqrt{N}$ as the sample size grows. To evaluate the statistical uncertainty of the estimation results, 1000 rounds of Bootstrap resampling are performed to compute the standard error of θ_k and the 95% confidence interval, completing the significance test for each sentiment dimension's effect.

Through DML computation, one can obtain the causal effect rankings of all 15 sentiment dimensions on the four engagement metrics respectively. This result provides causal evidence distinct from correlational analysis—it can distinguish between sentiment features that merely co-occur with high engagement and sentiment features that can actively boost interaction levels, thereby guiding the selection of content filters, configuration of topic hashtags, and setting of content push priorities. When the average treatment effect of a certain sentiment dimension takes a positive value and the confidence interval does not contain zero, enhancing the expression of that sentiment category can causally promote user engagement behaviors, providing quantitative support for practical operational interventions. This module and the aforementioned end-to-end prediction network together constitute a complete technical pipeline: the prediction network completes the fitting and output of engagement values, while the causal inference module excavates intervention-oriented decision support, realizing an extension from outcome prediction to strategic guidance.

3. EXPERIMENTAL RESULTS AND ANALYSIS

3.1 Experimental settings

This paper crawls publicly available UGC posts from the photography aesthetics topics on Rednote—a mainstream domestic visual community platform—spanning the period from January 2024 to June 2025. After deduplication, removal of pure-text and video posts, and privacy blurring, 12,847 valid samples are finally obtained. Each sample is independently annotated by three annotators with a photography aesthetics background across 15 sentiment categories; the Fleiss' Kappa coefficient is 0.78, indicating excellent annotation agreement. Dataset statistics are shown in Table 1. For sentiment classification, accuracy and macro-average F1 are adopted as evaluation metrics. For engagement prediction, root mean square error (RMSE) and Spearman rank correlation coefficient are adopted; the latter is insensitive to absolute numerical magnitudes and is more suitable for assessing ranking ability. All experiments adopt 5-fold cross-validation with mean $\pm$ standard deviation reported. Comparative methods cover the pure-vision models ResNet-50 (Deep Residual Network with 50 layers) and ViT-B/16 (Vision Transformer, Base variant, patch size 16×16), zero-shot vision-language CLIP-ZS (CLIP Zero-Shot), full-

parameter fine-tuned CLIP-FT (CLIP Full Fine-Tuning), simple multimodal concatenation Multimodal-Concat (Multimodal Concatenation), and the gated fusion method CGCT-ITMSA (Cross-modal Gated/Cross-attention Fusion for Image–Text Multi-modal Sentiment-aware Aggregation).

Table 1. Dataset statistics

Statistical Indicator	Value
Total number of samples	12,847
Number of sentiment label categories	15
Average caption length (number of words)	18.3 ± 12.7
Average number of likes	$1,247 \pm 3,856$
Average number of comments	89 ± 312
Average number of favorites	$356 \pm 1,023$
Average number of shares	42 ± 187
Entropy of sentiment label distribution	2.31

3.2 Sentiment Classification Performance Comparison

Figure 5 presents the classification performance of different methods across the 15 sentiment categories. The proposed adapter-based fine-tuning strategy achieves an accuracy of 74.63% while training only 0.52% additional parameters, significantly surpassing the 72.18% of full-parameter fine-tuning. On community-specific sentiment categories, the proposed method outperforms CLIP-FT by 5.2% and 4.7% respectively, verifying the effectiveness of the adapter layers in capturing community semantic shift. Compared with the 61.24% of pure-vision ViT-B/16, the semantic prior advantage brought by vision-language pre-training is extremely significant, indicating that relying purely on low-level textures is insufficient for modeling high-order abstract sentiments. The macro-average F1 of this paper is 73.89%; the minimal gap from accuracy indicates that the model does not bias toward high-frequency categories and maintains balanced recognition capability across all 15 sentiment categories.

To verify whether the model can effectively extract sentiment-related visual cues that are highly correlated with active engagement behaviors from aesthetically oriented UGC content, and to examine the progressive evolution of sentiment representations from scene perception to engagement-sensitive encoding, this paper selects a typical sunset landscape sample for visual analysis. As shown in Figure 6, at the general visual representation stage, although overall scene information—including sky, water surface, shoreline, and foreground rocks—is preserved, the visual responses remain relatively scattered, with some low-contribution regions also retained. After aesthetic sentiment discrimination, the model exhibits stronger selectivity toward regions with significant aesthetic value, such as the sunset subject, golden water reflections, and the silhouette of a figure on the horizon, while non-critical background information is notably suppressed. In the further sentiment feature enhancement stage, warm-light gradients, reflection continuity, and light-dark contour contrast are reinforced, making the visual structures that express aesthetic and healing sentiments more prominent. The final high-likes-sensitive visual representation further compresses peripheral redundant information, causing the sunset, reflection, and figure silhouette to form a more concentrated visual semantic center. This phenomenon indicates that EMO-Community Net does not rely on shallow attributes such as global color or brightness for engagement estimation; rather, it leverages the sentiment distribution to guide visual feature reweighting, transforming local visual

patterns with high sentiment discriminability into high-level representations required for engagement prediction. This is consistent with the design objective of the sentiment-aware cross-attention mechanism—using global sentiment semantics to focus on discriminative image regions. Therefore, Figure 6 qualitatively demonstrates that the high-likes tendency formed

by aesthetically oriented content is closely related to the stable extraction and reinforcement of its core aesthetic regions, providing interpretable visual evidence for identifying and recommending high-consensus content in community operations.

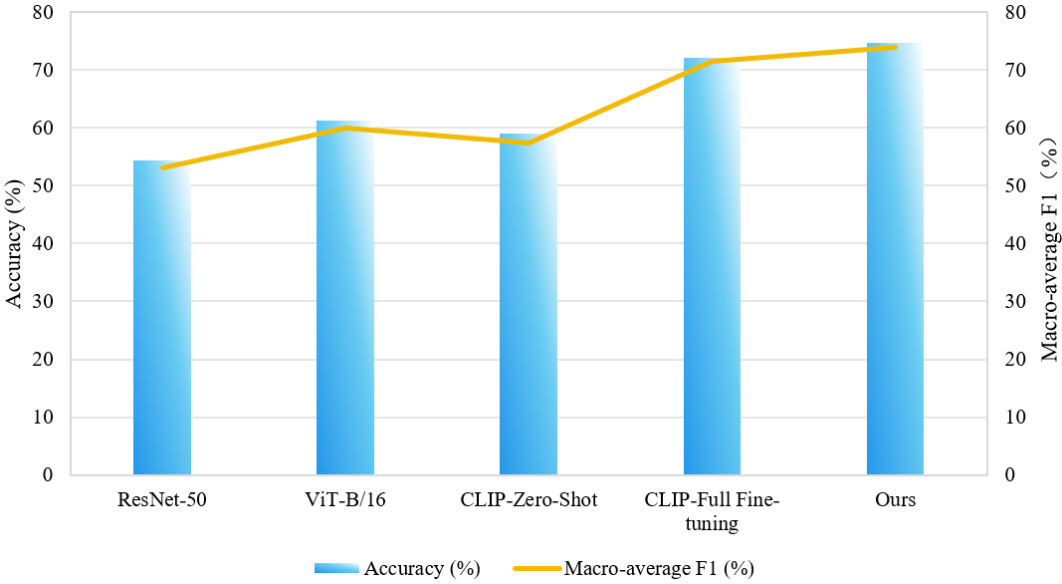


Figure 5. Performance comparison of different methods on the UGC image sentiment classification task
 Note: ResNet-50: Deep Residual Network with 50 layers; ViT-B/16: Vision Transformer, Base variant, patch size 16×16; CLIP-ZS: CLIP Zero-Shot; CLIP-FT: CLIP Full Fine-Tuning.



Figure 6. Visualization of the high-likes engagement pattern driven by aesthetic sentiment



Figure 7. Visualization of the high-comments engagement pattern driven by cyberpunk-conflict sentiment

To further test whether the model can capture high-order sentiment information that drives active discussion in complex UGC scenes with strong color conflict and cyberpunk visual characteristics, this paper selects a nighttime urban landscape sample to analyze the visual representation changes under the high-comments engagement pattern. Figure 7 shows that the raw image contains multi-source visual information—high-density buildings, cloud layers, vegetation, lighting, and water reflections. The general visual representation maintains a roughly balanced response to these elements, making it difficult to highlight the key structures that truly determine the cyberpunk-conflict sentiment. Upon entering the sentiment discrimination stage, the model significantly strengthens the contrast between blue-purple high-rises, cool-toned neon façades, and foreground orange-yellow illumination, while suppressing low-contribution regions such as the sky and peripheral vegetation, thereby concentrating the expression of warm-cool color conflict and the visual tension formed by artificial light sources. The sentiment feature enhancement result further elevates building outlines, neon color saturation, and local contrast between warm and cool regions. The final high-comments-sensitive visual representation further converges the visual center onto the highlighted building mass and the foreground warm-light band, thus retaining visual cues with greater conflict, heterogeneity, and discussion-provoking potential. This result shows that the model can separate emotional tension from general scene semantics in complex community images and transform it into discriminative features required for engagement prediction, which is consistent with the idea of treating sentiment inconsistency and sentiment intensity discrepancy as information sources for engagement modeling. At the same time, this visual pattern corroborates the causal analysis finding that cyberpunk sentiment exhibits a strong positive driving effect on comment behaviors—indicating that high-comment content does not rely merely on visual beauty, but is more likely triggered by salient emotional contrast and visual conflict that stimulate users' willingness to express themselves.

3.3 Comprehensive performance comparison on engagement prediction

Table 2 presents the prediction performance of each method on the four engagement metrics. The proposed method achieves Spearman correlation coefficients of 0.812, 0.785, 0.801, and 0.743 on likes, comments, favorites, and shares respectively, all significantly outperforming the best baseline CGCT-ITMSA. The gap between the comment prediction correlation and the likes prediction correlation of this paper is smaller than that of other baselines, indicating that the cross-modal inconsistency feature effectively captures the sentiment tension that drives user discussion. The proposed method also attains the lowest RMSE values, verifying the effectiveness of the log-normalization strategy in handling long-tail distributions. Under pure-vision models, comment prediction capability is far lower than likes prediction; after incorporating text and consistency modeling, the absolute gain on comment prediction exceeds that on likes prediction, strongly supporting the core hypothesis that image-text sentiment tension is a strong feature driving comment interactions.

3.4 Ablation study and module contribution analysis

To quantify the independent contribution of each innovation

point, six ablation variants are designed, with results shown in Figure 8. Removing the cross-modal inconsistency module causes the likes prediction correlation coefficient to drop from 0.812 to 0.758—a decrease of 5.4%, the most significant performance drop among all modules—verifying the cornerstone role of inconsistency modeling for the entire framework. Removing multi-task joint learning reduces the correlation to 0.769, indicating that relying solely on the regression task makes it difficult to learn shared representations that are both discriminative and predictive. Removing sentiment-aware attention also leads to a notable performance decline, proving that spatial reweighting guided by the sentiment distribution can effectively suppress background noise interference. Even after removing the adapter layer, the model still outperforms most baselines, yet the gain brought by the adapter layer is statistically significant, confirming the necessity of lightweight adaptation. The gated fusion removal yields the smallest drop, indicating that even simple concatenation of the inconsistency feature still works, while the gating mechanism further improves synergy efficiency through dynamic weighting.

Table 2. Performance comparison of different methods on the user engagement prediction task (Root Mean Square Error↓ / Spearman's ρ ↑)

Method	Likes	Comments	Favorites	Shares
<i>ViT-B/16</i>	0.432 / 0.614	0.401 / 0.536	0.389 / 0.581	0.447 / 0.493
<i>Multimodal-Concat</i>	0.378 / 0.702	0.352 / 0.658	0.341 / 0.683	0.392 / 0.612
<i>CLIP-FT + Simple Fusion</i>	0.361 / 0.731	0.338 / 0.691	0.325 / 0.712	0.374 / 0.644
<i>CGCT-ITMSA</i>	0.342 / 0.758	0.319 / 0.724	0.308 / 0.739	0.355 / 0.675
<i>Ours</i>	0.312 / 0.812	0.284 / 0.785	0.276 / 0.801	0.328 / 0.743

Note: ViT-B/16: Vision Transformer, Base variant, patch size 16×16 ; Multimodal-Concat: Multimodal Concatenation; CLIP-FT: CLIP Full Fine-Tuning; CGCT-ITMSA: Cross-modal Gated/Cross-attention Fusion for Image-Text Multi-modal Sentiment-aware Aggregation.

3.5 Stratified evaluation of cross-modal inconsistency

To directly verify the utility of the inconsistency feature under different scenarios, the test set is divided into three subsets—high, medium, and low—according to the value of $D_{KL}^{img \rightarrow txt}$, each occupying one third. The prediction performance of the full model versus the model with the inconsistency module removed is compared across the three groups, with results shown in Figure 9. On the high-inconsistency subset, the performance gap between the two models widens sharply to 0.087, while on the low-inconsistency subset the gap is only 0.012. This gradient change strongly proves that the stronger the image-text sentiment conflict, the richer the additional information provided by the inconsistency feature. In high-conflict samples, the model with the inconsistency module removed nearly fails, whereas the full model maintains high predictive power—fully demonstrating that explicitly modeling image-text tension is a key technical path for handling complex multi-modal UGC content.

3.6 Causal effect estimation and operation strategy analysis

The DML module is applied to estimate the average

treatment effect of the 15 sentiment categories on likes and comments, with results shown in Table 3. For likes-driving effects, aesthetic and surprise exhibit the strongest positive causal effects, while anger and depression significantly suppress likes—consistent with the basic tone of photography aesthetics communities that pursue visual enjoyment. The pattern for comments-driving is completely reversed: anger, cyberpunk, and depression are the strongest positive driving factors. This indicates that the sentiment paths driving passive appreciation and driving discussion are diametrically opposite: harmonious beauty prompts passive approval, while conflict and discomfort provoke active expression. The effect of healing sentiment on comments is negative, indicating that overly tranquil content lacks topicality in photography communities. This causal-level finding provides precise guidance for community operations: to increase likes, one should prioritize pushing content that reinforces aesthetic and surprise expressions; to ignite discussion, one should strategically introduce cyberpunk or slightly depressing contrastive compositions.

3.7 Model efficiency and complexity analysis

To verify the feasibility of the lightweight design in real deployment scenarios, the parameter count, FLOPs, and inference latency of each model are compared, with results shown in Table 4. Because the complete CLIP backbone is preserved, the total parameter count of the proposed method is roughly on par with full-parameter fine-tuning; however, the trainable parameters amount to only 1.58 M—0.52% of the total—reducing gradient update cost by nearly 99.5% compared with full-parameter fine-tuning. Inference latency is 132.5 ms, slightly higher than CLIP-FT's 124.3 ms; this is the additional computation overhead introduced by the inconsistency encoding module, yet it remains fully acceptable in community content stream processing. The minimal trainable parameters enable the model to quickly readapt at extremely low cost when concept drift occurs in the community data distribution—a property of great engineering value for dynamically evolving community operation scenarios.

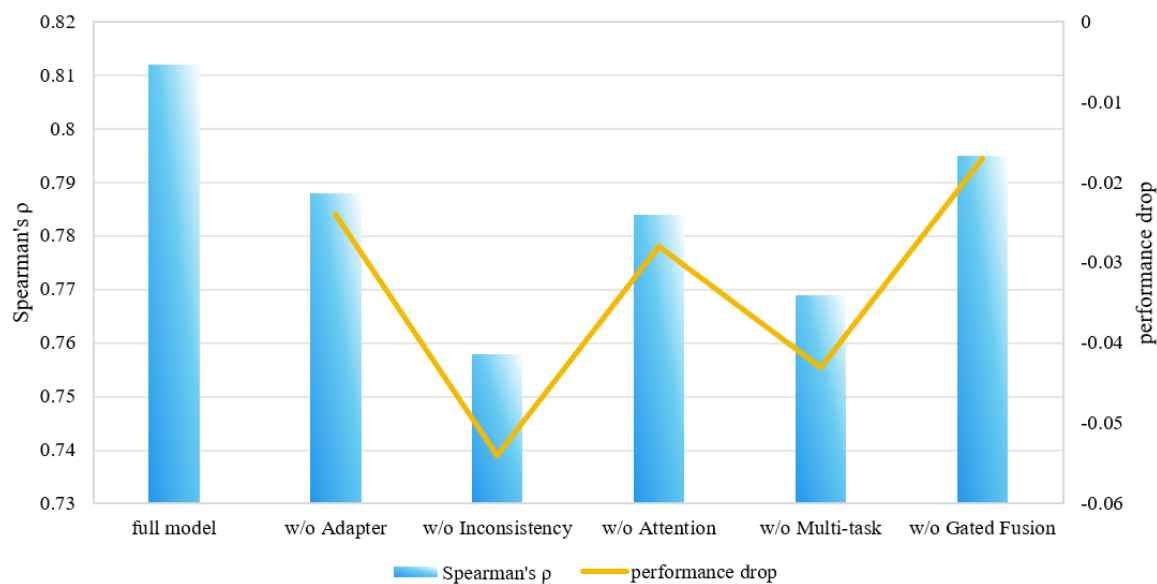


Figure 8. Ablation study on likes prediction (Spearman's $\rho \uparrow$)

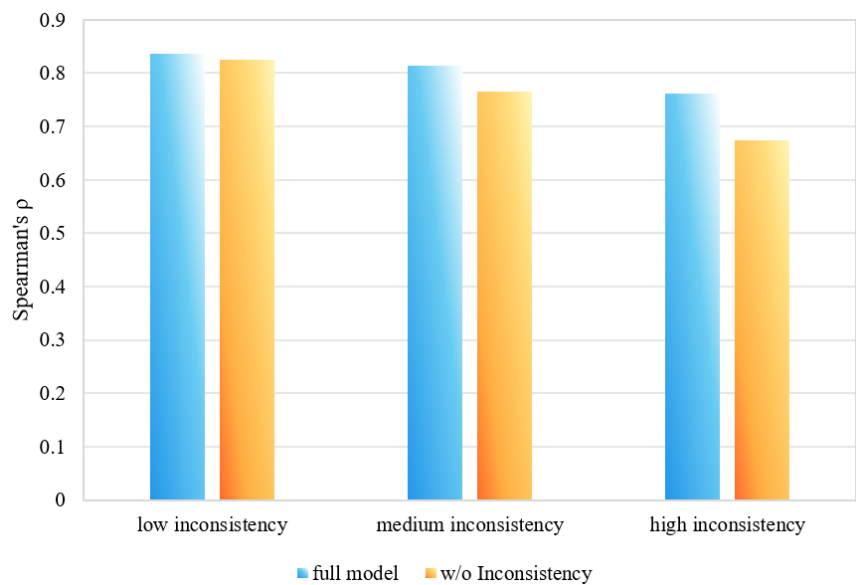


Figure 9. Model prediction performance under different levels of image-text inconsistency (Spearman's ρ)

Table 3. Average Treatment Effect (ATE) estimates of each sentiment dimension on engagement (with 95% Bootstrap confidence intervals)

Sentiment Dimension	ATE on Likes	ATE on Comments
Aesthetic	0.124 [0.112, 0.136]	0.031 [0.018, 0.044]
Surprise	0.118 [0.105, 0.131]	0.087 [0.072, 0.102]
Healing	0.067 [0.054, 0.080]	-0.015 [-0.028, -0.002]
Grandeur	0.085 [0.071, 0.099]	0.044 [0.031, 0.057]
Vintage	0.042 [0.029, 0.055]	0.028 [0.015, 0.041]
Cyberpunk	-0.018 [-0.031, -0.005]	0.152 [0.138, 0.166]
Depression	-0.087 [-0.102, -0.072]	0.134 [0.119, 0.149]
Anger	-0.112 [-0.128, -0.096]	0.198 [0.182, 0.214]
Joy	0.078 [0.065, 0.091]	0.035 [0.022, 0.048]

Table 4. Model complexity and inference efficiency comparison

Method	Total Params (M)	Trainable Params (M)	FLOPs (G)	Inference Latency (ms/sample)
CLIP-FT	302.5	302.5	98.7	124.3
CGCT-ITMSA	285.4	285.4	95.2	118.7
Ours	304.1	1.58 (0.52%)	100.1	132.5

Note: CLIP-FT: CLIP Full Fine-Tuning; CGCT-ITMSA: Cross-modal Gated/Cross-attention Fusion for Image–Text Multi-modal Sentiment-aware Aggregation.

4. CONCLUSION

Aiming at the problem of UGC image sentiment feature extraction and user engagement prediction in community operation scenarios, this paper proposes EMO-Community Net, an end-to-end joint learning framework. Starting from a community-adaptive vision-language sentiment encoder, the framework employs a lightweight adapter-based fine-tuning strategy to enable a general-purpose vision-language model to accurately adapt to the sentiment distribution of a specific community with approximately 0.5% trainable parameters, and constructs dynamic sentiment prototypes for multi-granularity sentiment distribution prediction. At the cross-modal modeling level, image-text sentiment inconsistency is elevated from an empirical phenomenon to a computable modeling object; quantification metrics are systematically constructed from three dimensions—distributional morphology, sentiment intensity, and sentiment polarity—and encoded into interpretable features for engagement prediction via an encoding network. At the prediction level, the global sentiment distribution serves as a query to guide cross-attention computation over the visual Transformer, realizing the reverse guidance of sentiment semantics on visual feature focusing; the multi-task gated predictor further enables sentiment classification and engagement prediction to mutually reinforce each other within the shared representation space.

Experiments on a real-world community UGC dataset show that the proposed method achieves state-of-the-art performance in both sentiment classification accuracy and

engagement prediction correlation. Ablation studies confirm the effective contribution of each module, and the stratified evaluation of cross-modal inconsistency reveals that the stronger the image-text sentiment conflict, the richer the information gain from the inconsistency feature. DML causal analysis further reveals that the sentiment paths driving likes and driving comments are diametrically opposite, providing community operations with causal-level decision support beyond correlational analysis. This paper offers technical support and theoretical reference for adapting vision-language models to community UGC scenarios, modeling multi-modal sentiment inconsistency, and analyzing the association between sentiment features and user behaviors. Future work will focus on the dynamic temporal extension of causal effect estimation and cross-platform community culture adaptive transfer.

REFERENCES

- [1] Shahbaznezhad, H., Dolan, R., Rashidirad, M. (2021). The role of social media content format and platform in users' engagement behavior. *Journal of Interactive Marketing*, 35(1): 47-65. <https://doi.org/10.1016/j.intmar.2020.05.001>
- [2] Kusuma, A.A., Afiff, A.Z., Gayatri, G., Hati, S.R.H. (2024). Is visual content modality a limiting factor for social capital? Examining user engagement within Instagram-based brand communities. *Humanities and Social Sciences Communications*, 11(1): 9. <https://doi.org/10.1057/s41599-023-02529-6>
- [3] Lin, H.H., Lin, J.D., Ople, J.J.M., Chen, J.C., Hua, K.L. (2021). Social media popularity prediction based on multi-modal self-attention mechanisms. *IEEE Access*, 10: 4448-4455. <https://doi.org/10.1109/ACCESS.2021.3136552>
- [4] Kim, J., Ahn, H., Park, E. (2024). Multi-pop: Enhancing user engagement with content-based multimodal popularity prediction in social media. *Expert Systems*, 41(12): e13707. <https://doi.org/10.1111/exsy.13707>
- [5] Rietveld, R., Van Dolen, W., Mazloom, M., Worring, M. (2020). What you feel, is what you like influence of message appeals on customer engagement on Instagram. *Journal of Interactive Marketing*, 49(1): 20-53. <https://doi.org/10.1016/j.intmar.2019.06.003>
- [6] Stsiampkouskaya, K., Joinson, A., Piwek, L., Ahlbom, C.P. (2021). Emotional responses to likes and comments regulate posting frequency and content change behaviour on social media: An experimental study and mediation model. *Computers in Human Behavior*, 124: 106940. <https://doi.org/10.1016/j.chb.2021.106940>
- [7] Zhao, S., Yao, X., Yang, J., et al. (2021). Affective image content analysis: Two decades review and new perspectives. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(10): 6729-6751. <https://doi.org/10.1109/TPAMI.2021.3094362>
- [8] Jeong, D., Son, H., Choi, Y., Kim, K. (2024). Enhancing social media post popularity prediction with visual content. *Journal of the Korean Statistical Society*, 53(3): 844-882. <https://doi.org/10.1007/s42952-024-00270-7>
- [9] Ortis, A., Farinella, G.M., Battiato, S. (2020). Survey on visual sentiment analysis. *IET Image Processing*, 14(8): 1440-1456. <https://doi.org/10.1049/iet-ipr.2019.1270>
- [10] Yang, J., Li, J., Wang, X., Ding, Y., Gao, X. (2021).

- Stimuli-aware visual emotion analysis. *IEEE Transactions on Image Processing*, 30: 7432-7445. <https://doi.org/10.1109/TIP.2021.3106813>
- [11] Yang, J., Li, J., Li, L., Wang, X., Ding, Y., Gao, X. (2022). Seeking subjectivity in visual emotion distribution learning. *IEEE Transactions on Image Processing*, 31: 5189-5202. <https://doi.org/10.1109/TIP.2022.3193749>
- [12] Yang, H., Fan, Y., Lv, G., Liu, S., Guo, Z. (2023). Exploiting emotional concepts for image emotion recognition. *The Visual Computer*, 39(5): 2177-2190. <https://doi.org/10.1007/s00371-022-02472-8>
- [13] Gandhi, A., Adhvaryu, K., Poria, S., Cambria, E., Hussain, A. (2023). Multimodal sentiment analysis: A systematic review of history, datasets, multimodal fusion methods, applications, challenges and future directions. *Information Fusion*, 91: 424-444. <https://doi.org/10.1016/j.inffus.2022.09.025>
- [14] Zhu, T., Li, L., Yang, J., Zhao, S., Liu, H., Qian, J. (2022). Multimodal sentiment analysis with image-text interaction network. *IEEE Transactions on Multimedia*, 25: 3375-3385. <https://doi.org/10.1109/TMM.2022.3160060>
- [15] Zhou, S., Yang, X., Wang, Y., Zheng, X., Zhang, Z. (2023). Affective agenda dynamics on social media: Interactions of emotional content posted by the public, government, and media during the COVID-19 pandemic. *Humanities and Social Sciences Communications*, 10(1): 797. <https://doi.org/10.1057/s41599-023-02265-x>
- [16] Gurung, M.I., Agarwal, N., Bhuiyan, M.M.I., Poudel, D. (2025). Symbolic signals on Instagram: How visual media shapes engagement, emotion, trust, and diffusion. *Social Network Analysis and Mining*, 15(1): 57. <https://doi.org/10.1007/s13278-025-01469-0>
- [17] Chandrasekaran, G., Antoanela, N., Andrei, G., Monica, C., Hemanth, J. (2022). Visual sentiment analysis using deep learning models with social media data. *Applied Sciences*, 12(3): 1030. <https://doi.org/10.3390/app12031030>
- [18] Zhang, H., Xu, D., Luo, G., He, K. (2022). Learning multi-level representations for affective image recognition. *Neural Computing and Applications*, 34(16): 14107-14120. <https://doi.org/10.1007/s00521-022-07139-y>
- [19] Zhou, K., Yang, J., Loy, C.C., Liu, Z. (2022). Learning to prompt for vision-language models. *International Journal of Computer Vision*, 130(9): 2337-2348. <https://doi.org/10.1007/s11263-022-01653-1>
- [20] Wu, C., Xu, Q., Wei, Y., Yuan, S., Wu, J., Wang, L. (2024). Towards visual emotion analysis via multi-perspective prompt learning with residual-enhanced adapter. *Knowledge-Based Systems*, 295: 111790. <https://doi.org/10.1016/j.knosys.2024.111790>
- [21] Xu, Q., Wei, Y., Yuan, S., Wu, J., Wang, L., Wu, C. (2024). Learning emotional prompt features with multiple views for visual emotion analysis. *Information Fusion*, 108: 102366. <https://doi.org/10.1016/j.inffus.2024.102366>
- [22] Deng, S., Wu, L., Shi, G., et al. (2024). Learning to compose diversified prompts for image emotion classification. *Computational Visual Media*, 10(6): 1169-1183. <https://doi.org/10.1007/s41095-023-0389-6>
- [23] Chen, D., Su, W., Wu, P., Hua, B. (2023). Joint multimodal sentiment analysis based on information relevance. *Information Processing & Management*, 60(2): 103193. <https://doi.org/10.1016/j.ipm.2022.103193>
- [24] Wang, J., Yang, Y., Jiang, Y., Ma, M., Xie, Z., Li, T. (2024). Cross-modal incongruity aligning and collaborating for multi-modal sarcasm detection. *Information Fusion*, 103: 102132. <https://doi.org/10.1016/j.inffus.2023.102132>
- [25] Wang, J., Yang, S., Zhao, H., Yang, Y. (2023). Social media popularity prediction with multimodal hierarchical fusion model. *Computer Speech & Language*, 80: 101490. <https://doi.org/10.1016/j.csl.2023.101490>
- [26] Liu, A.A., Du, H., Xu, N., et al. (2023). Exploring visual relationship for social media popularity prediction. *Journal of Visual Communication and Image Representation*, 90: 103738. <https://doi.org/10.1016/j.jvcir.2022.103738>