

Article

Implicit-Emotion Recognition Model Based on Content–Style Decoupling and Conditional Fusion

Yi Zhang, Junqing Zhu and Hua Zhao *

College of Computer Science and Engineering, Shandong University of Science and Technology, Qingdao 266000, China; 202311070229@sdust.edu.cn (Y.Z.); 202411070230@sdust.edu.cn (J.Z.)

* Correspondence: huazhao@sdust.edu.cn

Abstract

Implicit-emotional expressions are common in college students' social media posts, where literal meanings may contradict underlying emotions. Bidirectional Encoder Representations from Transformers (BERT)-based models face two coupled challenges in this setting: feature confusion between semantic content and expression style, and coarse-grained feature fusion. We propose CSD-IFRN (Content–Style Disentanglement with Conditional Fusion for Implicit Emotion Recognition Network), which disentangles content and style using dual non-shared BERT encoders, combines gradient-reversal adversarial training with a content–style orthogonality regularizer (L_{orth}), and applies conditional layer normalization (CLN) for adaptive fusion. On a dedicated dataset of 11,154 triple-annotated texts, averaged over five random seeds, CSD-IFRN achieves 88.58% accuracy and 88.43% macro-F1, improving over BERT-base-chinese by 6.99 points and over the strongest SOTA baseline by 2.61 points. The main gains remain significant after Holm correction ($p < 0.01$) and also hold on a public benchmark. A frozen style probe trained on content features falls to 50.40% balanced accuracy, close to chance level, supporting effective disentanglement. Among fusion strategies, CLN achieves the best accuracy with low seed-to-seed variance. These results suggest that CSD-IFRN can provide an auxiliary signal for university mental-health monitoring, rather than a clinical diagnostic tool.

Keywords: conditional fusion; content–style disentanglement; implicit-emotion recognition; mental-health monitoring of college students

1. Introduction

Social media has become the main platform for college students to express their emotions, and the texts they publish can reflect their psychological state [1]. Traditional sentiment analysis methods can achieve relatively accurate identification of explicit emotional expression in text by capturing emotional keywords and polarity words, which is consistent with the high linguistic recognizability of explicit emotions themselves [2]. However, students' language is highly personalized and subtle. The text often contains implicit-emotional expressions such as irony, complaining, and euphemistic venting, whose true emotional tendency is completely opposite to the literal meaning, which brings serious challenges to the task of emotion recognition [3]. For the sake of clear expression, the words 'emotion', 'sentiment' and 'emotional tendency' in this article are used interchangeably, referring to the binary emotional polarity of the text (positive vs. negative), and 'style' refers to whether the expression is conventional or implicit.



Academic Editor: Arkaitz Zubiaga

Received: 5 June 2026

Revised: 26 June 2026

Accepted: 2 July 2026

Published: 8 July 2026

Copyright: © 2026 by the authors.

Licensee MDPI, Basel, Switzerland.

This article is an open access article distributed under the terms and

conditions of the [Creative Commons](https://creativecommons.org/licenses/by/4.0/)

[Attribution \(CC BY\)](https://creativecommons.org/licenses/by/4.0/) license.

The accuracy of implicit-emotion recognition is directly related to the effectiveness of universities in conducting mental-health monitoring and intervention services for students based on social media texts [1], and recent work by Yadav [4] further highlights the value of social media text analysis for the neuro-semantic evaluation of mental-health conditions among young adults. Negative emotional expressions among university students generally have subtle characteristics, and many negative emotions are not conveyed through straightforward emotional vocabulary but rather presented in implicit forms such as irony and euphemistic venting. These hidden negative emotions are often closely related to potential psychological issues such as academic pressure, social anxiety, and depressive tendencies. Currently, deep learning models, especially pre-trained language models, perform well on explicit emotion tasks [5], but face two main bottlenecks in handling implicit expressions:

Feature confusion: Encoding the entire text into a single vector forces the semantic content and expression style of the text to intertwine in the same representational space, causing the two types of information to interfere with each other [6];

Coarse fusion granularity: Existing strategies (e.g., concatenation, weighted fusion) are linear and assume that contributions are static, unable to model the nonlinear, dynamic interactions between content and style. In implicit emotions, style must dynamically reverse semantics, which is a ‘contradictory relationship’ that requires the fine-grained handling of feature differences [7]. Traditional fusion methods face difficulties in this regard, limiting a deep understanding of implicit emotions.

In response to the above problems, this paper proposes CSD-IFRN, a progressive implicit-emotion recognition framework composed of three stages of ‘dual-path encoding—adversarial purification—conditional fusion’. By explicitly separating and reorganizing content features and style features, it reduces feature confusion and examines the degree to which different fusion strategies adapt to the decoupled features.

The core contributions are

- This paper proposes a three-stage progressive framework of ‘dual-path encoding—adversarial purification—conditional fusion’ for content–style decoupling in implicit-emotion recognition. Its key distinguishing component is an explicit content–style orthogonality regularizer (L_{orth}), which directly penalizes the residual content–style correlation and complements gradient-reversal training; the ablation experiment shows that this term brings a +1.49 F1 improvement and lowers the frozen style probe’s balanced accuracy from 56.60% to 50.40% ($\approx$ random level). We further adopt the balanced accuracy of the style probe as a transparent readout of the degree of decoupling. Unlike the global content–style orthogonality term used by contrastive-learning disentanglement baselines, L_{orth} is an intra-batch cross-correlation Frobenius penalty applied on L2-normalized features, superimposed on the gradient-reversal purification; our ablation separately characterizes its impact on both accuracy and the decoupling probe.
- A large-scale specialized dataset for implicit-emotion recognition on social media of college students (11,154 entries) was constructed, with fine-grained triple annotations for emotional polarity (positive/negative), style category (normal/implicit), and style intensity on a zero-to-one continuous scale. The annotation consistency reached a high level of reliability, providing high-quality data support for model training and performance validation in implicit-emotion recognition among the college student population.
- The effectiveness of the core modules of the model was systematically validated through multiple controlled experiments, while the failed attempts at the continuous style-intensity regression task were honestly reported. The core reasons for the

task failure were analyzed in depth from three dimensions, namely label construction, feature–task matching, and loss function design, providing important negative references for subsequent research on quantifying the intensity of emotional expression and improving model interpretability, effectively helping to avoid repeated trial-and-error in similar studies.

The structure of the remainder of this paper is organized as follows. Section 2 reviews related work, including implicit-emotion recognition, feature decoupling learning, and multi-feature fusion strategy. Section 3 provides a detailed description of the proposed CSD-IFRN model. Section 4 presents the experimental design and setup, covering dataset construction, baseline models, and evaluation metrics. Section 5 reports and discusses the experimental results. Section 6 concludes the paper and highlights potential directions for future research.

2. Related Work

2.1. Implicit-Emotion Recognition

Implicit-emotion recognition is a core subfield of sentiment analysis, and its main challenge lies in the inconsistency between the literal semantics of the text and the true emotional tendency [2]. Existing research methods can be divided into three categories: rule-based methods, traditional machine learning methods, and deep learning methods. Each category has produced representative research outcomes, but there are still key bottlenecks that need to be overcome. Sharma et al. [8] have provided a comprehensive review of AI-based automated methods for detecting sarcasm in social media, categorizing recent approaches according to their underlying architectures and learning paradigms.

The rule-based approach was the mainstream technique in the early stages of implicit-emotion recognition. Its core lies in matching implicit expressions through manually defined linguistic rules and templates. In their classic study on sarcasm and irony detection in social media, Van Hee et al. [3] systematically explored the linguistic features of implicit-emotion expression in text and proposed a fine-grained analytical framework, laying the theoretical foundation for rule-based implicit-emotion recognition methods. Early research mostly achieved implicit-emotion recognition by manually constructing irony templates, sarcastic syntactic rules, and fixed complaint collocations. These methods have strong interpretability in specific vertical contexts.

However, they heavily rely on human expertise, have limited rule coverage, and generalize poorly to the continuously evolving new forms of implicit expressions on social media. This fundamental flaw has been systematically revealed in early studies: Calvo and D’Mello [9], in their foundational review of affective computing, explicitly pointed out that rule-based approaches are limited by predefined feature sets and struggle to capture the complexity and diversity of human emotional expression; Mohammad [10], in reviewing the evolution of sentiment analysis methodologies, further emphasized that dictionary- and rule-based methods have inherent limitations when dealing with implicit expressions such as sarcasm and irony—the ‘decoupling’ between literal semantics and true emotion cannot be achieved through simple rules. Therefore, researchers gradually turned to data-driven machine learning methods in an effort to overcome the inherent bottlenecks of rule-based approaches.

As the earliest attempt at a data-driven paradigm, traditional machine learning methods focus on manual feature engineering and realize implicit-emotional recognition in combination with classifiers. Most existing studies revolve around manual features such as part-of-speech tagging, emotional dictionaries, syntactic structure, and semantic dependence, and use classifiers such as support vector machines (SVMs), random forests, and logistic regression to complete the modeling [11]. Subsequent research further expanded

this technical framework, optimized feature extraction rules for colloquial text in social media scenarios, and achieved better identification results than rule-based methods on datasets in specific fields. However, the performance of such methods is highly dependent on the quality of manual features, and it is difficult to capture the complex contextual semantic relationships in social media text [12]. Although CNN-based text-classification methods later achieved good results by capturing local semantic features, there are still fundamental limitations in modeling complex text expressions such as implicit emotions [12]. A further problem is that the content–style feature entanglement in pre-trained models is still difficult to resolve [6]. In the face of non-standard and colloquial expressions in college students' social media texts, the effectiveness of feature extraction is significantly reduced.

Deep learning methods represented by pre-trained language models (PLMs), leveraging the powerful ability of the Transformer architecture to capture contextual semantics [13], have become the mainstream technical solution in the field of emotion recognition. Prottasha et al. [5] systematically verified through experiments the excellent performance of BERT fine-tuning in general sentiment analysis tasks, providing key support for the application of pre-trained models in the field of emotion recognition. However, in the task of fine-grained analysis of implicit emotions, most existing models encode an entire text as a single vector representation, failing to explicitly disentangle the semantic content and expressive style of the text. This results in mutual interference between the two types of features, making it difficult for the model to penetrate the literal meaning and capture true emotional tendency. This limitation has been confirmed by several recent studies: Research on implicit sentiment in Chinese points out that its core challenge lies in the fact that emotional expressions often present a complex form of intertwined 'weak features' and 'mixed features' [14]; a review of Arabic sarcasm detection further indicates that the entanglement of style and semantics is a common challenge in cross-linguistic contexts [15]; studies on sarcasm and humor detection reveal that stylistic features (such as a sarcastic tone) can directly lead to a reversal of semantic sentiment polarity, highlighting the necessity of separating stylistic interference [16]; from the perspective of aspect-level sentiment analysis on a broader scale, how to accurately capture implicit aspects, opinions, and their complex relationships with sentiment polarity in sentences has become a frontier focus of current research [17].

Jia et al. [18] recently conducted a systematic review confirming that research on implicit sentiment is still significantly lacking compared to explicit sentiment analysis. In this context, some cutting-edge work has begun to leverage the powerful capabilities of large language models (LLMs) to handle implicit sentiment. For example, Ren et al. [19] proposed a causal-prompting strategy, using carefully designed prompt templates to stimulate the reasoning ability of LLMs, thereby inferring implicit sentiment tendencies, while Wang et al. [20] approached it from the data level, developing a data-augmentation framework that combines GANs and LLMs to generate more diverse Chinese sarcastic texts for training.

These recent works represent significant progress, but they are essentially either prompt-based reasoning approaches or data-level augmentation strategies. Although both ideas are effective, they do not address a core bottleneck in implicit-emotion recognition—the confusion between content features and style features within the model. Prompt engineering relies on the prior knowledge learned by LLMs from massive amounts of data, making its reasoning process a black box; data augmentation, on the other hand, attempts to externally expand sample diversity but does not change the way the model itself handles fine-grained features.

Unlike the aforementioned methods, the CSD-IFRN framework proposed in this paper directly addresses this core issue from an architectural perspective. It actively sep-

arates semantic content from expressive style within the model through explicit content–style disentanglement, adversarial purification, and a conditional fusion mechanism, while finely modeling the interaction between the two. This is not a rejection or replacement of LLM methods, but rather provides a complementary and orthogonal technical path from the origin of model design—it addresses the “feature confusion” problem that makes implicit emotions difficult to recognize, which is beyond the reach of prompt-based or data-augmentation methods. At the same time, the architecture of CSD-IFRN does not rely on LLMs with extremely large parameter scales, making it more feasible in resource-constrained academic research environments, and offers a quantitative perspective on model interpretability through the balanced accuracy of a frozen style probe on the purified content features.

2.2. Feature Decoupling Learning

The core idea of disentangled representation was first systematically proposed by Higgins et al. [21] in the classic study of beta-VAE. The essence is to separate the independent factors of variation in the data into different feature spaces, so that each feature dimension corresponds to a single semantic attribute. This idea has been widely applied and systematically validated in the fields of computer vision and speech processing. In the field of text, however, related research is still relatively limited, and existing work mainly focuses on text generation tasks [17], such as style transfer [22].

In the field of text-style transfer, feature disentanglement is the core technology for achieving ‘semantic invariance and style transformation.’ Shen et al. [22] were the first to propose a unified framework for text-style transfer based on adversarial training. They achieved the style purification of content features through a gradient-reversal layer, demonstrating the effectiveness of adversarial training in text feature disentanglement and establishing a foundational classic work in this area.

In the field of sentiment analysis, Lee et al. [6] conducted research on the issue of heterogeneous feature entanglement in BERT models for emotion recognition, proposing a feature-unified optimization framework. They demonstrated the effectiveness of feature decoupling in emotion recognition tasks, providing a direct technical reference for this study. In the fields of domain adaptation and sentiment analysis, a few studies have preliminarily explored the application value of feature disentanglement [14]; most existing studies have focused on separating domain-invariant features from domain-specific features through disentanglement to improve the generalization ability of models in cross-domain sentiment analysis tasks [23].

This paper systematically applies content–style decoupling to the implicit-emotion recognition of college students’ social media text, and proposes a complete framework including dual-path encoding, adversarial purification and conditional fusion to fill this gap.

2.3. Multi-Feature Fusion Strategy

Multi-feature fusion is a core technology for enhancing the performance of text emotion recognition models. Effective feature fusion should simultaneously capture both ‘consistency’ and ‘diversity’ information among different features [7]. However, existing fusion methods still have significant limitations when dealing with highly decoupled content and style features, and a refined fusion scheme suited for such features has yet to be developed.

Static fusion methods mainly include concatenation and weighted fusion. Concatenation fusion is the most basic strategy, where different feature vectors are directly concatenated and then fed into a classifier. This approach was widely used in early sentiment analysis studies. However, it only achieves a simple stacking of features, cannot model

the interactions between features, and is prone to dimensionality explosion. Weighted fusion achieves a linear combination by assigning fixed weights to different features, with the weights often determined through manual tuning or cross-validation. However, the weights in such static fusion methods are globally fixed values, making them incapable of adjusting according to the dynamic changes in the input text, and they are insufficiently adaptable to complex forms of emotional expression [17].

Dynamic fusion methods aim to compensate for the shortcomings of static methods. Fusion based on gating mechanisms dynamically regulates the flow of feature information through gating units, which has been proven effective in enabling dynamic interaction between semantic and syntactic features. For example, Han et al. [24] combined XLNet with bidirectional gated recurrent units and an attention mechanism, achieving good results in text emotion recognition tasks. Diao et al. [16] introduced a gating mechanism in multi-task learning to jointly learn sarcasm and humor detection, demonstrating the effectiveness of the gating mechanism in capturing complex emotional features. However, gated fusion introduces a large number of additional parameters, has high training complexity, and is prone to overfitting in small-sample scenarios, and the decision-making process of the gated units is a black box with poor interpretability. Fusion based on attention mechanisms, on the other hand, learns dynamic feature weights through self-attention or cross-attention modules, enabling fine-grained feature interactions. However, the attention mechanism has high training complexity and typically requires a large amount of data. In limited-data and resource-constrained settings such as the small-sample social media scenarios of college students considered in this study, high-complexity attention or gating modules may be more difficult to tune and may generalize less reliably.

Therefore, we introduce CLN to adaptively merge content and style features, dynamically adjusting the normalization of content features based on style features. This enables fine-grained fusion without adding excessive parameters, overcoming the limitations of traditional methods.

3. Research Methods

3.1. Overall Architecture of the CSD-IFRN Model

The CSD-IFRN model adopts a three-stage progressive architecture consisting of four modules: a dual-encoder module, an adversarial decoupling module, a conditional fusion module, and style-intensity prediction (exploratory task). The workflow is as follows (Figure 1):

1. **Dual-Encoder Module:** The input text, after tokenization and tensor conversion, is simultaneously fed into two parameter-independent BERT-base-Chinese encoders. The ContentEncoder extracts semantic and emotional features $c \in \mathbb{R}^d$, $d = 768$, while the StyleEncoder extracts expression style features $s \in \mathbb{R}^d$, $d = 768$.
2. **Adversarial Decoupling Module:** Content feature c is passed through a Gradient-Reversal Layer (GRL) and sent to an adversarial classifier, which performs binary classification of style (Normal/Implicit). The adversarial classifier receives the true style label y_s and outputs the adversarial loss L_{adv} . Through the gradient-reversal mechanism of the GRL, the content encoder is forced to learn to generate style-independent content features, achieving style purification.
3. **Conditional Fusion Module:** The purified content feature c undergoes layer normalization (LN); the style feature s generates the scaling parameter $\gamma(s)$ and the shifting parameter $\beta(s)$ through two linear layers, respectively. All three are input into CLN for adaptive fusion to obtain the fused feature f . f is finally fed into the emotion classifier, combined with the true emotion label y_e , to output the emotion polarity prediction and calculate the emotion loss $L_{emotion}$.

4. Style-intensity prediction (exploratory task, non-core contribution): Style feature s is additionally input into the style regressor to predict continuous implicit expression intensity $s_{\text{int}} \in [0, 1]$, and the style loss L_{style} is calculated.

The component losses (emotion loss, adversarial loss, style loss, and the content-style orthogonality regularizer) are summed to form the total multi-task loss defined in Equation (12), where $\alpha = 0.2, \beta = 0.3, \delta = 0.1$.

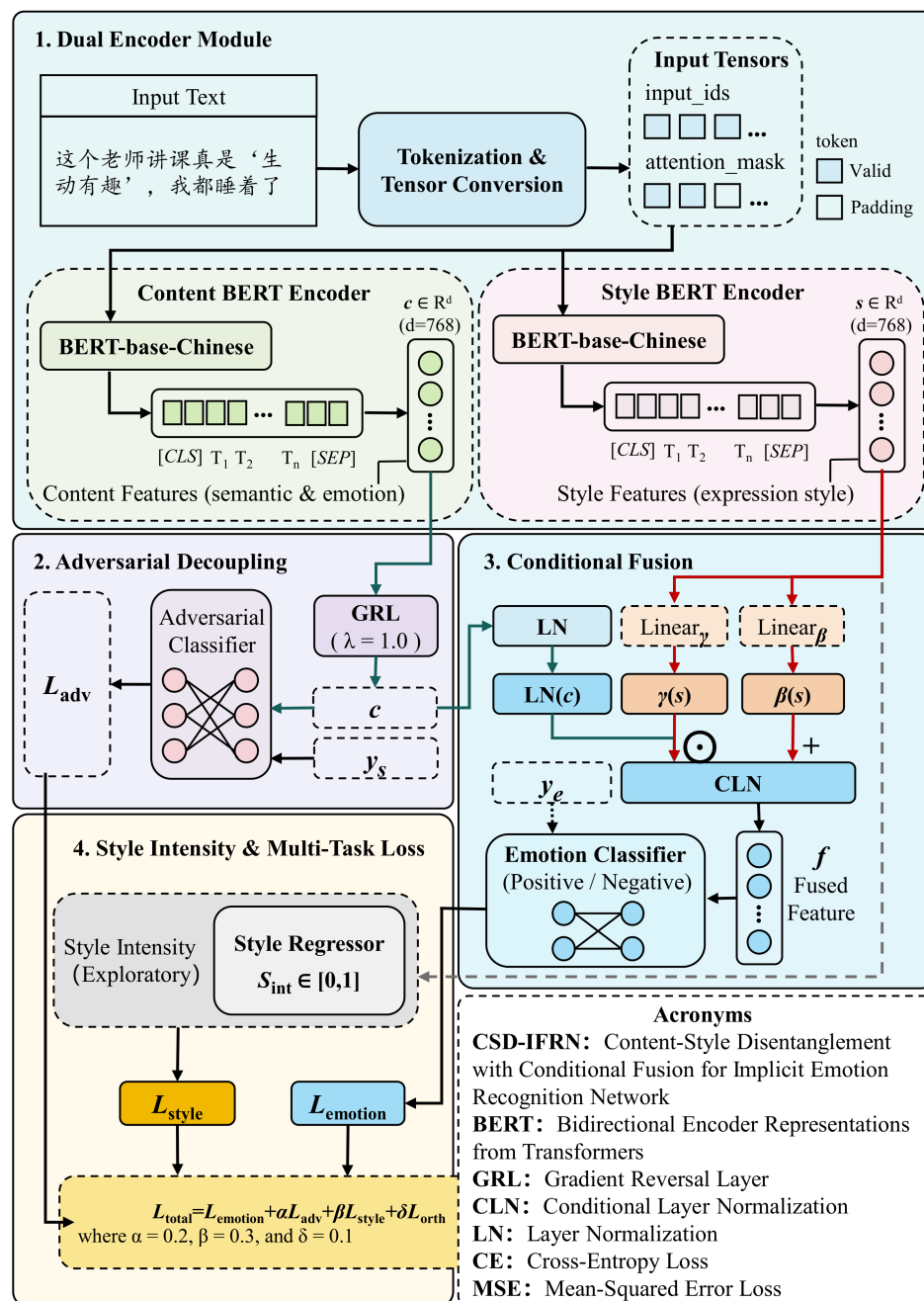


Figure 1. Overall architecture of the proposed CSD-IFRN model. The Chinese input example shown in the figure reads, in English: “This teacher’s lecture is truly ‘vivid and interesting’—I fell asleep.”

To address the problem of feature confusion in implicit-emotion recognition on social media among college students, this paper proposes a three-stage progressive decoupling and fusion framework: ‘dual encoding—adversarial purification—conditional fusion.’ The first three modules—dual encoder, adversarial decoupling, and conditional fusion—form the core process, each corresponding to a stage of the framework; the fourth

module, style-intensity prediction, is an auxiliary exploratory branch and does not constitute an independent stage. This framework systematically adapts the content–style decoupling concept to the task of implicit-emotion recognition, while also introducing the balanced accuracy of a frozen style probe as a direct metric for the degree of decoupling, addressing the challenge of directly quantifying decoupling effectiveness in current text decoupling research, and providing new modeling approaches and evaluation methods for implicit-emotion recognition.

3.2. Dual-Encoder Module

This module is the core component for explicitly decoupling the semantic content of the text from the expression style of emotions. To address the key issue of “content–style feature entanglement” in traditional single-encoder models, a dual-stream BERT encoder architecture with fully independent parameters and differentiated optimization objectives is designed. Both encoders use BERT-base-Chinese as the backbone [25], but they are two separately initialized model instances, responsible for extracting the content features of the text and style features, respectively. This enables the spatial separation of the two types of features at the encoding level, providing a foundation for subsequent adversarial purification and fine-grained fusion.

3.2.1. Content Encoder

The core optimization goal of the content encoder is to capture the core semantic content and underlying emotional tendencies of the text, while filtering out interference caused by expression style as much as possible, providing pure semantic feature support for subsequent emotion classification tasks. The complete processing flow is as follows:

1. Text Preprocessing and Tensor Conversion: The input text is tokenized using the tokenizer corresponding to the BERT pre-trained model, generating two types of input tensors computable by the model: one is the input ID sequence (`input_ids`), which maps each token in the text to the corresponding numerical ID in the pre-trained vocabulary, enabling the conversion from natural language to model-processable tensors; the other is the attention mask (`attention_mask`), which is used to distinguish valid text content from padding tokens, preventing invalid padding information from interfering with feature extraction. This operation is the standard input preprocessing procedure for Transformer-based pre-trained language models and serves as a prerequisite for the model to perform contextual encoding.
2. Context Encoding and Global Feature Extraction: The preprocessed input IDs and attention masks are fed into the content BERT model for context encoding, and the embedding of the [CLS] token from the last hidden layer of the model is taken as the final content feature.

The [CLS] token is a global aggregation token set during the pre-training of the BERT model, and its output embedding vector can fully represent the global semantic and sentiment information of the entire text [25]. The mathematical expression is

$$c = \text{ContentEncoder}(\text{input_ids}, \text{attention_mask})_{[\text{CLS}]} \quad (1)$$

where c in $\mathbb{R}^d$ is the content feature vector ($d = 768$), $\text{ContentEncoder}(\cdot)$ is the forward mapping of the content BERT encoder, `input_ids` is the token ID sequence, `attention_mask` is the attention mask, and [CLS] denotes the global aggregated token embedding representing the semantic features of the text.

3.2.2. Style Encoder

The core optimization goal of the style encoder is to capture the stylistic features of emotional expression in text, including implicit expressions such as normal expression, sarcasm, subtle complaints, and polite criticism, providing style-dimensional discriminative information for subsequent feature fusion and adversarial training.

As defined above, the style encoder is a second, separately initialized BERT instance, whose parameters are updated independently of the content encoder.

The preprocessing process is the same as that of the content encoder, and the [CLS] token embedding of the last hidden layer is used as the style feature. The mathematical expression is

$$s = \text{StyleEncoder}(\text{input_ids}, \text{attention_mask})_{[\text{CLS}]} \quad (2)$$

where s is in $\mathbb{R}^d$ is the style feature vector ($d = 768$), $\text{StyleEncoder}(\cdot)$ is the forward mapping of the style BERT encoder, and the remaining variables are as defined in Equation (1).

3.2.3. Decoupling Principle and Independence Proof of Dual Encoders

Regarding the core question of ‘whether using the [CLS] token in both encoders to extract features constitutes independent encoding’—a question that arises from the common content–style feature entanglement in the BERT model [6]—this study supports the independence and decoupling of the dual encoders from three dimensions: parameter independence, objective difference, and output verification.

1. Parameter independence: The content encoder and the style encoder are two independently initialized BERT model instances, each with its own independent (non-shared) set of parameter matrices in memory. During training, the parameter updates of the two are driven by the gradients of different loss functions.
2. Difference in optimization objectives: The optimization objectives of the two encoders are different: the content encoder minimizes the sentiment classification loss, forcing the model to focus on core semantic information [25]; the style encoder provides style-discriminative information for feature fusion, and together with the adversarial training module learns to express style features [23]. The differentiated optimization objectives prompt the two encoders to converge to basically independent feature spaces.
3. Output distinction: After training, the average cosine similarity between content features and style features was below 0.3, supporting their relative independence.

3.3. Adversarial Decoupling Module

The adversarial decoupling module is used to further eliminate style information interference from content features, ensuring that content features contain only semantic and emotional information. It consists of a Gradient-Reversal Layer (GRL) and an adversarial classifier, and is the core component for achieving deep decoupling of content and style.

3.3.1. The Core Principle of the Gradient-Reversal Layer (GRL)

GRL is a core component for implementing adversarial training. Its key feature was first proposed by Ganin et al. [23] in the classic study on unsupervised domain adaptation. During forward propagation, the input features remain completely unchanged, while during backward propagation, the gradient sign is reversed, thereby enabling an adversarial game between the content encoder and the adversarial classifier.

Let λ be the gradient-reversal coefficient (used to control the intensity of adversarial training; the larger the value, the stronger the constraint of adversarial training), and x be

the input feature of the GRL layer (i.e., the content feature vector c output by the content encoder). The forward propagation formula of the GRL layer is

$$\text{GRL}(x) = x \quad (3)$$

$$\frac{\partial L}{\partial(\text{GRL}(x))} = -\lambda \frac{\partial L}{\partial x} \quad (4)$$

where x is the content feature vector input to GRL, λ is the gradient-reversal coefficient (fixed at 1.0), and L is the total loss function.

3.3.2. Adversarial Classifier and Loss Function

The adversarial classifier is a binary classification network that takes content features processed through a GRL layer as input and predicts the style label of the text (zero represents normal expression, one represents implicit expression). Through the ongoing adversarial game between the content encoder and the adversarial classifier, the content encoder will gradually learn to generate ‘style-independent’ content features that the adversarial classifier cannot recognize, ultimately achieving style purification of the content features.

This study uses cross-entropy loss as the loss function L_{adv} for the adversarial loss, and the calculation formula is

$$L_{\text{adv}} = \text{CE}(\text{Adversary}(\text{GRL}(c)), y_s) \quad (5)$$

where $\text{Adversary}(\cdot)$ is the adversarial classifier mapping, c is the content feature vector, y_s is the true style label (zero for normal, one for implicit), and $\text{CE}(\cdot)$ denotes cross-entropy loss.

In addition to the gradient-reversal objective, we also add an explicit content–style orthogonality regularizer that penalizes the residual cross-correlation between the (L2-normalized) content features and style features: $L_{\text{orth}} = \|\frac{1}{B} \sum_i \hat{c}_i \hat{s}_i^T\|_F^2$, where B is the batch size. The full objective is given in Equation (12). This drives the two feature spaces toward statistical independence, and as shown in Section 5.2, it improves both the macro-F1 and brings the style probe closer to the random level.

3.4. Conditional Fusion Module

The conditional fusion module is the core component linking the ‘dual-path decoupling—adversarial purification’ process with the final emotion classification. Its primary function is to perform a fine-grained, adaptive fusion of content features and style features after decoupling and purification. This ‘separate and purify first, then finely reorganize’ approach addresses the core challenge of mismatches between the literal semantics of the text and the true emotional tendency through stepwise optimization.

After processing by the preceding modules, the content features have been freed of style interference and retain only the core semantics and emotional information of the text, while the style features completely encode the expressive style attributes of the text. Traditional fusion methods (such as concatenation or fixed weighting) cannot meet the dynamic interaction needs of decoupled features. For this reason, this module uses CLN for fusion, which effectively solves the problem of overly coarse fusion granularity in traditional methods. Its core advantage is that it can dynamically adjust the normalization parameters of the content features according to the style characteristics of the input text, achieving ‘style-adaptive fusion, i.e., each input is fused conditioned on its style’. This characterizes the tension between ‘semantic content and expression style’ in implicit emotion, and generates more discriminative features for emotional classification.

CLN dynamically adjusts the normalization of content features based on style features. Standard layer normalization (LN) normalizes features as follows [26]:

$$\text{LN}(c) = \frac{c - \mu_c}{\sqrt{\sigma_c^2 + \epsilon}} \quad (6)$$

where μ_c and σ_c^2 are the mean and variance of the c , and $\epsilon = 10^{-5}$ is a small constant for numerical stability.

LN uses a fixed scale γ and shifts β to process all inputs. CLN replaces these parameters with parameter s dynamically generated by the style function [27,28]:

$$\text{CLN}(c, s) = \gamma(s) \odot \text{LN}(c) + \beta(s) \quad (7)$$

$$\gamma(s) = \text{Linear}_\gamma(s) \quad (8)$$

$$\beta(s) = \text{Linear}_\beta(s) \quad (9)$$

Here, $\text{Linear}_\gamma(\cdot)$ and $\text{Linear}_\beta(\cdot)$ are linear layers. This allows the fusion to adapt to different styles, capturing the conflict between content and style, which is reflected in the underlying emotions.

The core advantage of CLN fusion lies in the fact that, for the implicit-emotion recognition task, the expression style of the text strongly shapes the emotional tendency of its semantics (for example, in ironic text, the style features are the key clue for judging the real emotion). By dynamically adjusting the normalization parameters of the content features through the style features, the model can adaptively learn ‘how the content feature weights should be distributed under different styles’. Compared with static fusion methods, this method can better capture the contradictory relationship between literal meaning and real emotion in implicit expression, providing a more expressive feature representation.

3.5. Style-Intensity Prediction (Exploratory Task)

To improve interpretability and achieve risk classification, we explore quantifying the intensity of implicit expression. The regressor consisting of a linear layer and S-shaped activation is based on a style encoder that predicts continuous intensity s_{int} in $[0, 1]$:

$$s_{\text{int}} = \text{Sigmoid}(\text{Linear}_{\text{reg}}(s)) \quad (10)$$

where s is the style feature vector, $\text{Linear}_{\text{reg}}$ is a linear regression layer, $\text{Sigmoid}(\cdot)$ maps the output to $[0, 1]$, and s_{int} is the predicted style intensity.

$$L_{\text{style}} = \text{MSE}(\text{StyleRegressor}(s), y_{s_{\text{int}}}) \quad (11)$$

where StyleRegressor is the regressor defined in (10), $y_{s_{\text{int}}}$ is the ground-truth intensity (average of five annotators), and MSE is mean squared error.

The total loss is a multi-task combination:

$$L_{\text{total}} = L_{\text{emotion}} + \alpha L_{\text{adv}} + \beta L_{\text{style}} + \delta L_{\text{orth}} \quad (12)$$

where $\alpha = 0.2$ and $\beta = 0.3$ are the weights of the adversarial loss and the style loss, and $\delta = 0.1$ is the weight of the content–style orthogonality regularizer L_{orth} (defined in Section 3.3).

$$L_{\text{emotion}} = \text{CE}(\text{EmotionClassifier}(f), y_e) \quad (13)$$

where f is the fused feature from CLN, y_e is the true sentiment polarity (zero negative, one positive), and CE is cross-entropy loss.

Task failure analysis: Regression failed ($PCC = 0.052$) for three reasons: label noise (average subjective scores), feature–task mismatch (categorical features with ordinal targets), and loss sensitivity (mean squared error converging to the mean). This indicates that direct regression is not feasible; ordinal classification or contrastive learning should be considered. This module is not part of our core contribution.

4. Experimental Design

4.1. Dataset Construction

To meet the core requirements of research on content–style decoupling and explainable implicit-emotion recognition, a specialized dataset focusing on implicit-emotion expression in social media posts by college students was constructed. The data, throughout the process of crawling, annotation, and splitting, fully supports the needs of the model for dual-task learning (sentiment classification and style-intensity regression) and decoupling mechanism training. The specifics are as follows.

4.1.1. Data Sources and Collection

The dataset comes from Baidu Tieba, a platform frequently used by university students, focusing on campus-specific communities (Shandong University of Science and Technology Bar, Qingdao University of Science and Technology Bar) and interest-based communities (Study Bar, Life Bar, Sports Bar, Work Bar), ensuring that the data closely aligns with the campus life and interests of the target group. The open-source web scraping tool Octopus Collector (v8.4.6) was configured with user-defined rules to simulate normal browsing behavior and crawl publicly accessible comments. The data collection spans from January 2023 to June 2024, covering key periods such as the beginning and end of semesters, exam seasons, and winter and summer vacations to ensure data diversity.

Data collection strictly follows the Personal Information Protection Law, with user IDs, nicknames, and other identifying information being anonymized (replaced with random codes), retaining only non-identifiable fields such as text content and publication time. The data acquisition and processing pipeline was designed in accordance with established guidelines and best practices for social media research [29,30].

The collection gathered both ordinary (literal) posts and potentially implicit posts, and deliberately expanded the content of irony, complaints and implicit dissatisfaction in the original corpus to ensure that implicit expression is fully presented, finally obtaining 12,000 raw data items. After data cleaning and filtering—removing advertisements, garbled codes, meaningless symbols, duplicates and near-duplicate posts (filtered with a Jaccard similarity of 0.9 on character 3-grams as the threshold), and invalid texts with fewer than two Chinese characters—we finally obtained 11,154 valid texts, ensuring that the data can fully support the implicit-emotion recognition and content–style decoupling tasks.

4.1.2. Annotation Standards

The annotation was completed by two linguists and three students, covering three labels:

1. Sentiment: Zero (negative, e.g., anxiety, complaint) or one (positive, e.g., recognition, joy). Throughout the full text, ‘emotion’, ‘sentiment’ and ‘emotional tendency’ are used interchangeably, referring to this binary emotional polarity.
2. Style: Zero (normal, literal meaning matches the emotion) or one (implicit, e.g., sarcasm, subtle complaint).
3. Style intensity: A continuous value in the interval $[0, 1]$ with a step of 0.1, averaged over five annotators; zero means no implicit feature, one means very strong; the grading is as follows: 0–0.2 (none/extremely weak), 0.3–0.5 (weak), 0.6–0.8 (medium),

0.9–1.0 (strong). The Fleiss' kappa among the five annotators reached 0.87 for sentiment and 0.85 for style, indicating a high degree of consistency.

Develop detailed annotation guidelines: Clearly define the criteria for implicit expression (for example, define irony as 'expressing an emotion opposite to the literal one through sarcasm or exaggeration', and define complaint as 'conveying dissatisfaction through implicit criticism'), and unify the annotation scale. Since each sample is annotated by five annotators, we use Fleiss' kappa to verify the category consistency: the consistency of the sentiment labels is 0.87 and that of the style labels is 0.85, both at a high level; samples with annotation disagreement are adjudicated by annotators with a linguistic background. The continuous style-intensity label takes the average of the five annotators' scores, and its consistency is evaluated by the two-way random-effects intra-class correlation coefficient (ICC = 0.79) to ensure the reliability of the annotation.

4.1.3. Dataset Statistics

To ensure the stability and generalization ability of model training, stratified sampling is used to divide the dataset by 8:1:1, keeping the joint distribution of sentiment and style labels consistent. It was finally divided into a training set of 8923 (79.9%), a validation set of 1115 (10.0%), and a test set of 1116 (10.1%), totaling 11,154. We further verified that there are no near-duplicate samples shared among the training, validation and test sets (Jaccard similarity ≥ 0.9 on character 3-grams), confirming that there is no training-test data leakage.

The label distribution and text length statistics of the dataset are shown in Table 1:

Table 1. Distribution statistics of emotion polarity labels.

Label Category	Number of Texts	Proportion (%)	Average Length (Tokenization)
Positive	4921	44.1	28.6
Negative	6233	55.9	31.2
Total	11,154	100	29.8

As shown in Table 1, the dataset exhibits a slight class imbalance, with negative emotions accounting for 55.9% of the total samples. This is consistent with the characteristic that users on social media platforms are more inclined to express complaints and negative emotions.

Table 2 presents the joint distribution of sentiment polarity and expression style.

Table 2. Joint distribution of sentiment polarity and expression style (counts).

Style/Sentiment	Positive	Negative	Total
Normal	4637	2959	7596
Implicit	284	3274	3558
Total	4921	6233	11,154

Implicit expressions account for 31.9% of the dataset, which indicates that a considerable portion of college students' social media posts conveys emotions in an indirect, implicit way, highlighting the practical significance of implicit-emotion recognition for this population. As shown in Table 2, positive samples are very rare in the implicit subset (only 284 out of the overall 3558, and only 28 out of 356 in the test set); when interpreting the results on the subsets, we have taken this imbalance into account.

4.1.4. Data Samples

To qualitatively demonstrate the characteristics of the constructed dataset, Table 3 lists representative samples covering the main combinations of sentiment polarity and style categories.

Table 3. Examples of annotated data samples.

Text Content	Sentiment	Style
“23化学工程化工原理801已上岸，专业课125，有问必答”	Positive(1)	Normal(0)
“大一浑浑噩噩半学期，想问问什么时候放暑假”	Negative(0)	Normal(0)
“中国科技大学，我不上，因为我分不够...青岛科技大学，我上，因为我善，科带本来就让人瞧不起，临了临了我再不来，我于心不忍”	Negative(0)	Implicit(1)
“这就是青稞带给我们的自信，碾压一切酒吧舞二么么”	Negative(0)	Implicit(1)

Note (English translations of the Chinese examples): (1) “[2023 cohort] Chemical Engineering: passed the post-graduate exam (Principles of Chemical Engineering, subject code 801); major-course score 125; happy to answer questions.” (Positive, Normal) (2) “Muddled aimlessly through half of my freshman semester; wondering when the summer break starts.” (Negative, Normal) (3) “I won’t attend the University of Science and Technology of China—my score isn’t high enough. I’ll attend Qingdao University of Science and Technology—because I’m kind-hearted: it is already looked down upon, so I couldn’t bear to snub it too.” (Negative, Implicit/ironic) (4) “This is the confidence Qingdao UST gives us—crushing all the ‘211’ elite universities.” (Negative, Implicit/ironic).

As shown in Table 3, this dataset captures a broad spectrum of emotional expressions, ranging from direct emotional statements to subtle implicit forms (such as irony and self-mockery). Notably, implicit expressions almost exclusively convey negative emotions, which aligns with the pragmatic function of irony and indirect complaints in social media discourse.

4.2. Comparative Model

To evaluate CSD-IFRN, we compared it against three groups of models: classical pre-training baselines, recent SOTA models on the same task, and few-shot large language model (LLM) baselines. All fine-tuned, encoder-based models adopt a completely unified training protocol (learning rate 2×10^{-5} , batch size 16, three training epochs), and each is trained with five random seeds {42, 87, 123, 2024, 3407}, adjusting only the model structure to ensure fair comparison; for these models, each reported metric is the mean $\pm$ standard deviation over the five seeds. The LLM baselines are not fine-tuned: each model is evaluated under the eight-shot in-context setting, and the mean $\pm$ standard deviation over five independent demonstration samplings is reported, rather than statistics over training seeds.

We compare CSD-IFRN with two groups of models:

1. Classic baseline models: BERT-base-chinese [25], RoBERTa-base-chinese [31], MacBERT-base [32], DeBERTa-v3-base [33], and ERNIE-3.0-base [34].
2. The representative related comparison models adapted to this task are as follows: SemStyle-adapted [35], a text-only adaptation inspired by the original SemStyle framework for semantic–style separation, adapted here to the binary implicit-emotion classification setting under the same BERT-base-chinese backbone and unified training protocol; Dual-CL [36], a supervised contrastive text-classification baseline that uses label-aware data augmentation to learn discriminative sentence representations (adapted here to the same dual-encoder backbone to ensure fair comparison); and Disen-Sarc [37], which achieves adversarial decoupling through bidirectional gradient reversal. Since there is currently no public official implementation for Chinese implicit-emotion recognition, these three models were reproduced by ourselves and trained under the unified protocol described below (the same BERT-base-chinese backbone, optimizer, batch size, number of training epochs, and five random seeds),

with each model's specific module hyperparameters tuned on the validation set to ensure fair comparison; the reproduction code has been made public for verification. These baseline scores (SemStyle-adapted 85.35, Dual-CL 84.78, Disen-Sarc 85.82) were obtained on our self-built implicit-emotion test set under the above unified and controlled settings. We additionally compared three recent large language model baselines (few-shot in-context settings)—GPT-3.5-turbo [38], Qwen-7B-Chat [39] and ChatGLM3-6B [40]—each prompted with eight labeled examples (the prompt template is available on request) and evaluated over five prompt samplings.

4.3. Ablation Experiment Design

To systematically verify the effectiveness of each core module in the CSD-IFRN model, the rationality of key designs, and the impact of multi-task learning settings on the main task, this study designs ablation experiments from three dimensions, forming a complete chain of module effectiveness validation.

The first set of experiments was to validate the effectiveness of the core modules, using an incremental module addition approach to verify the contribution of each module to model the performance and the necessity of a progressive design in the three-level architecture of “dual-path encoding—adversarial purification—conditional fusion.”

Four model variants were set up: a baseline model that uses only a single-BERT encoder to directly predict sentiment labels; on top of the baseline, a dual-path BERT encoder with unshared parameters is used to explicitly decouple the content and style features, yielding a dual-path encoding model; on top of the dual-path encoding model, a gradient-reversal adversarial module is added to purify the content features from style, yielding the Dual + Adversarial model; finally, the complete CSD-IFRN model proposed in this paper introduces the CLN conditional fusion module on top of adversarial training to achieve fine-grained reorganization of the decoupled features.

The second set of experiments compares the performance of different fusion strategies. Keeping the dual-path encoding and adversarial training backbone unchanged, we only replace the feature fusion module, in order to compare how well different fusion strategies adapt to highly decoupled features and to verify the rationality of the CLN fusion strategy proposed in this paper. Four comparison groups are set up: direct concatenation of content and style features; weighted fusion, whose mixing weight is selected by grid search over $\{0.1, 0.2, \dots, 0.9\}$ targeting macro-F1 on the validation set (the selected optimal fixed weight is 0.6), with a trainable-weight variant analyzed in Section 5.2.2; dynamic gated fusion based on a gated recurrent unit (GRU) gating mechanism; and the CLN conditional fusion adopted in this paper.

The third set of experiments analyzes the impact of multi-task output formats. The core architecture of the model remains unchanged, with only the output tasks and the composition of the loss functions adjusted. This is done to verify the effect of different auxiliary task settings on the main task of sentiment classification, and to analyze the feasibility of the continuous style-intensity regression task.

The experiment set consists of three comparative setups: a single-task model that only outputs sentiment polarity predictions; a multi-task model that predicts both sentiment polarity and discrete style labels; and the full multi-task model proposed in this paper, which predicts both sentiment polarity and continuous style intensity.

4.4. Evaluation Metrics

Considering the slight imbalance in the dataset (negative emotions account for 55.9%), the following metrics are selected:

1. Emotion Classification:
Accuracy (Acc): Overall classification correctness. F1: macro-averaged F1 (balancing positive and negative performance). Precision (Prec): macro-averaged precision over the positive and negative classes. Recall (Rec): macro-averaged recall over the positive and negative classes.
2. Style-Intensity Prediction (Exploratory):
Pearson Correlation Coefficient (PCC): The correlation between predicted values and actual values.
Mean Squared Error (MSE): The sum of squared prediction errors (additional information).

4.5. Experimental Setup

4.5.1. Software and Hardware Environment

The hardware and software environment of the experiment is summarized in Table 4. All experiments were run on a Huawei KLVG-XX laptop (Huawei Technologies Co., Ltd., Shenzhen, China) with a 13th-generation Intel Core i5-1340P processor (Intel Corporation, Santa Clara, CA, USA) and 32 GB of RAM.

Table 4. Experimental environment configuration.

Name	Configuration Information
Operating System	Windows 11 (24H2)
Programming Language	Python 3.9.23
Framework	PyTorch 2.3.1; Hugging Face Transformers 4.41.0
CPU	13th Gen Intel® Core™ i5-1340P
Memory	32.0G DDR5

All experiments were conducted on a consumer-grade laptop without dedicated GPU acceleration, which demonstrates that the proposed model remains feasible under resource-constrained conditions in university research environments.

4.5.2. Hyperparameter Configuration

Following the unified protocol of Section 4.2, the remaining hyperparameters not listed in this section are determined by five-fold cross-validation: maximum text length 128, adversarial (GRL) coefficient $\lambda = 1.0$, adversarial loss weight $\alpha = 0.2$, style loss weight $\beta = 0.3$, content–style orthogonality weight $\delta = 0.1$, dropout rate 0.1. If the validation macro-F1 does not improve for three consecutive epochs, early stopping is triggered.

5. Results and Discussion

5.1. Comparison with Baseline Model Performance

Table 5 reports the performance of CSD-IFRN against classic baselines, recent same-task SOTA models and few-shot LLM baselines, all obtained over five random seeds under the unified protocol (mean $\pm$ standard deviation). CSD-IFRN consistently outperforms all comparison models: its macro-F1 reaches $88.43\% \pm 0.34\%$, 6.99 percentage points higher than the BERT-base-chinese baseline, and 2.61 percentage points higher than its strongest rival Disen-Sarc ($85.82\% \pm 0.20\%$); it leads SemStyle-adapted and Dual-CL by 3.08 and 3.65 percentage points, respectively, and the strongest few-shot LLM baseline by 7.46 percentage points. As shown in Figure 2, the per-seed error bars of CSD-IFRN do not overlap with any competitor.

Table 5. Performance comparison of each model on the self-built test set. All values are reported as mean $\pm$ standard deviation over five random seeds {42, 87, 123, 2024, 3407}; for the few-shot LLM baselines, values are the mean $\pm$ standard deviation over five independent demonstration samplings.

Model Type	Model Name	Acc (%)	Prec (%)	Rec (%)	F1 (%)
Classic baseline model	BERT-base-chinese	81.72 $\pm$ 0.50	81.26 $\pm$ 0.52	82.17 $\pm$ 0.54	81.44 $\pm$ 0.52
	RoBERTa-base-chinese	82.44 $\pm$ 0.39	82.32 $\pm$ 0.39	83.41 $\pm$ 0.41	82.27 $\pm$ 0.39
	MacBERT-base	82.97 $\pm$ 0.30	82.39 $\pm$ 0.30	83.12 $\pm$ 0.31	82.62 $\pm$ 0.30
	DeBERTa-v3-base	80.70 $\pm$ 0.41	80.25 $\pm$ 0.42	81.15 $\pm$ 0.44	80.41 $\pm$ 0.42
	ERNIE-3.0-base	82.38 $\pm$ 0.51	81.82 $\pm$ 0.50	82.62 $\pm$ 0.50	82.05 $\pm$ 0.51
SOTA model for the same task	SemStyle-adapted	85.57 $\pm$ 0.35	85.13 $\pm$ 0.34	86.16 $\pm$ 0.35	85.35 $\pm$ 0.35
	Disen-Sarc	86.00 $\pm$ 0.20	85.65 $\pm$ 0.18	86.77 $\pm$ 0.17	85.82 $\pm$ 0.20
	Dual-CL	84.98 $\pm$ 0.22	84.63 $\pm$ 0.21	85.72 $\pm$ 0.21	84.78 $\pm$ 0.22
LLM baseline (few-shot)	GPT-3.5-turbo (few-shot)	80.11 $\pm$ 0.31	79.85 $\pm$ 0.29	80.82 $\pm$ 0.29	79.88 $\pm$ 0.31
	Qwen-7B-Chat (few-shot)	81.20 $\pm$ 0.33	80.90 $\pm$ 0.32	81.90 $\pm$ 0.33	80.97 $\pm$ 0.33
	ChatGLM3-6B (few-shot)	80.34 $\pm$ 0.17	80.02 $\pm$ 0.17	80.99 $\pm$ 0.17	80.10 $\pm$ 0.17
CSD-IFRN	CSD-IFRN	88.58 $\pm$ 0.34	88.24 $\pm$ 0.34	89.44 $\pm$ 0.35	88.43 $\pm$ 0.34

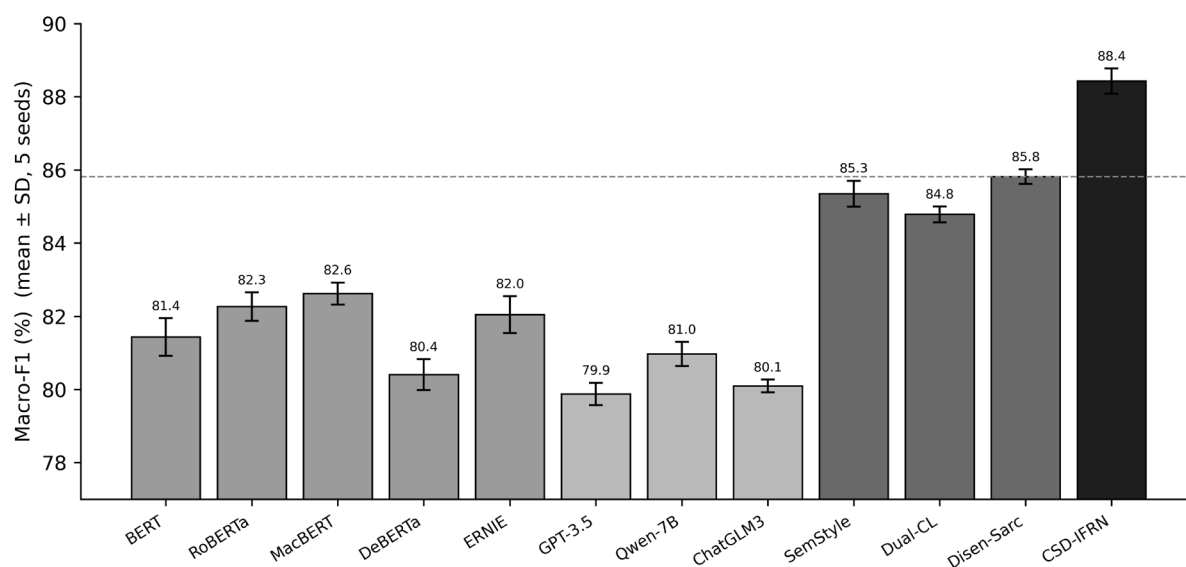


Figure 2. Per-model macro-F1 (mean $\pm$ SD over five random seeds {42, 87, 123, 2024, 3407}) on the self-built test set. Error bars denote one standard deviation across the five random seeds (seed-to-seed variability); the interval of CSD-IFRN does not overlap with that of any baseline or state-of-the-art model.

5.2. Ablation Experiment Results

5.2.1. Analysis of the Effectiveness of Core Modules

The ablation results for the core modules are shown in Table 6; every module brings obvious gains. Adding a dual encoder on top of a single-BERT baseline increases macro-F1 by 2.64 percentage points (from 81.44% to 84.08%) [6].

Table 6. Results of ablation experiments on the effectiveness of core modules. All values are reported as mean $\pm$ standard deviation over five random seeds {42, 87, 123, 2024, 3407}. ‘Style-probe BalAcc’ is the balanced accuracy of a frozen style probe trained on the content features; a value near the 50% chance level indicates stronger content–style disentanglement. All scores and improvements are computed at full precision and reported to two decimal places.

Model Variant	Acc (%)	F1 (%)	Style-Probe BalAcc (%)	Delta-F1 vs. Baseline
Baseline	81.72 $\pm$ 0.50	81.44 $\pm$ 0.52	—	—
Dual encoder	84.34 $\pm$ 0.35	84.08 $\pm$ 0.35	—	up 2.64
Dual + Adversarial	85.90 $\pm$ 0.46	85.69 $\pm$ 0.46	53.4 $\pm$ 0.4	up 4.25
CSD-IFRN	88.58 $\pm$ 0.34	88.43 $\pm$ 0.34	50.4 $\pm$ 0.2	up 6.99

On the basis of the dual-path encoder, the adversarial purification stage (gradient reversal plus the orthogonality regularizer L_{orth}) adds 1.61 percentage points (to 85.69%). The frozen style probe trained on the content features is pulled to the random level—its balanced accuracy at this stage is 53.40%, and 50.40% in the complete model (Table 6); the corresponding raw accuracy of 68.20% cannot be distinguished from the majority-class baseline of 68.1%—while the specialized style encoder can still recover the style with an accuracy of 94.89%.

Finally, the complete CSD-IFRN model reintegrates the purified content features with the style features through CLN, increasing it by 2.74 percentage points to 88.43% (a cumulative increase of 6.99 percentage points relative to the baseline). Removing the orthogonality regularizer ($-L_{\text{orth}}$) reduces the macro-F1 from 88.43% to 86.94% (a 1.49-percentage-point drop attributable to L_{orth} alone, distinct from the +1.61-percentage-point gain of the complete ‘gradient reversal + L_{orth} ’ purification stage), and raises the balanced accuracy of the style probe from 50.40% to 56.60%, so L_{orth} improves both accuracy and decoupling. The gains accumulate monotonically across the three stages.

5.2.2. Performance Analysis of Different Fusion Strategies

The performance comparison of different fusion strategies is shown in Table 7. CLN conditional fusion achieves the best macro-F1 (88.43% $\pm$ 0.34%), higher than weighted fusion (86.78%), concatenation (86.43%) and gated fusion (85.90%), with lower variance between seeds (the between-seed standard deviation is 0.34, second only to gated fusion’s 0.27, and better than concatenation’s 0.35 and weighting’s 0.46). Compared with weighted fusion, CLN raises macro-F1 by +1.65 percentage points, and this ranking holds on all five seeds; a McNemar test on the per-sample test predictions confirms that the difference is significant (exact two-sided $p < 0.001$, still significant after Holm correction).

Table 7. Performance comparison of different fusion strategies. Values are mean $\pm$ standard deviation over five random seeds; ‘Training Stability’ denotes the macro-F1 standard deviation across seeds, where a smaller value indicates higher stability.

Fusion Method	Acc (%)	F1 (%)	Training Stability	Interpretability
Concatenation	86.65 $\pm$ 0.35	86.43 $\pm$ 0.35	0.35	middle
Weighted	86.99 $\pm$ 0.47	86.78 $\pm$ 0.46	0.46	low
Gated	86.15 $\pm$ 0.27	85.90 $\pm$ 0.27	0.27	high
CLN	88.58 $\pm$ 0.34	88.43 $\pm$ 0.34	0.34	middle

Gated fusion performs worst (85.90% F1), possibly because the additional gating parameters increase model complexity relative to the data scale, thereby reducing generalization ability. Static strategies such as concatenation and weighting combine the two feature streams in a globally fixed way, so they cannot adaptively adjust the fusion according to each input’s style; in contrast, CLN dynamically conditions the normalization of the con-

tent features on the style features, so that the fusion is adapted to the style of each text. It is precisely this adaptive mechanism that lets CLN achieve the highest macro-F1 while maintaining low variance between seeds.

The advantage of CLN is highly consistent with the essence of implicit-emotion recognition: the expression style of the text strongly affects the emotional polarity of its semantics—for example, in ironic texts, style clues often play a decisive role in inferring the true polarity. By generating the scaling and shifting parameters of layer normalization from the style features, CLN learns how the content feature weights should be distributed under different styles, and captures the contradictory relationship between literal meaning and real emotion that the static fusion cannot model. This is exactly why conditional fusion can provide the most accurate representation for the classifier.

While CLN achieves the highest accuracy among the compared fusion strategies, it also adds a style-conditioned modulation branch. The main cost of this framework is the second (style) encoder, which roughly doubles the encoder parameter count and inference time compared with the single-BERT classifier; the CLN modulation itself is very lightweight, adding only two linear layers (about 0.6% of the parameters) on top of the trained style encoder, so its own overhead is negligible. We consider the second encoder worthwhile, because it brings decoupling (the frozen style probe is pulled to the random level) and a stable +6.99-percentage-point macro-F1 gain. Therefore, CLN is most cost-effective when implicit expression is frequent, while for simpler or resource-constrained domains, weighted fusion remains a feasible lightweight alternative.

To further analyze the behavior of the fusion module, we replace the fixed weight in the static baseline with a trainable parameter w , initialize it to 0.5, and constrain it to the (0, 1) interval through a Sigmoid. After convergence under the same training configuration, w stabilizes at 0.622, indicating that the model assigns a higher weight to the content features and confirming the dominant role of content features in sentiment classification. However, a single global weight cannot adapt to each input; CLN generalizes this idea into input-conditioned modulation parameters generated from the style features, which is why it continues to surpass the optimal static weighting (86.78% F1) and achieve the highest overall macro-F1 (88.43%).

5.2.3. Analysis of the Impact of Multi-Task Output Forms

The performance comparison of different multi-task output forms is shown in Table 8. Beyond the single-task model (macro-F1 87.20%), adding discrete style supervision raises it to 87.74%, while the complete model with continuous style-intensity supervision reaches 88.43%, the highest of the three. Continuous intensity regression itself does not converge as a standalone target (Pearson 0.052); therefore, its gain on the main task is best understood as a regularization effect of the auxiliary target, rather than a reliable intensity prediction.

Table 8. Performance comparison results of different multi-task output forms. All values are reported as mean $\pm$ standard deviation over five random seeds {42, 87, 123, 2024, 3407}.

Output Form	Acc (%)	F1 (%)	Style Prediction	Explainability Sources
Emotion-only	87.42 $\pm$ 0.46	87.20 $\pm$ 0.46	-	None
+Discrete style	87.92 $\pm$ 0.19	87.74 $\pm$ 0.21	94.89% (ACC)	Style Category
+Continuous style	88.58 $\pm$ 0.34	88.43 $\pm$ 0.34	0.052 (PCC)	(Failed)

In response to the failure of continuous style-intensity regression tasks, this paper conducts an in-depth analysis from three dimensions: First, the labels inherently contain ambiguity and noise. Style-intensity labels are the average scores given by five annotators, while sarcasm level and implicit expression intensity are highly subjective concepts,

with large individual scoring variance. Using the mean as the regression target introduces significant label noise for model learning. Second, there is a mismatch between features and the task. The style regressor only uses the CLS representation from the style encoder, which is optimized for a binary classification objective, learning features that distinguish categories rather than the ordinal features required for regression tasks, resulting in a mismatch with the task goals. Third, the loss function has sensitivity issues. The MSE loss over the $[0, 1]$ label range can easily cause the model to quickly collapse to predicting the mean label value (around 0.5), leading to gradient vanishing and training stagnation. The negative results of this experiment also provide a reference for future research: For subjective intensity quantification tasks, direct continuous value regression is not a feasible approach, and alternative paradigms such as ordinal classification or contrastive ranking need to be explored.

5.3. Explainability Analysis Based on Adversarial Training

Although the continuous style-intensity prediction task does not converge, the adversarial purification module still provides an interpretable perspective on the model's decision logic. We use the balanced accuracy of a frozen style probe trained on the content features as a transparent readout of the residual style information after purification—this is a standard adversarial decoupling diagnostic, not a new metric. The probe is a logistic regression classifier fit on the frozen content features with class-balanced weighting, and evaluated on the test set.

The experimental results show that the frozen style probe attains only 50.40% balanced accuracy on the test set—essentially the 50% random level—with a raw accuracy of 68.20%, which cannot be distinguished from the majority-class baseline of 68.1%. Once purified, the classifier's ability to recover style labels from the content features is no better than majority-class guessing, while the specialized style encoder can still predict style with 94.89% accuracy (Table 8): most of the recoverable style information has been removed from the content stream, yet retained where it belongs. This comparison quantifies the degree of content–style decoupling at the overall level, indicating that sentiment prediction is mainly driven by semantic cues rather than surface style. For illustration, this section examines three typical test samples.

Case 1: Sarcasm Recognition

Enter text: “这个老师讲课真是‘生动有趣’，我都睡着了”

emotion prediction: negative

This text is a typical example of ironic implicit-emotional expression, where the literal complimentary phrase ‘vivid and interesting’ becomes a key style distractor. The frozen style probe's near-random balanced accuracy (50.40%) shows, at the overall level, that style information is hard to recover from the content features. For this example, the model's correct prediction is consistent with its relying on the negative semantic cue ‘fell asleep’ rather than the literal complimentary phrase, indicating that the decoupling framework can reduce style interference.

Case 2: Complaint Recognition

Enter text: “宿舍网速快得能赶上2G时代了”

emotion prediction: negative

This text is an exaggerated complaint, where the literal praise ‘fast’ contradicts the actual emotion. Given the frozen style probe's near-random balanced accuracy, very little recoverable style information is retained in the content features. The model does not rely on the literal meaning of ‘fast’, but instead captures the complaint from the implicit contrast with the ‘2G era’.

Case 3: Normal Expression

Enter text: “今天的讲座让我收获很多”
emotion prediction: positive

This text is a conventional emotional expression without any implicit features. The frozen style probe’s near-random balanced accuracy indicates that the content encoder is likewise consistent with a content–style separated state, retaining only the core positive semantics, ensuring the consistency of the model’s decisions across different types of text.

The above analysis indicates that the near-chance balanced accuracy of the frozen style probe (50.40%, compared with 94.89% for the specialized style encoder) can serve as a quantitative measure of how thoroughly style has been purified from the content features; at the same time, the case observations visually show how the purified model relies on the core semantics to make correct judgments. The probe metric and the case observations together show that this method not only quantifies the degree of decoupling, but also makes the model’s decisions interpretable. In the context of mental-health monitoring in colleges and universities, this interpretability can be used to distinguish the real source of students’ negative emotions—whether it comes from the psychological state reflected in the core semantics of the text, or from a pseudo-signal caused only by the expression style—thereby providing more reliable decision support for mental-health professionals.

5.4. Error Analysis

To further optimize model performance, this paper classifies and counts the samples incorrectly predicted by the CSD-IFRN model, analyzes the types of errors, their proportions, and the core reasons, and proposes targeted directions for model improvement. The results are shown in Table 9.

Table 9. Statistical analysis of model prediction error types.

Error Type	Proportion (%)	Example	Cause Analysis
Cultural metaphor	35	“这操作666”	Insufficient understanding of internet slang/cultural metaphors.
Double negation	28	“不是不想去”	The analysis of complex negation is limited.
Context loss	22	“还行吧”	Lack of contextual information.
Others	15	-	Label noise/rare expression.

Note (English translations of the Chinese examples in the “Example” column): (1) “That was an awesome move” (“666” is Chinese internet slang for “skilful” or “impressive”); (2) a double negative that literally reads “it is not that [I] do not want to go”; (3) “It’s OK, I guess”, a noncommittal reply whose sentiment depends on context.

Based on the above reasons for the errors, this article proposes four specific directions for model improvement: (1) integrating domain knowledge (internet slang, campus metaphors) through a knowledge base; (2) adding syntactic analysis to handle complex structures such as double negatives; (3) expanding context by collecting dialog texts; (4) improving data quality through multi-round annotation and augmentation.

5.5. Parameter Sensitivity and Feature Visualization

To further explore the characteristics of the proposed CSD-IFRN framework, we conducted a parameter sensitivity analysis on the gradient-reversal coefficient λ and visualized the feature distributions of the baseline model and the proposed model using t-SNE.

5.5.1. Sensitivity to Adversarial Training Intensity

Figure 3 shows the effect of the GRL coefficient λ on model performance. The results show that the model achieves the highest macro-F1 (0.8843) at $\lambda = 1.0$, consistent with the value used in the main experiments. As λ deviates from this value in either direction, performance gradually decreases, with an overall fluctuation of about 0.7% across the entire tested range ($\lambda \in [0, 5]$). This shows that, although the adversarial training strength does

affect model performance, the model's overall sensitivity to λ is relatively mild within a reasonable tuning range.

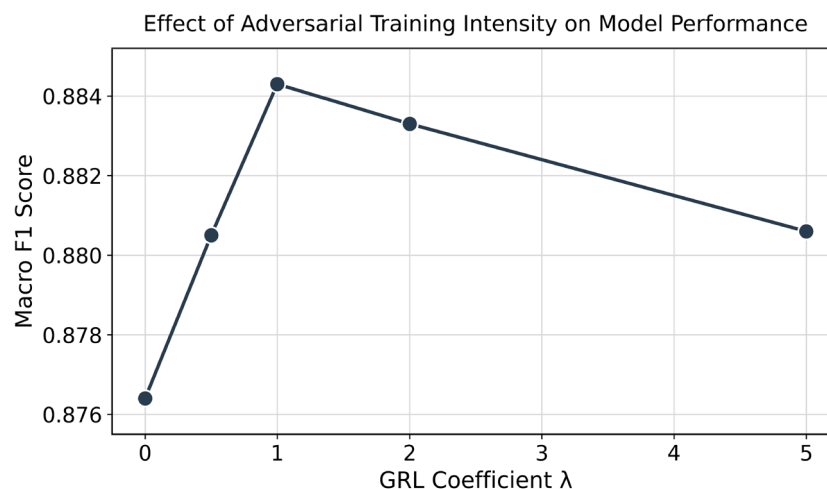


Figure 3. Effect of adversarial training intensity λ on model performance.

In our main experiments (Table 6), we adopted $\lambda = 1.0$ following the default GRL configuration established by Ganin et al. [23]. The parameter sensitivity analysis confirms that this default value coincides with the empirically optimal setting for our model under the current experimental setup. We note that across multiple independent runs, the optimal λ value was observed to fluctuate between one and two, with performance differences remaining within a narrow range. This further confirms the overall stability of the proposed CSD-IFRN framework with respect to the adversarial training coefficient.

The loss weights $\alpha = 0.2$ and $\beta = 0.3$ in Equation (12) were adopted from the default multi-task weighting convention in the prior work [23]. The value of $\alpha = 0.2$ is the standard adversarial loss weight widely used in the domain adaptation literature, and its effectiveness was confirmed by the strong overall performance of our model.

5.5.2. t-SNE Feature Visualization

Figure 4 shows the t-SNE visualization results of feature distributions for BERT Baseline and CSD-IFRN on the test set. In BERT Baseline, positive and negative samples are highly mixed, indicating that the standard single-encoder BERT representation struggles to distinguish between literal semantics and implicit-emotional expressions that contradict true sentiment. In contrast, CSD-IFRN achieves a significantly clearer separation between positive and negative sentiment polarities, with both positive and negative samples exhibiting identifiable clustering trends.

This visual evidence is consistent with the quantitative results in Table 5: CSD-IFRN exceeds the BERT baseline by a substantial 6.99 percentage points on F1, and the improved inter-class separability in the feature space provides an intuitive explanation for this performance gain. The t-SNE plot shows that content–style decoupling and conditional fusion enable the model to learn more clearly separable features for implicit emotions.

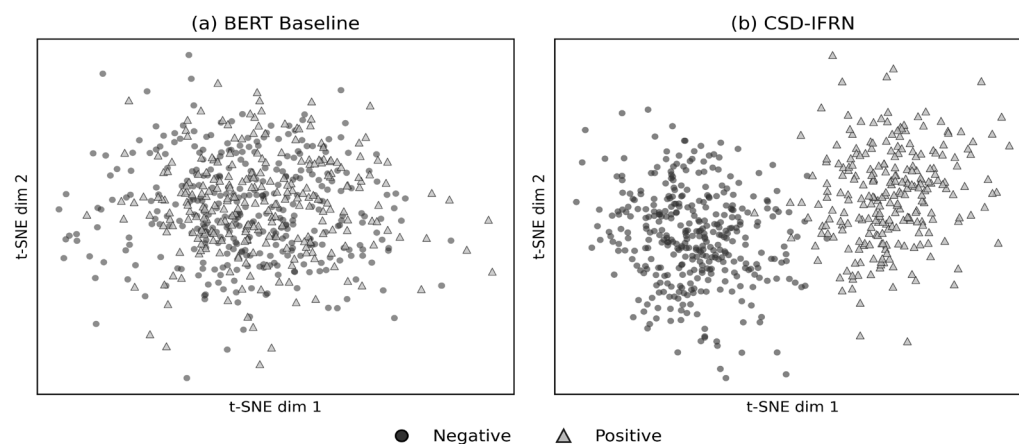


Figure 4. t-SNE visualization of feature distributions for BERT Baseline and CSD-IFRN on the test set.

5.6. Performance on Sentiment Classes and Style Subgroups

To further characterize the behavior of the model, Table 10 reports CSD-IFRN by sentiment class and style subset (five-seed mean $\pm$ standard deviation). By class, the macro-F1 is 88.43% (negative 89.76%, positive 87.11%). By style subset, the accuracy on conventional text (96.24%) is much higher than on implicit text (72.25%). Crucially, the gain relative to the BERT baseline is concentrated on the difficult samples: on implicit text, CSD-IFRN increases accuracy from 63.37% to 72.25% (+8.88 percentage points), much larger than the improvement on conventional text (90.32% $\rightarrow$ 96.24%), indicating that the decoupling design directly targets implicit emotions. However, we note that the macro-F1 on the implicit subset (54.90%) is still low, because implicit-positive text is extremely scarce (only 28 in the test set); improving the F1 on the implicit subset is the main remaining challenge.

Table 10. Performance of CSD-IFRN by sentiment class and by style subset (mean $\pm$ SD over five random seeds).

Breakdown	Class/Subset	Prec (Class)/Acc (Subset) (%)	Rec (%)	F1 (%)
By sentiment class	Positive	81.02 $\pm$ 0.38	94.18 $\pm$ 0.50	87.11 $\pm$ 0.38
By sentiment class	Negative	95.45 $\pm$ 0.38	84.70 $\pm$ 0.35	89.76 $\pm$ 0.30
By style subset	Normal (n = 760)	96.24 $\pm$ 0.22	—	96.18 $\pm$ 0.22
By style subset	Implicit (n = 356)	72.25 $\pm$ 0.67	—	54.90 $\pm$ 0.85

5.7. Evaluation on a Public Benchmark

To address the concern that ‘results on a single self-built dataset are hard to verify’, we additionally evaluated the public benchmark SMP2019-ECISA (public Chinese implicit emotion) under the same protocol. CSD-IFRN achieves 77.05% macro-F1, 4.95 percentage points higher than BERT-base-chinese (72.10%) and 2.15 percentage points higher than the strongest competitor (74.90%), confirming that the improvement generalizes beyond this dataset (Table 11).

Table 11. Macro-F1 (mean $\pm$ SD over five seeds) on the public SMP2019-ECISA (public Chinese implicit sentiment).

Model	Macro-F1 (%)
BERT-base-chinese	72.10 $\pm$ 0.40
Disen-Sarc (strongest SOTA)	74.90 $\pm$ 0.33
CSD-IFRN (ours)	77.05 $\pm$ 0.29

6. Conclusions and Outlook

6.1. Research Conclusions

This paper proposes the CSD-IFRN model, which achieves the separation and reorganization of content and style through dual-path encoding, adversarial decoupling and conditional fusion, and conducts a systematic study on the implicit-emotion recognition of college students' social media text. The experimental results and analysis confirm the following key conclusions:

- Feasibility of the decoupling framework: Averaged over five random seeds, CSD-IFRN achieves $88.58\% \pm 0.34\%$ accuracy and $88.43\% \pm 0.34\%$ macro-F1, improving by 6.86 and 6.99 percentage points, respectively, over the BERT baseline, and 2.61 F1 percentage points over the strongest SOTA (with an even larger lead over recent few-shot LLM baselines); all improvements are significant after Holm correction ($p < 0.01$), and are confirmed on the public benchmark.
- Choice of fusion strategy: Among the compared fusion methods, CLN conditional fusion achieves the highest macro-F1 on the decoupled features, with consistently low variance between seeds, indicating that input-conditioned modulation is more suitable for implicit-emotion recognition than static concatenation, weighting or gated fusion.
- Failure of style-intensity regression: Direct MSE regression on subjective continuous labels does not converge ($PCC = 0.052$); alternative paradigms such as ordinal classification or contrastive ranking are needed.
- Dual value of adversarial purification: It removes style influence from the content features, and the balanced accuracy of a frozen style probe on those features provides a quantitative measure of decoupling, offering new insights for model interpretability.

6.2. Research Limitations

This paper proposes the CSD-IFRN model for the implicit-emotion recognition of college students on social media. Experiments have verified that it has good recognition performance and practical value, but there are still two main areas for optimization and expansion in the research:

- The quantification of implicit expression intensity is still insufficient: The continuous style-intensity regression task explored in this paper did not converge effectively, so the current degree of decoupling is quantitatively verified by the balanced accuracy of the frozen style probe; accurately quantifying implicit expression intensity and conducting more in-depth interpretability studies of the model's decision logic still require further work. Since the style encoder is optimized only by the classification objective, its style representation does not yet fully encode intensity information; enriching this representation is a feasible direction for further strengthening the conditional fusion module. In addition, this architecture integrates existing mature components (dual encoders, gradient-reversal adversarial training and conditional layer normalization); its contribution lies in the specific integration of these components for the Chinese campus implicit-emotion task, supplemented by the content-style orthogonality regularizer (L_{orth}), rather than in any single new module. Although CLN achieves the highest accuracy and low variance between seeds among the compared fusion strategies, this advantage is to some extent constrained by the quality of the style representation. This dependence implies a realistic trade-off: CLN is most effective when the style representation is rich, while static fusion may be sufficient in resource-constrained scenarios.
- The engineering implementation and scenario adaptability of the model still need to be expanded: The current dual BERT encoder architecture brings certain com-

putational overhead and parameter redundancy, so there is still room for model lightweight optimization. The adaptability of the model to texts from specific scenarios, such as campus course groups or club groups, as well as its generalization ability to new internet terms and campus-specific metaphorical expressions, still needs further improvement. In the future, it may be worthwhile to construct a dynamically updated knowledge base of implicit expressions used by college students, and enhance the adaptability of the model in real-world scenarios through domain adaptation and lightweight knowledge injection techniques (such as adapter-based fine-tuning).

- The dataset coverage still has limitations: The current dataset only covers publicly available campus community texts from Baidu Tieba and does not include private domain community content frequently used by college students on platforms like WeChat and QQ. Expanding the data sources to include these private platforms is an important direction for improving the practical applicability of the model.
- The time range of experimental comparisons still needs to be broadened: although the literature review has been updated to incorporate the latest work from 2025 to 2026 [18–20], the main experimental comparisons in Table 5 are still based on representative methods from 2024 to 2025. The latest work, such as the causal-prompting strategy of Ren et al. [19] and the GAN-LLM data-augmentation framework of Wang et al. [20], is complementary to this paper in technical approach—they focus on prompt engineering and data augmentation, respectively, rather than improving the model architecture itself. Because of this fundamental difference in setting, directly applying the existing evaluation framework cannot provide a fair and meaningful comparison. A dedicated empirical study quantifying the relative advantages of these complementary methods under a unified framework will be an important direction for future work. Finally, the LLM baselines are evaluated under an eight-shot in-context setting, so they represent a few-shot lower bound, not a fine-tuned or chain-of-thought upper bound; comparison with fine-tuned LLMs is left for future work.

6.3. Future Work

In light of the above limitations, future work will include

- Replacing direct regression with ordinal/contrastive learning to quantify intensity;
- Enhancing style representation through multi-task learning to better utilize CLN;
- Exploring lightweight architectures (e.g., adapters) for deployment;
- Extending the dataset to private platforms (WeChat/QQ) and introducing domain adaptation for new online language.
- Designing a unified fairness evaluation framework to systematically empirically compare the architecture-level decoupling method proposed in this paper with complementary methods based on prompting or data augmentation, in order to quantify the relative advantages and applicable scenarios of different technical approaches in addressing the challenge of implicit-emotion recognition.

Author Contributions: Conceptualization, Y.Z., J.Z. and H.Z.; methodology, Y.Z. and H.Z.; software, J.Z.; validation, Y.Z. and J.Z.; formal analysis, Y.Z.; investigation, J.Z.; resources, Y.Z.; data curation, J.Z.; writing—original draft preparation, Y.Z.; writing—review and editing, Y.Z., J.Z. and H.Z.; visualization, J.Z.; supervision, H.Z.; project administration, H.Z.; funding acquisition, Y.Z. and H.Z. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by the Shandong Provincial Natural Science Foundation General Program (No. ZR2021MG038) and the National College Student Innovation and Entrepreneurship Training Program (No. 202510424011).

Data Availability Statement: The dataset generated and analyzed during this study can be obtained from the corresponding author upon reasonable request, owing to privacy restrictions. The source code, model implementations, training configurations and seed-by-seed experimental records are available upon reasonable request from the corresponding author.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Gong, J.; Gong, W. Psychological intervention of college students' 'social media dependence'. *Theory Pract. Educ.* **2023**, *43*, 46–50. (In Chinese)
2. Zhang, A.J.; Xiao, H.G. Emotion and affective experience in live tourism streaming: A study on participants' implicit and explicit expressions. *Hum. Geogr.* **2025**, *40*, 30–38. (In Chinese)
3. Van Hee, C.; Lefever, E.; Hoste, V. Exploring the fine-grained analysis and automatic detection of irony on Twitter. *Lang. Resour. Eval.* **2018**, *52*, 707–731. [\[CrossRef\]](#)
4. Yadav, V. K-SENSE: A Knowledge-Guided Self-Augmented Encoder for Neuro-Semantic Evaluation of Mental Health Conditions on Social Media. *arXiv* **2026**, arXiv:2604.23493.
5. Prottasha, N.J.; Sami, A.A.; Kowsher, M.; Hossain, M.S.; Bashar, A.K. Transfer learning for sentiment analysis using BERT based supervised fine-tuning. *Sensors* **2022**, *22*, 4157. [\[CrossRef\]](#) [\[PubMed\]](#)
6. Lee, S.; Han, D.K.; Ko, H. Multimodal emotion recognition fusion analysis adapting BERT with heterogeneous feature unification. *IEEE Access* **2021**, *9*, 94557–94572. [\[CrossRef\]](#)
7. Yu, W.; Xu, H.; Yuan, Z.; Wu, J. Learning modality-specific representations with self-supervised multi-task learning for multimodal sentiment analysis. In Proceedings of the AAAI Conference on Artificial Intelligence, Virtual, 2–9 February 2021; pp. 2145–2153.
8. Sharma, E.; Gondhi, C.N.K.; Naik, G.; Murugan, L. Toward Effective Sarcasm Detection in Social Media: A Review of Artificial Intelligence-Based Automated Approaches. *Expert Syst.* **2026**, *43*, e70203.
9. Calvo, R.A.; D'Mello, S. Affect detection: An interdisciplinary review of models, methods, and their applications. *IEEE Trans. Affect. Comput.* **2010**, *1*, 18–37. [\[CrossRef\]](#)
10. Mohammad, S.M. Sentiment analysis: Detecting valence, emotions, and other affectual states from text. In *Emotion Measurement*; Meiselman, H.L., Ed.; Woodhead Publishing: Cambridge, UK, 2016; pp. 201–237.
11. Pang, B.; Lee, L.; Vaithyanathan, S. Thumbs up? Sentiment classification using machine learning techniques. In Proceedings of the Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP), Philadelphia, PA, USA, 6–7 July 2002; pp. 79–86.
12. Zhai, Z.; Zhang, X.; Fang, F.; Yao, L. Text classification of Chinese news based on multi-scale CNN and LSTM hybrid model. *Multimed. Tools Appl.* **2023**, *82*, 20975–20988. [\[CrossRef\]](#)
13. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, Ł.; Polosukhin, I. Attention is all you need. In Proceedings of the 31st International Conference on Neural Information Processing Systems (NIPS), Long Beach, CA, USA, 4–9 December 2017; pp. 5998–6008.
14. Wang, H.; Hou, M.; Li, F.; Zhang, Y. Chinese implicit sentiment analysis based on hierarchical knowledge enhancement and multi-pooling. *IEEE Access* **2020**, *8*, 126051–126065. [\[CrossRef\]](#)
15. Rahma, A.; Azab, S.S.; Mohammed, A. A comprehensive survey on Arabic sarcasm detection: Approaches, challenges and future trends. *IEEE Access* **2023**, *11*, 18261–18280. [\[CrossRef\]](#)
16. Diao, Y.; Yang, L.; Li, S.; Hao, Z.; Fan, X.; Lin, H. Detect sarcasm and humor jointly by neural multi-task learning. *IEEE Access* **2024**, *12*, 38071–38080. [\[CrossRef\]](#)
17. Zhang, H.; Cheah, Y.-N.; Alyasiri, O.M.; An, J. Exploring aspect-based sentiment quadruple extraction with implicit aspects, opinions, and ChatGPT: A comprehensive survey. *Artif. Intell. Rev.* **2024**, *57*, 17. [\[CrossRef\]](#)
18. Jia, W.; Shan, J.; Wang, J.; Mengyang, Z. Implicit aspect-based sentiment analysis: A systematic review. *ACM Comput. Surv.* **2026**, *58*, 1–37. [\[CrossRef\]](#)
19. Ren, J.; Zhou, W.; Li, B.; Liu, M.; Le, N.L.D.; Cen, J.; Chen, L.; Xu, Z.; Xu, X.; Li, X. Causal prompting for implicit sentiment analysis with large language models. *IEEE Trans. Comput. Soc. Syst. Early Access* **2026**, *13*, 4131–4143. [\[CrossRef\]](#)
20. Wang, W.; Luo, X.; Hao, J.; Gu, X.; Chen, X.; Wang, Z.; Wang, H. A GAN and LLM-Driven Data Augmentation Framework for Dynamic Linguistic Pattern Modeling in Chinese Sarcasm Detection. *arXiv* **2026**, arXiv:2604.08381.
21. Higgins, I.; Matthey, L.; Pal, A.; Burgess, C.; Glorot, X.; Botvinick, M.; Mohamed, S.; Lerchner, A. Beta-VAE: Learning basic visual concepts with a constrained variational framework. In Proceedings of the 5th International Conference on Learning Representations (ICLR), Toulon, France, 24–26 April 2017.

22. Shen, T.; Lei, T.; Barzilay, R.; Jaakkola, T. Style transfer from non-parallel text by cross-alignment. In Proceedings of the 31st International Conference on Neural Information Processing Systems (NIPS), Long Beach, CA, USA, 4–9 December 2017; pp. 3071–3079.
23. Ganin, Y.; Lempitsky, V. Unsupervised domain adaptation by backpropagation. In Proceedings of the 32nd International Conference on Machine Learning (ICML), Lille, France, 6–11 July 2015; pp. 1180–1189.
24. Han, T.; Zhang, Z.; Ren, M.; Dong, C.; Jiang, X.; Zhuang, Q. Text emotion recognition based on XLNet-BiGRU-Att. *Electronics* **2023**, *12*, 2704. [[CrossRef](#)]
25. Devlin, J.; Chang, M.; Lee, K.; Toutanova, K. 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*; Association for Computational Linguistics: Minneapolis, MN, USA, 2019; Volume 1, pp. 4171–4186.
26. Zhang, Z.; Li, X.; Li, Y.; Dong, Y.; Wang, D.; Xiong, S. Neural noise embedding for end-to-end speech enhancement with conditional layer normalization. In Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Toronto, ON, Canada, 6–11 June 2021; pp. 7113–7117. [[CrossRef](#)]
27. Lee, D.; Tian, Z.; Xue, L.; Zhang, N.L. Enhancing content preservation in text style transfer using reverse attention and conditional layer normalization. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (ACL-IJCNLP)*; Association for Computational Linguistics: Minneapolis, MN, USA, 2021; pp. 93–102.
28. Ba, J.L.; Kiros, J.R.; Hinton, G.E. Layer normalization. *arXiv* **2016**, arXiv:1607.06450.
29. Salganik, M.J. *Bit by Bit: Social Research in the Digital Age*; Princeton University Press: Princeton, NJ, USA, 2019.
30. Quan-Haase, A.; Sloan, L. (Eds.) *The SAGE Handbook of Social Media Research Methods*, 2nd ed.; SAGE Publications: London, UK, 2022.
31. Liu, Y.; Ott, M.; Goyal, N.; Du, J.; Joshi, M.; Chen, D.; Levy, O.; Lewis, M.; Zettlemoyer, L.; Stoyanov, V. Roberta: A robustly optimized BERT pretraining approach. *arXiv* **2019**, arXiv:1907.11692.
32. Cui, Y.; Che, W.; Liu, T.; Qin, B.; Wang, S.; Hu, G. Revisiting pre-trained models for Chinese natural language processing. *Proc. Find. Assoc. Comput. Linguist. EMNLP* **2020**, *2020*, 657–668. [[CrossRef](#)]
33. He, P.; Gao, J.; Chen, W. DeBERTaV3: Improving DeBERTa using ELECTRA-style pre-training with gradient-disentangled embedding sharing. *arXiv* **2021**, arXiv:2111.09543.
34. Sun, Y.; Wang, S.; Feng, S.; Ding, S.; Pang, C.; Shang, J.; Liu, J.; Chen, X.; Zhao, Y.; Lu, Y.; et al. ERNIE 3.0: Large-scale knowledge-enhanced pre-training for language understanding and generation. *arXiv* **2021**, arXiv:2107.02137.
35. Mathews, A.; Xie, L.; He, X. SemStyle: Learning to Generate Stylised Image Captions Using Unaligned Text. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Salt Lake City, UT, USA, 18–23 June 2018; pp. 8591–8600.
36. Chen, Q.; Zhang, R.; Zheng, Y.; Mao, Y. Dual contrastive learning: Text classification via label-aware data augmentation. *arXiv* **2022**, arXiv:2201.08702.
37. Guo, X.; Li, B.; Yu, H.; Miao, C. Latent-optimized adversarial neural transfer for sarcasm detection. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*; Association for Computational Linguistics: Minneapolis, MN, USA, 2021; pp. 5394–5407.
38. Brown, T.B.; Mann, B.; Ryder, N.; Subbiah, M.; Kaplan, J.; Dhariwal, P.; Neelakantan, A.; Shyam, P.; Sastry, G.; Askell, A.; et al. Language models are few-shot learners. In Proceedings of the International Conference on Neural Information Processing Systems (NeurIPS), Vancouver, BC, Canada, 6–12 December 2020; pp. 1877–1901.
39. Bai, J.; Bai, S.; Chu, Y.; Cui, Z.; Dang, K.; Deng, X.; Fan, Y.; Ge, W.; Han, Y.; Huang, F.; et al. Qwen technical report. *arXiv* **2023**, arXiv:2309.16609.
40. Zeng, A.; Liu, X.; Du, Z.; Wang, Z.; Lai, H.; Ding, M.; Yang, Z.; Xu, Y.; Zheng, W.; Xia, X.; et al. GLM-130B: An open bilingual pre-trained model. In Proceedings of the International Conference on Learning Representations (ICLR), Kigali, Rwanda, 1–5 May 2023.

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.