Picturized and Recited with Dialects: A Multimodal Chinese Representation Framework for Sentiment Analysis of Classical Chinese Poetry

Xiaocong Du , Haoyu Pei , Haipeng Zhang*
ShanghaiTech University
{duxc2023, peihy2024, zhanghp}@shanghaitech.edu.cn

Abstract

Classical Chinese poetry is a vital and enduring part of Chinese literature, conveying profound emotional resonance. Existing studies analyze sentiment based on textual meanings, overlooking the unique rhythmic and visual features inherent in poetry, especially since it is often recited and accompanied by Chinese paintings. In this work, we propose a dialect-enhanced multimodal framework for classical Chinese poetry sentiment analysis. We extract sentence-level audio features from the poetry and incorporate audio from multiple dialects, which may retain regional ancient Chinese phonetic features, enriching the phonetic representation. Additionally, we generate sentence-level visual features, and the multimodal features are fused with textual features enhanced by LLM translation through multimodal contrastive representation learning. Our framework outperforms state-of-the-art methods on two public datasets, achieving at least 2.51% improvement in accuracy and 1.63% in macro F1. We open-source the code to facilitate research in this area and provide insights for general multimodal Chinese representation.

1 Introduction

Classical Chinese poetry, as an essential component of Chinese literature, is marked by its concise expression, rich imagery, and structured tones and rhythms, conveying profound emotional resonance [Liu, 2022]. From ancient times to today, it remains widely spread and has profoundly influenced modern Chinese [Zheng, 1995]. Sentiment analysis of classical Chinese poetry has gained increasing attention recently, with researchers building and refining models on benchmark datasets of manually labeled sentence-sentiment or poem-sentiment pairs [Chen *et al.*, 2019; Wei *et al.*, 2024a].

Existing research mainly focuses on textual analysis [Wang *et al.*, 2023b; Wei *et al.*, 2024b; Xiang *et al.*, 2024], however, neglecting the fact that beyond the *text* itself, poetry's emotions are in the meantime conveyed through the vivid *picture* it evokes, and the *rhythm* expressed in its recitation (example in Figure 1). This is evidenced by the fact that classical Chinese poetry is often accompanied by Chinese paintings [Qian, 1985], and commonly recited [Liu, 1985], serving as a prevalent form of transmission.

Figure 1: Case (a) illustrates the rhythmic features of classical Chinese poetry, Case (b) shows differences in dialectal pronunciations, and Case (c) demonstrates the visual features.

We are developing a multimodal Chinese representation framework for sentiment analysis of classical Chinese poetry, while also hoping to deepen insights into general Chinese understanding. For this specific task, it addresses the problem of overlooking important visual and rhythmic information, as well as the lack of integration among these modalities. In the broader context of Chinese representation, most studies focus on multimodal fusion at the character level [Meng *et al.*, 2019; Sun *et al.*, 2021b; Mai *et al.*, 2022], while we represent semantics at the sentence level. This allows for the combination of different imagery and rhythmic elements to convey new and more complete information within the context of Chinese poetry. As a first effort, our framework also integrates dialects at the phonetic level, as their pronunciations may more closely resemble ancient forms from specific geographic areas, allowing for a more accurate representation of the emotional nuances. We elaborate on our motivations and design considerations as follows.

The *rhythmic* features of poetry are reflected in the fact that poetry is specially designed for 'recitation' [Liu, 1985]. Strict phonological constraints, such as rhyming schemes and tonal patterns (平仄), govern the structure of classical Chinese po-

*Corresponding author.

etry, allowing readers to experience emotions during poetry recitals. As shown in Figure 1(a), in the poetic sentence '江南可采莲，莲叶何田田。', in addition to the usual endline rhyming ('莲' lian2 and '田' tian2), the same rhyming words appear four times in the sentence, creating a cyclical, repetitive tone when reciting that conveys a positive emotion.

The helpful rhythmic features may be influenced by variations in pronunciation across regions and over time, which can be reflected in *dialects*. As shown in Figure 1(b), some sentences may not adhere to the tonal patterns in Mandarin (based on northern dialects), while they may conform to those in Cantonese (a southern dialect), and vice versa. For a certain poem written in a certain time and space, there may be a more suitable dialect for reciting it, since some regional dialects can preserve ancient pronunciations [Yuan, 2001], necessitating the need to consider multiple dialects.

Meanwhile, ancient Chinese poets excelled at transforming objective nouns into picturized imagery, and the combination of imagery words constructs the *visual* features of classical Chinese poetry [Yuan, 2009]. For example, in Figure 1(c), the sentence '大漠孤烟直，长河落日圆。' (Vast desert and straight single smoke, long river and round setting sun.) is composed solely of nouns and adjectives, while the picture created by these imagery words evokes in the reader a visual association, and the composition, hue, and other visual details of the picture lead to an emotional resonance.

Considering these characteristics of this task, we design our model regarding rhythmic, dialect, and visual features. The rhythmic features of poetry are derived from the phonetic features of Chinese characters, which have been proven effective in general Chinese language understanding tasks [Sun *et al.*, 2021b; Peng *et al.*, 2021; Mai *et al.*, 2022]. Previous work typically extracts phonetic features from the pinyin representation of Chinese characters, which is the letter-based phonetic representation. However, instead of pinyin, we use the audio of Chinese character pronunciations since it can provide fine-grained phonetic features [Mai *et al.*, 2022]. Besides, as poetry is meant to be recited for emotional delivery, it naturally motivates us to use audio rather than written symbols. Furthermore, in previous methods, phonetic features are employed as auxiliary to enhance the token-level character representations. Since there are far more Chinese characters than pronunciations, the variances in characters capture much of the semantics, which may dominate the phonetic information, making it less expressive. Given the critical role of phonetic features in our task, we propose to disentangle pronunciation embeddings from character embeddings and design a phonetic feature extraction module. The module ensures that phonetic information is more explicitly emphasized in the sentence-level features. Moreover, this disentangled design separates the audio modality from the text modality module, allowing our framework to be flexible and seamlessly integrate with other single-textual modality methods.

We then want to use dialect features. Besides **Mandarin**, we choose four dialects as an additional audio source, including **Cantonese** (粤语), **Wu** (吴语), **Minnan** (闽南话), and **Chaozhou** (潮州话). We use Cantonese as an example below.We employ the same audio feature extraction module to obtain the Cantonese features and combine them with the Mandarin features. A challenge in audio feature fusion is that variations in speaker timbre and volume across different data sources inevitably introduce noise [Chauhan *et al.*, 2024]. Since noise is more likely to occur in the non-overlapping parts of the two features, our solution is to emphasize the shared features and pay less attention to others. We use cross-attention to align relevant features between two inputs by adjusting the attention weights [Jian *et al.*, 2024], thereby fusing dialect audio features into the overall phonetic features.

The visual features presented in classical Chinese poetry are composed of imagery entities. Su *et al.* [2023] uses individual words from a poem as queries to find matching images from a search engine, extracting visual features from images to enhance the word embeddings. This approach is similar to looking up textual knowledge bases to enhance individual words [Wei *et al.*, 2024a]. However, it does not provide a complete picture of the entire sentence and only offers information from isolated words. Benefiting from recent advancements in text-to-image models [Rombach *et al.*, 2022] trained on large-scale text-image pairs, our visual module can associate text with images to generate more complete visual representations at the sentence level.

Besides emphasizing phonetic and visual features, it remains essential to understand the concise and ancient textual content. We draw on prior work [Xiang *et al.*, 2024] and use a Large Language Model (LLM) to generate modern translations. Then we feed the original and translated texts into the pre-trained language model to obtain the overall text features before we integrate all three modality features.

As shown in Figure 2, we employ a two-stage training strategy to fuse audio (in green), visual (in blue), and textual (in orange) modality features. In the first stage, we train the three modality modules using multimodal contrastive representation learning [Guzhov *et al.*, 2022]. In the second stage, we concat the text, audio, and visual features to obtain the overall sentence-level features and train the entire framework.

We summarize our contributions as follows:

1. We propose a tri-modal Chinese representation framework that fuses sentence-level audio, visual and textual features for classical Chinese poetry sentiment classification. Our method outperforms SOTA on two public datasets by 2.51% in accuracy and 1.63% in macro F1.

2. This study is a first effort to explore the phonetic (audio) features of classical Chinese poetry for sentiment analysis. Upon this, we incorporate features from four pairs of dialects into Chinese text representation through a cross-attention module. We further confirm the effects of dialectal features with an empirical analysis.

3. We demonstrate that our framework serves as a flexible tool, as popular text-only models for sentiment analysis achieve enhanced performance when integrated into it compared to their original capabilities. We open-source our framework[1] and hope it not only facilitates classical Chinese poetry sentiment analysis but also brings insights for general multimodal Chinese representation.

[1]https://github.com/ZhangDataLab/chinese_poetry_sentiment

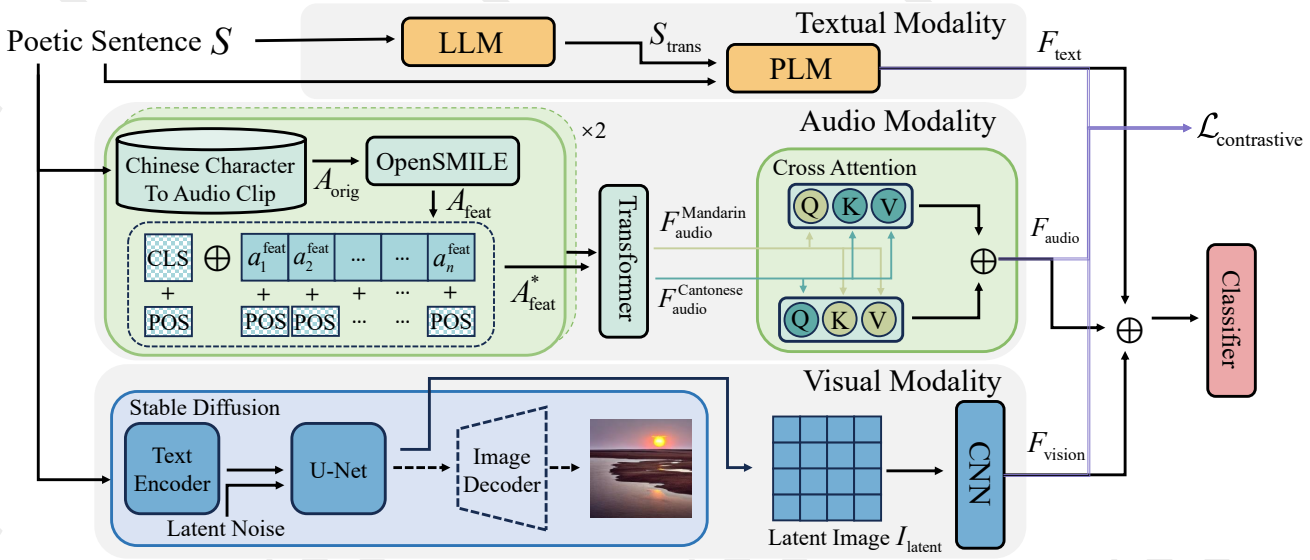

Figure 2: Tri-modal framework with textual (orange), audio (green), and visual (blue) modality modules. PLM, Transformer, Cross-Attention, CNN, and Classifier are trainable. In the first pre-training phase (the purple line), we use contrastive loss $\mathcal{L}_{contrastive}$ to train the three modality feature extractors. In the second phase, all trainable modules are jointly trained on specific downstream tasks.

2 Related Work

2.1 Classical Chinese Poetry Sentiment Analysis

A mainstream trend in sentiment analysis of classical Chinese poetry is to enhance the representation of poetic sentences with additional textual knowledge. CDBERT [Wang *et al.*, 2023b] introduces dictionary definitions for rare words in poetry to improve the representation of words that appear less frequently during pretraining. SAMCAP [Liu *et al.*, 2024] extracts knowledge features from the knowledge base of imagery entities and sentiment words. Others use the syntactic structure of poetic sentences [Wang *et al.*, 2023a; Zhang *et al.*, 2024] and the radical components of Chinese characters [Han *et al.*, 2023].

Recent studies interpret classical Chinese using modern Chinese, which may not always be easily understood nowadays. Poka [Wei *et al.*, 2024a] constructs a knowledge graph linking classical Chinese words with their modern Chinese equivalents. Other studies use large-scale classical-modern Chinese parallel corpora for pretraining, including different pretraining strategies, such as CCMC [Liu *et al.*, 2022], CP-ChineseBERT [Hong *et al.*, 2023], CGCLM [Xiang *et al.*, 2024] and CCDM [Wei *et al.*, 2024b].

However, most studies above focus on textual semantics, ignoring that classical Chinese poetry, which differs from other usual types of text, is also characterized by its rhythmic and visual features [Liu, 2022]. Although MISC [Su *et al.*, 2023] uses image features to enrich the imagery words, it still excludes the rhythmic features and does not fully integrate the phonetic, visual, and textual modalities. This motivates us to go beyond text to adopt a tri-modal perspective.

2.2 Multimodal Chinese Representation

Multimodal information can be used to enhance Chinese text representation. Many studies start with individual Chinese characters, which are the basic building blocks of the Chinese language. Glyce [Meng *et al.*, 2019] is the first to employ a CNN-based model to extract visual features from the glyph images of Chinese characters. Building upon glyph embeddings, DISA [Peng *et al.*, 2021] incorporates pronunciation embeddings. It employs a reinforcement learning framework to disambiguate the intonations of the Chinese characters. To improve the way that glyph and pronunciation features are fused, ChineseBERT [Sun *et al.*, 2021b] integrates character embeddings with glyph and pinyin embeddings into the BERT pretraining process and achieves SOTA performance on multiple Chinese NLP tasks. Following this idea, MPM-CNER [Mai *et al.*, 2022] further introduces four pretraining tasks specifically designed for this task.

However, unlike previous studies focusing on representation at the character level, we extract the representation of the sentence semantics. The rhythmic and visual characteristics of classical Chinese poetry make it inherently suitable for extracting sentence-level features. This idea may also be extended to texts with similar characteristics to classical Chinese poetry.

2.3 Multimodal Sentiment Analysis

The broad sentiment analysis has expanded from text classification to include other modalities such as images and audio. This evolution is aligned with the increasing availability of multimodal data, such as image-text comments on social media [Zhu *et al.*, 2022] and audiovisual content on video platforms [Hazarika *et al.*, 2020]. In most multimodal sentiment analysis studies, the data from different modalities naturally coexist, leading to a focus on aligning and fusing features across modalities [Gandhi *et al.*, 2023]. In our work, however, we emphasize that we "create" multiple modalities from the original single text modality data. We extend the input from a single textual modality to a multimodal approach, al-

lowing for a multi-perspective understanding. This involves extracting audio features from the Chinese character audio and obtaining visual features using a text-to-image model.

3 Method

The overall structure of our model is shown in Figure 2. First, we generate tri-modal features for the given poetic sentences. For the **audio modality** (3.1), we design a sentence-level audio feature extraction module to capture dialect-specific phonetic features. These dialectal features are then fused into a unified phonetic feature using a cross-attention module (3.2). For the **visual modality** (3.3), we extract intermediate features from a text-to-image model and process them with an image feature extractor to obtain the visual representation of the poetic sentences. For the **textual modality** (3.4), we employ LLM to generate modern Chinese translations of classical Chinese poetry. These translations are concatenated with the original text and jointly trained using a pre-trained language model. Finally, the features from all three modalities are fused through a warm-up training phase guided by multimodal contrastive representation learning (3.5), and the fused features are used for sentiment classification.

3.1 Audio Modality Feature Extraction

Commonly used audio feature extraction models, such as wav2vec2.0 [Baevski *et al.*, 2020] and AST [Gong *et al.*, 2021], are trained on complete, real audio recordings. While we do not have such recordings for the benchmark datasets, the audio of the sentences is alternatively composed of multiple character audio segments, making conventional audio feature extractors unsuitable. We therefore design a feature extraction model tailored for segment-level audio sequences. We first collect Mandarin audio clips[2] corresponding to individual Chinese characters. For a given poetic sentence $S = [c_1, c_2, \ldots, c_n]$, we map each character c_i to its corresponding audio clip a_i^{orig}, resulting in an audio sequence $A_{\text{orig}} = [a_1^{\text{orig}}, a_2^{\text{orig}}, \ldots, a_n^{\text{orig}}]$. Then, we use OpenSMILE [Eyben *et al.*, 2010] to extract acoustic features from each clip, obtaining a token-level audio feature sequence $A_{\text{feat}} = [a_1^{\text{feat}}, a_2^{\text{feat}}, \ldots, a_n^{\text{feat}}]$. To capture the rhyming position and tonal patterns of specific characters in the poetic sentence, we add position embedding p_i to the audio features of each token a_i^{feat}. Additionally, a class embedding a^{cls} is appended at the beginning of the audio feature sequence A_{feat} to aggregate the features of the entire sequence. Both position embedding and class embedding are learnable parameters. The final audio feature sequence $A_{\text{feat}}^* = [a^{\text{cls}} + p_0, a_1^{\text{feat}} + p_1, \ldots, a_n^{\text{feat}} + p_n]$ is sent to a transformer encoder to extract sentence-level phonetic features $F_{\text{audio}}^{\text{Mandarin}}$:

$$F_{\text{audio}}^{\text{Mandarin}} = \text{Transformer}\left(A_{\text{feat}}^*\right) \tag{1}$$

3.2 Dialect Feature Fusion

Mandarin (Putonghua) is based on northern dialects, while we also include Cantonese, a major southern dialect that may contain distinct regional and ancient phonetic features. In addition to Mandarin phonetic features $F_{\text{audio}}^{\text{Mandarin}}$, we collect Cantonese audio clips[3] and obtain their features $F_{\text{audio}}^{\text{Cantonese}}$ in the same way. The audio feature extractors for both dialects share the same parameters, aiming to map features from different dialects into a shared vector space. We then want to fuse their phonetic features. However, differences in voice characteristics among speakers in different audio sources can introduce noise. To mitigate this, we introduce a cross-attention module that enhances overlapping features across different audio sources while minimizing divergent features. We first transform the Mandarin phonetic features $F_{\text{audio}}^{\text{Mandarin}}$ into the query Q^{Mandarin}, key K^{Mandarin} and value V^{Mandarin}:

$$Q_{\text{Mandarin}} = \text{Linear}_Q\left(F_{\text{audio}}^{\text{Mandarin}}\right) \tag{2}$$

$$K_{\text{Mandarin}} = \text{Linear}_K\left(F_{\text{audio}}^{\text{Mandarin}}\right) \tag{3}$$

$$V_{\text{Mandarin}} = \text{Linear}_V\left(F_{\text{audio}}^{\text{Mandarin}}\right) \tag{4}$$

For the Cantonese phonetic features $F_{\text{audio}}^{\text{Cantonese}}$, we use the same method to obtain the query $Q_{\text{Cantonese}}$, key $K_{\text{Cantonese}}$ and value $V_{\text{Cantonese}}$. Subsequently, the query and key from different dialects are used to calculate dot-product attention to focus on the common features within the audio. This process is expressed as:

$$F_{\text{audio-att}}^{\text{Mandarin}} = \text{Softmax}\left(\frac{Q_{\text{Cantonese}} K_{\text{Mandarin}}^T}{d}\right) V_{\text{Mandarin}} \tag{5}$$

$$F_{\text{audio-att}}^{\text{Cantonese}} = \text{Softmax}\left(\frac{Q_{\text{Mandarin}} K_{\text{Cantonese}}^T}{d}\right) V_{\text{Cantonese}} \tag{6}$$

Finally, the enhanced features from different dialects are concatenated to obtain the audio modality features F_{audio}.

$$F_{\text{audio}} = \text{Linear}_{\text{audio}}\left(\left[F_{\text{audio-att}}^{\text{Mandarin}} \oplus F_{\text{audio-att}}^{\text{Cantonese}}\right]\right) \tag{7}$$

3.3 Visual Modality Feature Extraction

The visual modality features are extracted from a text-to-image model. We use the Chinese stable diffusion model 'Taiyi' [Zhang *et al.*, 2022a] with frozen parameters. The model consists of three components: a text encoder, an image information creator (U-Net), and an image decoder. The poetic sentence S is first input to the text encoder. Subsequently, the token embeddings and a latent noise tensor are fed into the U-Net, resulting in a latent image information tensor I_{latent}:

$$I_{\text{latent}} = \text{U-Net}\left(\text{TextEncoder}\left(S\right), \text{Latent Noise}\right) \tag{8}$$

We choose not to use the image decoder, as we do not require the generated images but only the intermediate visual features of the latent space. The intermediate features directly represent the high-level correspondence between text and images. These features preserve the core imagery described in the poetry while reducing the information loss or pixel-level noise introduced by image decoding.

We then use a three-layer convolutional network to convert the latent image I_{latent} into visual modality features F_{vision}:

$$F_{\text{vision}} = \text{CNN}\left(I_{\text{latent}}\right) \tag{9}$$

[2] https://chinese.yabla.com/chinese-pinyin-chart.php

[3] https://humanum.arts.cuhk.edu.hk/Lexis/lexi-can/

3.4 Textual Modality Feature Extraction

Understanding the text of the poetic sentence is also essential. Motivated by recent work in utilizing the alignment between classical Chinese and modern Chinese [Xiang *et al.*, 2024] for better comprehension of classical Chinese texts, we prompt[4] LLM GLM-4[5] to generate modern Chinese translations and explanations for classical Chinese poetry:

$$S_{\text{trans}} = \text{LLM}\left(S, \text{prompt}\right) \tag{10}$$

Subsequently, the original text and its translation are fed into GuwenBERT[6], a classical Chinese pre-trained language model, obtaining the final textual modality features F_{text}:

$$F_{\text{text}} = \text{PLM}\left(S, S_{\text{trans}}\right) \tag{11}$$

3.5 Multimodal Feature Fusion

Since our textual modality model is pre-trained while the audio and visual modality feature extractors are trained from scratch, the performance imbalance between them can lead to suboptimal learning efficiency in the overall framework's training. To address this, we implement a two-stage training strategy. In the first stage, we perform a warm-up training for the feature extractors of all three modalities using multimodal contrastive representation learning [Guzhov *et al.*, 2022], which aligns the modalities through a self-supervised approach. The training loss for the first stage is:

$$\begin{aligned}\mathcal{L}_{\text{contrastive}} =& \mathcal{L}_{\text{cs-ce}}\left(F_{\text{audio}}, F_{\text{vision}}\right) + \mathcal{L}_{\text{cs-ce}}\left(F_{\text{audio}}, F_{\text{text}}\right) \\ &+ \mathcal{L}_{\text{cs-ce}}\left(F_{\text{vision}}, F_{\text{text}}\right)\end{aligned} \tag{12}$$

Here $\mathcal{L}_{\text{cs-ce}}$ is the cosine similarity contrastive loss [Radford *et al.*, 2021]. It refers to calculating the cosine similarity between every pair of samples within a batch and then performing a binary classification for each pair to determine whether they match.

In the second stage, the features from three modalities are concatenated to form an overall feature representation of the poetic sentence, which is then classified using a linear layer.

$$\hat{y} = \text{Softmax}\left(\text{Linear}_{\text{fusion}}\left(\left[F_{\text{text}} \oplus F_{\text{audio}} \oplus F_{\text{vision}}\right]\right)\right) \tag{13}$$

For single-label classification tasks, we use cross-entropy loss $\mathcal{L}_{\text{CE}}$, and for multi-label classification tasks, we use binary cross-entropy loss $\mathcal{L}_{\text{BCE}}$. The training loss for the second stage on different downstream tasks is:

$$\mathcal{L}_{\text{CE}} = -\frac{1}{N}\sum_{i=1}^{N} y_i \log(\hat{y}_i) \tag{14}$$

$$\mathcal{L}_{\text{BCE}} = -\frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{M}\left[y_{ij}\log(\hat{y}_{ij}) + (1-y_{ij})\log(1-\hat{y}_{ij})\right] \tag{15}$$

[4]The prompt: 'Translate the following poem into modern Chinese and provide a detailed explanation of what it describes.'

[5]https://bigmodel.cn/dev/api/normal-model/glm-4

[6]https://github.com/ethan-yt/guwenbert

4 Experiments

4.1 Experimental Setup

Datasets We evaluate our model using two public datasets. The THU Fine-grained Sentimental Poetry Corpus (**FSPC**) dataset [Chen *et al.*, 2019] is designed for a single-label classification task with five categories. It consists of 5,000 poems, each containing four sentences. Every sentence is assigned a sentiment label from one of five categories: 'negative', 'implicit negative', 'neutral', 'implicit positive', and 'positive'.

The second is the Chinese Classical Poetry Dataset (**CCPD**) [Wei *et al.*, 2024a] for multi-label classification with 12 categories, which contains 17,103 poems. Each poem has multiple whole-poem-level labels. The task has 12 categories, including 'homesick', 'frustrated', etc. See details in Table 1.

Train/Test Split The split is consistent with previous studies to ensure a fair comparison. As detailed in Table 1, FSPC is divided into training, validation, and test in a ratio of 8:1:1, while for CCPD, the ratio is 7:2:1.

Implementation Details We use the base-sized GuwenBERT to obtain text embeddings. The audio feature extractor is a 3-layer transformer with hidden size $d = 768$ and attention heads $h = 8$. The visual feature extractor is a 3-layer CNN. We train the model using the Adam optimizer with a learning rate of $1e^{-6}$. We conduct our experiments with two RTX 3090.

Evaluation Metrics Being consistent with previous work, we use accuracy (Acc) as the evaluation metric on FSPC. For CCPD, we use micro F1 (Mic-F1) and macro F1 (Mac-F1).

Baselines Due to the availability of the code for the baselines in the original papers that work on the two datasets, we present the results from those papers. Each dataset includes six baselines, with only one baseline shared between them, resulting in a total of 11 baselines.

We begin by comparing our method with ChatGLM, a commonly used Chinese Large Language Model, on both the FSPC and CCPD datasets. On the FSPC dataset, we further compare it with two knowledge-enhanced models, CDBERT and RAC-BERT. The other three baselines, CCMC, CCDM, and CGCLM, use pretraining on large-scale classical-modern Chinese parallel corpora.

On the CCPD dataset, besides the fine-tuned ChatGLM, we evaluate two traditional multi-label classification models, HiAGM-TP [Zhou *et al.*, 2020] and LCM [Guo *et al.*, 2021], as well as two knowledge-enhanced text classification models: GreaseLM [Zhang *et al.*, 2022b], KPT [Hu *et al.*, 2022], and a modern-Chinese-enhanced method Poka.

Meanwhile, we test whether our framework can enhance the single textual modality models for this task. We do this by applying an open-source textual model first and comparing its result with our framework that uses this specific textual model as its textual modality. The models include RoBERTa [Cui *et al.*, 2020], ERNIE [Sun *et al.*, 2021a], ChineseBERT [Sun *et al.*, 2021b], and SikuRoBERTa [Dongbo *et al.*, 2021].

All experiments are conducted using base-sized models. We briefly review some of the key baselines that previously achieved SOTA performance on classical Chinese poetry sentiment classification:

Dataset	Text	Task	Samples	Categories	Train	Val	Test
FSPC	Sentence	Single-label classification	20,000	5	16,000	2,000	2,000
CCPD	Whole-Poem	Multi-label classification	17,103	12	11,971	3,420	1,712

Table 1: Details of the FSPC and CCPD datasets.

- **ChatGLM [Du *et al.*, 2022]:** ChatGLM is a Large Language Model fine-tuned for Chinese, demonstrating outstanding performance across various natural language processing tasks.
- **CDBERT [Wang *et al.*, 2023b]:** CDBERT incorporates external dictionary definitions for rare words and polysemous words, addressing the limitations of conventional pre-trained language models in modeling such words.
- **RAC-BERT [Han *et al.*, 2023]:** RAC-BERT introduces two radical-aware pre-training tasks to enhance the token-level representation of Chinese characters.
- **CCMC [Liu *et al.*, 2022]:** The Chinese Classical and Modern Corpus Model (CCMC) performs pre-training by replacing classical Chinese words with their modern counterparts and employs contrastive learning to bridge the language gap.
- **CCDM [Wei *et al.*, 2024b]:** The Cross-temporal Contrastive Disentangling Model (CCDM) proposes a disentanglement-reconstruction strategy to separately enhance the spatial representations of classical and modern Chinese texts.
- **CGCLM [Xiang *et al.*, 2024]:** The Cross-Guidance Cross-Lingual Model (CGCLM) utilizes LLM to generate large parallel corpora of classical and modern Chinese text pairs. Additionally, it designs a mutual masking pre-training strategy to achieve semantic alignment between the two language styles.
- **Poka [Wei *et al.*, 2024a]:** The Poetry Knowledge-augmented Joint Model (Poka) uses semantic knowledge graphs and a knowledge-guided mask-transformer to bridge the semantic gaps between classical and modern Chinese, enhancing both thematic and emotional understanding.

4.2 Experimental Results

Main Results The model performance on two datasets is shown in Table 2. Among all baselines, our framework achieves the best performance. On the FSPC dataset, our method outperforms the second-best by 2.51% in accuracy. On the CCPD dataset, our method yields an improvement of 1.19% in micro F1 and 1.63% in macro F1. The second-best methods on both datasets (CGCLM and Poka) both incorporate modern Chinese corpora, indicating the effectiveness of modern Chinese interpretations. Our model still outperforms them, and the above-mentioned margins come from audio and visual features and the tri-modal fusion with textual features.

We observe that although the CCPD dataset has more categories and appears to present a more challenging task, the performance metrics for the CCPD dataset are still significantly higher than those for the FSPC dataset. This suggests that the FSPC task may inherently be more difficult, as its labels, such as 'implicit negative' and 'implicit positive,' seem less straightforward compared to the clearer labels in the CCPD dataset, like 'homesick' and 'frustrated'.

FSPC		CCPD		
Method	Acc	Method	Micro F1	Macro F1
ChatGLM	59.48	ChatGLM	89.50	75.59
CDBERT	60.25	GreaseLM	84.44	69.39
RAC-BERT	61.40	HiAGM-TP	88.27	72.88
CCDM	61.45	LCM	90.19	80.27
CCMC	61.86	KPT	90.72	82.98
CGCLM	62.34	Poka	94.65	85.69
Ours	**64.85**	Ours	**95.84**	**87.32**

Table 2: Performance comparison on two datasets (%).

	Textual Model	+ Multimodal Framework
RoBERTa	59.05	**62.35** (+3.30)
ERNIE	59.55	**63.80** (+4.25)
ChineseBERT	59.00	**62.30** (+3.30)
SikuRoBERTa	59.90	**64.10** (+4.20)
GuwenBERT	60.10	**64.85** (+4.75)

Table 3: Performance comparison with different single-textual modality models on the FSPC dataset (%).

Interchange Textual Modality With Other Models We test the performance of our framework in combination with various single-textual modality models. The results on the FSPC dataset are presented in Table 3. By introducing multimodal features, our framework improves the performance of all baselines, resulting in an average improvement of 3.96% in accuracy. GuwenBERT (used in the main experiment) and SikuRoBERTa, both pre-trained on classical Chinese corpora, achieve the top two rankings.

Generated Visual Features vs. Image Features We compare the effect of extracting visual features from real images (Baidu Images) of imagery words [Su *et al.*, 2023] versus those generated by text-to-image models. We replace the visual feature generation module ($+trans\&audio\&vision$) with image features ($+trans\&audio\&image$). The model using image features outperforms the one without visual modality ($+trans\&audio$), demonstrating the effectiveness of visual features. However, it underperforms our model that uses generated visual features, indicating that our method provides more comprehensive visual features for the poetry.

Broader Dialect Exploration To validate the effectiveness of dialect features, besides Mandarin and Cantonese, we also experiment with other southern dialects that we can find, in-

Method	Acc
Ablation Study	
GuwenBERT	60.10
GuwenBERT$_{+trans}$	61.55
GuwenBERT$_{+trans\&audio}$	64.00
GuwenBERT$_{+trans\&audio\&vision}$	64.45
GuwenBERT$_{+trans\&audio\&vision\&dialect}$	**64.85**
Comparison with Image Feature	
GuwenBERT$_{+trans\&audio\&image}$	64.30

Table 4: Ablation study results (top) and comparison of generated visual features and image features (bottom) on the FSPC dataset.

Method	Mand	Cant	Wu	Minnan	Chaozhou
Single	+2.90	+3.10	+2.35	+2.85	+2.70
+ Mandarin	/	+3.30	+3.05	+3.00	+2.95

Table 5: Improvement over our model without audio (61.55% in Acc) using 5 different dialects on the FSPC dataset.

cluding Wu (Suzhou), Minnan, and Chaozhou. As shown in Table 5, on FSPC, incorporating these dialects alongside Mandarin yields an average Acc improvement of 3.08% over the model without any audio features and 0.18% over that using only Mandarin features. On CCPD, the corresponding F1 improvements are 0.70% and 0.24%.

4.3 Ablation Study

To validate the framework, we incrementally add our custom-designed modules to the GuwenBERT, including the LLM translation (*trans*), audio modality module (*audio*), visual modality module (*vision*) and dialect fusion module (*dialect*). Then we reconduct the experiment on the FSPC dataset.

As shown in Table 4, the results are in line with expectations. Among all modules, the audio features contribute the most significant improvement ($+trans$ vs. $+trans\&audio$, +2.45%). Following the audio modality module, the improvement brought by the LLM translation is notable (GuwenBERT vs. $+trans$, +1.45%), but still lags behind baselines specifically designed for aligning classical and modern Chinese texts ($+trans$ vs. CGCLM, -0.79%). This suggests that the overall performance of our framework could be further enhanced by integrating more advanced single-textual modality methods. The visual modality module and the dialect fusion module contribute 0.45% and 0.40% in accuracy, respectively, which is less than the two modules above. A possible explanation is that visual features are more effective for certain poetic sentences that contain rich imagery words, and dialect features are relevant to poems from their corresponding regions.

5 Southern Dialects for the Southern Poems?

Does the use of southern dialects, such as Cantonese, significantly improve accuracy for southern poems, and northern dialects for northern poems? To investigate this, we analyze the regional effects of different dialect features on classification results. Specifically, we use Mandarin audio and Cantonese audio as inputs for the audio modality module, training two separate models: *model-mandarin* and *model-cantonese* on the FSPC dataset.

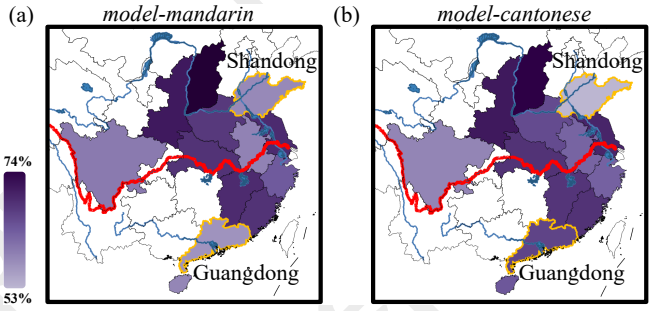

Figure 3: Comparison of *model-mandarin* (in (a), trained on Mandarin audio) and *model-cantonese* (in (b), trained on Cantonese audio) tested on poems from various provinces. The red line represents the Yangtze River, which divides northern and southern China.

The two models exhibit similar overall accuracy (64.45% for *model-mandarin* vs. 64.50% for *model-cantonese*), but they differ in their performance on northern and southern poems. A poem is categorized as northern or southern based on the province of origin of its poet. For northern poems, *model-mandarin* outperforms *model-cantonese* by 0.94%, while for southern poems, *model-cantonese* outperforms *model-mandarin* by 0.39%. This trend is even more pronounced in certain representative provinces: for Shandong Province in the north, *model-mandarin* surpasses *model-cantonese* by 4.44% (Shandong in Figure 3(a) vs. Shandong in Figure 3(b)), while in Guangdong Province, where Cantonese is most prevalent, *model-cantonese* outperforms *model-mandarin* by 8.11% (Guangdong in Figure 3(b) vs. Guangdong in Figure 3(a)). The observed differences between the North and South, as well as between provinces, underscore the role of dialects as effective features for understanding classical poetry within specific regional contexts.

6 Conclusion

We have proposed a multimodal Chinese representation framework for sentiment analysis of classical Chinese poetry. This framework extracts phonetic features, integrates regional dialect audio, generates visual features, and merges them with LLM-enhanced textual features using multimodal contrastive representation learning. It achieves state-of-the-art performance on two public datasets and seamlessly integrates with single-textual modality methods. Further analysis shows that dialectal features significantly aid in classifying poems from their corresponding regions. For future work, we intend to explore specialized approaches for reconstructing ancient pronunciations, extend the framework to support fusing more than two dialects, and delve deeper into sentence-level audio integration. Beyond classical poetry, we will expand our approach to a broader range of tasks in general Chinese representation and understanding.

Contribution Statement

Xiaocong Du and Haoyu Pei contributed equally in this work.

References

[Baevski *et al.*, 2020] Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. wav2vec 2.0: A framework for self-supervised learning of speech representations. *Advances in Neural Information Processing Systems*, 33:12449–12460, 2020.

[Chauhan *et al.*, 2024] Neha Chauhan, Tsuyoshi Isshiki, and Dongju Li. Enhancing speaker recognition models with noise-resilient feature optimization strategies. In *Acoustics*, volume 6, pages 439–469. MDPI, 2024.

[Chen *et al.*, 2019] Huimin Chen, Xiaoyuan Yi, Maosong Sun, Wenhao Li, Cheng Yang, and Zhipeng Guo. Sentiment-controllable chinese poetry generation. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence*. International Joint Conferences on Artificial Intelligence Organization, 2019.

[Cui *et al.*, 2020] Yiming Cui, Wanxiang Che, Ting Liu, Bing Qin, Shijin Wang, and Guoping Hu. Revisiting pre-trained models for Chinese natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: Findings*, pages 657–668, Online, November 2020. Association for Computational Linguistics.

[Dongbo *et al.*, 2021] Wang Dongbo, Liu Chang, Zhu Zihe, Liu Jiangfeng, Hu Haotian, Shen Si, and Bin Li. Sikubert and sikuroberta: Research on the construction and application of pre-training model of sikuquanshu for digital humanities. *Library Tribune*, pages 1–14, 2021.

[Du *et al.*, 2022] Zhengxiao Du, Yujie Qian, Xiao Liu, Ming Ding, Jiezhong Qiu, Zhilin Yang, and Jie Tang. Glm: General language model pretraining with autoregressive blank infilling. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 320–335, 2022.

[Eyben *et al.*, 2010] Florian Eyben, Martin Wöllmer, and Björn Schuller. Opensmile: the munich versatile and fast open-source audio feature extractor. In *Proceedings of the 18th ACM international conference on Multimedia*, pages 1459–1462, 2010.

[Gandhi *et al.*, 2023] Ankita Gandhi, Kinjal Adhvaryu, Soujanya Poria, Erik Cambria, and Amir Hussain. Multimodal sentiment analysis: A systematic review of history, datasets, multimodal fusion methods, applications, challenges and future directions. *Information Fusion*, 91:424–444, 2023.

[Gong *et al.*, 2021] Yuan Gong, Yu-An Chung, and James Glass. Ast: Audio spectrogram transformer. In *Interspeech 2021*, pages 571–575, 2021.

[Guo *et al.*, 2021] Biyang Guo, Songqiao Han, Xiao Han, Hailiang Huang, and Ting Lu. Label confusion learning to enhance text classification models. In *Proceedings of the AAAI conference on Artificial Intelligence*, volume 35, pages 12929–12936, 2021.

[Guzhov *et al.*, 2022] Andrey Guzhov, Federico Raue, Jörn Hees, and Andreas Dengel. Audioclip: Extending clip to image, text and audio. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 976–980. IEEE, 2022.

[Han *et al.*, 2023] Lifan Han, Xin Wang, Meng Wang, Zhao Li, Heyi Zhang, Zirui Chen, and Xiaowang Zhang. Racbert: Character radical enhanced bert for ancient chinese. In *CCF International Conference on Natural Language Processing and Chinese Computing*, pages 759–771. Springer, 2023.

[Hazarika *et al.*, 2020] Devamanyu Hazarika, Roger Zimmermann, and Soujanya Poria. Misa: Modality-invariant and-specific representations for multimodal sentiment analysis. In *Proceedings of the 28th ACM international conference on Multimedia*, pages 1122–1131, 2020.

[Hong *et al.*, 2023] Jie Hong, Tingting He, Jie Mei, Ming Dong, Zheming Zhang, and Xinhui Tu. A hybrid corpus based fine-grained semantic alignment method for pretrained language model of ancient chinese poetry. In *2023 IEEE International Conference on Big Data*, pages 4794–4801. IEEE, 2023.

[Hu *et al.*, 2022] Shengding Hu, Ning Ding, Huadong Wang, Zhiyuan Liu, Jingang Wang, Juanzi Li, Wei Wu, and Maosong Sun. Knowledgeable prompt-tuning: Incorporating knowledge into prompt verbalizer for text classification. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2225–2240, 2022.

[Jian *et al.*, 2024] Lihua Jian, Songlei Xiong, Han Yan, Xiaoguang Niu, Shaowu Wu, and Di Zhang. Rethinking cross-attention for infrared and visible image fusion. *arXiv preprint arXiv:2401.11675*, 2024.

[Liu *et al.*, 2022] Maofu Liu, Junyi Xiang, Xu Xia, and Huijun Hu. Contrastive learning between classical and modern chinese for classical chinese machine reading comprehension. *ACM Transactions on Asian and Low-Resource Language Information Processing*, 22(2):1–22, 2022.

[Liu *et al.*, 2024] Zhongbao Liu, Guangwen Wan, Xi Zuo, and Yingbin Liu. Sentiment analysis of chinese ancient poetry based on multidimensional knowledge attention. *Digital Scholarship in the Humanities*, page fqae069, 2024.

[Liu, 1985] Qifen Liu. Classical chinese poetry and music. *Chinese Music*, (4):3, 1985.

[Liu, 2022] James JY Liu. *The art of Chinese poetry*. Routledge, 2022.

[Mai *et al.*, 2022] Chengcheng Mai, Mengchuan Qiu, Kaiwen Luo, Ziyan Peng, Jian Liu, Chunfeng Yuan, and Yihua Huang. Pretraining multi-modal representations for chinese ner task with cross-modality attention. In *Proceedings of the fifteenth ACM international conference on Web Search and Data Mining*, pages 726–734, 2022.

[Meng *et al.*, 2019] Yuxian Meng, Wei Wu, Fei Wang, Xiaoya Li, Ping Nie, Fan Yin, Muyu Li, Qinghong Han, Xiaofei Sun, and Jiwei Li. Glyce: Glyph-vectors for chinese

character representations. *Advances in Neural Information Processing Systems*, 32, 2019.

[Peng *et al.*, 2021] Haiyun Peng, Yukun Ma, Soujanya Poria, Yang Li, and Erik Cambria. Phonetic-enriched text representation for chinese sentiment analysis with reinforcement learning. *Information Fusion*, 70:88–99, 2021.

[Qian, 1985] Zhongshu Qian. Chinese poetry and chinese painting. *Journal of Graduate School of Chinese Academy of Social Sciences*, 1, 1985.

[Radford *et al.*, 2021] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, pages 8748–8763. PMLR, 2021.

[Rombach *et al.*, 2022] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on Computer Vision and Pattern Recognition*, pages 10684–10695, 2022.

[Su *et al.*, 2023] Chang Su, Shupin Liu, and Chalian Luo. Misc: A multimodal approach for sentiment classification of classical chinese poetry. In *International Conference on Intelligent Computing*, pages 432–442. Springer, 2023.

[Sun *et al.*, 2021a] Yu Sun, Shuohuan Wang, Shikun Feng, Siyu Ding, Chao Pang, Junyuan Shang, Jiaxiang Liu, Xuyi Chen, Yanbin Zhao, Yuxiang Lu, et al. Ernie 3.0: Large-scale knowledge enhanced pre-training for language understanding and generation. *arXiv preprint arXiv:2107.02137*, 2021.

[Sun *et al.*, 2021b] Zijun Sun, Xiaoya Li, Xiaofei Sun, Yuxian Meng, Xiang Ao, Qing He, Fei Wu, and Jiwei Li. Chinesebert: Chinese pretraining enhanced by glyph and pinyin information. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 2065–2075, 2021.

[Wang *et al.*, 2023a] Ping Wang, Shitou Zhang, Zuchao Li, and Jingrui Hou. Enhancing ancient chinese understanding with derived noisy syntax trees. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 4: Student Research Workshop)*, pages 83–92, 2023.

[Wang *et al.*, 2023b] Yuxuan Wang, Jack Wang, Dongyan Zhao, and Zilong Zheng. Rethinking dictionaries and glyphs for chinese language pre-training. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 1089–1101, 2023.

[Wei *et al.*, 2024a] Yuting Wei, Linmei Hu, Yangfu Zhu, Jiaqi Zhao, and Bin Wu. Knowledge-guided transformer for joint theme and emotion classification of chinese classical poetry. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2024.

[Wei *et al.*, 2024b] Yuting Wei, Yangfu Zhu, Ting Bai, and Bin Wu. A cross-temporal contrastive disentangled model for ancient chinese understanding. *Neural Networks*, 179:106559, 2024.

[Xiang *et al.*, 2024] Junyi Xiang, Maofu Liu, Qiyuan Li, Chen Qiu, and Huijun Hu. A cross-guidance cross-lingual model on generated parallel corpus for classical chinese machine reading comprehension. *Information Processing & Management*, 61(2):103607, 2024.

[Yuan, 2001] Jiahua Yuan. *Outline of Chinese Dialects*. Beijing: Language and Culture Press, 2001.

[Yuan, 2009] Xingpei Yuan. *Research on Chinese Poetry Art*. BEIJING BOOK CO. INC., 2009.

[Zhang *et al.*, 2022a] Jiaxing Zhang, Ruyi Gan, Junjie Wang, Yuxiang Zhang, Lin Zhang, Ping Yang, Xinyu Gao, Ziwei Wu, Xiaoqun Dong, Junqing He, et al. Fengshenbang 1.0: Being the foundation of chinese cognitive intelligence. *arXiv preprint arXiv:2209.02970*, 2022.

[Zhang *et al.*, 2022b] X Zhang, A Bosselut, M Yasunaga, H Ren, P Liang, C Manning, and J Leskovec. Greaselm: Graph reasoning enhanced language models for question answering. In *International Conference on Representation Learning*, 2022.

[Zhang *et al.*, 2024] Shitou Zhang, Ping Wang, Zuchao Li, Jingrui Hou, and Qibiao Hu. Confidence-based syntax encoding network for better ancient chinese understanding. *Information Processing & Management*, 61(3):103616, 2024.

[Zheng, 1995] Min Zheng. Classical and modern chinese poetry. *Literary Review*, (6):79–90, 1995.

[Zhou *et al.*, 2020] Jie Zhou, Chunping Ma, Dingkun Long, Guangwei Xu, Ning Ding, Haoyu Zhang, Pengjun Xie, and Gongshen Liu. Hierarchy-aware global model for hierarchical text classification. In *Proceedings of the 58th annual meeting of the Association for Computational Linguistics*, pages 1106–1117, 2020.

[Zhu *et al.*, 2022] Tong Zhu, Leida Li, Jufeng Yang, Sicheng Zhao, Hantao Liu, and Jiansheng Qian. Multimodal sentiment analysis with image-text interaction network. *IEEE Transactions on Multimedia*, 25:3375–3385, 2022.