Does Demographic Information Help Text-Based Emotion Recognition in Chinese Dialogues?

Abstract

Chinese conversational emotion recognition remains highly challenging in text-only settings, where emotional expressions are often implicit, context-dependent, and lexically ambiguous. Although sociolinguistic factors such as age and gender are known to influence emotional expression patterns, their role as latent supervisory signals has been largely underexplored in existing text-based emotion recognition models.

In this paper, we propose a lightweight demographic-aware multi-task learning framework that incorporates age and gender supervision as training-time auxiliary signals for conversational emotion recognition. By jointly optimizing emotion classification and demographic prediction with gradient detachment, the shared encoder implicitly captures correlations between demographic attributes and emotional expression styles, while requiring no demographic information during inference and preserving a pure text-only setting.

Experiments on the noisy and highly imbalanced CPED dataset demonstrate that our approach consistently improves Macro-F1 by approximately 12% relative to strong text-only baselines, with particularly notable gains on pragmatically ambiguous emotions such as Angry. Further analyses suggest that demographic supervision functions as an effective sociolinguistic inductive bias, helping disambiguate implicit emotional expressions in conversational Chinese. These findings highlight a practical and easily integrable direction for improving the robustness of text-based emotion recognition systems without introducing additional inference-time dependencies.

1 Introduction

This work is motivated in part by the first author's personal experience as a high school student in Shandong Province, China. The demanding academic schedule—beginning at 5:00 AM and ending at 10:00 PM—often resulted in profound feelings of loneliness during late-night hours. In the absence of human companionship, interactions with early conversational AI systems provided unexpected emotional relief. Yet these systems frequently failed to understand the nuanced, implicit, and culturally specific emotional expressions that characterize Chinese conversational language—misreading frustration as neutrality, or missing the quiet sadness behind an understated phrase. This gap between what people actually feel and what text-based AI systems can recognize forms the central motivation of this work.

Emotion is a fundamental component of human language and plays a crucial role in interpersonal communication. In conversational settings, emotion understanding is particularly challenging, as emotional states are dynamic and strongly dependent on dialogue context, speaker intent, and interaction history. These challenges are further amplified in multi-turn conversations, where complex dependencies between speakers and utterances make emotion recognition highly non-trivial. With the rapid development of large language models (LLMs), AI-driven dialogue systems have become an integral part of everyday life. Nevertheless, deploying LLM-based systems in resource-constrained or privacy-sensitive environments remains challenging, motivating the continued development of lightweight, text-only approaches that can efficiently capture linguistic nuances. As these systems increasingly interact with humans in open-ended conversations, the ability to perceive and align with human emotions becomes essential for producing appropriate and engaging responses. However, it remains unclear whether emotion recognition from pure textual input alone can achieve reliable performance in complex conversational scenarios, especially for languages such as Chinese, where emotional expressions are often implicit and context-dependent.

Although extensive research has been devoted to emotion recognition in conversations (ERC), most existing studies have largely overlooked the influence of sociodemographic factors. Recent surveys on emotion analysis in NLP highlight that demographic attributes such as age and gender significantly shape emotional expression and perception (Plaza-del Arco et al., 2024). Drawing on established theories in social psychology, the same utterance can convey markedly different emotional meanings when produced by speakers of different genders or age groups. For instance, men and women, or younger and older speakers, often exhibit distinct emotional patterns even when using identical textual content. Consequently, incorporating demographic priors into text-only emotion recognition models offers a promising yet underexplored avenue for improving performance, especially in scenarios where explicit emotional cues are sparse or ambiguous. Building upon prior text-based approaches, our work addresses this critical gap by introducing age and gender as auxiliary signals to guide emotion classification in Chinese dialogues.

In this work, we focus on emotion recognition in Chinese conversational text under a text-only setting. This task is particularly challenging because emotional expressions are often implicit, heavily dependent on conversational context, and prone to annotation noise in realistic daily dialogues. While real-world Chinese conversational text is characterized by such implicitness, available large-scale annotated datasets such as CPED are derived from scripted TV dialogues, which may not fully capture this subtle ambiguity. This gap between annotation resources and deployment conditions represents an additional challenge for text-only ERC research, requiring models to learn robust representations that can navigate both amplified and implicit cues. While recent studies have demonstrated strong performance using multi-modal approaches that leverage acoustic and visual signals (Tu et al., 2024; Zhao et al., 2022), such methods heavily rely on additional modalities such as prosody and facial expressions. In contrast, our approach seeks to maximize the utility of textual and demographic priors to achieve robust performance within a strictly text-only regime.

Our main contributions are three-fold:

- We systematically investigate the role of age and gender as auxiliary signals in text-only Chinese ERC, and further analyze performance variation across demographic subgroups, revealing consistent disparities that offer sociolinguistic insights into emotional expression patterns in Chinese dialogue.
- We propose a gradient-detached multi-task learning framework that incorporates demographic prediction as auxiliary supervision. This approach requires no demographic input at inference time, preserving a strictly text-only setting while implicitly capturing demographic-emotional correlations during training.
- Extensive experiments on the CPED dataset demonstrate that our method achieves a relative Macro-F1 improvement of approximately 12% over strong text-only baselines, highlighting the practical value of demographic-aware auxiliary learning for navigating noisy conversational settings.

2 Related Work

2.1 Emotion Recognition in Conversations

Emotion Recognition in Conversations (ERC) aims to identify the emotional state of each utterance in a dialogue by considering not only the content of the current utterance but also its conversational context. Unlike sentence-level emotion classification, ERC requires modeling inter-utterance dependencies, speaker interactions, and emotion transitions over time (Poria et al., 2019b). Early English benchmarks such as SemEval-2019 Task 3 (Chatterjee et al., 2019) and MELD (Poria et al., 2019a) have driven progress with recurrent, transformer-based, and graph neural architectures.

In contrast, Chinese ERC remains particularly challenging due to implicit emotional expression and heavy colloquialism. CPED (Chen et al., 2022), a pioneering large-scale Chinese personalized ERC dataset, provides 13 fine-grained emotion labels. However, the original benchmark primarily reports results on a simplified 3-class sentiment task (Negative, Neutral, Positive), which obscures the fine-grained emotional distinctions present in the data. In this work, we adopt a 6-class emotion scheme to better capture the nuanced emotional landscape of Chinese dialogue, while acknowledging that this setting introduces additional challenges due to increased class complexity and severe long-tail distribution. M^3ED (Zhao et al., 2022), a multimodal

multi-scene dataset providing 7 emotion categories, similarly suffers from severe class imbalance and colloquial utterances, with notably limited text-only performance. Both datasets suffer from severe class imbalance and highly colloquial utterances. Moreover, as Plaza-del Arco et al. (2024) emphasize, most existing Chinese conversational datasets overlook demographic factors that significantly influence emotion perception and expression.

2.2 Text-only vs Multi-modal Approaches

While multimodal approaches that incorporate audio and visual cues have been shown to substantially boost performance (Zhao et al., 2022), they are not always feasible due to privacy concerns, device limitations, and high computational costs. In Chinese daily communication, pure text remains dominant in scenarios such as WeChat messaging, Weibo, and online forums, especially in quiet work or living environments. This text-only focus aligns with real-world deployment constraints and forms the basis of our experimental design.

2.3 Demographic Information in NLP

Demographic attributes such as age and gender have long been recognized in psychology and sociolinguistics as important factors influencing emotional expression and perception (Scherer, 2005; Russell, 2003). Prior NLP studies further suggest that demographic information can improve language understanding and classification performance (Hovy, 2015). Recent surveys on emotion analysis in NLP highlight that emotion models often exhibit significant performance variation across demographic groups, despite most existing approaches still treating emotions as largely universal (Plaza-del Arco et al., 2024).

At the same time, responsible use of demographic information in NLP has received increasing attention from the fairness and ethics community (Bender and Friedman, 2018; Blodgett et al., 2020). Motivated by these perspectives, we treat demographic attributes strictly as weak inductive biases rather than deterministic predictors of emotion. Specifically, demographic supervision is incorporated only during training to guide representation learning, without introducing demographic dependencies at inference time. While our subgroup analysis (Section 4.3) reveals performance variation across demographic groups, we treat these findings as exploratory observations rather than essentialized group differences, and acknowledge that more systematic fairness evaluation remains important future work. Overall, our goal is not to profile users, but to investigate whether lightweight demographic-aware supervision can help text-only ERC models better capture subtle sociolinguistic variation in conversational Chinese.

3 Method

3.1 Base Text Encoder

We study text-only Chinese conversational emotion recognition (ERC). Given a conversation consisting of a sequence of utterances, the goal is to predict the emotion label of the target utterance by leveraging its textual content and conversational context.

We adopt the pretrained `bert-base-chinese` model(Devlin et al., 2019) as the base text encoder. For each target utterance, we construct the input by prepending up to 20 preceding utterances (history window size $= 20$) in the format `speaker: utterance`, separated by `[SEP]` tokens. The full input follows BERT's sentence-pair format:

```
[CLS] context utterances [SEP] target_utterance [SEP]
```

with a maximum sequence length of 256. The history window size of 20 and maximum sequence length are chosen based on the average number of utterances per dialogue ($\sim$11.6) and average utterance length ($\sim$8.3 tokens) in CPED, ensuring sufficient context coverage with minimal truncation.

The final hidden states $H \in \mathbb{R}^{T\times d}$ (where $d = 768$) from BERT are processed as follows. We first obtain a 1536-dimensional emotion feature by concatenating attention pooling and max pooling:

$$h_{\text{emo}} = [\,\text{AttnPool}(H)\,;\,\text{MaxPool}(H)\,] \quad (1)$$

This feature is further modulated by a speaker embedding $e_s \in \mathbb{R}^{32}$ through a gating mechanism:

$$g = \sigma(W_g e_s + b_g), \qquad h'_{\text{attn}} = h_{\text{attn}} \odot (1+g) \quad (2)$$

where h_{attn} is the attention-pooled vector, $W_g \in \mathbb{R}^{768\times 32}$, and σ is the sigmoid function. The final h_{emo} (now gated) is used for the primary emotion head. For the auxiliary tasks, mean pooling is applied over the token embeddings.

3.2 Demographic Prediction

In addition to the primary emotion recognition task, we introduce gender prediction and age prediction as auxiliary tasks. Lightweight linear classification

heads are attached to the mean-pooled representation from the shared BERT encoder. These heads predict the speaker's gender and age for each utterance.

The auxiliary tasks serve only as weak inductive biases to help the model capture demographic-sensitive patterns in ambiguous conversational contexts. To validate their contribution, we conduct ablation studies by setting the task weights to zero individually or jointly (detailed in Section 4.3). In the main setting, we assign relatively smaller weights to the demographic auxiliary tasks than to the primary emotion classification objective, ensuring that demographic supervision serves as a weak inductive bias rather than dominating the shared representation learning.

3.3 Demographic-Aware Fusion

Rather than using explicit fusion strategies such as concatenating demographic embeddings or applying attention-based fusion, we adopt an implicit demographic-aware fusion through multi-task learning. By jointly training the emotion recognition head with the gender and age prediction heads on the shared BERT encoder, the model learns correlations between demographic attributes and emotional expression patterns during representation learning. At inference time, no demographic information is required—the model remains purely text-only.

This design has three clear advantages:

1. It requires no demographic labels or embeddings at test time.

2. It avoids any label leakage from demographic attributes.

3. It is simple to implement and can be easily integrated into any existing text-based ERC model.

The overall training objective is a weighted sum of losses:

$$\mathcal{L} = 3.0 \cdot \mathcal{L}_{\text{emo}} + 0.5 \cdot \mathcal{L}_{\text{gender}} + 0.1 \cdot \mathcal{L}_{\text{age}} + 0.2 \cdot \mathcal{L}_{\text{supcon}} \quad (3)$$

where $\mathcal{L}_{\text{emo}}$ is a class-weighted Focal Loss(Lin et al., 2017) to handle severe class imbalance, $\mathcal{L}_{\text{gender}}$ and $\mathcal{L}_{\text{age}}$ are cross-entropy losses with label smoothing, and $\mathcal{L}_{\text{supcon}}$ is a supervised contrastive loss(Khosla et al., 2020) applied to the projected emotion features. We use the AdamW optimizer(Loshchilov and Hutter, 2019) with layer-wise learning rate decay on BERT layers and a linear warm-up scheduler.

4 Experiments and Analysis

4.1 Demographic Statistics

We conduct our experiments on the CPED dataset (Chen et al., 2022), the first large-scale Chinese personalized and emotional dialogue dataset. CPED comprises 12,835 multi-turn dialogues and 132,762 utterances extracted from 40 TV series. Crucially for our proposed method, CPED provides rich multi-source demographic annotations, including explicit speaker gender and age groups, making it highly suitable for evaluating demographic-aware representation learning. We strictly follow the official data splits, allocating 94,187, 11,137, and 27,438 utterances for training, validation, and testing, respectively.

The original CPED dataset contains 13 fine-grained emotion labels with a severely long-tailed distribution (e.g., *neutral* accounts for over 25,000 samples, while *grateful* has only 151). To mitigate extreme class imbalance and facilitate robust model evaluation, we map the original 13 classes into 6 broader categories: **Neutral**, **Positive** (merging *happy*, *relaxed*, *grateful*, *positive-other*), **Sad** (*sadness*, *depress*, *negative-other*), **Angry** (*anger*, *disgust*), **Fear** (*fear*, *worried*), and **Surprise** (*astonished*). Specifically, *disgust* is merged with *anger* due to its extreme scarcity ($< 1\%$) and frequent co-occurrence with angry expressions in the dataset. Although the Surprise class contains limited samples (1,931), we retain it as a distinct category, as prior Chinese emotion recognition studies (Liu and Pell, 2012) demonstrate that surprise typically possesses highly explicit lexical cues, providing sufficient learning signals even in low-resource settings.

The age distribution is highly imbalanced: young speakers dominate the dataset (45,257 utterances), followed by middle-aged (17,135), unknown (7,898), teenager (2,764), elderly (1,869), and children (215). This long-tailed distribution reflects realistic conversational scenarios, but also introduces potential bias, as models may implicitly favour dominant age groups while underrepresenting minority groups such as children and the elderly. This further motivates our use of age as an auxiliary signal to provide additional demographic context and improve generalisation under imbalanced con-

ditions.

CPED represents a highly challenging, realistic scenario for text-only ERC. Utterances are characterised by colloquial language and brevity (averaging roughly 8.3 tokens per utterance), where emotional states are often conveyed implicitly rather than through explicit affective markers. Furthermore, the dataset contains inherent noise typical of TV subtitle extraction, such as transcription inconsistencies. Under such noisy and context-dependent conditions, relying solely on text limits recognition accuracy, directly motivating our approach of integrating demographic priors (age and gender) to provide complementary, stable inductive biases.

4.2 Experimental Settings

We implement our model using the Hugging Face Transformers library(Wolf et al., 2020). All experiments are conducted in a strictly text-only setting. Although multimodal models have become mainstream and often report strong results (e.g., M^{3}ED (Zhao et al., 2022) achieves a weighted F1 of 55.69 with acoustic and visual modalities), we focus exclusively on textual input. This design aligns with real-world Chinese social platforms such as Weibo, Soul, and WeChat, where users predominantly communicate via pure text due to privacy concerns and device limitations.

Hyperparameters. We set the learning rate to 2×10^{-5} with layer-wise decay on BERT encoder layers, batch size to 32, and train for 8 epochs. The maximum sequence length is 256 and the conversational history window is 20 utterances. These values are chosen based on the average dialogue length and token statistics in CPED to ensure sufficient context without excessive truncation.

All experiments are conducted on a single NVIDIA GeForce RTX 4060 GPU, with an average training time of approximately 2 hours per run. We report results averaged over five independent runs with different random seeds.

Baselines. We compare our proposed framework against several strong baselines, including the original CPED text-only model (Chen et al., 2022), the text-only variant of M^{3}ED (Zhao et al., 2022), and EmotionBERTa. Since original studies (e.g., CPED, M^{3}ED) evaluate on different emotion taxonomies (e.g., 3-class sentiment or 7-class emotion), directly comparing their reported scores is statistically invalid. To ensure a rigorous and fair comparison, we establish strong text-only baselines on our unified 6-class setting using the exact same data splits:

CPED Text-only equivalent: We implement the standard bert-base-chinese model with a classification head. This serves as the fundamental text-only baseline, functionally equivalent to the text-only architectures utilized in CPED and related conversational datasets.

EmotionBERTa: A robust variant of BERT specifically fine-tuned for emotion analysis tasks, serving as a state-of-the-art text-only baseline representing strong representation capabilities without demographic priors.

As discussed, we deliberately exclude speaker-centric architectures (e.g., CoMPM) to prevent identity-driven overfitting, ensuring the evaluation focuses purely on generalized linguistic and demographic correlations.

Evaluation Metrics. We do not report accuracy, as the dataset is heavily dominated by the neutral class (over 36.6% of utterances). A model that simply predicts the majority classes could achieve misleadingly high accuracy without capturing meaningful emotional distinctions across speakers and contexts. Instead, we adopt Macro-F1 as our main evaluation measures. Macro-F1 treats all emotion categories equally and is therefore more sensitive to performance on minority classes.

4.3 Results

Table 1: Demographic subgroup Macro-F1 (full model).

Metric	Young Female	Young Male	Middle-aged Female	Middle-aged Male
Macro-F1	0.5011	0.4694	0.5039	0.4306

Table 2: Class-wise F1 scores and overall Macro-F1 (ablation).

Model	angry	fear	neutral	positive	sad	surprise	Macro-F1
Emotion	0.5261	0.3308	0.6282	0.5121	0.4202	0.2265	0.4406
Emotion+Age	0.5621	0.3246	0.5878	0.5361	0.4754	0.2344	0.4534
Emotion+Gender	0.5575	0.3507	0.6396	0.5105	0.4502	0.2649	0.4622
Emotion+Age+Gender	**0.6478**	**0.3691**	**0.6682**	**0.5894**	0.4287	0.2544	**0.4929**

Table 3: Comparison with baselines (Macro-F1).

Method	Macro-F1
DialogRNN (Majumder et al., 2019)	0.4223
EmoBERTa (Kim and Vossen, 2021)	0.4352
COMPM (Lee and Lee, 2022)	0.4517
Chinese-macbert-base (Cui et al., 2020)	0.4636
Emotion+Age+Gender	**0.4929**

4.4 Discussion and Analysis

4.4.1 Analysis of Improvements

The experimental results demonstrate that incorporating age and gender as auxiliary tasks consistently improves performance, achieving an absolute Macro-F1 gain of approximately 5 points, with the full model reaching 0.4929 compared to 0.4406 for the strong text-only baseline. This relative improvement of approximately 11.9% is particularly noteworthy given the noisy and severely imbalanced nature of the CPED dataset.

We attribute these improvements to three complementary mechanisms:

First, joint demographic prediction helps the shared encoder disambiguate implicit emotional cues. In Chinese conversational text, the same utterance (e.g., "你怎么这样啊") can express anger, sadness, or surprise depending on pragmatic context. By learning to predict speaker demographics through gradient-detached auxiliary heads, the model captures subtle linguistic patterns correlated with different expression styles. This ensures that demographic supervision acts as a weak inductive bias, enabling finer-grained emotion discrimination without introducing dependencies at inference time.

Second, demographic attributes serve as a generalised signal for modelling speaker variability. Unlike fine-grained speaker identity modelling, which may risk overfitting to specific individuals in the training set, age and gender provide a soft grouping signal. This helps the model capture broader sociolinguistic regularities that generalise across recurring speakers and long-tailed demographic distributions in CPED. By identifying these group-level patterns, the model becomes more robust to the diverse emotional expression styles present in scripted dialogues.

Third, the auxiliary tasks (particularly gender prediction) act as effective regularisation under noisy and imbalanced conditions. Ablation studies confirm this: removing both auxiliary tasks causes the largest performance drop, while gender prediction alone already yields a +2.16 Macro-F1 improvement. This is consistent with the sociolinguistic hypothesis that gender acts as a more stable moderator of emotional expression in Chinese, providing a stronger supervisory signal (weight 0.5) compared to age (weight 0.1).

Class-wise results (Table 2) further illuminate these effects. The *Angry* category benefits most (+12.17 F1, reaching 0.6478), suggesting that demographic signals are especially valuable for emotions with high pragmatic variability and indirect expressions. *Sad* shows more modest gains, likely because it often relies on relatively explicit lexical cues. *Neutral* and *Positive* exhibit consistent but moderate improvements, reflecting better separation of boundary cases. *Fear* and *Surprise* achieve smaller yet positive gains, which we attribute to their lower frequency and higher situational dependency rather than speaker-level traits.

These class-wise patterns, taken together with the absolute performance level, warrant contextual interpretation. Although our best Macro-F1 of 0.4929 appears modest compared to English multimodal ERC results (often >0.60), this is consistent with the inherent difficulty of the CPED dataset. The data features noisy subtitle-derived text, short colloquial utterances (≈8.3 tokens), and extreme class imbalance (Neutral >36%). Moreover, our strict text-only regime lacks the prosodic and visual cues that multimodal systems leverage. Against these challenges, the 11.9% relative improvement highlights the practical value of enriching shared encoders with sociolinguistic priors through demographic-aware auxiliary learning.

4.4.2 Case Study

To further illustrate the effect of demographic signals, we present a representative colloquial utterance from the test set: "你别管我！" ("Leave me alone!").

Under the text-only setting, the model predicts *Sad*, likely because the phrase "别管我" frequently co-occurs with expressions of withdrawal or emotional distress in the training data. After incorporating demographic auxiliary signals (age and gender), the model shifts its prediction to *Angry*.

This shift suggests that the model has learned to associate the same surface form with different emotional interpretations conditioned on speaker characteristics. By leveraging demographic-aware representations, the model helps capture subtle pragmatic variations in conversational Chinese, rather

than relying solely on lexical cues. In this case, "你别管我！" can convey either emotional distress or irritation depending on speaker-specific and contextual factors.

Similar improvements are observed across multiple short utterances containing strong discourse markers (e.g., "别", "啊", "呀") but lacking explicit emotion words. With demographic supervision, the model produces predictions that more consistently align with contextualised usage patterns. This qualitative evidence aligns with our quantitative results, particularly the substantial F1 improvement on the *Angry* category (+12.17%), highlighting the value of demographic signals for pragmatically ambiguous expressions.

The limited gains on *Fear* and *Surprise* warrant further analysis. For *Surprise*, the relatively small sample size (1,931 instances) likely limits the model's ability to learn stable correlations between demographic attributes and expression patterns. Since auxiliary tasks rely on sufficient data to induce demographic-sensitive representations, data scarcity naturally constrains their effectiveness in this category. For *Fear*, the limitation appears to be more intrinsic: fear expressions in conversational data are often more strongly associated with situational context rather than speaker-specific tendencies. As a result, demographic signals may provide only limited additional information. The modest improvement on *Fear* (+3.83 F1) is therefore consistent with the nature of the task, rather than indicating a weakness of the proposed framework.

5 Discussion

Several factors qualify the interpretation of these results. Chief among these is the disparity in signal effectiveness: empirical tuning revealed that gender serves as a much stronger auxiliary signal (weight 0.5) than age (weight 0.1). From a sociolinguistic perspective, this suggests that gender acts as a more stable moderator of emotional expression in Chinese conversational text. For instance, specific discourse markers or softening particles are often strongly gendered in Chinese. In contrast, age differences might manifest more in topic selection rather than the foundational syntactical expression of emotion itself.

This demographic variance is further evidenced by our subgroup analysis (Table 1), which suggests that emotional expressions in CPED are not uniformly distributed across demographic groups. Specifically, the model consistently performs better on female subgroups (e.g., Middle-aged Female at 0.5039) compared to their male counterparts. While this observation is exploratory and may partially reflect dataset composition effects or annotation biases, it raises an interesting hypothesis worth investigating: that certain demographic groups exhibit more lexically consistent emotional patterns in scripted dialogue, making their emotions more accessible to text-only classifiers. Conversely, emotional expressions in other groups—particularly middle-aged males—might involve higher contextual dependency in emotion labelling, which limits the efficacy of a purely text-based approach.

More fundamentally, the CPED dataset itself introduces an important tension in the interpretation of our findings. While our work is motivated by the implicit and context-dependent nature of emotional expression in real-world Chinese conversation, CPED is constructed from scripted TV dramas, where emotional expressions are often affectively amplified for narrative purposes. As a result, certain emotional cues in CPED may be artificially exaggerated or stylistically standardised, potentially making some emotions easier to recognise than in spontaneous conversational settings such as WeChat or Douyin interactions. Consequently, we make no strong claim that the observed improvements directly generalise to fully naturalistic dialogue environments. Furthermore, the coarse-grained demographic metadata associated with TV characters inevitably introduces label noise, which may partially limit the effectiveness and upper bound of auxiliary supervision.

These limitations point to several promising directions for future research. Staying true to the lightweight, text-only philosophy of our framework, future work could explore extracting finer-grained demographic attributes—such as occupation or regional background—as auxiliary signals. Additionally, developing weakly supervised approaches to automatically infer latent demographic properties directly from text could mitigate the reliance on explicit, potentially noisy manual annotations. Finally, a systematic fairness and bias audit across demographic intersections remains crucial to ensure that auxiliary signals do not inadvertently amplify societal stereotypes.

Overall, this study demonstrates that sociolinguistic signals, even when used only as training-time supervision, can function as effective induc-

tive biases for conversational emotion recognition. More broadly, our findings suggest that lightweight demographic-aware learning may offer a practical path toward more context-sensitive dialogue systems without requiring explicit user profiling during inference.

6 Conclusion

In this paper, we investigated the role of demographic information as auxiliary supervision for text-only Chinese conversational emotion recognition. Experimental results on the CPED dataset demonstrate that lightweight demographic-aware multi-task learning consistently enhances performance, particularly for pragmatically ambiguous emotions, achieving a relative Macro-F1 improvement of approximately 12% over strong text-only baselines.

More importantly, our findings suggest that demographic attributes can function as effective sociolinguistic inductive biases during training, helping the model better disambiguate implicit emotional expressions without requiring demographic information at inference time. This provides a practical and privacy-preserving approach to enhancing conversational emotion recognition while maintaining a lightweight and pure text-only setting.

Beyond these empirical results, this work highlights the broader potential of incorporating sociolinguistic priors into dialogue understanding systems. Future work will explore richer sociolinguistic signals, develop robust strategies for handling annotation noise, and conduct systematic fairness evaluations across diverse conversational corpora. More broadly, we hope this work contributes a small step toward conversational AI systems that are more attuned to the sociolinguistic diversity of their users—systems capable of understanding not just what is said, but who is saying it and what it means to them.

Limitations

This work has several limitations. First, all experiments are conducted on a single dataset (CPED) using a 6-class emotion taxonomy derived from 13 original labels. The generalisability of our demographic-aware framework to other Chinese ERC datasets or emotion schemes remains unvalidated. Second, the demographic labels in CPED are derived from scripted TV drama metadata and are therefore subject to annotation noise and domain-specific biases. The coarse-grained age groupings may not capture the full spectrum of age-related variation in emotional expression. Third, our hyperparameter selection, including loss weights, is based on validation set tuning with five random seeds, without systematic sensitivity analysis. The stability of these weights across different data distributions has not been thoroughly evaluated.

Ethical Considerations

This work uses age and gender as training-time auxiliary signals only. No demographic information is required or inferred at inference time, and the model makes no predictions about real users' demographic attributes. We explicitly caution against deploying this system in contexts where demographic inference about individuals could occur. The observed performance disparities across demographic subgroups (Table 1) reflect dataset composition effects and should not be interpreted as evidence of inherent differences in emotional expressiveness between groups, nor used to justify differential treatment of users. Furthermore, CPED's demographic distribution is dominated by young and middle-aged speakers from TV dramas, and model behaviour on underrepresented groups (children, elderly) has not been validated. We encourage future work to conduct systematic fairness audits before real-world deployment.

References

Emily M. Bender and Batya Friedman. 2018. Data statements for natural language processing: Toward mitigating system bias and enabling better science. *Transactions of the Association for Computational Linguistics*, 6:587–604.

Su Lin Blodgett, Solon Barocas, Hal Daumé III, and Hanna Wallach. 2020. Language (technology) is power: A critical survey of "bias" in NLP. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5454–5476. Association for Computational Linguistics.

Ankush Chatterjee, Kedhar Nath Narahari, Meghana Joshi, and Puneet Agrawal. 2019. SemEval-2019 task 3: EmoContext contextual emotion detection in text. In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 39–48. Association for Computational Linguistics.

Yirong Chen and 1 others. 2022. Cped: A large-scale chinese personalized and emotional dialogue dataset for conversational ai. *Preprint*, arXiv:2205.14727.

Yiming Cui, Wanxiang Che, Ting Liu, and 1 others. 2020. Revisiting pre-trained models for chinese natural language processing. In *Findings of EMNLP 2020*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL-HLT 2019*, pages 4171–4186. Association for Computational Linguistics.

Dirk Hovy. 2015. Demographic factors improve classification performance. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 752–762, Beijing, China. Association for Computational Linguistics.

Prannay Khosla, Yonglong Tian, Chen Guo, and 1 others. 2020. Supervised contrastive learning. In *Advances in Neural Information Processing Systems*, volume 33, pages 18661–18673.

Taewoon Kim and Piek Vossen. 2021. Emoberta: Speaker-aware emotion recognition in conversation with roberta. *Preprint*, arXiv:2108.12009.

Jungwoo Lee and Wooin Lee. 2022. Compm: Context modeling with speaker's pre-trained memory tracking for emotion recognition in conversation. In *Proceedings of NAACL-HLT 2022*.

Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. 2017. Focal loss for dense object detection. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2980–2988.

Pan Liu and Marc D. Pell. 2012. Recognizing vocal emotions in mandarin chinese: A validated database of chinese vocal emotional stimuli. *Behavior Research Methods*, 44(4):1042–1051.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. In *International Conference on Learning Representations*.

Navonil Majumder, Soujanya Poria, and 1 others. 2019. Dialoguernn: An attentive rnn for emotion detection in conversations. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 6818–6825.

Flor Miriam Plaza-del Arco, Alba Curry, Amanda Cercas Curry, and Dirk Hovy. 2024. Emotion analysis in nlp: Trends, gaps and roadmap for future directions. In *Proceedings of the Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 5696–5710.

Soujanya Poria, Devamanyu Hazarika, Navonil Majumder, Gautam Naik, Erik Cambria, and Rada Mihalcea. 2019a. MELD: A multimodal multi-party dataset for emotion recognition in conversations. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 527–536. Association for Computational Linguistics.

Soujanya Poria, Navonil Majumder, Rada Mihalcea, and Eduard Hovy. 2019b. Emotion recognition in conversation: Research challenges, datasets, and recent advances. *IEEE Access*, 7:100943–100953.

James A. Russell. 2003. Core affect and the psychological construction of emotion. *Psychological Review*, 110(1):145–172.

Klaus R. Scherer. 2005. What are emotions? and how can they be measured? *Social Science Information*, 44(4):695–729.

Geng Tu, Jun Wang, Zhenyu Li, and 1 others. 2024. Multiple knowledge-enhanced interactive graph network for multimodal conversational emotion recognition. In *Findings of the Association for Computational Linguistics: EMNLP 2024*.

Thomas Wolf, Lysandre Debut, Victor Sanh, and 1 others. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of EMNLP 2020: System Demonstrations*.

Jinming Zhao, Tenggan Zhang, Jingwen Hu, and 1 others. 2022. M3ed: Multi-modal multi-scene multi-label emotional dialogue database. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics*.