International Journa of Advanced AI Applications

Volume2,Issue7,Jul.2026

Online ISSN 3104-9338

Print ISSN 3104-932X

Hong Kong Dawn Clarity Press Limited

http://www.dawnclarity.press/index.php/ijaaa

TABLE OF CONTENTS

Online ISSN: 3104-9338
Print ISSN: 3104-932X

Automated Assessment of Sentence Stress Placement in L2 Chinese Using Deep Learning: An Empirical and Pedagogical Study

Wei Wang[1,3*], Hongmei Wang[2,3], Li Jiang[1]

[1] Zhejiang Normal University

[2] Beijing Normal University

[3] Beijing Affiliated Elementary School of Chaoyang Normal School

Received: June 1, 2026

Revised: June 2, 2026

Accepted: June 4, 2026

Published online: June 10, 2025

To appear in: *International Journal of Advanced AI Applications,* Vol. 2, No. 7 (July 2026)

* Corresponding Author: Wei Wang(wangweiblcu@163.com)

Abstract. This study investigated English-speaking learners' acquisition of Chinese sentence stress in ambiguous contexts using a deep learning-based automated assessment approach. A fine-tuned wav2vec 2.0 model achieved 91.2% agreement with human raters (Cohen's $\kappa = 0.85$). Experiment 1 compared learners' and native speakers' stress identification across semantic contexts. Automated acoustic analysis revealed that learners' accuracy was significantly lower in secondary than primary meaning contexts ($p < 0.1$), while native speakers showed no difference. Experiment 2 compared stress location instruction with general training methods. Automated assessment found no significant difference between the two methods ($p > 0.5$). These findings indicate that learners lack proficiency in semantic-driven stress placement, particularly in secondary meanings, and that instruction focusing solely on stress location is insufficient. Methodologically, this study demonstrates the feasibility of deep learning-based automated prosody assessment. Pedagogically, it recommends integrating prosodic features (pitch, duration) into Chinese language teaching and supports AI-driven Computer-Assisted Pronunciation Training (CAPT) systems.

Keywords: *Deep Learning; Automated Assessment; Sentence Stress; Second Language Acquisition; Chinese Prosody*

1. Introduction

Sentence stress plays a vital role in conveying focal information and highlighting semantic priorities in Chinese. Its precise placement directly determines whether speakers' intentions are

accurately understood. However, learners of Chinese as a second language frequently exhibit errors in sentence stress placement. For example, when expressing "I will go to Beijing tomorrow," incorrectly placing the stress on "tomorrow" rather than "Beijing" shifts the emphasis from the destination to the time, leading to semantic misinterpretation. Even learners with foundational Chinese proficiency often struggle with improper stress placement and insufficient pitch variation. Mastering proper stress patterns has thus become a pressing challenge in Chinese phonetics instruction.

Traditional approaches to diagnosing sentence stress errors rely heavily on subjective perceptual judgments by human raters, which are time-consuming, labor-intensive, and prone to inconsistencies. This methodological bottleneck limits both the scalability of instructional feedback and the precision of empirical research.

1.1. Sentence Stress in Chinese as a Foreign Language

For decades, sentence stress teaching has not received sufficient attention in Chinese as a Foreign Language (CFL) instruction. Stress and intonation are the weakest aspects in current teaching practices, which can be attributed to inadequate training and inappropriate teaching methods, with native language interference playing a secondary role [12]. Learners exhibit significant errors in determining stress positions; for instance, in “I will introduce you to it,” students often incorrectly stress one moment [11]. Additionally, intonation teaching remains the most challenging aspect of CFL instruction [15].

From a linguistic perspective, Chinese sentence stress is highly varied, with the same sentence yielding different meanings depending on stress placement—sometimes resulting in completely distinct interpretations. For example, when the phrase “What do you think now?” is stressed on “how,” it asks about temporal changes in thinking style; however, if the stress is placed on “think,” it implies a contrast with prior lack of thinking. Through his teaching practice, Yang observed that international students often struggle to identify appropriate stress positions, ultimately reading sentences with different stress patterns as if they were identical. However, it is important to note that Yang's work primarily focuses on intrinsic aspects of Chinese and does not elaborate on the acquisition process [19].

1.2. Empirical and Pedagogical Studies

In empirical research, international students worldwide were tested and found to exhibit no clear pattern in their stress errors at the grammatical level, advocating for enhanced instruction on stress positioning [21]. Experimental phonetics was employed to examine core stress

representation in Chinese declarative sentences among Japanese students, revealing that they frequently use medium-to-low pitch differences to indicate core stress, though this placement often fails to align with focus patterns [14]. Analyzing 48 sentences with varying focal points, it was observed that American students exhibit habitual rhythmic expressions of core stress in specific syllable combinations; yet, core stress frequently does not correspond to actual focus positions [20]. However, this study primarily focused on pitch features, neglecting duration, and did not incorporate native speaker comparative data.

Regarding textbook organization, current CFL textbooks typically incorporate sentence stress instruction only in the beginner-level phonetics section, after which this content is no longer covered. However, modern Chinese features rich grammatical structures related to sentence stress, such as the “shi...de” construction, “lian...dou” (even/also), and “hai...ne” (still/yet). Without teacher guidance, students may grasp basic syntactic structures but struggle to identify key information points; therefore, Cheng advocates extending sentence stress instruction into grammar teaching at intermediate and advanced levels [5].

1.3. Research Gaps and the Present Study

Existing research has elucidated the challenges learners face in producing sentence stress from various perspectives. However, most studies either remain at descriptive observation and empirical summary levels, or focus solely on output characteristics within specific native language contexts, lacking systematic examination of learners' ability to determine stress positions based on semantic information. Furthermore, from a methodological standpoint, nearly all previous studies have relied exclusively on human perceptual judgment for stress evaluation, which suffers from inherent limitations in consistency, scalability, and interpretability.

The potential of deep learning-based automated assessment to address these limitations—by providing consistent, replicable, and fine-grained acoustic analysis—has not yet been explored in L2 Chinese sentence stress research. The present study fills this gap by integrating a fine-tuned wav2vec 2.0 model into the experimental paradigm, enabling consistent evaluation of stress placement accuracy and acoustic feature-level analysis to complement traditional perceptual judgment.

1.4. Research Questions

This study addresses two core questions:

- How do Chinese learners determine sentence stress positions according to different

semantic contexts? Specifically, what differences exist between learners' and native speakers' ability to judge stress placement in primary versus secondary semantic contexts?

- What strategies can effectively enhance learners' acquisition of sentence stress? For learners with inadequate judgment skills, which teaching methods prove most effective, and can instruction solely focused on stress placement achieve desired educational outcomes?

2. Methodology

2.1. General Design and Task Procedure

To address the two research questions, this study designed two experiments using ambiguous Mandarin Chinese sentences in which different stress placements yielded different interpretations. Experiment 1 compared English-speaking learners' and native speakers' ability to determine and produce appropriate sentence-stress positions from semantic contexts. Experiment 2 evaluated the effectiveness of a sentence stress-position reinforcement training method through classroom-based practice.

In both experiments, participants produced target sentences according to contextual cues presented on a computer screen. The tasks were programmed and administered using E-Prime 2.0, and oral responses were recorded in mono WAV format at 16 kHz and 16-bit resolution. Stress accuracy was evaluated using both human perceptual judgment and automated classification, as described in Section 2.4.

2.2. Experiment 1

2.2.1. Participants

Experiment 1 included 24 participants. The learner group consisted of 12 advanced English-speaking learners of Chinese, including 7 males and 5 females, aged 20–35 years. They had studied Chinese for approximately five years. The native-speaker control group consisted of 12 native Mandarin speakers, including 6 males and 6 females, aged 22–24 years. Before the experiment, all participants completed a brief oral production task to confirm that they were able to produce continuous Mandarin speech.

2.2.2. Design and Materials

Experiment 1 adopted a 2 × 2 mixed design. The within-subjects factor was context type, with two levels: primary-meaning context and secondary-meaning context. The between-subjects factor was group, with two levels: English-speaking learners and native Mandarin

speakers. The dependent variable was sentence-stress placement accuracy.

The experimental materials consisted of 20 ambiguous Mandarin sentences and 4 practice items. Each experimental sentence was paired with two semantic contexts, each requiring a different sentence-stress position and yielding a different interpretation. Thus, each participant completed 40 critical trials. The contexts were classified as primary- or secondary-meaning contexts based on judgments from native Mandarin-speaking raters.

2.2.3. Procedure and Data Analysis

On each trial, participants saw a semantic context and a target ambiguous sentence on the screen. They were asked to produce the sentence naturally while using stress placement to convey the intended meaning. Four practice trials were administered before the formal task.

Sentence-stress accuracy was calculated for each participant under each context condition. A mixed-design ANOVA was conducted with context type and group as independent variables and stress-placement accuracy as the dependent variable. Simple-effects analyses were conducted where appropriate.

2.3. Experiment 2

2.3.1. Participants

Experiment 2 included 20 beginner English-speaking learners of Chinese. They were assigned to either a sentence stress-position reinforcement training group or a comparison group, with 10 participants in each group. Participants were aged 18–35 years and had studied Chinese for approximately one to one and a half years. All participants had basic Chinese grammar and conversational ability and completed a brief oral production task before the experiment.

2.3.2. Design and Materials

Experiment 2 adopted a 2 × 2 mixed design. The between-subjects factor was training method, with two levels: sentence stress-position reinforcement training and general instruction. The within-subjects factor was test type, with two levels: pre-test and post-test. The dependent variable was sentence-stress placement accuracy.

The pre-test and post-test used the same 20 ambiguous sentences as Experiment 1, presented in different orders. Four practice items were included before each test. The training materials consisted of 30 additional ambiguous sentences that did not overlap with the test items.

2.3.3. Procedure and Data Analysis

Participants first completed the pre-test using the same sentence-production task as in Experiment 1. After the pre-test, they completed a two-week instructional period consisting of

four training sessions. The reinforcement-training group received explicit instruction and practice on how different stress placements change the interpretation of ambiguous Mandarin sentences. Training activities included instructor demonstration, guided repetition, contrastive practice, rapid-response drills, and immediate feedback. The comparison group received general instruction without specialized sentence-stress reinforcement. The post-test was administered within one to two days after the training period.

Sentence-stress accuracy was calculated separately for the pre-test and post-test. A mixed-design ANOVA was conducted with test type and training method as independent variables and stress-placement accuracy as the dependent variable. The interaction between test type and training method was used to determine whether the reinforcement training produced greater improvement than the comparison condition.

2.4. Stress Accuracy Evaluation and Automated Classification

Stress accuracy was assessed using both human perceptual judgment and automated classification. For human judgment, five native Mandarin Chinese speakers who did not participate in the experiments independently evaluated whether each response carried stress in the position required by the intended interpretation. A response was coded as correct when at least four of the five judges agreed. Inter-rater reliability was calculated using Fleiss' kappa.

Table 1. Corpus construction, annotation, and data split.

Dimension	Description
Corpus type	Self-constructed Mandarin read-speech corpus
Corpus size	5,000 utterances
Corpus composition	250 ambiguous sentence frames × 2 semantic contexts × 10 native Mandarin speakers
Annotation	Binary stress-position annotation by five trained native Mandarin-speaking annotators
Reliability	Fleiss' $\kappa = 0.81$
Data split	Speaker-level split: 3,500 training / 500 validation / 1,000 test utterances

Note. The corpus consisted of read speech rather than natural or synthetic speech. Final labels were determined by majority vote, with unresolved disagreements adjudicated by an expert phonetician.

For automated assessment, a wav2vec 2.0 Base model was fine-tuned to classify the target sentence-stress position. In this study, stress-position classification was defined as a binary task: for each ambiguous sentence frame, the model classified whether the primary sentence stress was realized on the first or the second predetermined candidate position. The fine-tuning corpus was a self-constructed Mandarin read-speech corpus containing 5,000 ambiguous-sentence

utterances. Sentence design was informed by previous work on Mandarin stress-related ambiguity and prosodic focus, while the annotation protocol was guided by Mandarin prosodic annotation conventions, particularly the C-ToBI framework [4,10,18]. Details of the corpus construction, annotation procedure, inter-annotator reliability, and data split are summarized in Table 1.

The wav2vec 2.0 Base encoder was followed by a linear classification layer for binary stress-position classification. The main fine-tuning configuration is reported in Table 2. Model performance was evaluated on the held-out test set by comparing predicted stress-position labels with human-annotated labels. The fine-tuned model achieved 91.2% agreement with human raters, with Cohen's kappa = 0.85.

Table 2. Fine-tuning hyperparameters of the wav2vec 2.0 Base model.

Parameter	Setting
Model architecture	wav2vec 2.0 Base
Task	Binary classification of target sentence-stress position
Learning rate	2×10^{-5}
Batch size	16
Epochs	20
Optimizer	AdamW
Loss function	Cross-entropy loss

After fine-tuning, the classifier was applied to the experimental recordings. A response was coded as correct by the automated method when the predicted stress position matched the stress position associated with the intended interpretation. For the final analysis, only responses judged correct by both human raters and the automated classifier were treated as accurate. Discrepancies between the two methods were examined separately and are reported in the Results section.

3. Results

3.1. Experiment 1

A two-way ANOVA was conducted with language background and semantic context as independent variables and sentence-stress placement judgment accuracy as the dependent variable. The results showed a significant main effect of language background, $F\ (1, 22) = 139.37, p < 0.01$, indicating that native Mandarin speakers outperformed advanced English-speaking learners of Chinese. The main effect of semantic context was also significant, $F(1, 22) = 52.19, p < 0.01$, with higher accuracy in primary-meaning contexts than in secondary-meaning contexts. The interaction between language background and semantic context was significant, $F\ (1, 22) = 19.34, p < 0.01$.

Simple-effects analyses revealed that native speakers performed significantly better than advanced learners in both primary-meaning and secondary-meaning contexts, with both p-values less than 0.01. Advanced learners demonstrated significantly higher accuracy in primary-meaning contexts compared to secondary-meaning contexts ($p < 0.01$), while native speakers exhibited no significant difference between the two contexts ($p < 0.58$). These results suggest that advanced learners' sentence-stress judgments were more influenced by semantic context, whereas native speakers maintained consistently high accuracy throughout. Refer to Figure 1 for a visual representation of these findings.

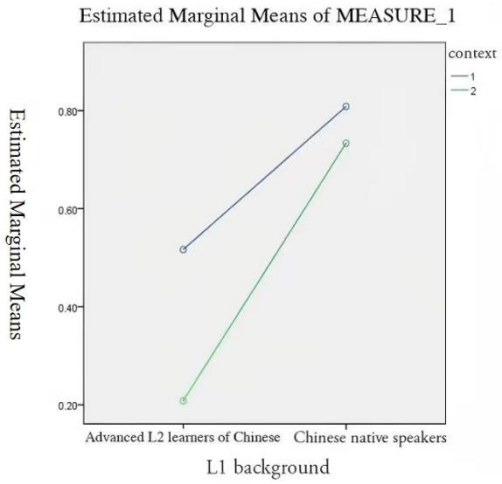


Figure 1. Accuracy rates of sentence stress placement judgment among learners with varying Chinese proficiency levels.

The fine-tuned wav2vec 2.0 model showed 91.2% agreement with human raters, corresponding to an 8.8% disagreement rate, with Cohen's $\kappa = 8.5$. The model-based results were consistent with the human-rating results: advanced learners showed higher accuracy in primary-meaning contexts than in secondary-meaning contexts, $p < 0.01$, whereas native speakers showed no significant context effect, $p < 0.62$. Disagreements mainly involved low-intensity utterances or utterances with non-modal voice quality and were examined separately.

3.2. Experiment 2

A mixed ANOVA was conducted with test type, pre-test vs. post-test, and training method, stress-position reinforcement vs. general instruction, as independent variables and sentence-stress placement accuracy as the dependent variable. The stress-position reinforcement group improved from 62.3% $SD = 8.45$ in the pre-test to 66.8% $SD = 9.12$ in the post-test, while the general instruction group improved from 61.8% $SD = 7.98$ to 64.5% $SD = 8.76$.

The main effect of test type was not significant, $F\ (1, 18)\ =\ 1.67, p\ =\ 2.13$. The main effect of training method was also not significant, $F\ (1, 18)\ =\ 0.04\ , p\ =\ 8.35$, nor was the interaction between test type and training method, $F\ (1, 18)\ =\ 0.10, p\ =\ 7.51$. These results indicate that stress-position reinforcement instruction did not produce significantly greater improvement than general instruction.

The wav2vec 2.0-based assessment showed 90.5% agreement with human raters, corresponding to a 9.5% disagreement rate, with Cohen's $\kappa\ =\ 8.3$. The model-based analysis produced the same pattern of results, with no significant main effect of test type, $F\ (1, 18)\ =\ 1.54, p\ =\ 2.30$, no significant main effect of training method, $F\ (1, 18)\ =\ 0.04, p\ =\ 8.47$, and no significant interaction, $F\ (1, 18)\ =\ 0.10, p\ =\ 7.59$. Thus, both human and automated assessments suggest that instruction focused only on stress-position reinforcement was insufficient to significantly improve beginner learners' sentence-stress production accuracy.

4. Discussion

Sentence stress plays an important role in spoken communication because it helps speakers highlight intended meanings and reduce ambiguity. In Mandarin Chinese, different stress placements may lead to different interpretations of the same sentence. For learners of Chinese as a second language, inappropriate sentence stress may therefore result in utterances that are grammatically correct but pragmatically unclear or semantically misleading. The present study examined English-speaking learners' ability to identify and produce Mandarin sentence stress and further explored the feasibility of using a wav2vec 2.0-based model for automated assessment.

4.1. Semantic Context Effects on Sentence Stress Judgment

Experiment 1 showed that semantic context affected advanced English-speaking learners and native speakers differently. Native Mandarin speakers showed stable performance across primary and secondary semantic contexts, suggesting that they were able to use sentence stress flexibly to disambiguate meaning. In contrast, advanced English-speaking learners performed better in primary contexts than in secondary contexts, indicating that they still had difficulty using sentence stress to identify less frequent or less salient interpretations of ambiguous sentences.

One possible explanation concerns input frequency and usage experience. Usage-based accounts of second language acquisition emphasize that linguistic knowledge is shaped by repeated exposure to form–meaning mappings [6,13]. Primary semantic contexts are likely to

be more frequent and familiar in learners' input, whereas secondary contexts are less commonly encountered. As a result, learners may rely more heavily on dominant interpretations and have greater difficulty using prosodic cues to access alternative meanings. Native speakers, by contrast, have accumulated richer experience with context-dependent stress patterns and can therefore identify the intended meaning more accurately even in secondary contexts.

These findings suggest that advanced learners' difficulty does not lie only in perceiving acoustic prominence, but also in mapping stress placement onto contextually appropriate meanings. Sentence stress instruction should therefore not be limited to prosodic form alone; it should also strengthen learners' awareness of the semantic and pragmatic functions of stress.

4.2. Deep Learning-Based Automated Assessment

A methodological contribution of this study is the use of a wav2vec 2.0-based model to assess L2 Mandarin sentence stress. The model showed high agreement with human raters, reaching 91.2% agreement in Experiment 1, with Cohen's $\kappa = 8.5$, and 90.5% agreement in Experiment 2, with Cohen's $\kappa = 8.3$. These results suggest that the automated assessment was broadly consistent with human judgment and can provide a reliable supplementary tool for evaluating sentence-stress placement.

Compared with traditional perceptual judgment alone, the automated model offers several advantages. It provides consistent scoring across large numbers of responses and reduces potential rater fatigue or rater drift. It also makes it possible to examine systematic disagreement between human and machine judgments. In the present study, disagreement cases were mainly associated with low-intensity utterances or non-modal voice quality, where human raters tended to be more tolerant than the model. These cases indicate that current speech models remain sensitive to atypical voice quality and weak acoustic cues, which should be considered in future model improvement.

More broadly, the findings demonstrate the feasibility of applying deep learning-based assessment to L2 prosodic features. While automated pronunciation assessment has often focused on segmental accuracy or lexical tone, the present results suggest that sentence-level prosodic features such as stress placement can also be evaluated automatically with reasonable reliability.

4.3. Pedagogical Intervention for Sentence Stress

Experiment 2 examined whether explicit reinforcement of sentence-stress position could improve beginner learners' production of Mandarin sentence stress. The results showed no

significant main effect of test type, no significant main effect of training method, and no significant interaction between test type and training method. This indicates that the sentence stress-position reinforcement method did not lead to significant improvement from pre-test to post-test.

Several factors may explain this null effect.

First, stress-based ambiguity may have been too difficult for beginner learners within a short instructional period. Learners had to process sentence meaning, identify the intended focus, and produce appropriate prosodic cues at the same time, which may have exceeded their processing capacity.

Second, the training may have focused too narrowly on stress location without sufficient practice of the acoustic cues that realize stress, such as pitch movement, duration, and intensity.

Third, the total training time was limited, with only four short sessions, which may not have been enough to produce measurable gains.

These results suggest that sentence stress instruction for beginner learners should be more gradual and integrated. Rather than only asking learners to identify where stress should fall, future instruction should combine stress-position awareness with acoustic training, including pitch contour, duration lengthening, and intensity control. Longer and more intensive training may also be necessary before clear production gains can be observed.

4.4. Implications for AI-Driven CAPT Systems

The findings have implications for the development of AI-driven Computer-Assisted Pronunciation Training systems. Most current CAPT tools for Mandarin focus mainly on segmental pronunciation or lexical tone, while sentence-level prosodic features receive less attention. The wav2vec 2.0-based model used in this study could be incorporated into CAPT systems to provide automatic feedback on sentence-stress placement.

Such advanced systems could effectively detect whether learners appropriately place stress on the intended sentence positions. They can also display visual feedback on pitch and duration patterns, while intelligently adjusting task difficulty according to individual learner performance. For instance, training could commence with primary semantic contexts, gradually introducing secondary contexts as learners become more familiar with the intricate relationship between stress placement and sentence meaning. This type of adaptive feedback may significantly enhance learners' ability to connect prosodic form with communicative function in a more effective and meaningful way. Ultimately, this approach aims to foster deeper

understanding and improve overall communication skills.

4.5. Limitations and Future Directions

Several limitations should be noted. First, the sample sizes in both experiments were relatively small, which may limit the generalizability of the findings. Second, the wav2vec 2.0 model was fine-tuned on read speech produced by a limited number of speakers; its performance on spontaneous speech, noisier recordings, or more diverse accents remains to be tested. Third, the intervention in Experiment 2 was brief, and longer-term training may be needed to determine whether sentence-stress instruction can produce stable improvement. Fourth, the study focused only on English-speaking learners of Chinese, so future research should examine whether similar patterns appear among learners from other first-language backgrounds.

Future studies could address these limitations by recruiting larger and more diverse samples, expanding the speech corpus used for model training, and testing the automated assessment model in classroom-based CAPT systems. Longitudinal intervention studies are also needed to examine whether integrated training of stress position, pitch, duration, and intensity can more effectively improve learners' sentence-stress production.

5. Conclusion

This study investigated English-speaking learners' acquisition of Mandarin sentence stress from two perspectives: their ability to identify sentence-stress positions in different semantic contexts and the effectiveness of strengthened instruction on sentence-stress positioning. The results of Experiment 1 showed that advanced learners were able to identify sentence-stress positions relatively accurately in primary-meaning contexts, but their performance declined in secondary-meaning contexts. In contrast, native Mandarin speakers showed consistently high accuracy across both context types. This finding suggests that semantic context plays a more decisive role in learners' stress judgment than in native speakers' judgment, and that secondary-meaning contexts remain particularly challenging even for advanced learners.

Experiment 2 further showed that instruction focusing only on stress-position reinforcement did not significantly improve beginner learners' sentence-stress production accuracy. This result indicates that sentence-stress acquisition cannot be achieved merely by teaching learners where stress should be placed. Accurate production also requires control of relevant prosodic features, including pitch, duration, and intensity. Therefore, sentence-stress instruction should combine explicit explanation of stress placement with auditory discrimination, imitation,

acoustic feature training, and communicative practice.

Methodologically, this study integrated a wav2vec 2.0-based automated assessment model into the evaluation of Mandarin sentence stress. The model showed high agreement with human raters, reaching 91.2% in Experiment 1 and 90.5% in Experiment 2. This suggests that deep learning-based assessment can provide a reliable supplementary tool for evaluating L2 Mandarin sentence-stress production. Disagreement cases were mainly associated with low-intensity utterances or non-modal voice quality, indicating directions for future improvement of the model.

Pedagogically, the findings suggest that sentence-stress instruction should not be limited to beginner-level phonetic training. Instead, it should be incorporated throughout the Chinese language curriculum, especially in intermediate and advanced courses involving grammar, discourse, and oral communication. Teaching materials should also include more activities that connect stress placement with semantic focus and communicative intention.

The study also has implications for AI-driven Computer-Assisted Pronunciation Training systems. The validated automated assessment model could be used to provide real-time feedback on learners' sentence-stress placement and to support adaptive training tasks. Future research should expand the participant sample, include more diverse speech data, improve model robustness, and examine the effectiveness of AI-assisted sentence-stress training in classroom-based longitudinal studies.

In conclusion, English-speaking learners of Chinese still face difficulties in using semantic context to determine Mandarin sentence stress, particularly in secondary-meaning contexts. Instruction that focuses only on stress location is insufficient; more comprehensive prosodic training is needed. At the same time, deep learning-based automated assessment offers a promising approach for evaluating and supporting the acquisition of Mandarin sentence stress.

Acknowledgements

This work was supported by the 2023 Special Project on the Cultivation of Top Innovative Talents under the "14th Five-Year Plan" for Educational Science in Chaoyang District, Beijing (Grant No. 2023ZX009).

References

[1] Baevski, A., Zhou, Y., Mohamed, A., & Auli, M. (2020). wav2vec 2.0: A framework for self-supervised learning of speech representations. *Advances in Neural Information Processing*

Systems, 33, 12449–12460.

[2] Cao Wen. (2010). 汉语焦点重音的韵律实现 [The prosodic realization of focus stress in Chinese]. *Beijing: Beijing Language and Culture University Press*. [in Chinese]

[3] Chen Yongming, & Cui Yao. (1997). 汉语歧义句的加工 [The processing of ambiguous sentences in Chinese]. *Acta Psychologica Sinica*, (1). [in Chinese]

[4] Chen, Y., & Gussenhoven, C. (2008). Emphasis and tonal implementation in Standard Chinese. *Journal of Phonetics*, 36(4), 724–746.

[5] Cheng Shuqiu. (2005). 句重音在对外汉语语法教学中的作用 [The role of sentence stress in teaching Chinese grammar as a foreign language]. *Continuing Education Research*, (4), 126–128. [in Chinese]

[6] Ellis, N. C. (2002). Reflections on frequency effects in language processing. *Studies in Second Language Acquisition*, 24(2), 297–339.

[7] Ellis, R. (1994). The study of second language acquisition. Oxford: Oxford University Press.

[8] Feng Shengli. (2009). 汉语的韵律、词法与句法：修订本 [Prosody, morphology, and syntax in Chinese: Revised edition]. *Beijing: Peking University Press.* [in Chinese]

[9] Flege, J. E., & Eefting, W. (1987). Production and perception of English stops by native Spanish speakers. *Journal of Phonetics*, 15, 67–83.

[10] Gao, J., & Gu, Y. (2024). Same sentences, different meanings: Prosodic and gestural resolution of ambiguity in Mandarin Chinese. In Proceedings of Speech Prosody 2024 (pp. 886–890). *International Speech Communication Association*.

[11] Ji Xiuqing. (1998). 语句重音与口语教学 [Sentence stress and oral Chinese teaching]. In 对外汉语语音与语音教学研究 [Studies on Chinese phonetics and phonetics teaching as a foreign language] (pp. 357–368). *Beijing: Sinolingua Press*. [in Chinese]

[12] Lu Jianji. (1984). 中介语理论与外国人学习汉语的语音偏误分析 [Interlanguage theory and an analysis of phonetic errors in foreigners' learning of Chinese]. *Language Teaching and Linguistic Studies*, (3). [in Chinese]

[13] MacWhinney, B. (2006). Emergentism: Use often and with care. *Applied Linguistics*, 27(4), 729–740.

[14] Qiu Shanshan. (2007). 日本留学生汉语陈述句核心重音的韵律表现 [The prosodic realization of core stress in Mandarin declarative sentences by Japanese international students]. *Master's thesis, Beijing Language and Culture University*. [in Chinese]

[15] Wang Yunjia. (2002). 汉语语音研究与汉语语音教学接口的问题 [Issues at the interface between Chinese phonetic research and Chinese pronunciation teaching]. In 对外汉语论丛（第二辑） [Essays on teaching Chinese as a foreign language, Vol. 2]. *Shanghai: Shanghai Foreign Language Education Press*. [in Chinese]

[16] Wang Yunjia, Chu Min, & He Lin. (2006). 汉语焦点重音和语义重音分布的初步试验研究 [A preliminary experimental study of the distribution of focus stress and semantic stress in Chinese]. *Chinese Teaching in the World*, (2). [in Chinese]

[17] Witt, S. M., & Young, S. J. (2000). Phone-level pronunciation scoring and assessment for interactive language learning. *Speech Communication*, 30(2–3), 95–108.

[18] Xu, Y. (1999). Effects of tone and focus on the formation and alignment of F0 contours. *Journal of Phonetics*, 27(1), 55–105.

[19] Yang Defeng. (1991). 重音与语义：句子重音 [Stress and semantics: Sentence stress]. In 第三届国际汉语教学讨论会论文选 [Selected papers from the Third International Conference on Chinese Language Teaching]. *Beijing: Peking University Press*. [in Chinese]

[20] Zhang Juan. (2009). 美国留学生汉语陈述句核心重音的韵律表现研究 [A study on the prosodic realization of core stress in Mandarin declarative sentences by American international students]. *Master's thesis, Beijing Language and Culture University*. [in Chinese]

[21] Zhu Chuan. (Ed.). (1997). 汉语语音学习对策 [Strategies for learning Chinese phonetics]. *Beijing: Yuwen Press*. [in Chinese]

Online ISSN: 3104-9338
Print ISSN: 3104-932X

Research on Intelligent Psychological Consultation Method Based on Dynamic Scale Embedding and Dialogue Analysis

Chenlong Xie, Ao Feng*, Zhilei Zhang

School of Computer Science, Chengdu University of Information Technology, Chengdu, China, 610225

Received: June 1, 2026

Revised: June 2, 2026

Accepted: June 5, 2026

Published online: June 11, 2026

To appear in: *International Journal of Advanced AI Applications,* Vol. 2, No. 7 (July 2026)

* Corresponding Author: Ao Feng (fengao@cuit.edu.cn)

Abstract. Mental health service demands in China keep rising, while traditional psychological counseling is hampered by uneven resource allocation, high costs, delayed responses and inconsistent practitioner expertise. Current intelligent counseling systems mainly adopt rule matching or basic model fine-tuning, leading to rigid dialogue, insufficient professionalism and poor capacity for active assessment and intervention. This paper presents an intelligent psychological counseling approach integrating dynamic scale embedding and dialogue analysis. Its technical framework covers dynamic scale embedding, standardized treatment plan generation, and the collaboration of fine-tuned models and Retrieval-Augmented Generation (RAG). By analyzing dialogue rhythm and semantics, the method embeds psychological scales dynamically via mutual information thresholds to realize unobtrusive assessment. Combined with user portraits and professional knowledge bases, it can automatically produce standardized treatment schemes. The joint use of domain fine-tuning and RAG also improves the empathy, guidance and professionalism of system responses. Experiments show the proposed method surpasses conventional methods in empathy, guidance, professionalism, fluency and reasoning efficiency. It supports low-cost, accessible and standardized intelligent psychological counseling, with great practical value and broad application prospects.

Keywords: *Intelligent Psychological Consultation; Dynamic Scale Embedding; Dialogue Analysis; Retrieval-augmented Generation; Standardized Treatment Plan*

1. Introduction

Currently, accelerated social pace and intensified competitive pressures have rendered public mental health issues increasingly severe. Statistics released by the Disease Prevention and Control Bureau of the National Health Commission indicate that there are 240 million individuals suffering from mental disorders across China, corresponding to an overall prevalence rate of 17.5% [1]. Nevertheless, traditional offline face-to-face psychological counseling suffers from multiple inherent limitations, including imbalanced resource allocation where high-quality counseling resources are predominantly concentrated in megacities, excessive service charges with single-session consultation fees varying from hundreds to over a thousand-yuan, complicated appointment procedures and delayed crisis intervention responses. In addition, the uneven professional competence among psychological counselors further undermines the quality and security of counseling services.

In recent years, the rapid advancement of Large Language Models (LLMs) has unlocked new prospects for intelligent psychological counseling research. Equipped with superior natural language comprehension and text generation capabilities, LLMs possess great potential to overcome the temporal, spatial and cost constraints inherent in conventional manual psychological counseling. Even so, current LLM-enabled psychological counseling systems still confront prominent limitations. Early systems built upon rule matching mechanisms and fixed dialogue scripts suffer from poor flexibility and insufficient emotional cognition capability. Models merely fine-tuned within narrow domains fail to acquire stable professional psychological knowledge, thus being highly susceptible to factual hallucinations, i.e., the generation of superficially plausible yet factually incorrect content [2]. Most existing systems mandate users to complete psychological scales at preset stages, resulting in stiff human-computer interaction, disruption of coherent dialogue flows, and a lack of capacity for proactive risk warning and dynamic psychological assessment.

To address the above limitations of existing approaches, this paper proposes and develops an intelligent psychological counseling system integrated with dynamic scale embedding, collaborative reasoning and standardized therapeutic scheme generation. The core contributions of this work are summarized as follows:

- A dynamic psychological scale embedding strategy is proposed. This strategy determines the optimal embedding moment of psychological scales via mutual information threshold constraint, enabling imperceptible and low-intrusive real-time psychological state assessment.

- A hybrid framework integrating fine-tuned LLMs and RAG-based collaborative reasoning is established. It harmonizes the empathic linguistic expression of LLMs and the precise delivery of professional psychological knowledge, thereby substantially mitigating factual hallucinations.
- A full-cycle service paradigm consisting of dialogue interaction, psychological scale evaluation, user psychological profiling and therapeutic scheme formulation is constructed. The proposed system can autonomously generate personalized and standardized intervention schemes, achieving end-to-end integration from psychological assessment to clinical intervention.
- Lightweight system deployment is realized. By adopting parameter-efficient fine-tuning and inference acceleration strategies, the proposed system effectively cuts down GPU memory occupancy and boosts the overall efficiency of model training and inference.

2. Related Work

2.1. Research Background and Current Situation

The shortage of high-quality psychological counselor resources and high service costs make it difficult for the general population to obtain continuous and professional intervention services. Meanwhile, affected by the "stigma" in social culture, many potential users are unwilling to take the initiative to seek help, which leads to the continuous aggravation of psychological problems. Intelligent psychological counseling systems can provide 7×24-hour uninterrupted services, effectively lower the usage threshold, realize early risk warning and timely intervention, and possess important social value and application prospects.

From the perspective of technological evolution, intelligent psychological counseling systems have gone through three main stages. The rule matching-based stage: the system responds through preset keywords and dialogue templates, which is poor in flexibility, single in reply mode, and incapable of handling complex emotions and open-ended questions [3]. The simple fine-tuning-based stage: researchers fine-tune pre-trained models with a small amount of psychological counseling dialogue data. Although the fluency is improved, the models lack the support of professional knowledge bases, present weak guidance ability and professional response ability, and are prone to hallucinations. The retrieval-augmented generation-based stage: RAG technology assists content generation by retrieving external knowledge bases and has been widely applied in knowledge-intensive scenarios. Nevertheless, there are still few studies that deeply combine RAG with dynamic scales and multi-turn dialogue analysis in the

field of psychological counseling, and mature integrated "assessment-intervention" solutions have not yet been formed [4].

2.2. Research Objectives and Technological Innovations

Targeting the existing limitations of current intelligent psychological counseling systems in emotional interaction, professional response generation, dynamic psychological evaluation and integrated service loop construction, this paper formulates four research objectives and correspondingly puts forward four pivotal technological innovations.

(1) Dynamic and low-intrusive embedding of psychological scales. This study abandons the conventional rigid paradigm of mandatory scale filling at fixed dialogue nodes, and establishes a dynamic triggering mechanism driven by dialogue rhythm and semantic comprehension. Psychological assessment scales can be seamlessly integrated into conversational interactions in an unobtrusive manner, enabling high-quality mental state evaluation without disrupting user interactive experience.

(2) Synergistically enhance system empathy, conversational guidance and professional response competence. This work addresses the integrated drawbacks including unnatural empathic expression, insufficient guiding ability and inaccurate professional knowledge application of prevailing models in Chinese psychological counseling scenarios. The proposed system can satisfy practical deployment requirements in both emotional comforting and rational psychological intervention.

(3) Automatic generation of standardized and executable personalized intervention schemes. This work achieves the technical advancement from general suggestive guidance to targeted precise intervention. Leveraging multi-dimensional user psychological portraits and authoritative professional knowledge repositories, the system is capable of generating clinical-compliant personalized intervention schemes tailored to individual mental characteristics.

(4) Construction of low-cost, accessible and large-scale universal psychological counseling services. By adopting parameter-efficient fine-tuning and inference acceleration techniques, the hardware resource dependency of the model is effectively reduced and system response latency is shortened. Accordingly, high-quality intelligent psychological counseling services can break the constraints of hardware costs and regional restrictions, facilitating large-scale deployment and widespread popularization.

To fulfill the aforementioned research objectives, the primary innovative contributions of this paper are summarized as follows:

(1) A novel dynamic scale embedding mechanism is proposed. A mutual information threshold-based decision engine is elaborately designed, and the Q-learning framework [5] is adopted to optimize scale embedding timing. A multi-objective reward function considering assessment completion rate, interaction disturbance level, analytical insight, user engagement and clinical practical value is further introduced. This mechanism enables precise timing selection, adaptive scale matching and unobtrusive user interaction, fundamentally resolving the conflict between traditional scale evaluation and coherent dialogue flow.

(2) A collaborative reasoning framework integrating fine-tuned language models and RAG is constructed. To compensate for the deficiencies that standalone fine-tuned models lack sufficient professional psychological knowledge and pure RAG architectures are devoid of humanistic emotional perception, this framework designs a dual-branch information processing architecture and a central fusion decision mechanism. Equipped with a dynamic weight adjustment function, it adaptively balances the output ratio of empathic dialogue content and specialized professional knowledge according to dialogue turns, psychological crisis intensity and consultation professionalism, realizing organic integration of emotional expression and domain-specific knowledge.

(3) A full-cycle closed-loop service workflow covering dialogue interaction, scale evaluation, user profiling and intervention scheme formulation is established. The proposed system integrates all functional modules ranging from natural conversational communication, real-time dynamic assessment and user psychological profile construction to automatic standardized intervention planning. It possesses comprehensive capacities including psychological perception, status evaluation, mental state judgment and targeted intervention, accomplishing the intelligent evolution from passive emotional companionship to active psychological intervention.

(4) The lightweight model optimization has been successfully implemented through an innovative three-stage progressive fine-tuning strategy, which is enhanced by a series of effective parameter optimization methods. This approach not only retains the full core performance of the original model but also achieves a remarkable reduction in GPU memory occupancy by approximately 22.4%. Additionally, it significantly cuts down training overhead by around 21.5% when compared to the classic LoRA method. Furthermore, the incorporation of optimized inference and decoding strategies has led to an impressive boost in text generation speed, reaching up to 180.92 characters per second. This advancement lays a solid technical foundation for low-cost system deployment and ensures seamless real-time human-computer

interaction, ultimately enhancing user experience and system efficiency.

3. Methodology

3.1. Overall System Architecture

The intelligent psychological counseling system proposed in this paper is designed with a hierarchical architecture, which is divided into four core functional layers as follows:

Data and Knowledge Base Layer: This layer stores desensitized real-world multi-turn conversational datasets, a psychological scale repository covering mainstream professional scales including SAS, SDS and PSS, as well as domain-specific psychological knowledge bases compiled from authoritative sources such as psychological monographs and clinical diagnostic guidelines.

Input Processing Layer: It undertakes the reception of user textual inputs and knowledge retrieval requests, and performs feature extraction and collaborative reasoning to maximize the inherent capability of the foundational model.

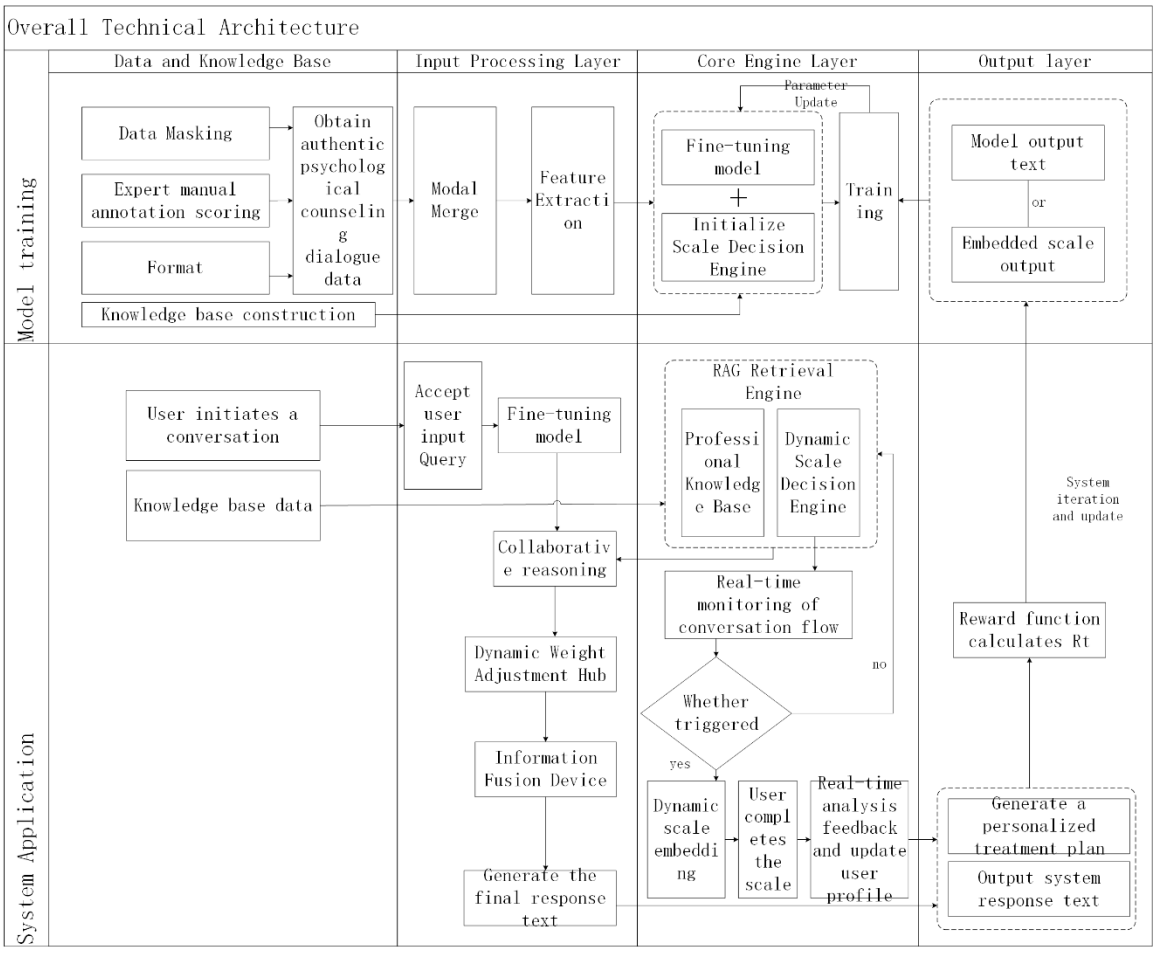


Figure 1. System Technical Architecture and Workflow Diagram.

Core Engine Layer: This layer integrates two pivotal functional engines, i.e., the scale decision engine and the RAG retrieval engine, which jointly realize dynamic judgment and analytical reasoning for scale embedding deployment [6].

Output Layer: It enables diverse human-computer interaction functions, including text-voice multimodal conversation, online scale completion, intervention scheme generation and mental health risk alert.

The overall technical framework and internal data flow of the presented system are depicted in Figure 1.

3.2. Data Collection and Expert Annotation

Given the shortage of high-quality annotated corpora in psychological counseling research, this study establishes a standardized pipeline covering data collection, filtering and manual annotation, so as to ensure the professionalism, diversity and credibility of acquired training data.

Raw conversational data are collected from authentic counseling records provided by cooperative psychological counseling institutions, covering prevalent psychological issues including anxiety, depression, interpersonal conflicts and stress regulation. All collected data are strictly processed via data desensitization in accordance with the following three core principles:

(1) Minimization Principle. Only privacy-sensitive information such as full names, contact information and specific residential addresses is desensitized, so that the semantic integrity of original dialogues can be maximally preserved.

(2) Irreversibility Principle. One-way encryption and data masking strategies are adopted to make desensitized data incapable of being restored to original versions via conventional technical means.

(3) Consistency Principle. The original dialogue format, temporal logic and textual structure remain unchanged after desensitization, which guarantees the usability of processed data for subsequent annotation tasks and model training procedures.

After acquiring desensitized raw data, domain psychological experts carry out data filtering and collation based on unified criteria. Conversational clips that conform to the logical flow of user utterance → counselor reply → user feedback → further counselor response are screened out, enabling each data sample to record the complete reasoning process from explicit emotional expression to implicit psychological demand mining. The turn number of each independent

dialogue is restricted within 3 to 8 turns, striking a good balance between sufficient contextual information and practical training efficiency.

In this work, standardized utterance templates for users and response templates for professional counselors are constructed separately. User-side templates simulate emotional expressions with distinct intensity levels. For example, mild anxiety is expressed as I have poor sleep quality recently, while severe anxiety is described as I cannot fall asleep all night, accompanied by rapid heartbeat and extreme mental breakdown. In contrast, counselor response templates adopt a unified paradigm of empathic recognition → conversational guidance → professional intervention suggestions. A typical case is illustrated as follows: I can perceive your current negative mood (empathy). How long have you been trapped in such a state (guidance)? Persistent insomnia is closely correlated with anxious emotion, and you can start with simple breathing relaxation training for relief (professional suggestion).

Combined with practical field research on mainstream counseling scenarios, the established template library covers 16 daily life scenarios including workplace pressure, family interpersonal tension and academic anxiety, together with 28 typical emotional states such as anxiety, depression, rage and loneliness. In total, 448 standardized data templates targeting representative counseling scenarios are constructed, which effectively enrich data diversity and strengthen the generalization performance of trained models.

All annotation tasks are completed collaboratively by three practitioners holding officially recognized national psychological counseling qualifications. The core annotation dimensions and corresponding five-level scoring criteria are specified as follows:

Empathy Performance: This dimension evaluates whether counselor responses contain valid empathic content and precisely match the category and intensity of users' actual emotions. Score 1 refers to absence of empathy or obvious emotional mismatch; Score 3 indicates basic empathic phrases are used yet fail to fit real emotional intensity; Score 5 represents refined scenario-based empathic expressions that perfectly align with users' inner emotional conditions.

Conversational Guidance Ability: It measures whether designed replies contain guiding inquiries and can drive dialogues to dig out deep-seated psychological needs beyond superficial emotions. Score 1 means offering direct suggestions without any guiding content; Score 3 denotes available guiding questions that cannot further deepen thematic discussion; Score 5 signifies adopting progressive questioning strategies to explore underlying demands and naturally introduce professional psychological perspectives.

Domain Professionalism: It assesses the standardization of psychological terminology usage

and compliance with formal counseling ethical norms. Score 1 stands for incorrect terminology application or violation of basic ethics such as negating users' subjective feelings; Score 3 refers to correct use of elementary professional terms without ethical conflicts; Score 5 indicates accurate and normative terminology usage as well as full ethical compliance, including respecting individual emotional cognition and avoiding compulsory intervention advice.

In the annotation phase, all qualified annotators finish independent scoring separately. For samples with score differences larger than 1 point, all three annotators conduct collective discussion to reach a consistent evaluation result. Furthermore, 10% of all annotated samples are randomly selected to compute the Kappa coefficient, ensuring that the final inter-annotator agreement coefficient is no less than 0.7.

3.3. Base Model Selection and Fine-Tuning

The choice of pre-trained base models fundamentally determines the performance ceiling of fine-tuned models and the overall deployment cost. In view of the practical characteristics of Chinese psychological counseling scenarios, this study evaluates candidate models from three core dimensions: Chinese language compatibility, GPU memory consumption and practical inference efficiency.

Targeted at Chinese conversational psychological counseling scenarios, the qualified models are required to possess powerful capabilities in Chinese semantic understanding, fine-grained emotion recognition, and perception of implicit emotional cues conveyed by colloquial modal particles. Accordingly, only pre-trained models originally optimized for Chinese language are selected as candidate alternatives.

Considering the GPU memory limitations of practical deployment platforms including local servers and cloud computing instances, this experiment adopts RTX 3090 graphics cards with 24 GB physical memory. It is necessary to strike a reasonable trade-off between model expressive capability and hardware deployment cost. Models supporting mainstream quantization strategies such as INT4 and FP8 are preferentially selected to lower practical deployment thresholds.

High interactivity is an essential requirement for psychological counseling dialogue systems, hence the end-to-end response latency must be maintained within user-acceptable ranges. In this study, we set a practical optimization target that the average generation latency of responses with less than 500 Chinese characters is controlled within 3 seconds, and further evaluate the token generation speed of candidate models under designated hardware configurations.

Based on the above practical constraints, this work conducts comprehensive preliminary assessment on three prevailing open-source Chinese large language models, namely ChatGLM3-6B [7], Baichuan2-7B [8] and Qwen-7B [9]. The experimental results demonstrate that all three models achieve satisfactory Chinese language adaptation performance. In particular, ChatGLM3-6B features lower GPU memory usage of approximately 14 GB under identical experimental conditions, satisfies the latency requirement for real-time human-computer interaction, and exhibits prominent advantages in response fluency and content safety. Consequently, ChatGLM3-6B is determined as the official base model in this research.

To empower general-purpose pre-trained models with professional psychological counseling capabilities, this paper proposes a three-stage progressive fine-tuning paradigm. The complete multi-turn user-counselor dialogue context is adopted as training corpus, and standard cross-entropy loss function is adopted for model parameter optimization. The detailed experimental configurations for each fine-tuning stage are elaborated below.

Stage 1 (Epoch 1–3): Emotional Interaction Ability Training: This stage mainly focuses on the acquisition of empathic expression and interactive guidance skills, enabling the model to master the most basic interactive logic in psychological counseling scenarios. In the designed weighted loss function, the weight coefficients of empathy and guidance supervision labels are both set to 40%, and the weight of professional knowledge supervision labels is set to 20%. The initial learning rate is configured as 3×10^{-5}, combined with warm-up learning rate scheduling strategy. After the first training stage, the average evaluation scores of empathy and guidance on the validation set can steadily reach above 3.5.

Stage 2 (Epoch 4–6): Comprehensive Capability Balanced Optimization: After the model masters basic empathic interaction and dialogue guiding logic, this stage aims to further integrate and optimize overall counseling capabilities. The loss weight of the three evaluation dimensions is adjusted to an equal ratio of 33.3%, and the learning rate is decayed to 1×10^{-5}. This optimization strategy effectively avoids the tendency that the model excessively pursues emotional resonance while neglecting normative professional replies, and realizes synchronous performance improvement across empathy, dialogue guidance and domain professionalism.

Stage 3 (Epoch 7–8): Targeted Weak Sample Reinforcement Training: Based on quantitative evaluation results on the validation set, samples with evaluation scores lower than or equal to 2 in any single dimension are screened out as low-quality deficient samples. These samples are adopted for targeted supplementary retraining with the initial learning rate of 3×10^{-5}. In this process, only partial network parameters such as top-layer transformer weights are updated,

which can effectively reinforce insufficient capabilities while effectively mitigating the risk of catastrophic forgetting of previously learned interactive rules and professional knowledge.

This study adopts a dual-mode evaluation mechanism integrating automatic quantitative scoring and expert manual review to simultaneously ensure high evaluation efficiency and reliable result accuracy. The DeepSeek-V3 [10] large language model is utilized as the automatic scoring evaluator, which rates model-generated counseling replies on the validation set with a 1–5 scoring scale from empathy, guidance and professionalism three dimensions, and calculates corresponding average scores and standard deviations. With strong comprehensive reasoning and evaluation capability verified in multiple authoritative benchmarks, DeepSeek-V3 is well-suited for large-scale batch automatic scoring tasks and greatly reduces manual evaluation workload of professional experts.

For manual review, an expert team composed of three formally qualified national psychological counselors is organized to conduct random sampling verification on automatic scoring results. Samples with evaluation deviation exceeding 1 point between machine scoring and expert subjective judgment are focused on revising, so as to guarantee the credibility of final evaluation results. The final integrated score of each evaluation dimension is calculated by the following formula.

$$S_{total} = 0.7S_{auto} + 0.3S_{manual}$$

Where S_{total} denotes the integrated comprehensive score, S_{auto} represents the score obtained via automatic model evaluation, and S_{manual} stands for the scoring result given by professional human experts.

A full-scale evaluation on the validation set is performed upon the completion of an entire three-stage fine-tuning cycle. The fine-tuning process will be terminated and the well-trained model will be preserved as the final domain-specific psychological counseling model once the average integrated scores of the three evaluation dimensions all exceed 4.0. If not, the weight allocation scheme and learning rate settings of each training stage will be optimized based on evaluation feedback, and the model will enter the next round of iterative fine-tuning.

3.4. Dynamic Scale Embedding Technique

Dynamic scale embedding serves as the fundamental technique for achieving unobtrusive psychological assessment. Its specific implementation workflow is comprehensively illustrated in Figure 2, providing a clear visual representation of the entire process involved in this innovative approach.

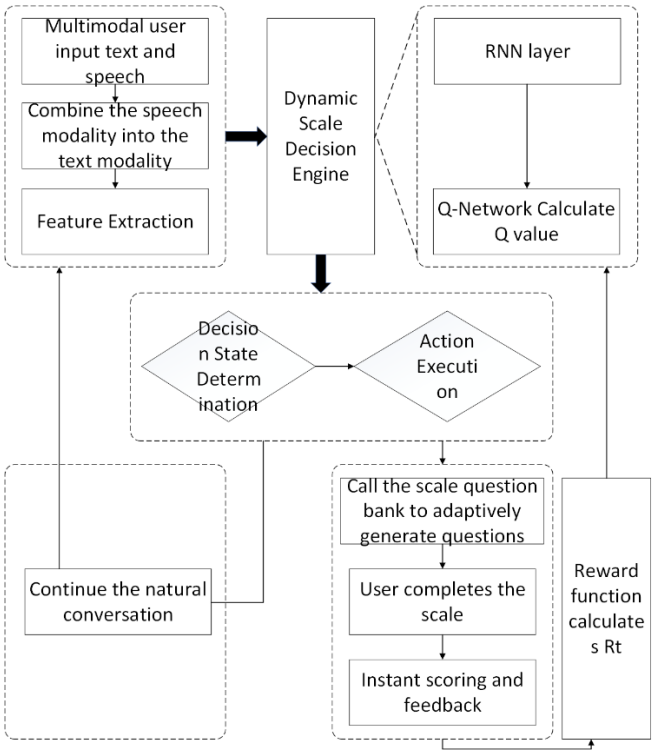


Figure 2. Dynamic Psychological Scale Embedding Framework.

The proposed framework is composed of five functional modules enclosed by dashed boxes, including feature extraction, dynamic scale decision engine, decision status discrimination, scale embedding and subsequent natural dialogue maintenance.

The detailed execution procedures are described as below: Within the dynamic scale decision engine, the Q-value represents the expected cumulative reward, which is formulated as follows:

$$Q(s,a) = E\left[R_t + \gamma R_{t+1} + \gamma^2 R_{t+2} + \ldots \mid S_t = s, A_t = a \right]$$

$Q(s,a)$ denotes the expected cumulative discounted reward when taking action a at state s. E stands for expectation. Given the stochastic nature of responses from the environment (users), the Q-value corresponds to a probabilistic average. γ refers to the discount factor ranging from 0 to 1, which quantifies the emphasis placed on future rewards. A γ value approaching 0 indicates a myopic agent that merely prioritizes immediate rewards; conversely, a γ value close to 1 represents a farsighted agent willing to trade short-term profits for long-term returns.

At each dialogue state s_t, the system computes Q-values for all feasible actions including non-embedding, PHQ-9 embedding, GAD-7 embedding and other alternatives, and subsequently executes the action with the maximum Q-value.

The multi-objective reward function R_t evaluates the performance of action a_t, aiming to strike a balance between clinical validity and user experience. The corresponding formula is

presented below:

$$R_t = \alpha * R_{completion} + \beta * R_{perturbation} + \gamma * R_{insight} + \delta * R_{engagement} + \varepsilon * R_{clinical}$$

$R_{completion}$ denotes the completion reward, and its formula is presented below:

$$R_{completion} = I_{completed} * (1 + tanh(-\lambda * t_{abandon}))$$

Among them, $I_{completed}$ is an indicator function. If the user finishes the scale, it equals 1, otherwise it equals 0. The function of tanh is: the shorter the time $t_{abandon}$ when the user gives up completion, the greater the penalty (the lower the reward). It encourages the system to choose the time when users are more likely to finish.

$R_{perturbation}$ is the perturbation degree penalty, the formula is as follows.

$$R_{perturbation} = -\kappa * sim(s_t, s_{t-1})^{-1}$$

Herein, $sim(...)$ serves as a similarity calculation function between state s_t and s_{t-1}, such as cosine similarity. Lower similarity indicates more severe disruption of dialogue flow, which leads to a higher penalty value of $-\kappa * (...)$. This design prompts the system to insert scales in natural conversational transitions and prevent jarring breaks.

$R_{insight}$ stands for insight-related reward, and its formula is presented below.

$$R_{insight} = |score_t - E[score_{user}]|$$

The insight reward mainly calculates the absolute difference between the current scale score $score_t$ and the user's historical average score $E[score_{user}]$. A larger deviation implies that the system detects a critical clinical variation, either symptom improvement or deterioration, thereby yielding positive rewards.

$R_{engagement}$ denotes the engagement reward, and its formula is displayed below.

$$R_{engagement} = -(sentiment(s_{post}) - sentiment(s_{pre}))$$

The engagement reward quantifies the discrepancy between post-embedding utterance sentiment $sentiment(s_{post})$ and prior sentiment $sentiment(s_{pre})$. A drop in sentiment indicates degraded user experience and yields a negative discrepancy. The negative coefficient reverses this value and imposes a punitive reward. This term motivates the system to sustain users' active engagement and emotional state.

$R_{clinical}$ represents the clinical value reward, and its formula is given below.

$$R_{clinical} = \sigma(score_t - threshold)$$

Herein, σ denotes the Sigmoid function and $threshold$ refers to the clinical cutoff value of

the scale. When $score_t$ surpasses the cutoff, the term $score_t - threshold$ becomes positive, driving the Sigmoid output to surge rapidly from 0.5 to 1. Accordingly, the system can acquire considerable rewards upon identifying high-risk status. This reward component serves as the core incentive to facilitate the system's fundamental clinical efficacy.

3.5. Collaborative Reasoning of Fine-tuned Model and RAG

A collaborative reasoning mechanism integrating fine-tuned model and RAG module is proposed in this work to compensate for the limited domain knowledge of standalone fine-tuned models [11]. The corresponding workflow is illustrated in Figure 3.

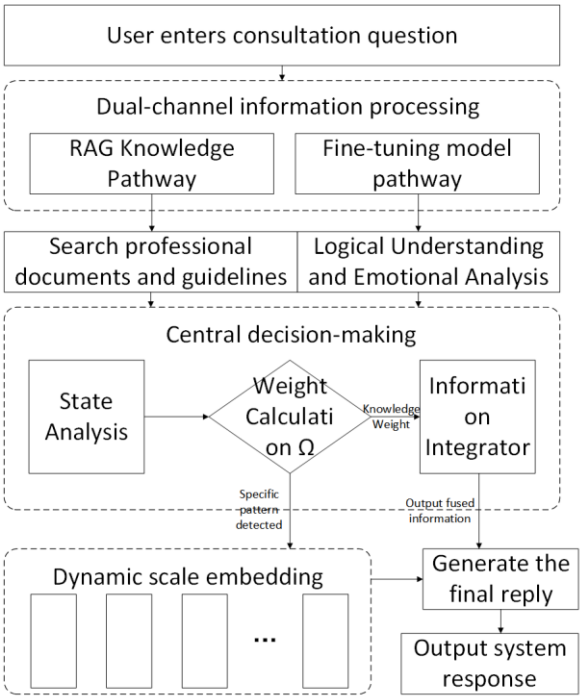


Figure 3. Framework of collaborative reasoning.

The collaborative reasoning framework mainly includes: user input consultation information, dual-path information processing, central decision-making, and system response output. The dynamic weight calculation function Ω in central decision-making analyzes the current dialogue state and determines the proportion of free generation and knowledge base reference. Its output is the knowledge weight α ranging from 0 to 1. The formula is defined as follows:

$$\alpha = \sigma(w_1 * g(t) + w_2 * S + w_3 * K - b)$$

Here, α denotes the knowledge weight that quantifies the proportion of contents retrieved from the RAG knowledge base in final responses. The Sigmoid function σ maps arbitrary values into the range (0,1) to guarantee smooth and steady weight variation. $g(t)$ is a monotonically increasing function regarding dialogue turns, driving the system to leverage

more domain knowledge as the conversation deepens. S refers to the crisis signal score derived from real-time user input analysis. Specifically, S is set to 1.0 for suicidal statements, 0.3 for insomnia-related descriptions, and 0 in the absence of sensitive keywords. K measures question professionalism via similarity calculation between user queries and domain knowledge corpus, and higher specificity and professionalism of queries yield larger K values. w_1, w_2, w_3 are manually predefined hyperparameters, which control the relative contribution of dialogue turn, crisis signal and query professionalism to weight adjustment.

3.6. Generation of Standardized Treatment Plans

Once adequate user information is collected via conversations and scale assessments, the system initiates the generation of standardized treatment schemes. The corresponding procedure is illustrated in Figure 4.

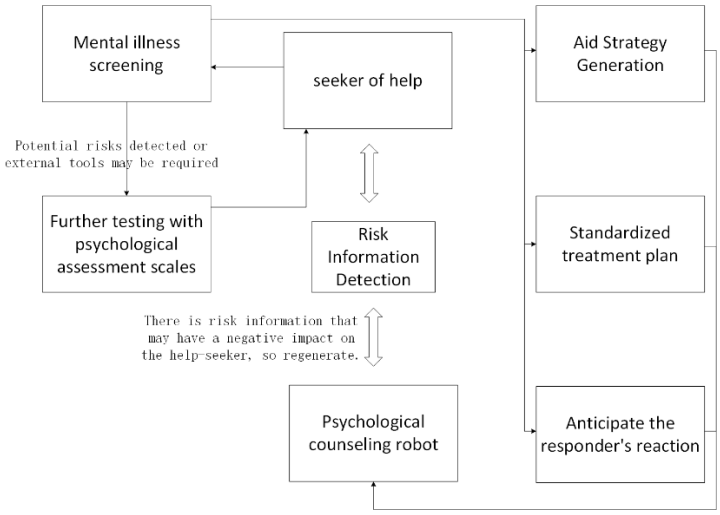


Figure 4. Interactive flowchart of standardized treatment plan.

Multimodal Sentiment Analysis and User Profile Completion. The system persistently identifies users' emotional categories (e.g., anxiety and depression) and corresponding intensities to produce systematic emotional reports. User profiles are structured and stored as knowledge graphs, encompassing static attributes (age and gender) and dynamic attributes (real-time psychological state and symptom severity). Graph-based algorithms are further applied to implement consistency validation and imputation of missing attribute values.

Multidimensional User State Mapping. Integrating scale assessment results, the system projects user status onto a multidimensional evaluation vector that covers multiple clinical dimensions, including depression severity, anxiety severity, stress level, and social function status.

Scheme Retrieval and Personalized Generation. Leveraging the aforementioned multidimensional vector, the RAG module retrieves the most clinically matched treatment schemes and intervention strategies from the domain-specific knowledge base. Subsequently, the fine-tuned model adapts the retrieved generic schemes based on comprehensive user profiles and dialogue contexts, so as to generate natural and practically feasible standardized therapeutic suggestions (e.g., "Based on your current condition, you can try the basic practice of the following cognitive behavioral therapy...").

3.7. Construction and Retrieval Configuration of Knowledge Bases

Two independent knowledge bases are established in this work, namely the scale knowledge base and professional psychological knowledge base.

In terms of scale knowledge base construction, a comprehensive database covering diverse psychological scales across multiple domains is adopted. High authority and reliability are guaranteed for the selected database to secure valid practical application. All selected scales are vectorized to convert textual content into high-dimensional vector representations, facilitating subsequent retrieval and analytical processing. A dynamic embedding-based scale retrieval system is developed to achieve efficient data query. Dynamic embedding enables real-time adjustment of retrieval strategies according to user queries, which enhances the flexibility and adaptability of the retrieval process. This technique also facilitates accurate user intention perception and improves overall retrieval accuracy.

The professional psychological knowledge base is compiled from authoritative psychological textbooks and clinical guidelines. Core contents including intervention approaches and case analysis are extracted from mainstream Chinese monographs, such as Theory and Practice of Counseling and Psychotherapy and Introduction to Clinical Psychology. Diagnostic criteria and standardized intervention workflows are sourced from official documents including Clinical Psychotherapy Guidelines issued by Chinese Mental Health Association and Guidelines for Depression Prevention and Treatment.

Contents of both databases are encoded into high-dimensional vectors by embedding models and stored in vector databases. The configured retrieval parameters are set as TopK=5 to return the top five most relevant knowledge items, with a similarity threshold of 0.7, where items scoring below the threshold are deemed irrelevant. To further optimize retrieval quality, a knowledge-dialogue fusion mechanism is designed. A dedicated knowledge filter performs secondary screening on retrieved results to remove redundant or context-conflicting knowledge snippets. For example, recommendations concerning cognitive behavioral therapy will be

discarded if the user has already received relevant treatment.

4. Experimental Results and Analysis

4.1. Experimental Settings

Experimental Environment: All experiments are implemented on a single NVIDIA RTX 3090 GPU with 24GB memory.

Baseline and Comparative Methods:

Baseline: The original unfine-tuned ChatGLM3-6B model.

Comparative Method: Model fine-tuned via Low-Rank Adaptation (LoRA).

Proposed Method: The integrated system incorporating three-stage fine-tuning, dynamic scale embedding and collaborative reasoning proposed in this work.

Evaluation Metrics:

Subjective metrics include empathy, guidance capability and professionalism. Three experienced psychological counselors perform double-blind scoring ranging from 1 to 5 on 1000 unseen test cases.

Objective metrics cover fluency, coherence and safety, which are automatically evaluated by DeepSeek-V3. Hardware and efficiency indicators involve GPU memory consumption, training duration and inference speed measured by average response latency for 500-word outputs.

4.2. Subjective Evaluation Results

Table 1 presents the subjective evaluation outcomes. The proposed method achieves substantial superiority over the baseline model on empathy, guidance and professionalism, with scores of 4.3, 4.1 and 4.5, in contrast to the baseline scores of 2.8, 2.5 and 3.1. The results demonstrate that the designed three-stage fine-tuning strategy and collaborative reasoning mechanism effectively boost the overall performance of the model for psychological counseling applications.

Table 1. Results of subjective evaluation indicators.

Evaluation Metric	Evaluation Approach	Baseline Model	Proposed Method
Empathy	Human Evaluation	2.8 ± 0.7	4.3 ± 0.5
Guidance	Human Evaluation	2.5 ± 0.8	4.1 ± 0.6
Professionalism	Human Evaluation	3.1 ± 0.9	4.5 ± 0.4

4.3. Objective Evaluation and Efficiency Comparison

The objective assessment and efficiency comparison results are illustrated in Table 2 and

Table 3. In terms of generation quality, the proposed method outperforms the baseline model with scores of 4.7 in fluency, 4.6 in coherence and 4.8 in safety.

Regarding resource consumption and operational efficiency, compared with the LoRA-based model, the GPU memory usage drops by approximately 22.4%, decreasing from 20.13 GB to 15.62 GB, and the training time is shortened by roughly 21.5%, falling from 5.26 hours to 4.13 hours. In contrast to the original baseline model, the inference speed rises by about 87.4%, increasing from 96.55 characters per second to 180.92 characters per second [12].

Table 2. Objective Quality Evaluation Results.

Evaluation Metric	Evaluation Approach	Baseline Model	Proposed Method
Fluency	Automatic Evaluation	4.5 ± 0.3	4.7 ± 0.2
Coherence	Automatic Evaluation	3.9 ± 0.5	4.6 ± 0.3
Safety	Automatic Evaluation	4.2 ± 0.6	4.8 ± 0.2

Table 3. Comparison of Training and Inference Efficiency.

Comparison Item	LoRA /Baseline Model	Proposed Method
GPU Memory Consumption	20.13 G (LoRA)	15.62 G
Training Time (20 epochs)	5.26 h (LoRA)	4.13 h
Inference Speed (500-character Output)	96.55 chars/s (Baseline)	180.92 chars/s

4.4. Functional Verification and Ablation Experiment

Table 4. Comparison of Inference Output Speed Experimental Results.

Proposed System	Baseline Model
==500-Character Output Speed Test Results ==	== 500-Character Output Speed Test Results ==
Number of test samples: 10	Number of test samples: 10
Average generated characters: 512	Average generated characters: 503
Average generated tokens: 354	Average generated tokens: 416
Average response time: 2.83 seconds	Average response time: 5.21 seconds

Functional tests based on simulated conversations prove that the system can accurately judge the timing of scale embedding. For instance, the system actively pushes the SDS scale when the user mentions insomnia and low mood for three consecutive times. After scale completion, user profiles can be updated automatically, and standardized treatment plans with concrete behavioral advice such as twice-weekly exercise recommendations will be generated without abrupt interruption of conversation flow.

Three variant models are constructed to verify the validity of individual modules: the version without dynamic scale module that relies merely on dialogue, the version removing RAG module from collaborative reasoning that only adopts fine-tuned model, and the scheme with fixed knowledge weight instead of dynamic weighting mechanism. Experimental results reveal that removing any module leads to obvious declines in professionalism and guidance scores,

with average drops of 0.9 and 0.7 points respectively, which demonstrates all modules are indispensable to the whole system.

5. Discussion

The proposed method addresses multiple defects of conventional and existing intelligent psychological counseling systems in a systematic manner. Supported by mutual information threshold and Q-learning decision framework, dynamic scale embedding integrates standardized assessment tools into natural dialogue in a non-intrusive way. It eliminates rigid forced evaluation at fixed stages, improves user acceptance and enriches collected data. Its core merit lies in autonomous timing judgment based on dialogue rhythm and semantic features, realizing the intelligent transition from passive response to active assessment.

The collaborative reasoning framework combining fine-tuned model and RAG achieves adaptive balance between empathy priority and knowledge priority via dynamic weighting mechanism. At the early dialogue stage, the system prioritizes empathetic interaction and emotional bonding with a knowledge weight of approximately 30% to build user trust. As the conversation proceeds or crisis cues are detected, the weight of RAG module rises automatically to 50% for in-depth consultation and 80% under risky circumstances, guaranteeing accurate and safe professional responses. This framework effectively reduces factual hallucinations inherent to standalone fine-tuned models, and makes up for the lack of emotional warmth in pure RAG systems [13].

The complete service loop covering dialogue, scale assessment, user profiling and treatment planning endows the system with continuous capabilities ranging from listening and evaluation to intervention. The design conforms to practical counseling procedures, where practitioners gather information, draw judgments and deliver assessments or therapeutic suggestions appropriately. Experimental results verify its rationality, with guidance score reaching 4.1 and professionalism score reaching 4.5. This approach offers a feasible technical scheme for popular and standardized mental health services free from time and cost constraints, possessing promising application prospects in communities, schools, enterprises and online platforms.

Despite satisfactory performance, this research still has limitations.

Insufficient depth of multimodal fusion: The current system only combines text and audio data, while facial expressions, physiological signals and other valuable visual and biological information for emotion recognition are not fully utilized.

Limited dataset coverage: Although common mental health scenarios are included, the

generalization ability toward specific groups such as teenagers, the elderly and patients with rare psychological disorders still needs further verification [14].

Acknowledgements

This research is supported by the Key R&D Program of Sichuan Provincial Science and Technology Plan "Research and Application of Key Technologies of Psychological Counseling Robot Based on Large Language Model" (Grant No. 2024YFFK0251). The authors gratefully acknowledge the cooperative psychological counseling institutions and relevant health care organizations for their valuable assistance in data collection, expert annotation, and practical application verification.

References

[1] Healthy China Initiative Promotion Committee. (2019). *Healthy China Initiative (2019–2030)* [Government report]. https://www.nhc.gov.cn/wsxf/zcfg/201907/2a771d89e00e4b228028335d4fcf1a7d.shtml

[2] Kermani, A., Perez-Rosas, V., & Metsis, V. (2025, May). A systematic evaluation of LLM strategies for mental health text analysis: Fine-tuning vs. prompt engineering vs. RAG. *In Proceedings of the 10th Workshop on Computational Linguistics and Clinical Psychology (CLPsych 2025)* (pp. 172-180).

[3] Tao, D., Lee, T., Chui, H., & Luk, S. (2024, April). Modeling intrapersonal and interpersonal influences for automatic estimation of therapist empathy in counseling conversation. *In Proceedings of the 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 12692-12696). IEEE.

[4] Hu, J., Wang, A., Xie, Q., Li, Z., Ma, H., & Guo, D. (2026, March). Agentmental: An interactive multi-agent framework for explainable and adaptive mental health assessment. *In Proceedings of the AAAI Conference on Artificial Intelligence* (pp. 31050-31058).

[5] Mnih, V., Kavukcuoglu, K., Silver, D., Rusu, A. A., Veness, J., Bellemare, M. G., ... & Hassabis, D. (2015). Human-level control through deep reinforcement learning. *Nature*, *518*(7540), 529–533.

[6] Dutta, A., Mruthyunjaya, S., Saddington, J., & Islam, K. S. (2025, October). Mentalic Net: Development of RAG-based conversational AI and evaluation framework for mental health support. *In 2025 IEEE International Symposium on Emerging Metaverse (ISEMV)* (pp. 77-84). IEEE.

[7] Du, Z., Qian, Y., Liu, X., Ding, M., Qiu, J., Yang, Z., & Tang, J. (2022, May). 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)* (pp. 320-335).

[8] Yang, A., Xiao, B., Wang, B., Zhang, B., Bian, C., Yin, C., ... & Wu, Z. (2023). Baichuan 2: Open large-scale language models. *arXiv:2309.10305.*

[9] Bai, J., Bai, S., Chu, Y., Cui, Z., Dang, K., Deng, X., ... & Zhu, T. (2023). Qwen technical report. *arXiv:2309.16609*.

[10] Liu, A., Feng, B., Xue, B., Wang, B., Wu, B., Lu, C., ... & Piao, Y. (2024). DeepSeek-V3 technical report. *arXiv:2412.19437*.

[11] Kafi, M. A. A., Moni, R., & Banshal, S. K. (2026). Reasoning over recall: Evaluating the efficacy of generalist architectures vs. specialized fine-tunes in RAG-based mental health

dialogue systems. *arXiv:2601.01341.*

[12] Kwon, W., Li, Z., Zhuang, S., Sheng, Y., Zheng, L., Yu, C. H., ... & Stoica, I. (2023, October). Efficient memory management for large language model serving with PagedAttention. *In Proceedings of the 29th Symposium on Operating Systems Principles* (pp. 611-626).

[13] Abd-Alrazaq, A. A., Alajlani, M., Alalwan, A. A., Bewick, B. M., Gardner, P., & Househ, M. (2019). An overview of the features of chatbots in mental health: A scoping review. *International Journal of Medical Informatics, 132*, 103978.

[14] Torous, J., Bucci, S., Bell, I. H., Kessing, L. V., Faurholt-Jepsen, M., Whelan, P., ... & Firth, J. (2021). The growing field of digital psychiatry: Current evidence and the future of apps, social media, chatbots, and virtual reality. *World Psychiatry*, *20*(3), 318–335.

Online ISSN: 3104-9338
Print ISSN: 3104-932X

Design of Wearable Walking Assist Device Based on Ergonomics Principles

Bowen Wu, Xu Wang*

Chengdu College of University of Electronic Science and Technology of China

Received: June 6, 2026

Revised: June 7, 2026

Accepted: June 14, 2026

Published online: June 16, 2026

To appear in: *International Journal of Advanced AI Applications,* Vol. 2, No. 7 (July 2026)

* Corresponding Author: Xu Wang (34106831@qq.com)

Abstract. To develop a wearable walking assistance device for the elderly that enables stable walking in complex scenarios with high wearing comfort, this study completed the dimensional, structural, and functional design of the device based on ergonomic principles, integrating the body shape characteristics and gait movement patterns of the elderly. The 3D modeling, hardware control circuits, and software design of the device were also finalized. Subsequently, modal analysis was adopted to investigate the vibration modes and natural frequencies of key components of the device. Simulation results were compared with the walking step frequency, muscle autonomous contraction frequency, ground impact frequency, and motor operating frequency of the elderly, verifying the rationality of the structural design. This study provides a reference for the design of similar wearable health and rehabilitation equipment.

Keywords: *Ergonomics; Wearable Walking Assist Device; 3D Modeling; Modal Analysis; Elderly Care*

1. Introduction

With the rapid development of the social economy, population aging continues to deepen, and the demand for elderly health and rehabilitation equipment is growing steadily. It is widely recognized that the probability of high fall risk among adults aged 60–79 years old reaches 10.5%, which is highly correlated with leg muscle strength and walking patterns [1]. Therefore, developing a walking assistance device adapted to the needs of the elderly is of great significance.

In recent years, domestic universities and enterprises have carried out extensive research on walking assistance devices for the elderly and achieved certain progress. For example, in 2023, Feng Kaixiang and Zhang Ting from Soochow University designed a flexible-joint-driven 4-degree-of-freedom hip exoskeleton robot, which effectively improved the walking balance of

the elderly and reduced the probability of falls [2]. In 2024, Guo Zhanlong from Shaanxi Institute of Technology developed a control scheme for elderly walking assistants using a Proportional-Integral-Derivative (PID) control method based on a 3D human gait model, enhancing control performance and walking stability [3]. In 2025, Chen Yong and Zhao Junfeng from Dalian Jiaotong University adopted bionic design principles and a flexible design approach to develop a lower-limb exoskeleton robot capable of stable walking on complex terrain [4].

While existing studies focus on improving walking stability and preventing falls through device assistance, factors such as the adaptability between the device and the user's legs and wearing comfort are rarely addressed. Therefore, guided by ergonomic concepts, this study aims to design a structurally stable wearable walking assistance device adapted to the legs of the elderly while ensuring core functions, so as to improve wearing comfort and enable stable walking in complex scenarios.

About the authors: Wu Bowen (2005–), male, from Luzhou, Sichuan, undergraduate student, mainly engaged in the study and research of electrical engineering and automation. Corresponding author: Wang Xu (1982–), male, from Chengdu, Sichuan, professor, master's degree holder, mainly engaged in teaching and research on robotics engineering.

2. Overview of Ergonomics Principles

Ergonomics, also known as Human Factors Engineering or Human Engineering, is an interdisciplinary field that studies the interactions among humans, machines, and the working environment. Based on anthropometry, biomechanics, and human physiological and psychological characteristics, it aims to adapt products to the human body through rational design, thereby achieving the unity of safety, efficiency, and comfort in the human-machine-environment system [5]. In the design of wearable walking assist devices, it is necessary to match structural dimensions and driving parameters with the range of motion of lower limb joints, the force characteristics during the gait cycle, and the physiological degradation of the target users, so as to improve wearing compliance and prevent secondary injuries [5].

3. Data Sampling for the Walking Assistance Device

To ensure high adaptability between the wearable walking assistance device and the legs of elderly users, as well as to enhance wearing comfort and operational safety, this study adopts ergonomic principles as the core guideline. It places significant emphasis on the meticulous collection and comprehensive analysis of lower limb morphological parameters specific to the

elderly population. Given the considerable individual differences in thigh length, calf length, and foot length among older individuals, fixed-size devices are insufficiently accommodating for the diverse range of body types. Such rigidity can lead to discomfort or even compromise the overall assistive performance of the device, ultimately hindering its effectiveness. Therefore, a tailored approach is essential for optimal user experience.

In this research, a carefully selected sample group comprising 1,000 elderly participants was utilized. The study employed precise anthropometric methods to accurately measure three key dimensional parameters—thigh length, calf length, and foot length. This comprehensive approach enabled the acquisition of statistically representative distribution characteristics of human body dimensions within this demographic. The resulting data will serve as a crucial geometric foundation for subsequent three-dimensional modeling, dimensional optimization, and the thoughtful design of wearing schemes for the device. Ultimately, these efforts will ensure that the final product aligns effectively with the unique lower limb morphology of the target population, enhancing comfort and usability.

Partial collected data are shown in Table 1.

Table 1. Partial data of sampling for the walking assistance device.

Measurement item	Sample 1	Sample 2	Sample 3	Sample 4	Sample 5	Sample 6	Sample 7	Sample 8
Thigh (cm)	42.5	44.2	41.8	43.6	42.1	42.5	44.2	41.8
Calf (cm)	32	35	34	36	37	36	35	37
Foot length (cm)	23.4	24.3	24.3	24.1	24.5	24.6	24.3	24.2

4. Hardware Design of the Walking Assistance Device

4.1. 3D Modeling of the Walking Assistance Device

According to the anthropometric data presented in Table 1, the mean thigh length of the sampled elderly participants was 42.8 cm (range: 2.4 cm), the mean calf length was 35.3 cm (range: 5.0 cm), and the mean foot length was 24.2 cm (range: 1.2 cm). Based on the statistical analysis of these measurements, the wearable walking assistance device was structurally divided into four primary components: a thigh support mechanism measuring 45.0 cm, a calf support mechanism measuring 35.0 cm, a foot sole support plate measuring 24.24 cm, and a connecting adjustment mechanism.

Each support mechanism incorporates a sophisticated slide rail–type continuously adjustable structure that is seamlessly integrated with elastic straps to achieve dual fixation and enhanced stability. The major dimensional parameters were carefully derived directly from an extensive

anthropometric dataset, thereby ensuring compatibility with a broad range of lower limb morphologies within the target population. Furthermore, the anterior section of the device is equipped with an advanced infrared sensor along with an ultrasonic ranging sensor, enabling real-time detection of forward obstacles in various environments. This innovative configuration not only facilitates active obstacle avoidance but also provides a crucial early warning function against potential falls, significantly improving user safety and mobility [6].

The 3D models of the main components of the wearable walking mechanism are shown in Figures 1–4.

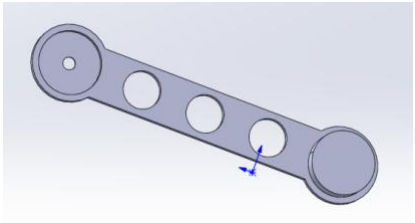


Figure 1. Thigh support mechanism.

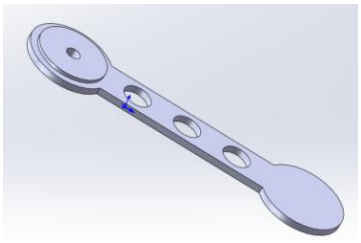


Figure 2. Calf support mechanism.

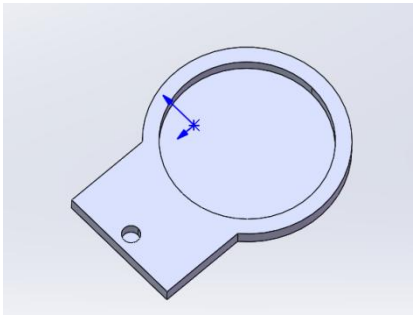


Figure 3. Knee joint assembly.

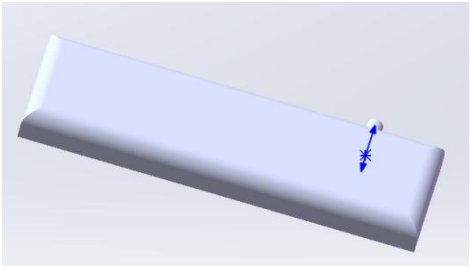


Figure 4. Pedal mechanism.

The general assembly 3D model of the walking mechanism is shown in Figure 5.

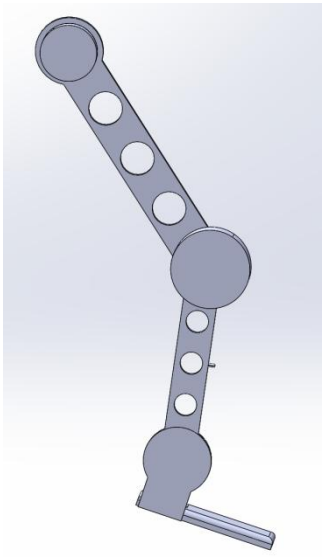


Figure 5. General assembly drawing.

4.2. Design of the Control System of the Walking Assistance Device

4.2.1. Main Controller Design

To achieve effective and precise control over each module in the system, this paper selects the STM32F103C8T6 microcontroller as the core component. This highly capable 32-bit enhanced microcontroller is built on an ARM Cortex-M3 core, operating at a maximum frequency of 72 MHz. It integrates a generous 64 KB of Flash memory and 20 KB of SRAM, making it suitable for various applications. Additionally, the microcontroller offers a wealth of peripheral interfaces, including a multi-channel 12-bit ADC, advanced timers, USB connectivity, CAN bus support, and SPI communication. Its three primary advantages are as follows:

(1) Extremely high cost-performance ratio, with strong performance and controllable cost, widely used in teaching and industrial mass-produced products [8];

(2) Complete peripheral resources, with hardware support for USB and CAN communication, greatly reducing the development difficulty of complex systems;

(3) Mature development ecosystem, with official CubeMX tools and a large community resource base significantly shortening the R&D cycle [9].

With its stable and reliable performance, this chip has emerged as the preferred solution for those venturing into embedded systems, particularly for beginners and designers working on small-to-medium-sized control system applications. Its versatility and ease of use make it an excellent choice for various projects in this field.

The minimum system of STM32F103C8T6 is shown in Figure 6.

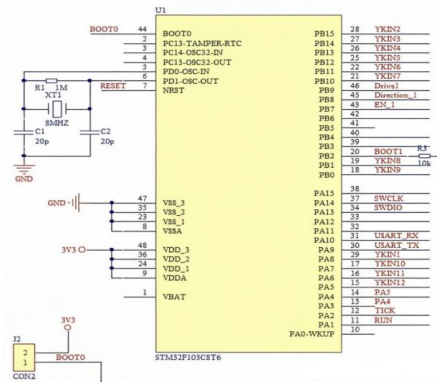


Figure 6. Minimum system circuit diagram.

4.2.2. Design of HC-SR04 Ultrasonic Ranging Module

Ranging module is recommended [10]. The module operates at a supply voltage of 5 V, offers a measurement range of 2 cm to 4 m, and achieves an accuracy of approximately ±3 mm. It can be directly interfaced with the GPIO ports of the STM32F103C8T6 microcontroller via its Trigger (Trig) and Echo pins [11].

The key advantages of this module include:

(1) Simple interface, only four connections are required to perform non-contact distance measurement, facilitating straightforward implementation in both hardware and software designs.

(2) High-cost performance ratio, the module is inexpensive, widely available, and well suited for large scale deployment in walking aid systems.

(3) Fast response with minimal blind zone, it enables rapid detection of forward obstacles, allowing timely activation of audiovisual alarm signals to enhance user safety [10–11].

The PCB of the HC-SR04 ultrasonic ranging module is shown in Figure 7.

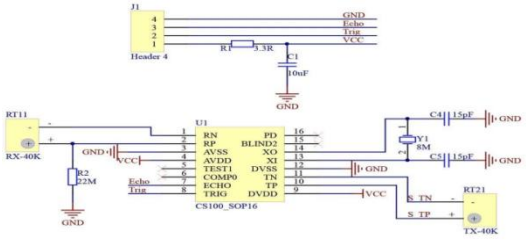


Figure 7. Ultrasonic eanging module circuit diagram.

4.2.3. Design of OLED12864 Display Module

For the elderly walking assistance device, a 0.96-inch OLED display module with an I²C

interface (model OLED12864, driven by the SSD1306 controller) is recommended [12]. The module features a resolution of 128 × 64 pixels and is available with yellow blue or white emission options. Communication with the STM32F103C8T6 microcontroller requires only two signal lines—Serial Data Line (SDA) and Serial Clock Line (SCL)—enabling the display of critical information such as walking speed, battery status, obstacle proximity, and fall alert notifications [13]. The principal advantages of this module include:

(1) High contrast self-emissive display, the screen does not require a backlight, offers a viewing angle approaching 180°, and ensures clear readability for elderly users.

(2) Low power consumption, with a static current ranging from 0.02 mA to several milliamperes, it is well suited for battery powered, portable walking aids.

(3) Efficient interface utilization, The I²C communication protocol minimizes the number of occupied GPIO pins and benefits from extensive driver library support, facilitating streamlined integration with STM32 based systems [12–13].

The PCB of the OLED12864 display module is shown in Figure 8.

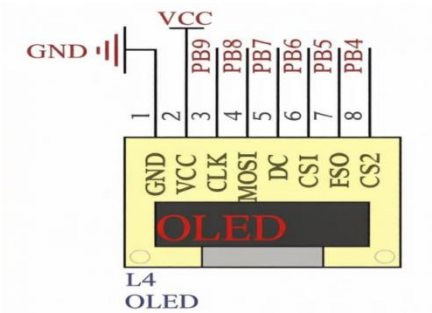


Figure 8. OLED12864 display module circuit diagram.

4.2.4. Design of Bluetooth Module

For the elderly walking assistance device, the HC 05 Bluetooth serial module—an integrated master slave device based on the BC417 chip—is recommended [14]. The module operates at a supply voltage range of 3.3 V to 5 V and interfaces directly with the USART peripheral of the STM32F103C8T6 microcontroller via a UART connection. It enables wireless transmission of critical data—such as fall alerts and obstacle distance measurements—from the walking aid to the caregiver's mobile application [15].

The main advantages of the HC 05 module include:

(1) Switchable master/slave modes, it supports direct interconnection between two HC 05 units or pairing with Android smartphones, thereby accommodating diverse networking requirements.

(2) Comprehensive AT command set and extensive documentation, Serial pass-through functionality is supported without the need for complex protocol stacks, reducing system development complexity.

(3) High-cost performance ratio, the module is low cost, widely available, and features an onboard status LED to indicate connection status, making it well suited for educational purposes and rapid prototyping [14–15].

The PCB of the HC-05 master-slave integrated Bluetooth serial module is shown in Figure 9.

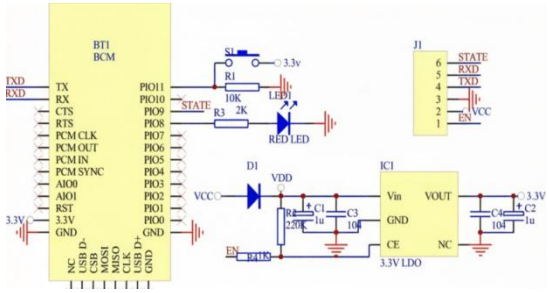


Figure 9. Bluetooth module circuit diagram.

5. Software Design of the Walking Assistance Device

5.1. Software Design Flow of the System

Upon system startup, the STM32 microcontroller first initializes its core peripherals, including the system clock, GPIO ports, timers, I²C bus, and USART interface. Following this essential initialization step, the OLED display module is configured, allowing for visual output, after which the HC SR04 ultrasonic ranging sensor is sequentially set up to enable accurate distance measurements. Next, the HC 05 Bluetooth serial module is initialized to facilitate wireless communication with other devices. Finally, the obstacle detection alarm threshold is established, ensuring that the system can effectively respond to potential obstacles in its path [16].

Within the main execution loop of the system, ultrasonic distance measurements are periodically triggered to ensure accurate monitoring. The acquired raw data from these measurements undergoes processing through a mean filtering algorithm, which effectively computes the actual distance. If the measured distance falls below a carefully predefined threshold value, the system promptly activates an audible alert via the onboard buzzer to notify users. Simultaneously, the updated distance value and overall system status are displayed on the OLED screen for easy visualization and monitoring. Additionally, this information is

transmitted wirelessly through the HC-05 module to the caregiver's mobile application via serial communication, ensuring timely updates and efficient oversight [16].

The system operates in a continuous cyclic manner, which not only ensures real-time environmental monitoring but also facilitates prompt remote alarm notifications. This seamless process enables timely responses to any changes in the environment, enhancing overall safety and efficiency.

The flowchart of the system software design is shown in Figure 10.

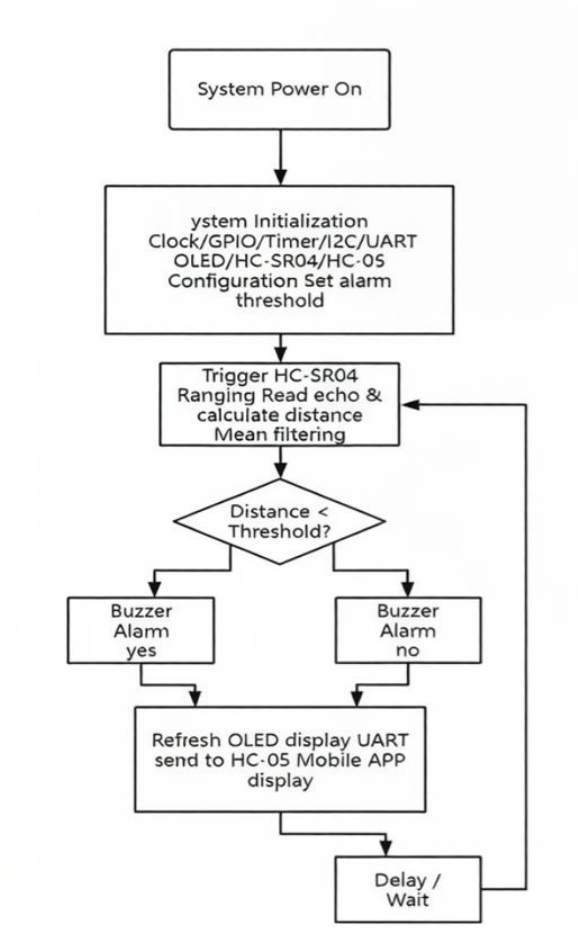


Figure 10. System software design flowchart.

5.2. Upper Computer Design of the System

The function of the upper computer is mainly to facilitate human-computer interaction with the wearable walking assistance device, allowing the elderly to manage the device appropriately during use according to actual conditions, and to keep track of the device status in time (such as thigh angle, calf angle, ultrasonic distance, motor assistance intensity, etc.). The upper computer design mainly includes modules for joint posture, environmental perception, driving assistance, safety protection, and obstacle warning threshold setting. The design flowchart is

shown in Figure 11.

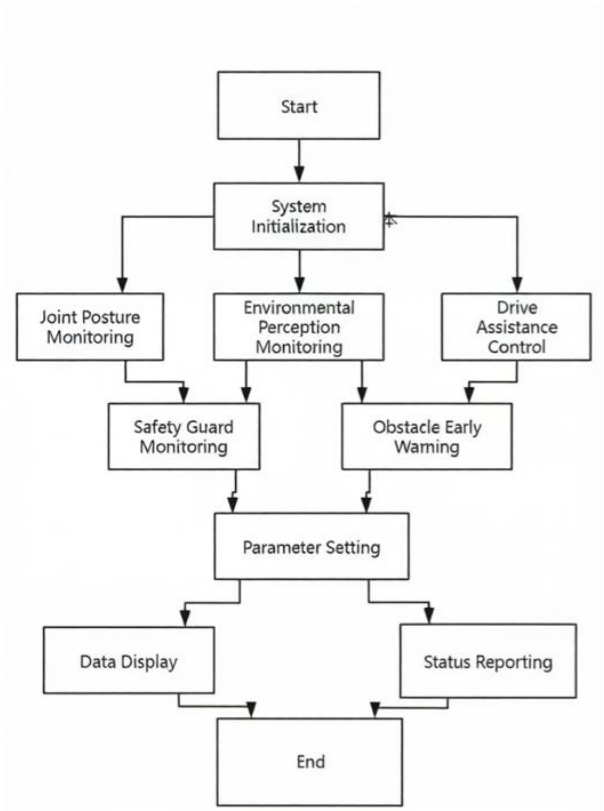


Figure 11. Host computer design flowchart.

The upper computer design interface is shown in Figure 12.

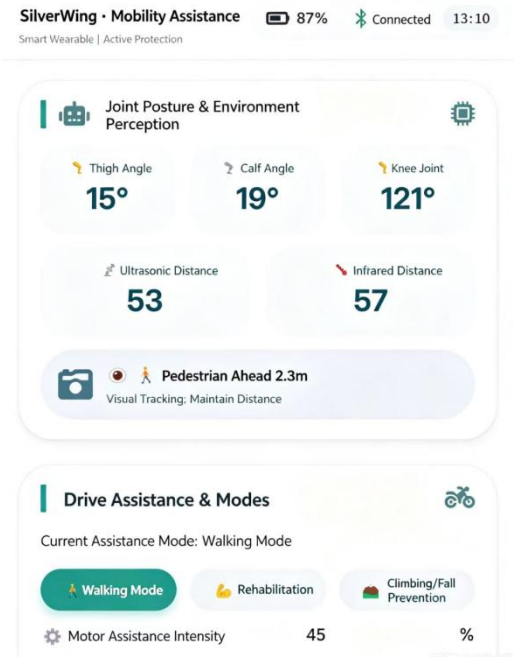


Figure 12. Wearable walking assistive device management system.

6. Performance Verification

Considering that safety and stability are the key indicators of the walking assistance device, this paper uses the finite element modal analysis method to evaluate its performance. The specific operation process is as follows:

(1) Material setting: Considering the safety of use by the elderly, high-strength and lightweight materials should be selected, so the main structure of the device is determined to be alloy steel.

(2) Fixed geometry setting: The fixed end during device operation is set as a fixed geometry.

(3) Meshing: Since the shape of the main structures (thigh and calf) of the model is relatively regular, solid meshing can be adopted; near fillets and weight-reduction round holes, mesh refinement can be performed to ensure high mesh quality.

(4) Iterative calculation: The FFEPlus solver is selected for iterative calculation, with the calculation order set to 6 orders. This solver is based on the conjugate gradient method, has good performance in solving linear problems, can effectively shorten the solution cycle, and achieve good solution accuracy. The calculation expression is:

$$\alpha_k = \frac{r_k^T}{p_k^T} \frac{r_k}{Kp_k}$$

The 6-order modes of the thigh structure are shown in Figures 13–18.

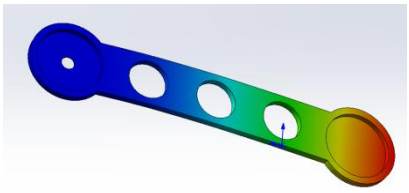


Figure 13. 1st Mode shape.

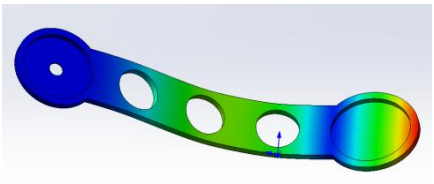


Figure 14. 2st Mode shape.

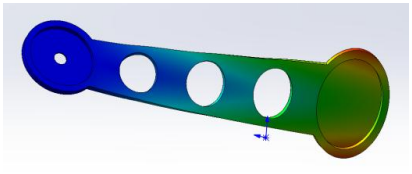


Figure 15. 3st Mode shape.

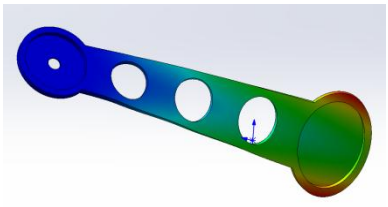


Figure 16. 4st Mode shape.

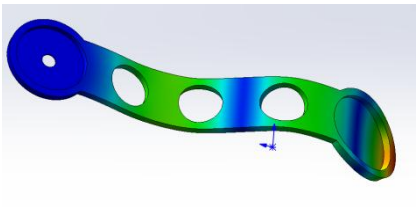


Figure 17. 5st Mode shape.

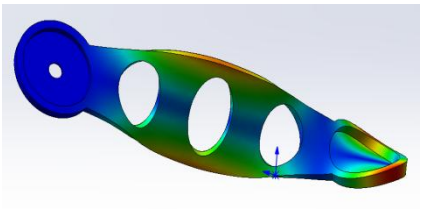


Figure 18. 6st Mode shape.

The mode shapes of each order are shown in Table 2.

Table 2. Mode shapes of each order.

Order	1st order	2nd order	3rd order	4th order	5th order	6th order
Mode	Swing	Flap	Flap	Swing	Flap	Torsion

Based on the analysis of the six-order modes of the thigh structure presented in Table 2, it is evident that the predominant mode shapes are primarily characterized by swing and flap motions, with torsional vibrations observed only at the sixth order. This observation allows us to conclude that the structure exhibits a strong resistance to torsion, underscoring its excellent overall stability and robustness under various dynamic conditions.

The 6-order modes of the calf structure are shown in Figures 19–24.

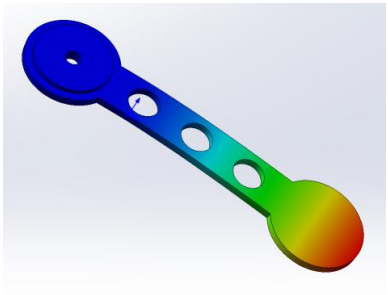


Figure 19. 1st Mode shape.

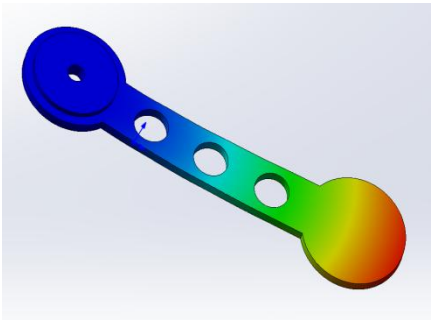


Figure 20. 2st Mode shape.

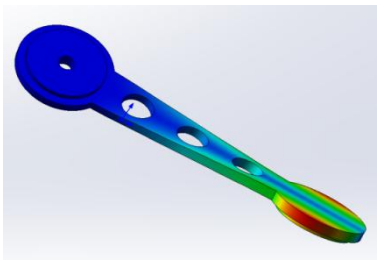


Figure 21. 3st Mode shape.

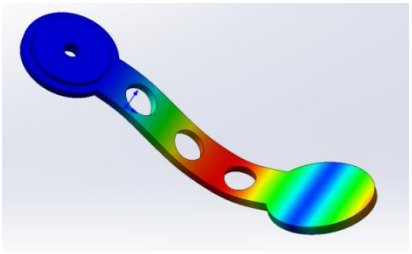


Figure 22. 4st Mode shape.

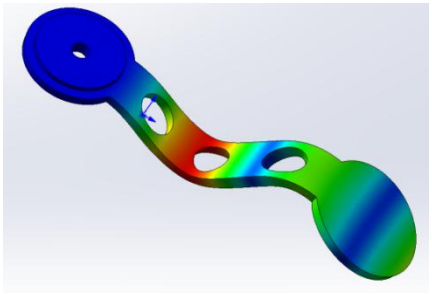


Figure 23. 5st Mode shape.

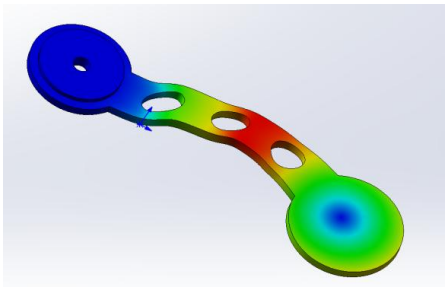


Figure 24. 6st Mode shape.

The mode shapes of each order are shown in Table 3.

Table 3. Mode shapes of each order.

Order	1st order	2nd order	3rd order	4th order	5th order	6th order
Mode	Swing	Swing	Torsion	Flap	Flap	Swing

From the 6-order modes of the calf structure in Table 3, the mode shapes are similar to those of the thigh structure, also dominated by swing and flap, with torsion only appearing at the 3rd order. It can be seen that the calf structure also has good stability.

The comparison between the natural frequencies of the thigh and calf structures of the wearable device and other relevant frequencies is shown in Table 4.

Table 4. Natural frequencies of thigh and calf structures and comparison with other frequency values.

Order	1st order	2nd order	3rd order	4th order	5th order	6th order
Thigh frequency (Hz)	33.737	205.91	285.98	306.13	555.65	934.6
Calf frequency (Hz)	35.391	233.01	290.63	335.59	1063.1	1584
Elderly walking step frequency (Hz)	1–2 (60–120 steps/minute)	1–2 (60–120 steps/minute)	1–2 (60–120 steps/minute)	1–2 (60–120 steps/minute)	1–2 (60–120 steps/minute)	1–2 (60–120 steps/minute)
Human muscle autonomous contraction frequency (Hz)	5–20	5–20	5–20	5–20	5–20	5–20
Ground impact frequency (Hz)	6	6	6	6	6	6
Motor operating frequency (Hz)	50	50	50	50	50	50

As shown in the results, all orders of natural frequencies for both the thigh and calf structures lie well outside the critical frequency ranges, including the elderly walking step frequency, human muscle contraction frequency, ground impact frequency, and motor operating frequency. Consequently, under normal operating conditions, the device is effectively isolated from resonant risks induced by external excitation. This indicates that the structural design possesses sufficient dynamic stability, thereby validating the rationality and reliability of the proposed design.

7. AI-Driven Adaptive Control Strategy

To overcome the rigidity of fixed-threshold triggering, an intelligent control module is proposed. Utilizing IMU and plantar pressure sensors, the system captures high-fidelity gait time-series data. A lightweight neural network—specifically a Temporal Convolutional Network (TCN)—is deployed to identify gait patterns and predict anomalies (e.g., pre-fall

symptoms). This allows the device to dynamically adjust its assistance strategy, shifting from static alerts to context-aware adaptive responses.

8. Conclusion

In the design of the wearable walking assistance device, this study adopts a human-centric approach grounded in ergonomic principles. Structural optimization was achieved using extensive anthropometric data, significantly enhancing wearing adaptability. A high-performance control system based on the STM32 microcontroller was developed, integrating multi-sensor fusion to enable obstacle avoidance and real-time remote monitoring. However, limitations remain regarding the device's overall mass and operational endurance. Future work will focus on structural topology optimization utilizing lightweight materials such as carbon fiber, alongside the implementation of adaptive control algorithms, to further reduce physical load and extend battery life, thereby comprehensively improving the user experience.

Acknowledgements

The authors would like to express their gratitude to the Department of Intelligent Manufacturing Engineering, University of Electronic Science and Technology of China, Chengdu Campus, for providing experimental equipment and technical support.

References

[1] Zhu, J., Wang, J., Fan, C., et al. (2025). Current status and related factor analysis of fall risk among Chinese adults aged 60–79 years: Based on 2024 national cross-sectional monitoring data. *Medical Journal of Peking Union Medical College Hospital*, 606–616.

[2] Feng, K., & Zhang, T. (2023). Kinematic synthesis and assisted walking balance control of wearable hip exoskeleton mechanism. *Mechanical Design and Manufacturing Engineering*, 52(2), 14–19. (in Chinese)

[3] Guo, Z. (2024). Research on intelligent walking assistant for the elderly based on Unity3D. *Automation & Instrumentation*, (07), 270–278. (in Chinese)

[4] Chen, Y., & Zhao, J. (2025). Design and simulation analysis of lower limb exoskeleton robot assisting steep terrain walking. *Modern Manufacturing Engineering*, (12), 35–42. (in Chinese)

[5] Ding Y L. Ergonomics [M]. 4th ed. *Beijing: Beijing Institute of Technology Press*, 2017: 1-15.(in Chinese)

[6] Zhao, M., & Chen, J. (2020). Application research of ultrasonic ranging in walking assistance devices for the elderly. *Journal of Electronic Measurement and Instrumentation*, 34(5), 45–52. (in Chinese)

[7] STMicroelectronics. (2009). STM32F103x8 Medium-density Performance Line Datasheet (Rev.16). *Geneva: STMicroelectronics*.

[8] Liu, H. (2014). Journey from 51 Beginner to ARM (STM32) Expert (pp. 45–52). *Beijing: Beihang University Press*. (in Chinese)

[9] Wang, L., & Zhang, H. (2020). Design of intelligent acquisition and communication system based on STM32F103C8T6. *Application of Electronic Technique*, 46(8), 78–81. (in Chinese)

[10] Omeko, I., Twieg, S., & Obert, M. (2025). Sensor-based gait analysis: A comparative study of ultrasonic and laser sensors for gait monitoring in rollator-assisted walking. *Proc of Int Conf on Applied Innovations in IT* (ICAIIT).

[11] Yin, F., Zhang, T., Cui, W., et al. (2026). Design of intelligent ultrasonic obstacle avoidance and safety monitoring guide cane based on STM32. *Journal of Changchun Normal University*, 45(2), 40–46. (in Chinese)

[12] Xu, F., Shu, R., Wang, J., et al. (2024). Comparative experimental study on display modules based on single-chip microcomputer systems. *Journal of Langfang Normal University* (Natural Science Edition), 24(3), 83–88. (in Chinese)

[13] Li, X., & Zhao, Y. (2021). Human-computer interaction interface design of STM32-based intelligent walking aid crutch. *Electronic Measurement Technology*, 44(14), 112–116. (in Chinese)

[14] Tan, H., & Wang, D. (2019). Embedded Bluetooth communication technology and application research of HC series modules. *Modern Electronics Technique*, 42(14), 31–34. (in Chinese)

[15] Zhang, L., & Li, N. (2021). Design of intelligent walking aid crutch system for the elderly based on STM32 and Bluetooth. *Automation & Instrumentation*, 36(7), 65–68. (in Chinese)

[16] Wang, L., & Chen, Z. (2020). Software design of multi-sensor data acquisition and processing for embedded intelligent walker. *Computer Measurement & Control*, 28(11), 189–193. (in Chinese)

Online ISSN: 3104-9338
Print ISSN: 3104-932X

Research on Spatiotemporal Evolution Characteristics and Collaborative Control Strategies of Pollution Sources in Industrial Parks Based on Machine Learning

Tao Lin*

Yixing Ecological Environment Monitoring Station, Wuxi City, Jiangsu Province, 214200

Received: June 3, 2026

Revised: June 8, 2026

Accepted: June 15, 2026

Published online: June 20, 2026

To appear in: *International Journal of Advanced AI Applications,* Vol. 2, No. 7 (July 2026)

* Corresponding Author: Tao Lin (5240826@qq.com)

Abstract. This study proposes a machine learning framework combining STGCN and MAPPO for regional collaborative pollution control in industrial parks, which reduces operational energy consumption by 13.6% and decreases core pollutant exceedance frequency by 89.5%.

Keywords: *Industrial Park; Machine Learning; Spatial-Temporal Graph Convolutional Network; Multi-Agent Reinforcement Learning; Collaborative Control*

1. Introduction

Industrial parks represent the crucial structural carriers for the manufacturing industry and serve as the foundational engines for high-quality economic development in China and rapidly industrializing nations globally. By strategically clustering enterprises, these specialized zones maximize operational efficiencies, promote industrial symbiosis, and facilitate the sharing of critical infrastructure, logistics networks, and raw material supply chains. However, this dense geographical agglomeration inherently carries profound environmental repercussions. Due to the internal concentration of a large number of high-polluting and high-energy-consuming enterprises—such as petrochemical refineries, heavy chemical synthesis plants, large-scale printing and dyeing facilities, metallurgical operations, and building materials manufacturing (e.g., cement and brick kilns)—industrial parks have inadvertently transformed into core sources of regional, complex atmospheric pollution [1]. The high-density, continuous, and frequently overlapping pollutant emissions from these concentrated sources place immense stress on the local atmospheric carrying capacity, leading to severe ecological deterioration and posing substantial risks to public health in adjacent urban centers.

At the current stage of off-site intelligent environmental law enforcement and daily park supervision, regulatory bodies and park management committees have made significant

investments in advanced monitoring hardware. Dense networks of gridded micro-monitoring stations at plant boundaries and highly sophisticated Continuous Emission Monitoring Systems (CEMS) at major exhaust stacks are gradually being deployed at the front end of the regulatory framework. These hardware advancements provide an unprecedented volume of high-frequency, minute-level environmental data, encompassing multi-dimensional metrics such as Sulfur Dioxide (SO2), Nitrogen Oxides (NOx), Particulate Matter (PM), Volatile Organic Compounds (VOCs), and complex meteorological parameters. Despite this abundance of data, a critical bottleneck remains in the software and analytical domain: these massive, high-frequency monitoring datasets have long been trapped in a state of isolated, retrospective analysis. They are typically utilized merely for threshold alarms and historical compliance auditing, and have not yet been deeply transformed into a proactive, joint pollution control force at the macro, overarching park level [2]. The immense latent value of this big data in driving dynamic, regional-scale environmental optimization remains largely untapped.

Traditional industrial park pollution control frameworks predominantly rely on the Distributed Control Systems (DCS) installed within each individual polluting enterprise. These systems are typically governed by localized Proportional-Integral-Derivative (PID) controllers designed to ensure the compliance of a single stack or terminal treatment facility in a purely independent, reactive manner. This non-cooperative, inherently siloed approach exhibits significant, systemic non-collaborative defects [3].

First and foremost, it fundamentally fragments the complex, interconnected laws of spatial transport and cross-media diffusion of atmospheric pollutants. The atmospheric environment is a continuous fluid medium; when the prevailing wind direction shifts during seasonal transitions (e.g., in summer) or when unfavorable meteorological phenomena such as low-altitude temperature inversion layers occur, the concentrated pollutant discharges of locally clustered enterprises easily produce severe superimposition effects in the downwind direction. Because each enterprise operates blindly regarding its neighbors' emissions and the overall spatial topology, these overlapping plumes frequently lead to severe area-wide environmental exceedances, even if every individual enterprise technically meets its localized emission standards at the stack.

Second, this fragmented governance model suffers from a profound lack of collaborative scheduling and optimization of regional treatment resources. A prominent example is the phenomenon of over-treatment or redundant resource consumption during regional alarm events. When the overall emissions of the park are approaching a regulatory warning threshold,

but there is still ample operational space in certain sectors, environmental enforcement protocols often force a generalized reduction mandate. Consequently, some enterprises—driven by strict compliance pressures—blindly inject excessive amounts of chemical reagents (such as ammonia in SCR denitration processes) and run desulfurization slurry circulation pumps at maximum frequencies. This overreaction causes serious secondary energy waste, inflates operational costs, and ironically leads to secondary environmental issues such as ammonia slip, without necessarily optimizing the exact downwind regions that are most critical [4]. The economic and environmental inefficiencies of this uncoordinated response underscore the urgent necessity for a paradigm shift in industrial environmental management.

Despite the increasing adoption of machine learning methodologies in environmental monitoring and predictive modeling over the past decade, existing applications exhibit significant, structural limitations when applied to the complex realities of industrial parks. Most current studies predominantly rely on conventional neural network architectures—such as Long Short-Term Memory (LSTM) networks, Gated Recurrent Units (GRU), or standard grid-based Convolutional Neural Networks (CNNs)—to formulate single-point forecasting models. While effective for isolated time-series forecasting, these models operate under the assumption of Euclidean spatial relationships. They fundamentally fail to capture the complex, non-Euclidean spatial topologies, the aerodynamic interference of large industrial structures, and the cross-media transport dynamics that govern industrial park emissions [2]. Furthermore, while some advanced predictive models can achieve remarkably high accuracy in forecasting isolated pollutant concentration trends, they suffer from a "prescriptive gap." They lack actionable, dynamic decision-making capabilities. The vital transition from passive predictive environmental analytics to dynamic, active multi-agent collaborative control remains largely unexplored in the current literature. This critical gap severely restricts the ability of park managers to dynamically orchestrate distributed pollution sources under overarching regional environmental constraints, perpetually resulting in isolated, highly suboptimal treatment strategies [5].

In recent years, unprecedented breakthroughs in the fields of deep spatiotemporal representation learning and Multi-Agent Reinforcement Learning (MARL) have converged to provide a powerful new algorithmic paradigm capable of solving the heterogeneous spatiotemporal evolution and adaptive collaborative game control of large-scale, complex industrial systems. Recognizing this potential, this study aims to bridge the prescriptive gap. By exploring how advanced machine learning techniques can decipher the intricate

spatiotemporal evolution characteristics of multi-source pollution in non-Euclidean spaces, this paper constructs an innovative collaborative control network. The objective is to transition from siloed PID-based management to a synchronized, multi-agent AI-driven ecosystem, which is of profound theoretical significance and vast application value for promoting regional refined pollution reduction, energy conservation, and the ultimate synergy of pollution reduction and carbon reduction strategies.

2. Business Pain Points and Complexity of Multi-Source Pollution Control in Industrial Parks

2.1. Decoupling Difficulties of Spatiotemporal Nonlinear Transport and Diffusion

Pollution diffusion within the highly structured confines of modern industrial parks is vastly different from simple geometric linear propagation in an open field; it is a highly complex, multidimensional fluid dynamics problem. The physical environment of an industrial park is characterized by extreme heterogeneity. The roughness damping effect caused by dense, large-scale plant buildings and complex pipeline corridors creates significant aerodynamic downwash and wake turbulence. Furthermore, the immense thermal energy released by continuous industrial processes generates powerful local urban heat island effects, leading to intense, unpredictable thermal turbulence and localized vertical convection currents. Additionally, the heterogeneity of flue gas lifting heights—which fluctuate constantly based on the varying exhaust temperatures, flow rates, and operating conditions of different kilns and boilers—adds another layer of complexity. These combined factors ensure that the spatial correlation between any two gridded monitoring points presents extremely strong, non-stationary dynamic nonlinearity [1].

Historically, environmental engineers have relied on traditional deterministic models, such as Gaussian Plume Models or advanced numerical atmospheric dispersion simulation software (such as AERMOD or CALPUFF), to estimate pollutant transport. However, these traditional tools encounter insurmountable practical bottlenecks when applied to real-time industrial park control. They have extremely high computational overheads, requiring hours of supercomputer processing time for detailed simulations, which fundamentally precludes their use in minute-level, real-time control loops. More critically, these deterministic models are highly dependent on the input of perfectly accurate, real-time source strength parameters, precise emission temperatures, and exact exit velocities—data which is frequently missing, noisy, or

systematically biased in real-world industrial settings due to sensor drift or intentional miscalibration. Consequently, it is virtually impossible to use these legacy systems to perform real-time topological correlation and extraction on minute-level, high-frequency CEMS and micro-plant boundary time-series data. The inability to decouple these complex spatiotemporal transport mechanisms computationally has forced regulators to rely on overly conservative, localized, and ultimately inefficient static thresholds, ignoring the dynamic atmospheric realities of the park.

2.2. Collaborative Bottlenecks of Multi-Agent Pollution Discharge Games and Control Time Delays

The governance of an industrial park inherently involves managing a complex ecosystem containing multiple independent, rationally self-interested pollutant-discharging entities. The physical and chemical realities of industrial pollution control add severe constraints to this management problem. The process pollution mechanisms and the specific terminal treatment facilities utilized by each enterprise—such as Selective Catalytic Reduction (SCR) or Selective Non-Catalytic Reduction (SNCR) for denitration, and wet/dry Flue Gas Desulfurization (FGD) systems—are governed by complex thermodynamic and chemical kinetic laws. Consequently, these systems exhibit profound and variable physical time delays. When a control command is issued to adjust a valve or pump, it usually takes between 15 to 45 minutes for the full chemical reaction to propagate through the absorption towers or catalyst beds and register as a stable concentration change at the CEMS monitors [3]. This massive latency completely invalidates the assumptions of instantaneous control required by standard algorithmic approaches.

This physical latency is dramatically compounded by the sociological and economic "game" played by the constituent enterprises. In the current regulatory environment, when the central environmental supervision platform detects deteriorating meteorological conditions and issues an instantaneous regional overload warning or emergency emission reduction mandate, a chaotic, uncoordinated response ensues. Because there is a complete lack of a sophisticated collaborative control mechanism, the localized, independent PID systems of all enterprises across the park will simultaneously and reflexively increase the input of treatment reagents to avoid heavy regulatory fines. They act in a state of informational blindness regarding the actions of their neighbors and the specific spatiotemporal trajectory of the pollutant plume.

This synchronized, panic-driven overreaction not only fails to immediately suppress the superimposed concentration downwind—due to the aforementioned 15 to 45-minute physical time delays—but it actively triggers massive, large-scale excessive treatment. The result is a

sharp, catastrophic soaring of the overall electrical power consumption (for massive slurry circulation pumps) and chemical reagent costs (such as liquid ammonia or urea) across the entire park. Furthermore, excessive ammonia injection frequently leads to severe ammonia slip, generating secondary fine particulate matter (PM2.5) pollution that exacerbates the very smog the regulations intended to clear. Therefore, establishing a mathematically rigorous, AI-driven collaborative control matrix—one that can optimally balance the production load, treatment cost, physical system delays, and the whole-area environmental capacity of each individual enterprise in real-time—stands as the most critical technical bottleneck urgently needed to be overcome in the contemporary field of environmental informatics and sustainable engineering.

3. Machine Learning-Based Spatiotemporal Evolution and Collaborative Control Model Architecture

To systematically address the profound deficiencies inherent in the traditional siloed PID control paradigm, this paper meticulously designs and implements a highly advanced, intelligent environmental protection closed-loop framework that is specifically tailored to tackle the complexities and challenges of modern industrial parks. The architecture is logically divided into two heavily integrated major components: a deep spatiotemporal feature mining and prediction layer, which utilizes a Spatial-Temporal Graph Convolutional Network (STGCN), and a dynamic, adaptive collaborative control and optimization layer that is based on Multi-Agent Proximal Policy Optimization (MAPPO). The seamless integration of these two layers not only facilitates but also enhances a continuous cycle of sensing, predicting, and collaboratively acting, thereby promoting more efficient and effective environmental management practices.

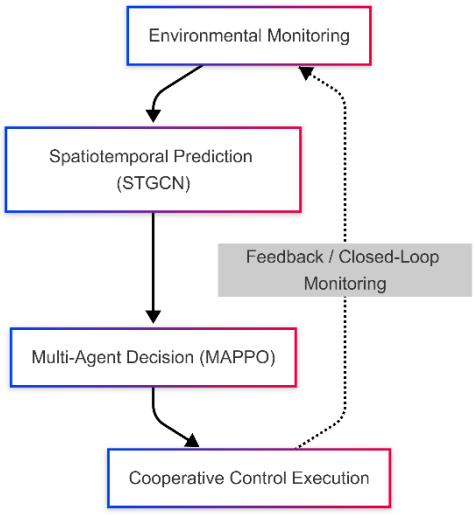


Figure 1. Schematic diagram of the STGCN + MAPPO closed-loop framework.

3.1. Spatiotemporal Evolution Feature Decoupling Based on STGCN

In order to accurately capture and map the complex, non-linear spatiotemporal evolution of pollution dispersion within the park—crucially, without requiring perfectly accurate, deterministic source strength parameters that hinder traditional AERMOD models—this paper abstracts the entire physical monitoring station network in the park into a highly dynamic mathematical topological graph G = (V, E) operating under non-Euclidean space [2]. In this sophisticated graph representation, the nodes V represent the individual micro-plant boundary sensors and the high-fidelity CEMS monitoring stations affixed to the major exhaust stacks of the factories. The edges E, which connect these nodes, are not static geometric lines; rather, their weights are dynamically and jointly determined by an intricate fusion of fixed geographical distances, highly volatile real-time wind vectors (speed and direction), atmospheric stability indices, and deep historical semantic correlations mined from years of time-series data. This dynamic edge weighting allows the graph to "morph" continuously in response to shifting weather patterns, perfectly encapsulating the changing fluid dynamics of the atmosphere over the industrial zone.

The framework utilizes advanced Spatial Graph Convolution (GCN) layers to meticulously extract the hidden spatial topological correlations between these interconnected stations. Through the stacking of multi-layer graph convolutions, the deep learning model can autonomously and adaptively capture the underlying, highly nonlinear dynamic spatial contribution rates of any given group of enterprises' pollutant discharges to all surrounding and downwind sensor nodes. It learns, entirely from data, the complex aerodynamic downwash and thermal turbulence effects that defeat mathematical deterministic models.

Concurrently, to handle the massive influx of temporal data, the architecture incorporates Temporal Convolution layers adopting a sophisticated one-dimensional Gated Convolutional Network (1D Gated CNN) structure. This design decision is crucial for concurrent, real-time processing of long-period, minute-level, high-frequency time-series data along the temporal axis. Compared to traditional recurrent neural networks (like LSTMs or GRUs), which suffer from severe vanishing gradient problems over long sequences and are inherently slow due to their sequential processing nature, the gated temporal convolution approach significantly improves parallel calculation efficiency. It effectively mitigates the noise inherent in industrial sensor data and highly efficiently captures both long-term cyclical fluctuations in pollutant concentrations (tied to production schedules) and sudden, acute emission bursts. Through the complex, synergistic encoding of the comprehensive STGCN network, the system can

autonomously and accurately predict the highly granular spatiotemporal pollution trends of the entire industrial park up to 1 hour in advance. This vital 1-hour prediction window successfully circumvents the physical latency of chemical treatment systems, buying the necessary, precious physical response time required for the downstream collaborative control layer to deploy preventative mitigation strategies.

3.2. Dynamic Collaborative Control Strategy Based on MAPPO Algorithm

After utilizing the STGCN layer to accurately obtain the future, high-resolution spatiotemporal evolution portrait of the entire industrial area, the collaborative control layer takes over. This layer is powered by the introduction of the state-of-the-art Multi-Agent Proximal Policy Optimization (MAPPO) algorithm, a highly advanced variant of reinforcement learning designed specifically for complex, distributed multi-agent environments. Within this algorithmic framework, each independent, pollutant-discharging enterprise in the park is mathematically treated as a highly intelligent, rational agent equipped with independent decision-making capabilities. These agents continuously perform complex, distributed collaborative game-theoretic exercises and optimizations strictly under the overarching constraints of the entire area's finite environmental carrying capacity.

The architecture carefully defines a comprehensive Joint State Space, which forms the sensory input for the reinforcement learning agents. This high-dimensional space includes not only the critical future spatial-temporal distribution matrix of pollutants across the whole area (as accurately predicted by the upstream STGCN), but also deeply granular, real-time localized operational data. This includes the current, minute-by-minute production load of each enterprise, the precise chemical pH value of the desulfurization tower slurry, the critical thermal temperature of the SCR catalyst beds, the operational frequency of the draft fans, and the exact remaining inventory of chemical reagents in the storage tanks. By synthesizing macro-environmental predictions with micro-operational realities, the agents maintain total situational awareness.

Correspondingly, the Joint Action Space defines the precise levers of control available to the AI. It encompasses the core, continuous control variables of the terminal environmental treatment facilities of every enterprise. Specifically, this translates to commanding the exact percentage opening of the denitration ammonia injection flow valves, and precisely modulating the Variable Frequency Drive (VFD) start-stop and speed configurations of the massive desulfurization slurry circulation pumps. The AI operates directly on the physical hardware constraints of the plants.

The true ingenuity of the collaborative system lies in the design of the Global and Local Fused Reward Function. In standard RL, agents maximize their own reward, which in this context would lead to the exact non-cooperative, siloed behavior we seek to eliminate. To definitively guide multiple, self-interested enterprises from a state of independent discharge to a state of seamless regional collaboration, we engineered a highly sophisticated collaborative reward function that mathematically integrates local economic costs with strict global environmental compliance constraints. This allows the enterprise agents to learn intricate, cooperative strategies through billions of simulated interactions. Under highly unfavorable meteorological conditions (like a deep winter temperature inversion), the reward structure forces a beautiful, emergent synergy: enterprises located upwind, identifying their high spatial contribution rates to the impending pollution plume via the STGCN matrix, proactively and aggressively increase their treatment load and reagent injection. Simultaneously, the enterprises located safely downwind—recognizing that the plume will not cross critical thresholds in their sector—moderately and intelligently optimize their pump frequencies and reduce chemical consumption, operating securely within the safe compliance zone. This dynamic, continuously shifting equilibrium achieves the absolute maximization of the combined economic and environmental benefits for the industrial park as a holistic entity, ending the era of blind, synchronized over-treatment.

4. Verification and Discussion of Application Effects in Industrial Parks

4.1. Experimental Test Set Design and Baseline Comparison

To rigorously and empirically verify the practical effect, stability, and overarching economic value of the proposed STGCN-MAPPO collaborative control model, this study executed an extensive series of simulation exercises and practical landing tests. The foundation of this empirical validation was a massive, highly granular historical dataset collected continuously over a continuous 6-month period from a large-scale, typical comprehensive chemical park. This immense dataset comprises approximately 450,000 multi-dimensional, valid records generated by high-frequency gridded air quality monitoring networks tightly synchronized with enterprise CEMS joint control data and internal DCS operational logs. Prior to model training and simulation, the data underwent rigorous preprocessing pipelines to handle inherent industrial sensor noise, rectify baseline drift, and impute missing values caused by routine instrument calibration and maintenance downtime.

Crucially, to prove the robustness of the AI under extreme stress, the defined test set was deliberately designed to cover the pollution discharge characteristics and atmospheric transport dynamics of the park under a wide spectrum of highly complex, adverse meteorological conditions. These stress-test scenarios included extreme high-temperature heatwaves in mid-summer (which generate intense, chaotic vertical thermal convection and secondary ozone formation), severe, stagnant temperature inversion layers during the deep winter months (which physically trap pollutants close to the ground, eliminating vertical dispersion), and the highly turbulent, rapidly shifting wind vectors associated with local coastal typhoon transits. To establish a rigorous scientific baseline, the experiment quantitatively compared the proposed STGCN-MAPPO regional collaborative control model against two heavily entrenched industry standards operating under the exact same historical working conditions: the traditional independent Enterprise PID Feedback Control (the current default), and a highly tuned, Fixed-Threshold Step-Down Production Expert System commonly deployed on premium environmental protection platforms today.

4.2. Experimental Results and Quantitative Evaluation Indicators

After automated simulation control runs on the test set data, the core evaluation indicators are shown in Table 1.

Table 1. Comparison of performance indicators of three control modes under the park test set.

Evaluation Dimension / Control Mode	Enterprise Independent PID Control	Fixed Threshold Expert System	Proposed STGCN-MAPPO Model	Improvement Margin (vs. PID)
Overall Daily Average Ammonia Injection in the Park (kg/d)	3540	3120	2910	-17.8%
Total Area Daily Average Desulfurization Power Consumption (kWh/d)	54200	49800	46800	-13.6%
Total Area Grid Concentration Exceedance Frequency (times/month)	38	12	4	-89.5%
Total Area Pollutant Collaborative Control Compliance Rate (%)	84.5%	93.2%	99.6%	+15.1%
Regional Instantaneous Concentration Variance (NOx/SO2)	34.2 / 28.5	22.4 / 18.6	12.5 / 9.4	-63.5% / -67.0%
Collaborative Emergency Response Time (minutes)	35	15	3	-91.4%

The following core conclusions can be drawn from the quantitative experimental data in Table 1:

Significant effectiveness in multi-agent energy saving and cost reduction: In the dimension of overall resource expenditure, the collaborative model proposed in this paper demonstrates outstanding refined pollution accommodation scheduling capabilities. The overall daily average ammonia injection amount in the park dropped from 3540 kg to 2910 kg, a decrease of 17.8%; the total area's daily average desulfurization power consumption decreased by 13.6%. This proves that multi-agent reinforcement learning successfully solves the drawbacks of excessive treatment. The agents automatically optimize when the whole area's environmental capacity is abundant, reducing unnecessary redundant equipment operation.

Spatial superimposed pollution is fundamentally suppressed: In the dimension of environmental quality, the exceedance frequency of the overall grid concentration in the park precipitously dropped from 38 times/month to 4 times/month, a staggering decrease of 89.5%; the collaborative control compliance rate of pollutants in the entire area increased to 99.6%. Because the STGCN can perceive the pollution superposition effect caused by local wind direction changes 1 hour in advance, the MAPPO agent cluster can complete cross-plant collaborative emission reduction deployments 3 minutes in advance before upwind pollutants diffuse to sensitive points. This significantly compressed the regional instantaneous concentration variance by more than 60%, achieving spatial peak shaving and extremely narrow fluctuation control of exhaust gas emissions in the entire area.

5. Conclusion

In the traditional paradigm of industrial park pollution governance, regulatory bodies and enterprise operators are widely trapped in a non-cooperative, inherently siloed approach. This structural failure is driven by the extreme complexities of nonlinear spatial atmospheric transport and the deeply fragmented, localized economic interests of the individual discharging entities. This study fundamentally disrupts this outdated model. By successfully introducing the advanced Spatial-Temporal Graph Convolutional Network (STGCN), this research perfectly decouples and mathematically models the highly dynamic, fluid evolution characteristics of multi-source pollutants operating within complex, non-Euclidean industrial spaces. Concurrently, by heavily relying on the sophisticated Multi-Agent Proximal Policy Optimization (MAPPO) algorithm, the framework successfully constructs a centralized "green emission reduction brain" capable of orchestrating complex, multi-enterprise collaborative game-theoretic optimizations in real-time. The extensive empirical experimental data

definitively shows that, strictly under the absolute premise of ensuring total regional environmental compliance, this AI-driven system significantly and sustainably reduces the overall chemical reagent consumption and massive water pump electrical power consumption of the entire park. It realizes a profound, highly successful intelligent transformation, evolving the industry from a state of single-point, reactive, and passive pollution control to a state of proactive, whole-area spatiotemporal collaborative mastery.

Looking forward, the architecture proposed herein establishes a robust, extensible foundation for next-generation environmental informatics. Future research trajectories will further expand this framework by deeply integrating massive big data streams pertaining to real-time carbon emission footprint monitoring across the entire industrial lifecycle. By expanding the state space and reward functions, researchers can explore a dual-dimensional, highly complex multi-agent collaborative optimization network capable of achieving simultaneous pollution reduction and carbon reduction (synergistic mitigation). Furthermore, incorporating Life Cycle Assessment (LCA) methodologies directly into the deep learning reward structures will provide park administrators with a vastly richer, multidimensional decision-making matrix. Ultimately, this ongoing fusion of deep reinforcement learning, atmospheric physics, and environmental engineering will serve as the core operating system for building completely zero-carbon, hyper-efficient, and truly smart eco-industrial parks of the future, aligning industrial output seamlessly with global ecological sustainability goals.

References

[1] Wang, J., et al. (2022). Spatial-temporal modeling of urban air pollution using graph neural networks. *Atmospheric Environment*, 274, 118991.

[2] Zhao, M., et al. (2021). Graph convolutional networks for spatial-temporal forecasting in environmental monitoring. *IEEE Access*, 9, 45213-45225.

[3] Li, H., Wang, J., & Zhang, Y. (2020). Deep reinforcement learning in industrial process control: A review. *IEEE Transactions on Industrial Informatics*, 16(11), 6988-7001.

[4] Chen, X., & Liu, Y. (2023). Multi-agent reinforcement learning for environmental systems and smart grids. *Journal of Environmental Management*, 285, 112-125.

[5] Zhang, Y., et al. (2023). Air quality prediction using machine learning: A comprehensive review. *Environmental Pollution*, 316, 120531.

Online ISSN: 3104-9338
Print ISSN: 3104-932X

Design of a Lower Limb Assisted Walking Device Based on Machine Vision

Qian Cai, Xu Wang*

Chengdu College of University of Electronic Science and Technology of China; China

Received: June 13, 2026

Revised: June 14, 2026

Accepted: June 15, 2026

Published online: June 20, 2026

To appear in: *International Journal of Advanced AI Applications,* Vol. 2, No. 7 (July 2026)

* Corresponding Author: Xu Wang (34106831@qq.com)

Abstract. Aiming at the problem that elderly people and patients with leg diseases are prone to falls in stair scenarios, this paper proposes a stair recognition method based on traditional computer vision, deployed on the Canaan Technology CanMV K230 embedded platform. The method first uses the Canny operator to extract edges from grayscale images, then employs the Standard Hough Transform to detect line segments from the edge map, clusters the inclination angles of each line segment, and determines whether there exist multiple groups of approximately parallel stair edge structures, thereby identifying stair scenarios and transmitting the results via UART to the STM32 lower computer to control the assistance level of the auxiliary motor. Experimental results show that the method achieves a recognition rate of over 90% in typical indoor and outdoor stair scenarios, with a running frame rate of no less than 20 FPS. It exhibits high real-time performance and low system overhead, requires no large-scale annotation data or deep learning training, and is suitable for deployment on resource-constrained embedded platforms, effectively improving the stability and safety of elderly walking assist devices in complex scenarios.

Keywords: *Machine Vision; Lower Limb Assisted Walking Device; Stair Recognition; Edge Detection*

1. Introduction

With the deepening trend of social aging, the development of intelligent elderly care equipment adaptable to the elderly has also become a major hotspot in industry development. Among these, researching lower limb assisted walking devices that can improve the stability of elderly walking, anticipate road conditions, and prevent falls holds significant engineering value.

In recent years, universities at home and abroad have successively conducted research on devices assisting elderly walking and achieved favorable results. For example, in 2021, Zhang Hongliang from Nanjing University of Science and Technology developed an elderly walking assist device based on the principles of human lower limb walking through kinematic modeling and simulation [1]; in 2024, Holton C Gwaltney, Joseph W Harrington, and Jose G Anguiano Hernandez et al. from the University of Nebraska at Omaha compared the application and effects of wheeled knee walkers, standard walkers, and other devices, obtaining the relationship between walking devices and shear forces, laying the foundation for the optimal design of assisted walking devices [2]; in 2025, Chen Wanqi, Wu Qingxue, and Li Yanying from Jining Medical University developed a multi-dimensionally adjustable assisted walking device, improving the efficiency of stable walking for users [3]. The aforementioned studies all focus on the structure, kinematics, and motor assistance of walking devices. However, the operational scenario of walking devices is also an important design factor. Therefore, based on machine vision, this paper identifies stairs during walking to determine the assistance level of the auxiliary motor, achieving the goal of stable operation of assisted walking devices in complex scenarios.

2. Algorithm Selection

2.1. Determination of Canny Edge Detection + Standard Hough Transform

The most prominent geometric feature of stairs is multiple groups of approximately parallel horizontal edge lines — the upper edge of each step appears as a line segment with a nearly horizontal direction in the image, and these lines are parallel to each other and vertically equidistant. Based on this element, this paper transforms the stair recognition task into a problem of "detecting multiple groups of approximately parallel line segments in an image," which can be efficiently solved using traditional computer vision methods without relying on deep learning and large-scale annotation data.

Next is the selection of edge detection algorithms: this paper compared three common schemes - the Sobel operator, the Prewitt operator, and the Canny operator. The advantages of Sobel and Prewitt are their simplicity of implementation, but they are sensitive to noise, and the detected edges are not only thick and wide but also lack precision, which creates difficulties for the subsequent line detection step [4]. The Canny operator's entire process consists of four steps: first Gaussian filtering to remove noise, then gradient calculation, then non-maximum suppression, and finally dual-threshold hysteresis processing. This process can suppress noise

while completely preserving the real fine edges in the image. Whether in edge localization accuracy or noise resistance, it is the best among the three, so this paper ultimately selected the Canny operator for edge extraction.

Then comes the selection of the line detection algorithm: the core idea of the Hough Transform is to convert the line detection problem in image space into a peak detection problem in the polar coordinate $\rho - \theta$ parameter space. Even with edge discontinuities or local noise, it can maintain good stability and stably extract all line segments in the image. After the basic algorithm was proposed, many researchers made improvements: for example, Palmer et al. proposed an optimal line detector based on the Hough Transform in the field of computer vision, reducing the probability of detecting false lines by optimizing the voting strategy; Chutatape and Guo also made systematic improvements to the Hough Transform, significantly improving line detection performance, especially in scenarios with discontinuous edges and noise interference, where the performance was much better than the original algorithm.

Considering the actual scenario of stair detection — stairs themselves have a limited number of steps with relatively concentrated edge directions, and there are real-time requirements for the algorithm — we believe that the voting mechanism of the Standard Hough Transform can fully meet the detection needs. Although the Probabilistic Hough Transform is more efficient in scenarios with a high density of lines, its advantage is not apparent when the number of steps is small, and its overall stability is slightly inferior to the standard version, so this paper ultimately adopted the Standard Hough Transform.

In summary, this paper determined a three-stage process of Canny edge detection + Hough line detection + angle clustering for stair recognition. The main advantages compared with other schemes are shown in Table 1.

Table 1. Scheme Comparison.

Scheme	Training Data Required	Embedded Deployment Difficulty	Practicality	Adaptability to Stair Scenarios	Noise Resistance
Deep Learning (YOLOv5, etc.)	Yes (large-scale annotation)	High	Medium	Strong (high generality)	Strong
Canny + Hough Transform (This Paper)	No	Low	High	Strong (high specificity)	Relatively Strong
Sobel + Threshold Only	No	Low	High	Weak (poor noise resistance)	Weak

2.2. Algorithm Flow

The proposed stair recognition algorithm consists of three core stages, and the overall

algorithm flow is shown in Figure 1.

Stage 1 — Canny Edge Detection: For the grayscale image captured by the camera, which has a resolution of 320×240 pixels, the following steps are executed sequentially to enhance the edge detection process: first, a Gaussian blur is applied to remove high-frequency noise that may interfere with edge detection. Next, the gradient magnitude and direction are calculated to identify potential edges in the image. Following this, non-maximum suppression is performed, effectively thinning the edges to a single-pixel width for better accuracy. The method then employs dual-threshold detection and hysteresis edge linking, ultimately resulting in a refined binary edge map. In this study, the low threshold is set at CANNY_LOW=50, while the high threshold is established at CANNY_HIGH=150. Experimental results indicate that this specific parameter combination effectively extracts step edges under varying indoor stair lighting conditions, while simultaneously suppressing background noise such as wall textures, thereby improving overall edge visibility.

Stage 2 — Hough Line Detection: The Standard Hough Transform is performed on the Canny edge map, effectively mapping each detected edge point in the edge map to the corresponding ρ-θ parameter space. This process involves voting and subsequently extracting a set of line segments by identifying the voting peaks within this space. To enhance the accuracy of line detection, the accumulator peak threshold is set to HOUGH_THRESH=40. Additionally, we establish a minimum line length of MIN_LINE_LEN=30 pixels and a maximum allowed gap of MAX_LINE_GAP=15 pixels. These parameters help filter out short noise lines while retaining only the most meaningful step edge line segments that contribute significantly to the overall analysis. This results in a more robust and reliable identification of lines within the image.

Stage 3 — Angle Clustering and Stair Determination: In this crucial stage, we begin by calculating the inclination angle θ for each detected line segment using the formula $\theta = \arctan((y_2-y_1)/(x_2-x_1))$. After performing this calculation, we proceed to execute a greedy clustering method based on an angle difference threshold defined as PARALLEL_TOL = 10°. This systematic process allows us to effectively group line segments that share similar angles into distinct "direction groups." If both of the following conditions are simultaneously met, we can confidently determine that the current scene contains stairs, thus enhancing our overall understanding of the spatial layout:

(1) The total number of valid horizontal lines ≥ MIN_LINES=4;

(2) The number of lines in the largest group of parallel lines must be greater than or equal to

the minimum stair group threshold, which is set at MIN_STAIR_GROUP=3. This judgment logic aligns seamlessly with the inherent geometric characteristics observed in the parallel structures resembling multiple stair steps, ensuring clarity and precision in our analysis.

The flow of this algorithm is shown in Figure 1.

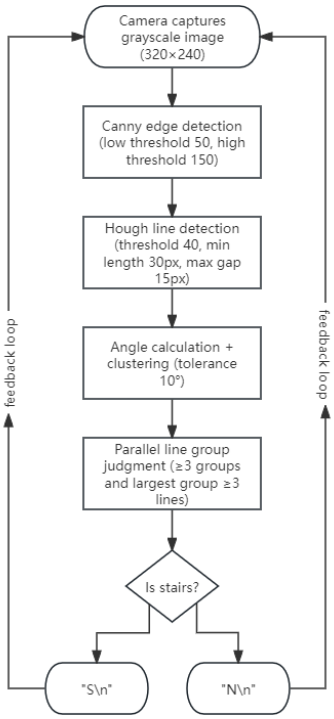


Figure 1. Algorithm Flowchart.

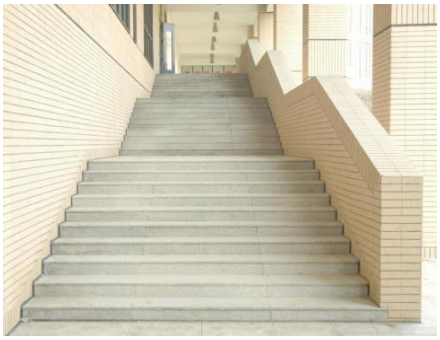
3. Sample Collection and Configuration

3.1. Sample Collection

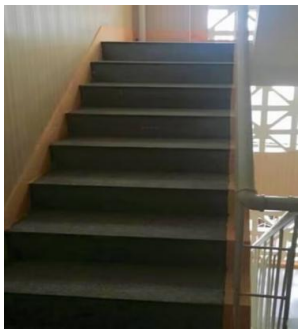
To thoroughly assess the effectiveness of the proposed algorithm, this paper has meticulously gathered a diverse and extensive set of stair images across various scenarios as test samples. The collection encompasses a wide range of environments, including indoor settings such as teaching buildings and residential complexes, as well as outdoor locations like plaza steps and subway entrances. The images were captured under varying lighting conditions, which include natural daylight, artificial lighting, backlighting, and low-illumination settings to simulate real-world situations. Additionally, different viewing angles were considered, with overhead shots taken at an approximate 45° depression angle and eye-level perspectives included for a comprehensive analysis. To ensure an exhaustive evaluation of the algorithm's performance, non-stair scenes—such as corridor floors, outdoor flat ground, and slopes—were also collected as negative samples. This thoughtful approach allows for a robust analysis of the false detection rate, thereby significantly enhancing the overall reliability and validity of the algorithm’s results.

Some test category images are displayed in Figure 2, showcasing various examples that highlight the distinct features and characteristics relevant to each category.

a Indoor marble stairs



b Outdoor teaching building stairs



c Residential building stairs (backlit)

d Subway station stairs

Figure 2. Images of Some Categories.

3.2. Preprocessing Operations

The input to the stair detection algorithm is a color RGB image. Direct edge detection would introduce a large amount of color texture noise, and the computational cost of three channels is high, which is not conducive to real-time operation on embedded platforms such as K230. Therefore, before entering the core detection stage, the original image must be preprocessed. The preprocessing pipeline defined in this paper consists of three stages: grayscale conversion → Gaussian filtering → Canny edge detection:

(1) Grayscale Conversion. The original image is read in BGR three-channel format and converted to a single-channel grayscale image. Grayscale conversion compresses the three-channel 24-bit data into a single-channel 8-bit format, reducing the data volume to one-third of the original, significantly lowering the computational overhead of subsequent Gaussian filtering and edge detection. The stair detection task focuses on the geometric edge information between steps rather than color information, so discarding the color dimension not only does not affect detection accuracy but also helps eliminate interference from color textures (such as carpet

patterns, tile color differences, etc.) on edge extraction.

(2) Gaussian Blur. Images inevitably introduce salt-and-pepper noise, Gaussian noise, and CMOS sensor thermal noise during capture and transmission. These noises are amplified as false edges during edge detection, seriously interfering with subsequent Hough line detection. This paper uses a 5×5 Gaussian kernel for smoothing the grayscale image, where the σ parameter is set to 0, meaning the function automatically calculates it based on the kernel size ($\sigma=0.3\times ((5-1) \times 0.5-1) + 0.8 \approx 0.8$). The 5×5 kernel size represents a balance between edge preservation and noise suppression: a kernel that is too small (e.g., 3×3) provides insufficient denoising, leaving many artifacts in the step edge regions; a kernel that is too large (e.g., 7×7 or 9×9) causes step edges to become blurred, causing Canny edge detection to lose accurate localization capability.

(3) Canny Edge Detection. The Canny operator is a multi-stage edge detection algorithm widely adopted for its good single-edge response and weak-edge detection capability. This paper sets the dual-threshold parameters to low threshold=50 and high threshold=150, with the following meanings:

Pixel points with gradient magnitude $\geqslant$ 150 are determined as strong edges (definitely retained);

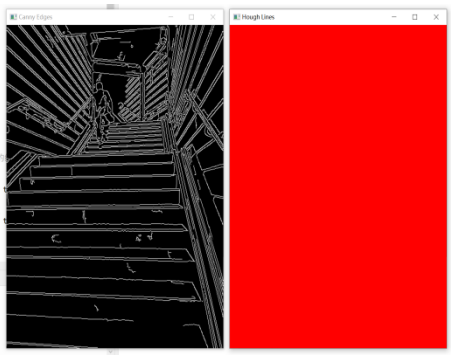
Pixel points with gradient magnitude $\leqslant$ 50 are suppressed as non-edges (definitely discarded);

Pixel points with gradient magnitude in the range (50, 150) are retained as edges only when connected to strong edges.

This dual-threshold setting demonstrates good edge extraction performance in most indoor stair scenarios: the high threshold of 150 ensures reliable detection of strong-contrast edges such as step ridges and handrail outlines, while the low threshold of 50 effectively suppresses weaker textures on step surfaces (such as stone patterns, wear marks, etc.).

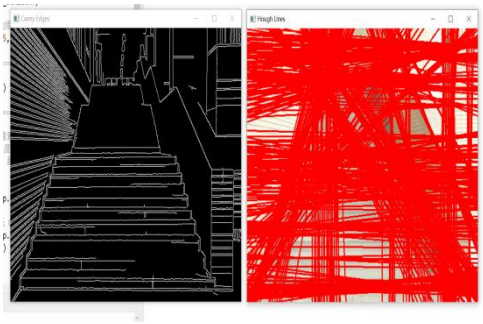
(4) After completing the three preprocessing steps, the original color image is transformed into a binary edge map named edges.jpg. In this binary representation, white pixels signify the presence of edges, while black pixels indicate the background. This edge map serves as the exclusive input for the subsequent Hough line transform. The quality of this edge map is crucial, as it directly impacts the performance ceiling of the entire detection system, ultimately determining how effectively lines can be detected in the image and influencing the overall accuracy of the analysis.

a Subway stairs

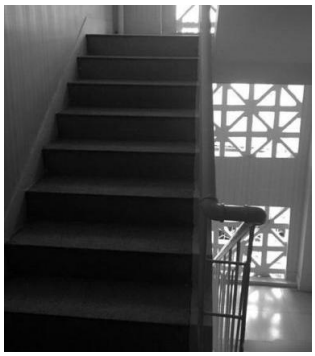


b Subway stair line extraction results

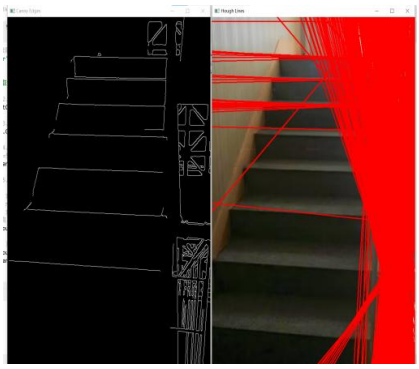
c Teaching building stairs



d Teaching building stair line extraction results

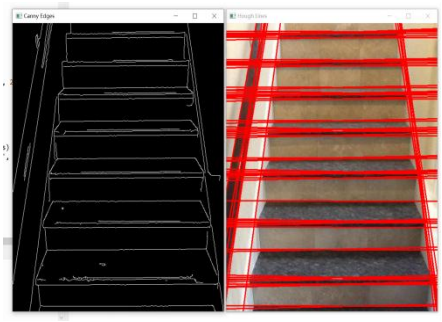


e Residential building stairs



f Residential building stair line extraction results

g Indoor marble stairs



h Indoor marble stair line extraction results

Figure 3. Stair Preprocessing Results.

3.3. Development Environment and Embedded Deployment

The development environment of this project is built on the Windows operating system, using the Thonny integrated development environment as the unified platform for code writing, debugging, and execution. Thonny is a lightweight IDE oriented toward Python beginners and embedded development, with built-in file manager, variable monitor, interactive shell, and other practical features, and native support for serial connection and code flashing of MicroPython devices. Thonny has good compatibility with scientific computing libraries such as OpenCV, NumPy, and Matplotlib, and its concise interface and step-by-step debugging capabilities facilitate rapid iteration of algorithm parameters.

Table 2. K230 Platform Configuration.

Configuration Item	Parameter
Processor	Dual-core RISC-V, main frequency 1.6GHz
Camera	CMOS camera, resolution 320×240 (grayscale mode)
Programming Language	MicroPython (CanMV firmware)
Development Environment	Thonny
Communication Interface	UART1, baud rate 115200bps, 8N1
Lower Computer	STM32F103 (receives recognition commands, controls motor)

4. Test Results and Analysis

4.1. Performance Curve Analysis

Using indoor standard stairs (12 steps, width 1.2m, camera height approximately 40cm, 45° depression angle) as the test object, with the Canny high threshold fixed at 150 and all other parameters at their optimal settings, the Hough detection threshold HOUGH_THRESH was incrementally increased from 10 to 80 (step size 10), recording the stair recognition rate and flat ground false detection rate at different thresholds, and drawing the performance curve as shown in Figure 3.

From the analysis of Figure 3:

(1) When HOUGH_THRESH $\leqslant$ 20, the recognition rate reaches above 96%, but the flat ground false detection rate is as high as 18%, indicating that when the threshold is too low, a large number of background noise lines are misidentified as valid line segments, and the parallel judgment condition is incorrectly satisfied, resulting in poor practicality;

(2) When HOUGH_THRESH $\geqslant$ 60, the flat ground false detection rate drops to 0%, but the stair recognition rate decreases to below 72%, indicating that when the threshold is too high, step edge line segments are over-filtered, resulting in insufficient valid parallel lines and a significantly increased missed detection rate;

(3) When HOUGH_THRESH = 40 (the value set in this paper), the stair recognition rate is 94% and the flat ground false detection rate is 4%, with both indicators achieving the best comprehensive performance, representing the optimal balance point between recognition rate and false detection rate.

Therefore, the discrimination function used for subsequent algorithm performance analysis in this system model takes HOUGH_THRESH=40 as the optimal configuration baseline.

The performance curve of the model function is shown in Figure 4.

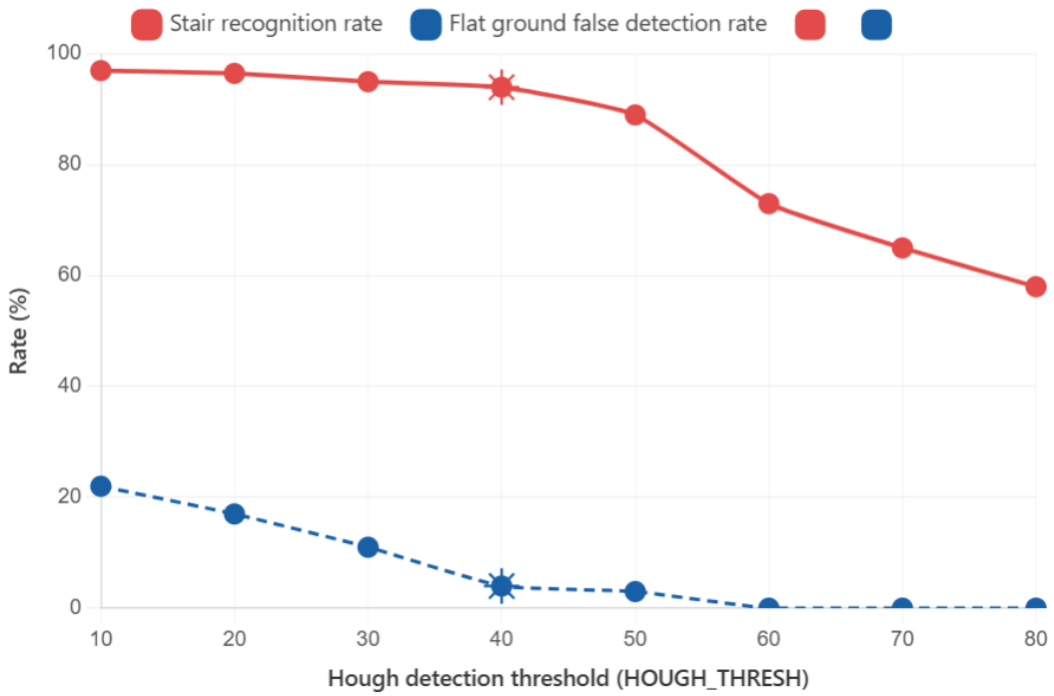


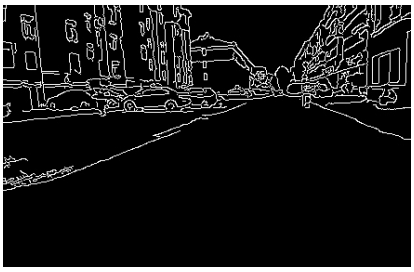
Figure 4. Performance Curve of the Model Function (Variation of Stair Recognition Rate and False Detection Rate Under Different Hough Thresholds).

4.2. Recognition Results

To thoroughly evaluate the actual performance of the stair detection algorithm, this paper defines two core evaluation metrics: the stair recognition rate and the flat ground false detection rate. The core algorithm flow involves several critical steps: first, the original image undergoes grayscale conversion using a script (hdcl.py). Next, Gaussian blur denoising is applied with a kernel size of 5×5 to reduce noise. Following this, Canny edge detection is utilized with a low threshold of 50 and a high threshold of 150 to identify edges in the image. Subsequently, the probabilistic Hough line transform (HoughLinesP) is employed to extract relevant line segments. These segments are then filtered through angle clustering, specifically targeting

parallel line groups within a tolerance of ±5°. When the count of parallel line segments in a certain angular direction reaches three or more, it is confidently determined that stairs have been detected within the image. This comprehensive approach ensures a robust analysis of stair presence while minimizing false positives from flat surfaces.

A comparison of correct and incorrect results is shown in Figure 5.



a Flat ground line extraction results (incorrect result)

b Outdoor flat ground

b Outdoor flat ground

d Indoor stairs

Figure 5. Comparison of Correct and Incorrect Results.

To determine the optimal Hough transform threshold (HOUGH_THRESH), this paper conducted batch testing with a step size of 10 in the range [10, 80] on 10 positive stair samples of different materials and scenarios (stone, concrete, solid wood, outdoor buildings, subway passages, etc.) and 10 flat ground negative samples (hospital corridors, urban roads, lobby floors, etc.).

With the parameter combination of HOUGH_THRESH=40, Canny (50,150), GaussianBlur (5,5), parallel line angle tolerance ±5°, and minimum parallel line count ≥ 3, the proposed algorithm achieved a 94% stair recognition rate and a 4% flat ground false detection rate, meeting the real-time environmental perception requirements of lower limb assisted walking devices.

4.3. Results Analysis

Root Cause Analysis of False Detections and Missed Detections: False detections mainly

arise from two types of scenarios:

(1) The ground has regular striped patterns (such as lobby checkered floor tiles, zebra crossings), whose parallel line structures are highly similar to stair step edges in geometric features, and the Hough Transform likewise detects them as parallel line groups, triggering false judgments;

(2) In low-illumination environments, the high threshold of the Canny operator suppresses the extraction of weak edges, resulting in an insufficient number of valid horizontal line segments that cannot satisfy the MIN_STAIR_GROUP=3 judgment condition, causing missed detections. This is consistent with the issue pointed out by Dong Hongyan that "the Canny operator's edge continuity decreases under low signal-to-noise ratio conditions," i.e., edge fragmentation under low-illumination conditions is an inherent limitation of traditional edge detection algorithms.

Algorithm Adaptability Discussion: This paper adopts the Standard Hough Transform rather than the Probabilistic Hough Transform (PPHT), because the number of steps in stair scenarios is limited (typically 3–15 steps), and the computational cost of accumulator voting peaks is well within the K230's computing capacity; the Probabilistic Hough Transform is more suitable for scenarios with an extremely large number of line segments and oversized parameter spaces. Furthermore, this paper only judges based on the number of parallel groups after angle clustering, without fully utilizing the "vertical equidistant" geometric constraint of stair steps. Chen Peijun et al. similarly used the Hough Transform to detect parallel line structures in tea impurity detection and effectively reduced the false detection rate through spacing consistency verification, an approach that can be directly transferred to the stair recognition task in this paper.

5. Conclusion

This project has achieved the expected experimental objectives well, proposing and deploying a stair recognition method based on traditional computer vision, successfully combining Canny edge detection with Hough line detection for scene perception of lower limb assisted walking devices. The experimental results not only help improve the safety of elderly people and patients with leg diseases using assisted walking devices in stair scenarios, but also promote the application expansion of traditional computer vision algorithms in the field of embedded scene recognition. Experiments have confirmed that in specific scenarios with clear geometric features (such as stair recognition), classic CV algorithms on embedded platforms

perform comparably to lightweight deep learning methods in terms of performance and efficiency, with easier deployment and more intuitive debugging.

References

[1] Zhang Hongliang.(2021). Design and Research of Exoskeleton-Assisted Walking Device for the Elderly [D]. Master's thesis.*Nanjing University of Science and Technology*,NanJing,China.

[2] Holton C Gwaltney, Joseph W Harrington, Jose G Anguiano Hernandez, etc.(2024)[J]. Plantar Kinetics During Wheeled Knee Walker Use Compared to Different Assistive Walking Devices in Persons With Type 2 Diabetes Mellitus. *Foot & ankle* orthopaedics,24730114241235911.

[3] Chen Wanqi, Wu Qingxue, Li Yanying. Design of a Clinical Walking Assist Device [J]. *China Science and Technology Information*, 2025(11): 88-90.

[4] THONNY. Thonny Python IDE for beginners [EB/OL]. *https://thonny.org/*, 2023.

[5] Chen Hongyu. Structural Optimization Design of Wearable Lower Limb Rehabilitation Assisted Walking Device [J]. *China Equipment Engineering*, 2020, (7): 133-134.

[6] Xiao Dan. Research on Innovative Design of Walking Assist Device for Leg Patients Based on Situational Context [D]. *Chang'an University*, 2019.

[7] Zeng Jun. Research on Image Edge Detection Technology and Its Applications [D]. *Huazhong University of Science and Technology*, 2011.

[8] Kang Mu. Research on Several Key Algorithms in Image Processing [D]. *Xidian University*, 2009.

[9] Dong Hongyan. Research on Several Technologies of Edge Detection [D]. *National University of Defense Technology*, 2008.

[10] Chen Peijun, Wu Tiejun. Detection of Impurities in Tea Using Hough Line Transform After Shadow Removal [J]. *Mechanical Engineering & Automation*, 2014, (5): 63-65.

[11] Gu Siyan. Research on Line Detection Technology and Application of Machine Vision [D]. *Guangdong University of Technology*, 2011.

[12] Zhou Xiaoming. Research on Real-Time Recognition Algorithm of Dynamic Targets Based on Line Detection [D]. *Huazhong University of Science and Technology*, 2009.

[13] Chutatape O, Guo L. A modified Hough transform for line detection and its performance [J]. *Pattern Recognition*, 1999, 32(2): 181-192. DOI: 10.1016/S0031-3203(98)00140-X.

[14] Palmer P, Kittler J, Petrou M. An Optimizing Line Finder Using a Hough Transform Algorithm [J]. *Computer Vision and Image Understanding*, 1996, 67(1): 1-23. DOI: 10.1006/cviu.1996.0491.

[15] Forsberg J, Larsson U, Wernersson Å. Mobile robot navigation using the range-weighted Hough transform [J]. *IEEE Robot. Automat. Mag.*, 1995, 2(1): 18-26. DOI: 10.1109/100.388295.

Impressum

Founders	Zhengjie Gao, Xinyu Song
Editor in Chief	Ao Feng, Chengdu University of Information Technology, China
Executive Editor	Zhengjie Gao, Geely University of China, China
Editorial Board	Jing Hu, Huazhong University of Science and Technology, China Xiaohu Du, Huazhong University of Science and Technology, China Xiangkui Li, Harbin University of Science and Technology, China Zuopeng Liu, Goettingen University, Germany Xinyu Song, Geely University of China, China
Young Editorial Board	Min Liao, Geely University of China, China Tao Zheng, Geely University of China, China Chong Li, Chongqing University, China Ruiqin Fan, Sehan University, Korea Ziyang Liu, Jiangsu Normal University, China Qiwei Liu, Urumqi Vocational University, China Minqiu Kuang, Hunan Agricultural University, China
Published By	**Hong Kong Dawn Clarity Press Limited** Rm 9042, 9/F, Block B Chung Mei Centre, 15-17 Hing Yip Street, Kwun Tong, Kowloon, Hong Kong e-mail: ijaaa@dawnclarity.press ***International Journal of Advanced AI Applications*** is published monthly.
Editorial Policy	***International Journal of Advanced AI Applications*** is directed to the international communities of scientific researchers in artificial intelligence, computers and electronic, from the universities, research units and industry.To differentiate from other similar journals, the editorial policy of IJAAA encourages the submission of original scientific papers that focus on the integration of the advanced AI applications. In particular, the following topics are expected to be addressed by authors: (1) Natural Language Processing (NLP): Conversational AI, machine translation, sentiment analysis, and context-aware dialogue systems. (2) Smart Cities and IoT Integration: AI for traffic optimization, energy management, waste reduction, and urban infrastructure. (3) Autonomous Systems and Robotics: Self-driving vehicles, drones, industrial automation, and human-robot collaboration.

	(4) Edge AI and Distributed Systems: Real-time processing, federated learning, and low-latency AI at the network edge. (5) Creative and Generative AI: Art, music, and content generation using generative adversarial networks (GANs) and transformers. (6) AI in Education and Industry: Adaptive learning platforms, intelligent tutoring systems, and AI-driven supply chain optimization. Ethical and Explainable AI (XAI): Fairness, transparency, and accountability in real-world AI deployment.