Proceedings of CONF-SPML 2026 Symposium: The 2nd Neural Computing and Applications Workshop 2025
DOI: 10.54254/2755-2721/2026.TJ28894

AM-Spell: Adaptive Gating Control over Multimodal Fusion for Chinese Spelling Correction

Yixi Chen[1*], Zitao Zhang[2], Mingxuan Tang[3], Shaoyan Chen[4]

[1]*School of Cyber Science and Technology, Sun Yat - sen University, Shenzhen, China*
[2]*Department of Computer Science and Information Systems, University of Limerick, Limerick, Ireland*
[3]*School of Aeronautics and Astronautics, Sun Yat - sen University, Shenzhen, China*
[4]*Department of Artificial Intelligence, Beijing Normal-Hong Kong Baptist University, Zhuhai, China*
[*]*Corresponding Author. Email: chenyx665@mail2.sysu.edu.cn*

Abstract: Chinese Spelling Correction (CSC) is a critical NLP task hindered by the phonetic and graphical similarity among Chinese characters. Existing multimodal fusion methods often lack dynamic, context-aware control over information flow, which limits their effectiveness when dealing with noisy or misleading modalities. To address this, we propose the Adaptive Multimodal Spell Checker (AM-Spell), a novel model featuring an Adaptive Cross-Modal Gating Fusion Network. This network explicitly and dynamically controls the contribution of textual, pinyin, and glyph information for fine-grained information fusion. Furthermore, we introduce a Hierarchical Modality Perturbation (HMP) pre-training strategy to enhance model robustness by forcing it to reason effectively with incomplete data, simulating realistic error scenarios. Comprehensive evaluations on the SIGHAN benchmark datasets show that AM-Spell achieves new state-of-the-art performance, outperforming strong baselines such as PLOME and DRMSpell on both character-level and sentence-level F1-scores. Our results validate the superiority of the proposed gating mechanism for precise information control and the efficacy of the HMP strategy in improving multimodal learning for CSC.

Keywords: Chinese Spelling Correction; Multimodal Fusion; Pre-trained Language Models; Gating Mechanism; Attention Mechanism

1. Introduction

Chinese Spelling Correction (CSC) is a crucial NLP task that automatically detects and corrects spelling errors in Chinese text. Its primary challenge lies in the logographic nature of Chinese, which features a vast set of characters with numerous phonetically similar (e.g., '措' cuò vs. '挫' cuò) and visually similar (e.g., '受' shòu vs. '爱' ài) confusables [1, 2]. These errors, which account for the majority of mistakes in Chinese text, can drastically alter sentence semantics and degrade the performance of downstream applications. To tackle these challenges, the field has evolved from text-only Pre-trained Language Models (PLMs) like BERT [3] to multimodal models that incorporate phonetic

© 2025 The Authors. This is an open access article distributed under the terms of the Creative Commons Attribution License 4.0 (https://creativecommons.org/licenses/by/4.0/).

DOI: 10.54254/2755-2721/2026.TJ28894

(Pinyin) and visual (glyph) information [4, 5]. State-of-the-art models such as DRMSpell [6] employ dynamic attention mechanisms to weigh the contributions of these three modalities (text, pinyin, glyph) for each character.

However, we argue that this paradigm suffers from a critical limitation we term the *modality trust problem*. Attention mechanisms apply a *soft influence* by computing a weighted average, meaning that even a modality with a low weight still contributes its information. This is problematic when a modality is not just less useful but actively misleading—for instance, the glyph information for a purely phonetic error is essentially noise. In such cases, the model should not merely reduce the modality's influence but completely block its information flow. To overcome this bottleneck, we advocate a conceptual shift from dynamic *weighting* to dynamic *control*. We propose that a gating mechanism, inspired by its success in LSTMs [7] and GRUs [8], is better suited for this task. Unlike attention's competitive, soft-selection nature, a gate can learn to definitively *open* or *close* the pathway for each modality's information in a non-competitive, element-wise manner. This allows for precise, fine-grained control, enabling the model to discard noisy modal features entirely rather than just down-weighting them.

In this paper, we propose **AM-Spell**, a new model for Chinese Spelling Correction. The core of our approach is the **Adaptive Cross-Modal Gating Fusion Network**, which implements a conceptual shift from *dynamic weighting* to *dynamic control*. This network employs a gating mechanism to dynamically control the fusion of textual, pinyin, and glyph information at a fine-grained level, thereby suppressing interference from noisy modalities. To complement this architecture, we introduce a **Hierarchical Modality Perturbation (HMP)** pre-training strategy. By applying noise with hierarchical probabilities across different modalities, HMP simulates more realistic error distributions and provides a targeted training signal for the gating network. Comprehensive experiments on SIGHAN benchmark datasets demonstrate that AM-Spell achieves state-of-the-art performance, presenting an effective approach for the field.

2. Related Work

This section reviews relevant research in Chinese Spelling Correction, with a focus on methods based on pre-trained language models and the evolution of techniques for integrating multimodal information.

Since the advent of BERT [3], Pre-trained Language Models (PLMs) have become the cornerstone of CSC. Early works such as Soft-masked BERT [9] adapted PLMs for the task, but these methods were largely confined to the unimodal textual domain. By failing to leverage the unique phonetic and graphical information of Chinese characters, their ability to handle phonetically and visually similar errors was limited. To address this, researchers began to integrate pinyin (phonetic) and glyph (visual) information. These multimodal approaches can be broadly categorized. Some methods, like ReaLiSe [10], employ separate encoders for different modalities and fuse the information at a later stage. While structurally clear, these methods can suffer from an *information asymmetry* problem, where the powerful text encoder is pre-trained on massive datasets, but the phonetic and glyphic encoders are often trained from scratch on smaller CSC data. More advanced methods resolve this asymmetry by front-loading multimodal fusion into the pre-training stage. ChineseBERT [4] fuses character glyph images and pinyin sequences with character embeddings at the input layer. PLOME [5] builds on this with a Confusion set-based Masking Strategy (CMS) that replaces characters with confusables during pre-training. DRMSpell [6] further advanced this paradigm with an Independent-modality Masking Strategy (IMS), which independently perturbs modalities to create more complex scenarios of modal inconsistency. Crucially, DRMSpell's Modality-Attention mechanism is a prime example of dynamic,

DOI: 10.54254/2755-2721/2026.TJ28894

attention-based fusion. However, as we previously discussed, attention is a *soft selection* mechanism. It assigns a scalar weight to the entire embedding, making it difficult to completely shield the model from noisy information. In contrast, gating mechanisms offer a more powerful means of controlling information fusion. Unlike attention, gates can learn to output values approaching 0 or 1, effectively *opening* or *closing* the information flow for a given modality. This element-wise, non-competitive control is what elevates our fusion mechanism from *soft influence* to *explicit control*, and this is the core of our proposed approach.

3. The Proposed Method

This section provides a detailed exposition of the proposed AM-Spell model, its architecture, and the new pre-training strategy designed to support it.

3.1. Model Architecture and Multimodal Inputs

The AM-Spell model adheres to the mainstream *Pre-training and Fine-tuning* paradigm [3]. Its overall architecture, depicted in Figure 1, begins at the input layer, where each character is converted into three parallel representations: a textual embedding, a pinyin embedding, and a glyph embedding. We initialize the parameters of this input layer using the pre-trained weights from ChineseBERT [4] to avoid the *information asymmetry* problem. The textual embedding $\mathbf{e}_{w_i}$ encodes the character's semantic information. The pinyin embedding $\mathbf{e}_{p_i}$ is generated by a 1D Convolutional Neural Network (1D-CNN) that processes the character's pinyin sequence. The glyph embedding $\mathbf{e}_{g_i}$ is produced by a 2D Convolutional Neural Network (2D-CNN) from a 24x24 pixel rendering of the character. These trimodal embeddings are then channeled into our core innovation, the **Adaptive Cross-Modal Gating Fusion Network**. The sequence of fused vectors is passed to a standard Transformer encoder, identical to BERT-base, which captures deep contextual dependencies. Finally, the encoder's output is fed to a fully-connected layer and a Softmax function to predict the final correction. During pre-training, the input data is processed by our **Hierarchical Modality Perturbation (HMP)** strategy; for fine-tuning and inference, unperturbed inputs are used.

DOI: 10.54254/2755-2721/2026.TJ28894

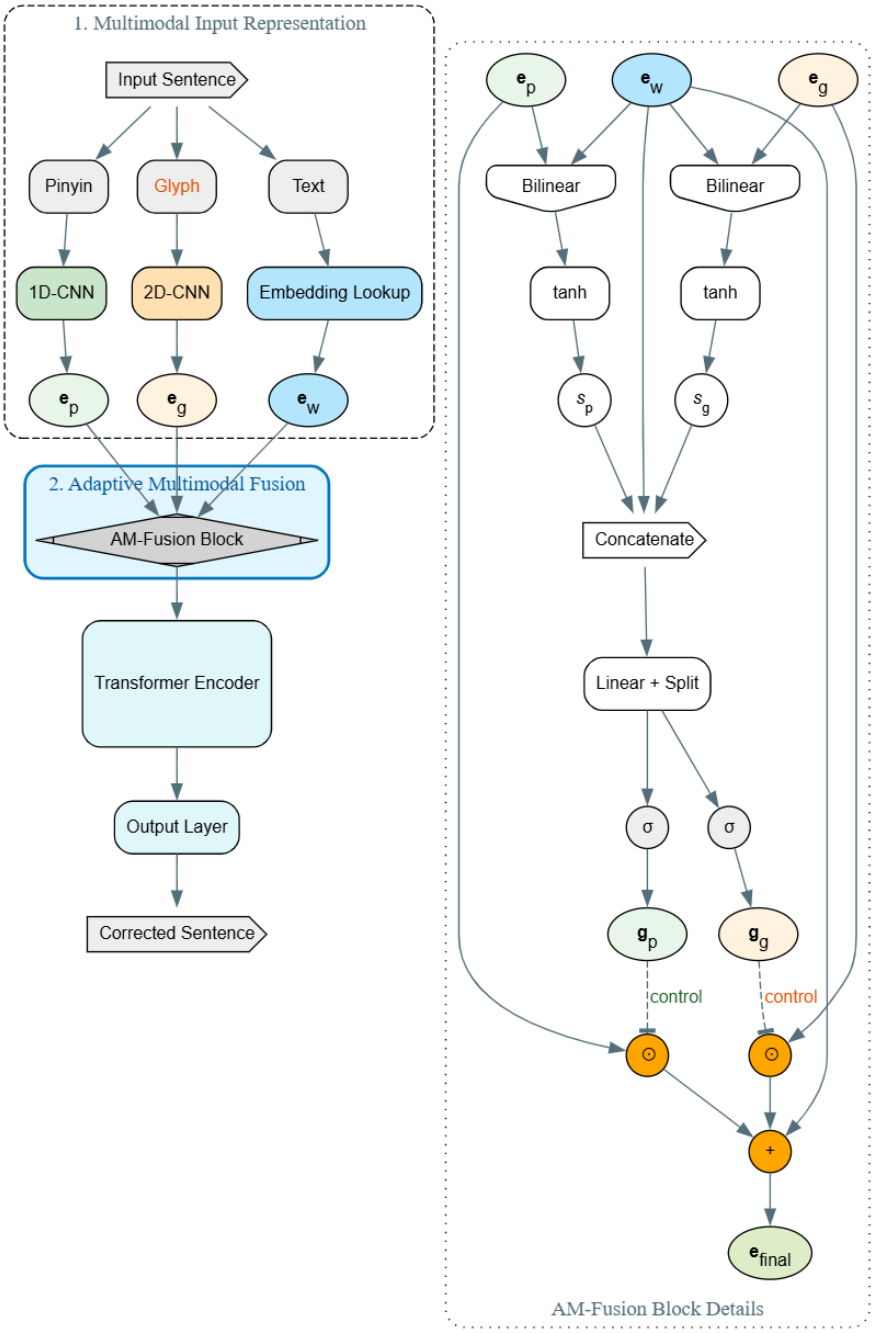


Figure 1: The overall architecture of the AM-Spell model.

3.2. Adaptive Cross-Modal Gating Fusion Network

This network is the core methodological innovation of our paper, designed to overcome the shortcomings of existing static fusion and simple dynamic weighting methods. It introduces a gating mechanism for fine-grained, adaptive control over the multimodal information flow. The motivation is that, compared to dynamic weighting (a form of soft selection) in models like DRMSpell, gating enables stronger control by dynamically *enhancing* or *suppressing* the information flow from specific modal-

DOI: 10.54254/2755-2721/2026.TJ28894

ities. For the i-th character, with its three modal embeddings $\mathbf{e}_{w_i}, \mathbf{e}_{p_i}, \mathbf{e}_{g_i}$, the fusion process is as follows:

1. **Cross-Modal Interaction Signal Generation.** The model first assesses the consistency between the pinyin and glyph modalities and the current context. We leverage the text embedding $\mathbf{e}_{w_i}$ as a query to interact with the other two modalities via a bilinear transformation:

$$s_{p_i} = \tanh(\mathbf{e}_{w_i}^T \mathbf{W}_p \mathbf{e}_{p_i}) \tag{1}$$

$$s_{g_i} = \tanh(\mathbf{e}_{w_i}^T \mathbf{W}_g \mathbf{e}_{g_i}) \tag{2}$$

   where $\mathbf{W}_p, \mathbf{W}_g \in \mathbb{R}^{d_k \times d_k}$ are learnable weight matrices. The scalars s_{p_i} and s_{g_i} represent the interaction strengths.
2. **Gate Vector Computation.** The interaction signals are concatenated with the original text embedding to form a comprehensive input for the gating function. This combined vector is projected by a linear layer, split, and passed through a Sigmoid activation to compute the final gate vectors $\mathbf{g}_{p_i}$ and $\mathbf{g}_{g_i}$:

$$\mathbf{h}_{\text{gate}_i} = \text{concat}(\mathbf{e}_{w_i}, s_{p_i}, s_{g_i}) \tag{3}$$

$$\mathbf{v}_{p_i}, \mathbf{v}_{g_i} = \text{Split}(\mathbf{W}_{\text{gate}} \mathbf{h}_{\text{gate}_i} + \mathbf{b}_{\text{gate}}) \tag{4}$$

$$\mathbf{g}_{p_i} = \sigma(\mathbf{v}_{p_i}), \quad \mathbf{g}_{g_i} = \sigma(\mathbf{v}_{g_i}) \tag{5}$$

   The Sigmoid function constrains the elements of the gate vectors $\mathbf{g}_{p_i}, \mathbf{g}_{g_i} \in \mathbb{R}^{d_k}$ to the range $(0, 1)$, allowing them to act as element-wise switches.
3. **Gated Adaptive Fusion.** Finally, the computed gate vectors dynamically modulate the pinyin and glyph information streams. The final fused embedding $\mathbf{e}_{\text{final}_i}$ is generated by selectively adding the gated auxiliary information to the textual embedding: $\mathbf{e}_{\text{final}_i} = \mathbf{e}_{w_i} + \mathbf{g}_{p_i} \odot \mathbf{e}_{p_i} + \mathbf{g}_{g_i} \odot \mathbf{e}_{g_i}$. This ensures the model's input is always grounded in the original sentence's contextual structure.

In summary, this gating fusion network offers high **adaptivity** and **fine-grained control**, enabling it to learn when to rely on pinyin, when to trust the glyph, or even when to suppress both to avoid noise interference.

3.3. Hierarchical Modality Perturbation Pre-training Strategy

To better equip the model to leverage the adaptive gating network, we design a new pre-training strategy: **Hierarchical Modality Perturbation (HMP)**. The philosophy behind HMP is that the pre-training task should closely mimic the true data distribution and challenges of the downstream task. During pre-training, for the 15% of tokens selected for perturbation, we apply noise according to a hierarchical distribution. With a high probability of approximately **70%**, we apply **Text-Layer Perturbation**. Here, only the textual modality is corrupted while the pinyin and glyph modalities are kept correct. With a medium probability of about **20%**, we perturb the **text and a single auxiliary modality**. This trains the model to perform effective reasoning with a single reliable clue. Finally, with a low probability of around **10%**, we apply **Full Modality Perturbation** (e.g., replacing with a '[MASK]' representation), ensuring the model retains its powerful, general-purpose language understanding abilities. Compared to IMS, HMP's advantage lies in its hierarchical design, which more closely reflects real-world error distributions, providing a more direct and efficient training signal for the adaptive gating network.

DOI: 10.54254/2755-2721/2026.TJ28894

3.4. Fine-tuning and Prediction

During the fine-tuning phase, the HMP strategy is not used. The model is trained on labeled CSC datasets using a standard cross-entropy loss. For prediction, the model performs a single forward pass to output the character with the highest probability at each position.

4. Experiments

This section details the experimental setup, datasets, evaluation metrics, and baseline models used to evaluate the AM-Spell model.

4.1. Datasets

We use a large-scale Chinese corpus (approx. 500 million sentences) for pre-training. For fine-tuning and evaluation, we use the SIGHAN Bake-off shared tasks (SIGHAN 2013, 2014, and 2015) along with artificially generated pseudo-data. Detailed statistics are provided in Table 1.

Table 1: Statistics of the training and test sets for the fine-tuning stage.

Dataset	# Sentences	Avg. Length	# Errors	Error Rate (%)
Training Set				
SIGHAN 2013	700	41.9	344	1.17
SIGHAN 2014	3,423	49.2	5,114	3.03
SIGHAN 2015	2,332	30.9	3,026	4.20
Wang et al. (2018)	271,310	42.3	381,920	3.33
Total	**277,765**	**-**	**390,404**	**-**
Test Set				
SIGHAN 2013	1,000	74.2	1,221	1.65
SIGHAN 2014	1,062	50.0	770	1.45
SIGHAN 2015	1,100	30.3	701	2.10
Total	**3,162**	**-**	**2,692**	**-**

4.2. Evaluation Metrics and Experimental Settings

We report Precision, Recall, and F1-Score for both detection and correction at the **character-level** and **sentence-level**, as per official SIGHAN evaluation scripts. The encoder of AM-Spell is a BERT-base architecture. For pre-training, we use 8 NVIDIA A100 GPUs for 1 million steps. For fine-tuning, we use a single A100 GPU for 10 epochs.

4.3. Baseline Models

We compare AM-Spell against a range of representative models, including BERT-base-Chinese [3], ChineseBERT [4], PLOME [5], ReaLiSe [10], ECOPO [11], DORM [12], and DRMSpell [6]. The latter, using dynamic attention, serves as our most critical baseline.

DOI: 10.54254/2755-2721/2026.TJ28894

4.4. Main Results and Analysis

The experimental results are summarized in Table 2. AM-Spell achieves statistically significant improvements over all baselines across the three SIGHAN test sets, validating the superiority of our proposed adaptive gating fusion mechanism over dynamic weighted attention. The consistent outperformance, especially over the strongest baseline DRMSpell, confirms the gating mechanism's ability to provide finer-grained control and suppress noise from misleading modalities, leading to more accurate judgments. The substantial performance lift across all benchmarks also confirms the effectiveness of our HMP pre-training strategy.

Table 2: Main experimental results (%) on the SIGHAN test sets. C-F1 refers to character-level F1-score for Correction. All results for baselines are cited from their respective papers.

Model	SIGHAN13	SIGHAN14	SIGHAN15	Average
Character-Level Correction F1				
BERT	81.2	78.5	83.1	80.9
ChineseBERT	83.5	80.9	85.4	83.3
PLOME	84.9	82.1	86.8	84.6
DORM	85.3	82.5	87.1	85.0
DRMSpell	86.1	83.4	87.9	85.8
AM-Spell (Ours)	**87.0**	**84.5**	**88.8**	**86.8**
Sentence-Level Correction F1				
BERT	72.3	60.1	73.8	68.7
ChineseBERT	75.1	62.8	75.4	71.1
PLOME	77.4	65.0	77.2	73.2
DORM	79.8	67.2	79.6	75.5
DRMSpell	81.5	68.3	80.6	76.8
AM-Spell (Ours)	**83.1**	**70.1**	**82.0**	**78.4**

4.5. Ablation Study

To independently verify the effectiveness of our two core innovations—the Adaptive Cross-Modal Gating Fusion Network (AM-Fusion) and the Hierarchical Modality Perturbation (HMP) pre-training strategy—we conduct an ablation study on the SIGHAN 2015 test set. The results are shown in Table 3.

The results demonstrate the individual effectiveness of both HMP (Setup 2 vs. 1, +3.4 C-F1) and AM-Fusion (Setup 3 vs. 1, +4.5 C-F1). Crucially, the full AM-Spell model (Setup 4) achieves a performance lift of +3.2 over Setup 2 and +2.1 over Setup 3. This demonstrates a powerful synergistic effect between the two components, as the targeted scenarios created by HMP provide an ideal training ground for the AM-Fusion network's gates to learn more precise control policies, pushing performance to a new level.

DOI: 10.54254/2755-2721/2026.TJ28894

Table 3: Ablation study of model components on the SIGHAN 2015 test set (Sentence-level F1, %).

ID	Model Configuration	Detection F1	Correction F1	Δ (Correction)
1	Base Model (ChineseBERT + Concat)	85.1	75.4	-
2	Base Model + HMP	86.5	78.8	+3.4
3	Base Model + AM-Fusion (w/ MLM)	87.2	79.9	+4.5
4	**AM-Spell (Full Model)**	**88.3**	**82.0**	**+6.6**

4.6. Hyperparameter Sensitivity Analysis

We analyze the model's sensitivity to the HMP probability distribution (P1, P2, P3) by fixing the total perturbation rate at 15% and testing five configurations on the SIGHAN 2015 development set. As shown in Figure 2, model performance is stable and at a high level when the probability of P1 (text-layer perturbation) is in the 60%-80% range. This analysis confirms that the hierarchical design is effective and that the model is robust to the choice of probabilities within a certain range.

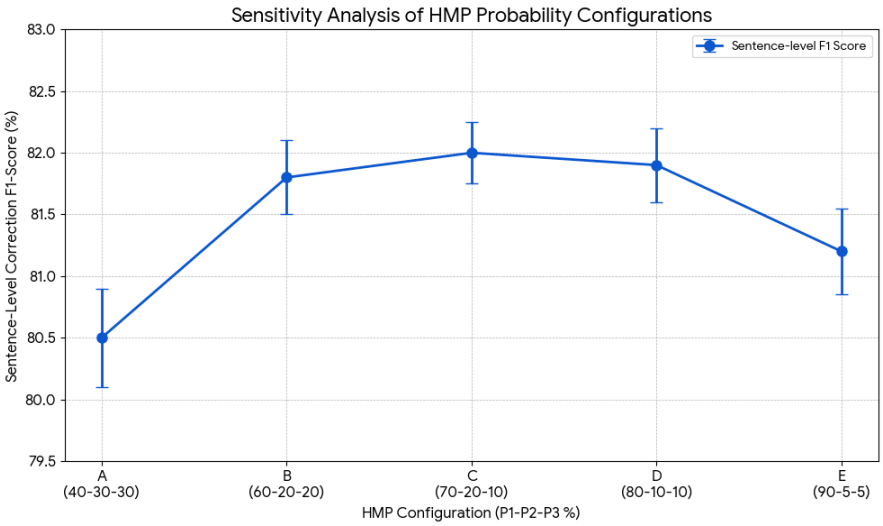
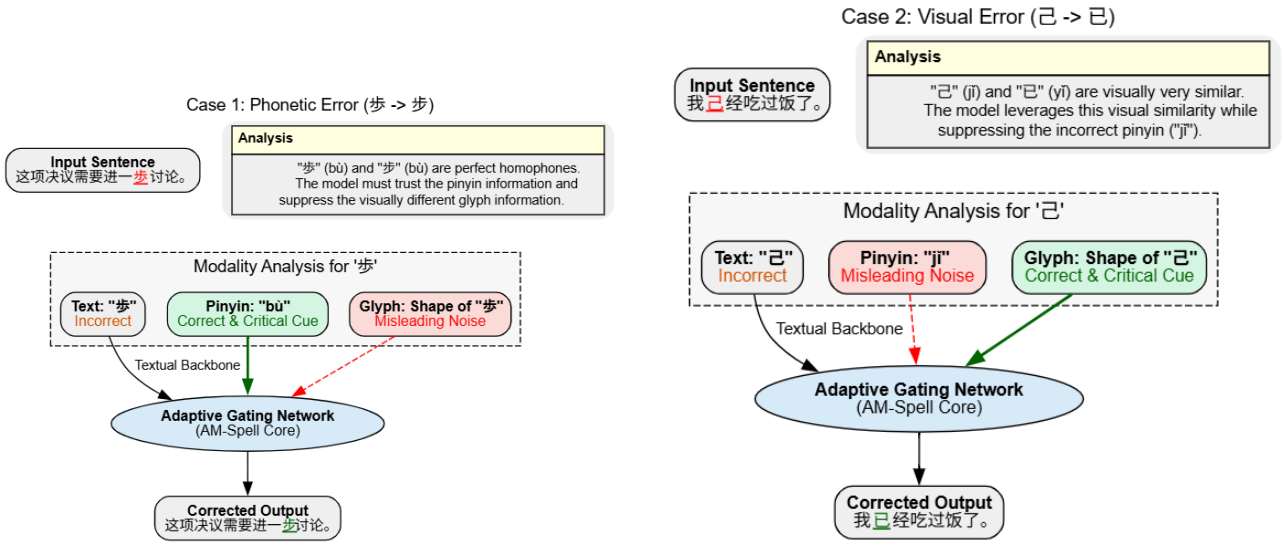


Figure 2: Sensitivity analysis for HMP probabilities.

4.7. Case Study and Discussion

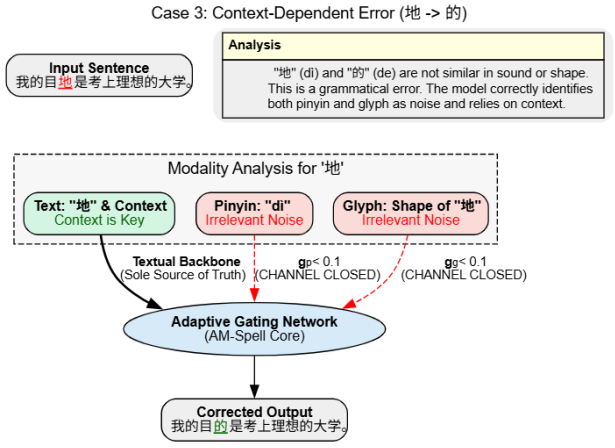
To provide a qualitative illustration of the quantitatively validated gating behavior, we conduct a case study of three distinct error types. Our broader statistical analysis confirms that the model systematically assigns higher gate values to the pinyin modality for phonetically similar errors and to the glyph modality for visually similar errors. Figures 3a, 3b, and 4a intuitively demonstrate how AM-Spell's adaptive gating network dynamically adjusts its reliance on different modalities according to the error type.

DOI: 10.54254/2755-2721/2026.TJ28894



(a) Case 1: Phonetic Error (b) Case 2: Visual Error

Figure 3: Case studies of phonetic and visual errors.



(a) Case 3: Context-Dependent Error

Figure 4: Case study of a context-dependent error.

As shown in Figure 3a, for a phonetic error like ‘’ → ‘步‘, the pinyin channel is given a high gate value, indicating a strong reliance on pronunciation information. Conversely, the glyph channel is suppressed, as the visual difference between the characters would be misleading. In contrast, Figure 3b illustrates a visual error like ‘己‘ → ‘已‘. Here, the model correctly identifies the glyph similarity as the critical cue and opens the glyph channel while suppressing the pinyin channel, which is irrelevant. Finally, in Figure 4a, for a grammatical error like ‘地‘ → ‘的‘, neither phonetic nor glyphic information is helpful. The model intelligently closes both auxiliary channels, relying solely on the textual backbone and contextual information to make the correct prediction. These cases provide strong qualitative evidence for the effectiveness of our adaptive gating mechanism.

DOI: 10.54254/2755-2721/2026.TJ28894

5. Conclusion

In this paper, we addressed the challenges of multimodal information fusion in the Chinese Spelling Correction task by proposing a novel pre-trained language model named AM-Spell. The core innovation is the design of an Adaptive Cross-Modal Gating Fusion Network, which moves beyond existing attention-based dynamic weighting methods to achieve explicit, fine-grained control over information flow. To fully unleash the potential of this network, we also introduced a Hierarchical Modality Perturbation (HMP) pre-training strategy that more closely simulates real-world error distributions. Comprehensive experimental results on the SIGHAN series of standard benchmark datasets demonstrate that AM-Spell achieves state-of-the-art performance across multiple metrics. In-depth ablation studies and case analyses further confirm the existence of a powerful synergistic effect between the gating fusion network and the hierarchical perturbation pre-training strategy, which together form the foundation of AM-Spell's superior performance. This work thoroughly validates the effectiveness of our proposed methods and offers a new perspective on the field of multimodal information fusion.

References

[1] Wenyu Zhao, Dong Zhou, Lin Li, and Jinjun Chen. Manifold learning-based word representation refinement incorporating global and local information. In Donia Scott, Nuria Bel, and Chengqing Zong, editors, *Proceedings of the 28th International Conference on Computational Linguistics*, pages 3401–3412, Barcelona, Spain (Online), December 2020. International Committee on Computational Linguistics.

[2] Yuzhong Hong, Xianguo Yu, Neng He, Nan Liu, and Junhui Liu. FASPell: A fast, adaptable, simple, powerful Chinese spell checker based on DAE-decoder paradigm. In Wei Xu, Alan Ritter, Tim Baldwin, and Afshin Rahimi, editors, *Proceedings of the 5th Workshop on Noisy User-generated Text (W-NUT 2019)*, pages 160–169, Hong Kong, China, November 2019. Association for Computational Linguistics.

[3] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In Jill Burstein, Christy Doran, and Thamar Solorio, editors, *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics.

[4] Zijun Sun, Xiaoya Li, Xiaofei Sun, Yuxian Meng, Xiang Ao, Qing He, Fei Wu, and Jiwei Li. Chinesebert: Chinese pretraining enhanced by glyph and pinyin information. *arXiv preprint arXiv:2106.16038*, 2021.

[5] Shulin Liu, Tao Yang, Tianchi Yue, Feng Zhang, and Di Wang. PLOME: Pre-training with misspelled knowledge for Chinese spelling correction. In Chengqing Zong, Fei Xia, Wenjie Li, and Roberto Navigli, editors, *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 2991–3000, Online, August 2021. Association for Computational Linguistics.

[6] Yinghao Li, Heyan Huang, Baojun Wang, and Yang Gao. DRMSpell: Dynamically reweighting multimodality for Chinese spelling correction. *Frontiers of Information Technology and Electronic Engineering*, 26(3):354–366, 2025.

[7] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural Computation*, 9(8):1735–1780, 1997.

[8] Kyunghyun Cho, Bart van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. Learning phrase representations using RNN encoder–decoder for statistical machine translation. In Alessandro Moschitti, Bo Pang, and Walter Daelemans, editors, *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1724–1734, Doha, Qatar, October 2014. Association for Computational Linguistics.

[9] Shaohua Zhang, Haoran Huang, Jicong Liu, and Hang Li. Spelling error correction with soft-masked BERT. In Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault, editors, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 882–890, Online, July 2020. Association for Computational Linguistics.

[10] Heng-Da Xu, Zhongli Li, Qingyu Zhou, Chao Li, Zizhen Wang, Yunbo Cao, Heyan Huang, and Xian-Ling Mao. Read, listen, and see: Leveraging multimodal information helps Chinese spell checking. In Chengqing Zong, Fei Xia, Wenjie Li, and Roberto Navigli, editors, *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 716–728, Online, August 2021. Association for Computational Linguistics.

[11] Yinghui Li, Qingyu Zhou, Yangning Li, Zhongli Li, Ruiyang Liu, Rongyi Sun, Zizhen Wang, Chao Li, Yunbo Cao, and Hai-Tao Zheng. The past mistake is the future wisdom: Error-driven contrastive probability optimization for Chinese spell checking. In Smaranda Muresan, Preslav Nakov, and Aline Villavicencio, editors, *Findings of the Association for Computational Linguistics: ACL 2022*, pages 3202–3213, Dublin, Ireland, May 2022. Association for Computational Linguistics.

[12] Zihong Liang, Xiaojun Quan, and Qifan Wang. Disentangled phonetic representation for chinese spelling correction. pages 13509–13521, 01 2023.