SCU_Mesclab at SemEval-2026 Task 3: An Adaptive Dual-Track Framework for Dimensional Aspect-Based Sentiment Analysis

Chia-Yun Lee[1]*, **Matus Pleva**[2], **Daniel Hladek**[2], **Ming-Hsiang Su**[1]
[1]Soochow University, Taiwan
[2]Technical University of Košice, Slovakia
huntfox.su@gmail.com

Abstract

This paper describes our system for SemEval-2026 Task 3, which focuses on predicting continuous valence and arousal scores. The task poses significant challenges due to variations in data scale and pragmatic ambiguities across languages. To address these disparities, we propose an Adaptive Dual-Track Framework that dynamically selects modeling strategies based on task characteristics. For semantically stable tasks, we apply a robust single baseline optimized with layer-wise learning rate decay (LLRD) to ensure stability. For high-ambiguity scenarios such as the Environmental Protection domain, we adopt a heterogeneous ensemble strategy to mitigate prediction variance. Experimental results demonstrate that our system consistently outperforms the initial standard baseline across all subtasks. Furthermore, our lightweight approach exhibits remarkable parameter efficiency, achieving highly competitive performance against newly introduced large language model (LLM) baselines. Additionally, ablation studies reveal that under regression settings, conventional regularization techniques, cross-lingual data transfer, and homogeneous ensemble learning can lead to negative transfer, confirming the necessity of strategically diverging approaches tailored to linguistic characteristics.

1 Introduction

This study participates in the official Dimensional Aspect-Based Sentiment Analysis (DimABSA) task (Lee et al., 2026), covering multilingual aspect-based affective and stance regression (Becker et al., 2026). The objective is to predict continuous scores for valence, arousal, and stance intensity for specific aspects within a text. This study encompasses both English and Chinese languages, covering various review topics and public issue domains. Compared to traditional categorical sentiment analysis, DimABSA requires models to master fine-grained emotional intensity, context-dependent stance expression, and pragmatic signals hidden within the text, placing higher demands on semantic understanding and inference capabilities. Particularly in public issue texts such as Environmental Protection, sarcasm, conditionals, and implicit stance expressions are prevalent. This makes models prone to unstable predictions when facing domain shift and label noise, serving as a critical testbed for model robustness.

Given the significant differences in data scale, noise levels, and semantic stability across languages and subtasks, we propose an Adaptive Dual-Track Framework. This framework dynamically selects appropriate modeling strategies based on task characteristics. For subtasks with limited data but relatively clear semantic boundaries, we adopt a single optimized baseline model to ensure training stability and avoid negative transfer. Conversely, for the Track B English subtask, characterized by high noise and significant pragmatic ambiguity, we introduce a heterogeneous ensemble strategy. This combines pre-trained models with differences in architecture and training objectives to effectively reduce prediction variance through error decorrelation. This framework aims to balance bias and variance across different subtasks, avoiding the potential risks associated with blindly increasing model complexity.

Evaluations on the final test set show that our method stably outperforms the initial standard baseline on all subtasks, indicating strong cross-lingual and cross-domain generalization. In the Track B English subtask, the proposed heterogeneous ensemble strategy achieved an RMSE of 1.5714, ranking 4th on the official leaderboard and confirming its effectiveness in highly subjective stance regres-

* The authors from Slovakia acknowledge financial support from the projects KEGA 049TUKE-4/2024 and VEGA 1/0685/26, and the authors from Taiwan acknowledge financial support from MOST under Grant No. 114-2221-E-031-002.

Proceedings of the The 20th International Workshop on Semantic Evaluation (2026), pages 1482–1488

sion. Further ablation experiments and error analysis reveal that in semantically explicit tasks, ensemble methods may lead to performance degradation due to negative transfer. Moreover, conventional data augmentation and regularization techniques led to performance drops, highlighting the sensitivity of emotion and stance regression tasks to subtle semantic variations and label noise.

2 Background

In SemEval-2026 Task 3 (Yu et al., 2026), given a sentence S and its corresponding aspect A, the system must predict an affective intensity tuple (v, a), where v represents Valence and a represents Arousal, quantifying the subjective emotional state expressed regarding the target.

2.1 Data

The task covers seven sub-datasets across English and Chinese. Track A primarily involves reviews in Restaurant, Laptop, and Finance domains, where semantics are relatively clear and expression styles are consistent. Track B focuses on highly subjective public issue discussions like Environmental Protection, often accompanied by sarcasm, implicit stance, and pragmatic ambiguity. We analyze the datasets along two dimensions: data scale and lexical complexity. Each sub-dataset is projected onto a "Resource–Complexity" plane. The horizontal axis represents the number of training samples, reflecting resource scale, where smaller datasets are associated with a higher risk of overfitting. The vertical axis measures lexical complexity using the Type–Token Ratio (TTR); higher TTR values indicate greater lexical diversity and increased semantic variability.

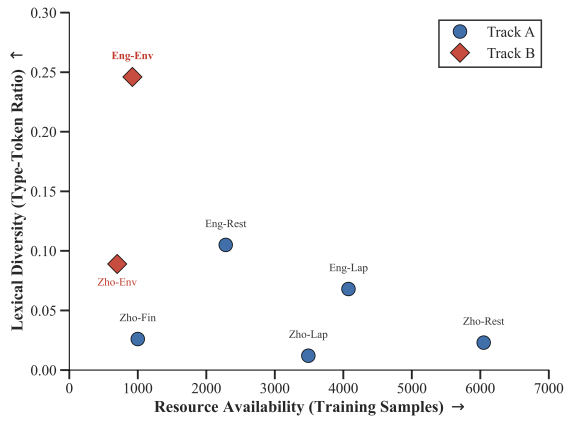

Figure 1: Dataset distribution in the Resource–Complexity space.

As shown in Figure 1, most datasets fall within the low-TTR region and exhibit relatively concentrated distributions, suggesting stable semantic boundaries. In contrast, the Track B English subtask shows both high lexical complexity (TTR = 0.246) and limited training data, making it the most challenging subtask. Based on these observations, datasets with higher lexical complexity (e.g., TTR > 0.20) and limited training samples tend to exhibit greater semantic ambiguity, where variance-reduction strategies such as heterogeneous ensembles become more effective.

2.2 Related Work

Aspect-Based Sentiment Analysis has shifted from discrete polarity classification to dimensional modeling that represents sentiment using continuous Valence and Arousal values (Lee et al., 2026). This enables quantitative descriptions of subjective states (Lee et al., 2024), laying a continuous numerical foundation for high-level tasks like stance regression. However, Hua et al. (2024) mention that current ABSA datasets are concentrated in semantically clear scenarios like commercial reviews, resulting in insufficient model generalization on public issues and high-subjectivity texts. Furthermore, the absence of unified frameworks and standardized continuous metrics widens interpretability gaps (Venkit et al., 2023). Consequently, Track B in this task belongs to a high-ambiguity scenario, where the uncertainty of continuous annotation is particularly significant under dimensional regression.

Methodologically, with the maturity of Transformer architectures, fine-tuned small language models (SLMs) and BERT-style models have become the mainstream approach for ABSA and sentiment classification tasks (Zhang et al., 2024; Apostol et al., 2025). LLMs have also emerged as a prominent trend in NLP, but medium-sized Transformers fine-tuned for specific tasks (e.g., RoBERTa, DeBERTa) continue to demonstrate advantages in stability and efficiency. Research also points out that LLMs are susceptible to internal biases and spurious correlations when handling implicit sentiment (Ren et al., 2025). To address these limitations, heterogeneous model collaboration has been proposed to reduce the learning of erroneous signals by combining models with different architectures or learning objectives (Apostol et al., 2025; Lai et al., 2025). Motivated by these findings, we adopted a robust modeling strategy

tailored to resource-scarce and high-ambiguity dimensional ABSA.

3 System Overview

The primary challenge of Task 3 stems not from the subtask definition but rather from significant differences in dataset scale, annotation completeness, and language-specific noise across different languages. Based on this observation, we propose a language-conditioned Adaptive Dual-Track Framework. The system uses pre-trained language models as the core and dynamically selects either a single strong baseline or a heterogeneous ensemble strategy based on language-level data characteristics to improve generalization stability. The overall architecture of the system is illustrated in Figure 2.

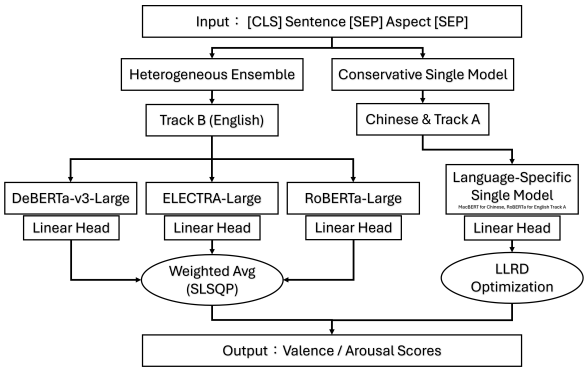

Figure 2: Overview of the proposed Adaptive Dual-Track Framework

3.1 Input Representation

We model DimABSA and DimStance as unified multi-dimensional regression tasks, providing a consistent framework for fine-grained affective intensity prediction across subtasks and languages. To handle potentially conflicting emotions within a text, we adopt an aspect-aware input format:

$$\text{Input} = [\text{CLS}] + S + [\text{SEP}] + A + [\text{SEP}] \quad (1)$$

This design explicitly models interactions between the text context S and the target aspect A via the Transformer's self-attention mechanism, thereby mitigating the impact of aspect ambiguity on prediction. The final hidden state of the `[CLS]` token is used as the semantic representation vector $\mathbf{h} \in \mathbb{R}^d$ and is fed into a linear regression head to predict the (v, a) tuple.

3.2 Model Architecture

Ensemble Strategy The Track B English subtask dataset contains complex semantics such as sarcasm, metaphor, and implicit stance, resulting in high prediction variance for single models. To reduce this instability (Mendes-Moreira and Mendes-Neves, 2024), we adopt a heterogeneous ensemble strategy that combines three complementary pre-trained language models. DeBERTa-v3-Large serves as the primary backbone, leveraging its Disentangled Attention to capture complex semantic dependencies (He et al., 2023). ELECTRA-Large is included to exploit its Replaced Token Detection (RTD) pretraining task, enhancing sensitivity to subtle semantic differences, while RoBERTa-Large provides structural diversity and training stability. Final predictions are obtained via weighted averaging, with ensemble weights optimized on the validation set.

Single Strong Baseline In contrast, the Track A English dataset is large and contains complete quadruple annotations, such as opinion terms and categories, providing clear and high-quality supervision. This reduces semantic ambiguity in stance and sentiment prediction, allowing a single model to reliably learn the relationship between stance and sentiment. Experiments showed that a single RoBERTa-Large with dynamic masking produced more stable gradient trajectories than DeBERTa during domain adaptation for stance detection. It avoided oscillations on the validation set and achieved optimal performance. As a result, ensemble strategies provide little additional benefit while incurring extra computational cost.

In the Chinese sector, we adopt a unified single-model strategy for different reasons. For Track A, despite sufficient data, preliminary experiments indicated that the MacBERT-Large model, with its Whole Word Masking (WWM) and N-gram masking, already captures precise semantic boundaries. Introducing heterogeneous ensembles fusing with XLM-R yielded negligible gains and even caused negative transfer due to semantic misalignment. For Track B, which suffers from severe data scarcity, we maintain this robust single-model approach. Following Occam's Razor, we prioritize the stability of MacBERT to minimize the risk of overfitting that complex ensemble architectures often induce in resource-scarce settings.

4 Experimental Setup

We use the official training, validation, and test splits without additional corpora or annotations. To avoid cross-language or domain bias, no extra text

cleaning or language-specific preprocessing is applied, except for model-required input truncation.

4.1 Implementation Details

Our system is implemented in PyTorch with Hugging Face Transformers. All models use a pre-trained Transformer backbone with a linear regression head. Hyperparameters are tuned per subtask. We use AdamW with a linear learning rate scheduler and select checkpoints based on development set performance. Task- and language-specific optimization and regularization techniques are applied to improve stability and generalization.

4.2 Track A Configuration

For Track A, which covers multiple review domains, we apply cross-domain joint training for both English and Chinese tasks to expand the sample size and learn generalized affective representations. All Track A subtasks adopt aspect-aware input formatting, with sequence lengths restricted to a reasonable range (128–256 tokens) to reduce interference from irrelevant background information in long texts. For Chinese subtasks, considering the higher information density of Chinese text and the need to balance efficiency and stability, we set the maximum sequence length to 128 tokens and use larger batch sizes to ensure stable gradient estimation during training. For English subtasks, the model already converges stably under cross-domain joint training, so we adopt a conservative fine-tuning strategy with a fixed learning rate of 1e-5 and early stopping to prevent overfitting. Additionally, we employ R-Drop regularization (α = 0.5) to enforce output consistency under stochastic dropout, further reducing overfitting and enhancing generalization. The maximum sequence length is set to 256 tokens to preserve the complete review context.

4.3 Track B Configuration

In contrast to the multi-domain nature of Track A, Track B focuses exclusively on the Environmental Protection domain. However, this domain specificity is characterized by high pragmatic noise, including pervasive sarcasm and implicit stances, which presents significant challenges for robust predictive modeling. To address these challenges in the Chinese subtask, we apply layer-wise learning rate decay (LLRD), assigning a higher learning rate (1e-4) to the regression head to accelerate task adaptation, while using a lower learning rate (2e-5) for the lower encoder layers to prevent catastrophic forgetting. This differential tuning balances the retention of pre-trained linguistic knowledge with specific task adaptation. Considering that environmental texts tend to be long and contain extensive background information, we set the maximum sequence length to 256 tokens. This can be viewed as a form of hard attention, helping to filter irrelevant context and improve convergence stability in resource-scarce settings. To further minimize variance arising from random initialization, we applied a seed ensemble technique, averaging predictions from three distinct random seeds. For English subtasks, single-model performance on the development set was relatively unstable. To mitigate this, we adopt a heterogeneous model ensemble strategy, fine-tuning multiple pre-trained models with different architectures to capture diverse semantic features and reduce model bias. To optimally integrate these models, we utilized the Sequential Least Squares Programming (SLSQP) algorithm to compute fusion weights based on validation set performance, thereby dynamically adjusting the contribution of each constituent model. Finally, a lightweight linear calibration is applied to correct observed systematic biases in the stance regression task, further improving overall prediction stability and accuracy.

5 Results

Table 1 summarizes the final evaluation results. Our Adaptive Dual-Track Framework consistently outperforms the official baseline across all subtasks, reducing RMSE by an average of 45.8%. Furthermore, our lightweight system achieves superior performance over the newly introduced heavy-weight LLM baselines in 6 out of 7 subtasks. Notably, our English Track B ensemble ranked 4th officially (RMSE 1.5714). Even in the Chinese Track B, where our model (0.7452) marginally trails Mistral-3 14B (0.7400), it relies on only 300M parameters (MacBERT-Large). This demonstrates remarkable parameter efficiency, proving that strategic architectural design reduces prediction variance in highly subjective tasks more effectively than simply scaling up parameters.

5.1 Ablation and Quantitative Analysis

To further quantify the contributions of our design choices, we conducted ablation experiments on the development and test sets. Detailed visualizations

Model/Language	Eng-lap	Eng-res	Zho-fin	Zho-lap	Zho-res	Eng-env	Zho-env
Standard Baseline†	2.8053	2.7910	1.1281	1.6583	1.7536	2.6985	1.2756
LLM Baseline (Max)‡	1.6440	1.8959	1.4707	2.1893	2.1461	1.6430	0.7400
Best result	1.2408	1.1035	0.4841	0.6103	0.9256	1.4734	0.5468
Ours (Proposed)	**1.3946**	**1.2277**	**0.6692**	**0.9222**	**1.1210**	**1.5714**	**0.7452**
Improvement (%)	+50.3%	+56.0%	+40.7%	+44.4%	+36.1%	+41.8%	+41.6%

Table 1: Official Test Set Performance Comparison (RMSE). †Initial standard benchmark provided by the organizers. ‡Best performance among LLM baselines (Kimi-K2, Qwen-3, or Mistral-3) introduced during the later evaluation phase.

for ablation studies and error analysis are provided in Appendix A for clarity. As shown in Figure 3(b) for English Track B, a homogeneous ensemble using seed averaging provided limited improvement. In contrast, introducing heterogeneous models significantly reduced RMSE, surpassing the performance ceiling of single models. This confirms that the error decorrelation effect provided by different pre-training objectives was pivotal for achieving robust performance in high-ambiguity tasks.

Conversely, in Track A and Chinese subtasks, increasing model complexity did not yield positive results. As illustrated in Figure 3(a) for the English Laptop task, the single-model strategy achieved the lowest RMSE. A similar trend was observed in Chinese tasks, where incorporating RoBERTa into the MacBERT baseline increased RMSE. These results suggest that increased architectural complexity does not necessarily translate into performance gains in semantically stable domains.

5.2 Failure Analysis & Error Distribution

We also explored several advanced training and data augmentation strategies to improve model performance. As illustrated by the failure analysis in Figure 4(a), most augmentation and regularization methods failed to reduce RMSE, highlighting the intrinsic characteristics of the task. Specifically, R-Drop resulted in the most severe performance degradation, increasing RMSE to 2.115 due to optimization collapse in small-batch settings. Similarly, pseudo-labels from self-training and external signals from cross-lingual training increased RMSE by 14–24%. These results suggest that, for environmental texts, the semantic boundaries of valence are highly language- and context-dependent, making cross-lingual alignment challenging.

Further error analysis is presented in Figure 4(b), which illustrates the relationship between the predictions of our best model and the ground-truth annotations along the valence dimension. Although the model exhibits high overall correlation (PCC = 0.78), the regression line slope reveals a conservative bias, indicating that the model struggles to predict extreme emotion values. Qualitative inspection further reveals that the model has difficulty capturing subtle pragmatic nuances. For example, in sentences with sarcastic meaning, such as "Great, another delay", the model incorrectly predicted a positive valence due to the lexical cue "Great", failing to recognize the implied dissatisfaction. Similarly, statements containing conditional or implicitly skeptical tones (e.g., "They won't be 100% until...") are often misinterpreted as factual praise. These observations suggest that, even under a stable regression setup, current pre-trained language models have structural limitations in handling expressions that require pragmatic reasoning and implicit stance.

6 Conclusion

In this paper, we propose an Adaptive Dual-Track Framework for DimABSA that dynamically selects between a single strong baseline and a heterogeneous ensemble strategy based on data scale and noise levels. Results show consistent superiority over the standard official baselines and competitive parameter efficiency against newly introduced large language models. Future work will focus on improving the understanding of complex pragmatic phenomena, such as sarcasm, through contrastive learning and instruction tuning of LLMs (Tang et al., 2024), aiming to enhance sensitivity to subtle pragmatic features without introducing excessive noise.

Acknowledgments

The authors from Slovakia acknowledge financial support from the projects KEGA 049TUKE-4/2024 and VEGA 1/0685/26, and the authors from Taiwan acknowledge financial support from MOST under

Grant No. 114-2221-E-031-002.

References

Lung-Hao Lee, Liang-Chih Yu, Natalia Loukashevich, et al. 2026. DimABSA: Building multilingual and multidomain datasets for dimensional aspect-based sentiment analysis. *arXiv preprint arXiv:2601.23022*.

Jonas Becker, Liang-Chih Yu, Shamsuddeen Hassan Muhammad, et al. 2026. DimStance: Multilingual datasets for dimensional stance analysis. *arXiv preprint arXiv:2601.21483*.

Liang-Chih Yu, Jonas Becker, Shamsuddeen Hassan Muhammad, et al. 2026. SemEval-2026 Task 3: Dimensional Aspect-Based Sentiment Analysis (DimABSA). In *Proceedings of the 20th International Workshop on Semantic Evaluation (SemEval-2026)*.

Pranav Venkit, Mukund Srinath, Sanjana Gautam, et al. 2023. The sentiment problem: A critical survey towards deconstructing sentiment analysis. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 13743–13763.

Lung-Hao Lee, Liang-Chih Yu, Suge Wang, and Jian Liao. 2024. Overview of the SIGHAN 2024 shared task for Chinese dimensional aspect-based sentiment analysis. In *Proceedings of the 10th SIGHAN Workshop on Chinese Language Processing (SIGHAN-10)*, pages 165–174.

Yan Hua, Paul Denny, Jörg Wicker, and Katerina Taskova. 2024. A systematic review of aspect-based sentiment analysis: Domains, methods, and trends. *Artificial Intelligence Review*, 57.

Wenxuan Zhang, Yue Deng, Bing Liu, Sinno Pan, and Lidong Bing. 2024. Sentiment analysis in the era of large language models: A reality check. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 3881–3906.

Jing Ren, Wenhao Zhou, Bowen Li, et al. 2025. Causal prompting for implicit sentiment analysis with large language models. *arXiv preprint arXiv:2507.00389*.

Elena-Simona Apostol, Alin-Georgian Pisică, and Ciprian-Octavian Truică. 2025. ATESA-BÆRT: A heterogeneous ensemble learning model for aspect-based sentiment analysis. *Knowledge-Based Systems*, 326:113987.

W. Lai, H. Xie, G. Xu, and Q. Li. 2025. RVISA: Reasoning and verification for implicit sentiment analysis. *IEEE Transactions on Affective Computing*, 16(3):1760–1771.

João Mendes-Moreira and Tiago Mendes-Neves. 2024. Towards a Systematic Approach to Design New Ensemble Learning Algorithms. *arXiv preprint arXiv:2402.06818*.

Pengcheng He, Jianfeng Gao, and Weizhu Chen. 2023. DeBERTaV3: Improving DeBERTa using ELECTRA-Style Pre-Training with Gradient-Disentangled Embedding Sharing. *arXiv preprint arXiv:2111.09543*.

Binghao Tang, Boda Lin, Haolong Yan, and Si Li. 2024. Leveraging Generative Large Language Models with Visual Instruction and Demonstration Retrieval for Multimodal Sarcasm Detection. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics*, pages 1732–1742.

A Supplemental Materials

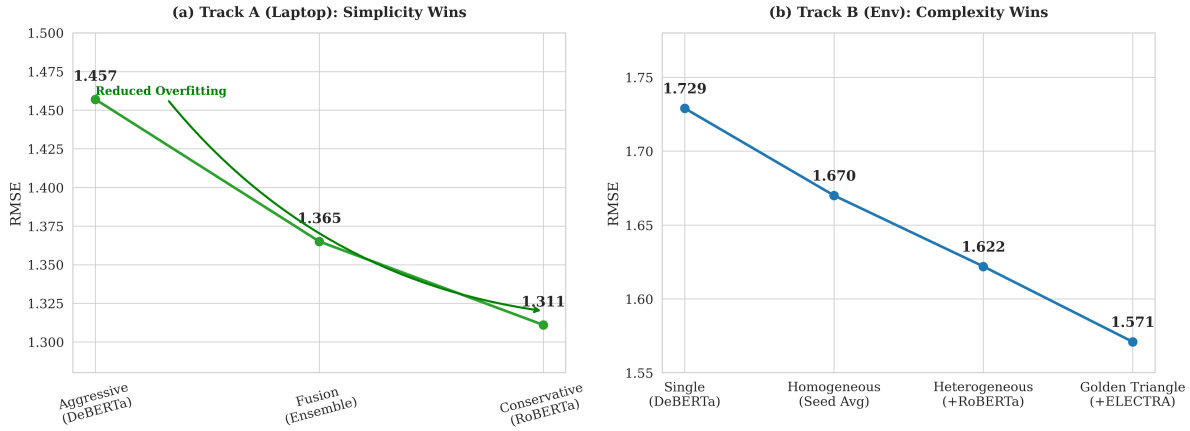

Figure 3: Ablation Study. (a) Track A: Ensembles degrade performance due to overfitting. (b) English Track B: Heterogeneous models (RoBERTa & ELECTRA) progressively lower RMSE.

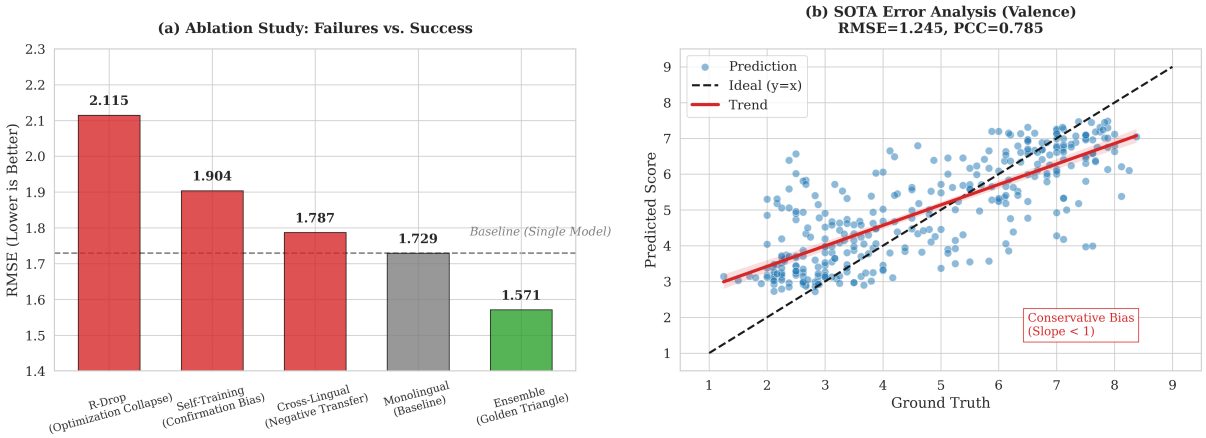

Figure 4: (a) Ablation study and negative result analysis. (b) Error analysis of the SOTA model.