Modules			MIntRec				MELD-DA			
SBMA	MAP	TCL	ACC (%)	WF1 (%)	WP (%)	R (%)	ACC (%)	WF1 (%)	WP (%)	R (%)
×	✓	✓	73.15	72.73	73.02	69.82	61.24	59.32	59.58	49.89
×	×	✓	72.67	72.31	72.81	69.78	60.40	58.69	59.78	49.63
✓	✓	×	72.13	71.80	72.50	68.80	61.26	59.54	59.97	50.09
✓	✓	✓	**73.62**	**73.31**	**73.72**	**70.50**	**61.75**	**59.77**	**60.33**	**50.14**

Table 3: Ablation experiments of modules in TCL-MAP on the MIntRec dataset and the MELD-DA dataset. SBMA stands for Similarity-Based Modality Alignment, MAP stands for Modality-Aware Prompt and TCL stands for Token-Level Contrastive Learning. With SBMA incorporated into MAP, there exist three distinct settings.

to extract video features. The training batch size is set to 16, while the validation and test batch sizes are both set to 8. For the total loss $\mathcal{L}$, we employ AdamW (**?**) to optimize the parameters.

Results

We conduct experiments on both the MIntRec dataset and the MELD-DA dataset, comparing our approach with state-of-the-art baselines. The results are presented in Table 1 with the optimal outcomes highlighted in bold, and the enhancements of our method over the top-performing baseline are indicated by Δ.

Firstly, we observe the overall performance. As indicated by the results, our approach has consistently outperformed the current state-of-the-art methods across all four metrics on both datasets, demonstrating significant advancements. Secondly, on the MIntRec dataset, our approach demonstrates enhancements of 0.97% on ACC, 0.93% on WF1 and 1.22% on R, which indicates the robust capability of our approach to effectively leverage multimodal information for the extraction and identification of intricate intents in real-world scenarios. Thirdly, on the MELD-DA dataset, our method also achieves notable improvements on both ACC and WP metrics, despite the presence of challenging "Others" label which is difficult to distinguish. This observation showcases the effectiveness of our method in recognizing ambiguous intents such as dialogue acts.

Discussion

Effectiveness of Each Module

To further analyze the individual contributions of the modules within TCL-MAP to the overall performance, we conducted the following ablation experiments and the results are illustrated in Table 3.

Similarity-Based Modality Alignment To assess the effectiveness of similarity-based modality alignment, we replace the alignment method with a CTC module (**?**) which aligns multimodal features solely from a temporal perspective and disregards the correlations. As indicated by the results, the performance of TCL-MAP exhibites a reduction of more than 0.50% across most metrics for both datasets. The most significant decrease, amounting to 0.75%, is observed in the WP metric for the MELD-DA dataset. These observations illustrate the effectiveness of our proposed similarity-based modality alignment in aligning multimodal features and facilitating the extraction of semantic information. Moreover, even without the presence of similarity-based modality alignment, TCL-MAP continues to achieve superior results on the MIntRec dataset, underscoring the efficacy of the other modules.

Modality-Aware Prompting In this setting, we remove modality-aware prompting module and directly concatenate the original text tokens with the [MASK]/Label token as the augmented pair. We note more substantial reductions on MIntRec, such as a 0.95% decrease on ACC and a 1.00% decrease on WF1. Meanwhile, the performance on MELD-DA experiences notable declines on ACC, WF1 and WP metrics. We attribute this to the fact that the optimal semantic environment created by our modality-aware prompting module aids in filtering out irrelevant semantics within the [MASK]/Label token, which makes the token-level contrastive learning more precise.

Token-Level Contrastive Learning In the absence of token-level contrastive learning, we exclude the contrastive learning loss $\mathcal{L}_{con}$ from the total loss $\mathcal{L}$ and proceed with the learning process guided by classification. In this experimental setup, all of the four metrics decrease by 1.49%, 1.51%, 1.22% and 1.70% on MIntRec and the ACC metric and the WP metric decrease by 0.49% and 0.36% on MELD-DA, indicating a significant decline on performance. The experimental results suggest that our introduced token-level contrastive learning effectively leverages the rich semantic information within the ground truth labels to guide the learning process of nonverbal modalities and simultaneously optimizes feature representations together with the classification guidance, leading to improved performance.

Performance of Fine-grained Intent Classes

To examine the performance of our method in each fine-grained intent classes, we compare the F1 scores of TCL-MAP and baseline methods for each intent class in MIntRec. As shown in Table 2, the results are obtained by averaging the scores over ten runs of experiments with different random seeds and for the scores of TCL-MAP we mark the best results in bold and the second best results with underlines within each class.

To begin with, we analyze the comprehensive results of our proposed TCL-MAP in comparison to the baseline methods.

Remarkably, across all 20 intent categories, our approach attains top-2 scores in 13 categories, comprising 7 highest scores and 6 second highest scores, which indicates that TCL-MAP achieves better performance than the majority of baselines across various classes. Specifically, in categories like"Complain", "Agree" and "Leave", TCL-MAP consistently outperforms the best baseline by over 1%. Significantly, the "Leave" category exhibits the most substantial improvement of 7.63%. The significant gains can be attributed to TCL-MAP's utilization of modality-aware prompts for better text representation, which in turn enhances video and audio learning through token-level contrastive learning. Nevertheless, in the "Taunt" and "Joke" categories, TCL-MAP seems to provide less assistance in recognizing the intent, which could be caused by a combination of factors, including the limited availability of data within these categories and the intricate nature of the intents themselves.

On the other hand, we evaluate the efficacy of TCL-MAP in comparison to the human performance. From the results, we can observe that humans achieve the best performance in the majority of intent categories, which confirms the strong ability of humans to process multimodal information and infer intents through them. However, TCL-MAP surpasses human performance in the "Apologize," "Thank," and "Agree" classes, showcasing the stability of our method when handling challenging samples where humans may make mistakes. In addition, TCL-MAP has approached human performance in intent categories (e.g. "complain", "praise" and "care") which involve distinct emotional aspects and also achieved comparable performance to humans in intent categories (e.g. "Inform", "leave" and "prevent") which require an understanding of actions. These findings further validate the capability of TCL-MAP to effectively extract fearures related to human intents from raw multimodal data, such as expressions, tone of speech and movements.

Comparison between Handcraft Prompt and Modality-Aware Prompt

To further analyze the superiority of our modality-aware prompt, we conduct experiments with handcrafted prompt and modality-aware prompt respectively. Concretly, we select the MIntRec dataset for our experiments, driven by the fact that certain labels (e.g. "Others") in the MELD-DA dataset do not strictly represent intent categories. To make comparison, we design two handcraft prompts aimed at expressing ideas or intents, "I want to" and "I intend to", which maintain the same positions and lengths with the modality-aware prompt. Besides, we conduct an additional set of experiments using [MASK] as the prompt to demonstrate the effectiveness.

As shown in Figure 3, we observe a substantial performance advantage in the model that employs the modality-aware prompt in comparison to models using handcrafted prompts, thanks to better integration of non-textual modalities enhancing textual intent semantics extraction. Conversely, the [MASK] prompt shows a notable performance decline compared to handcrafted prompts, highlighting the risk of inappropriate prompts misleading intent understanding. Our modality-aware prompt incorporates the instance-conditional prompt concept of CoCoOp (**?**), thereby mitigating this drawback.

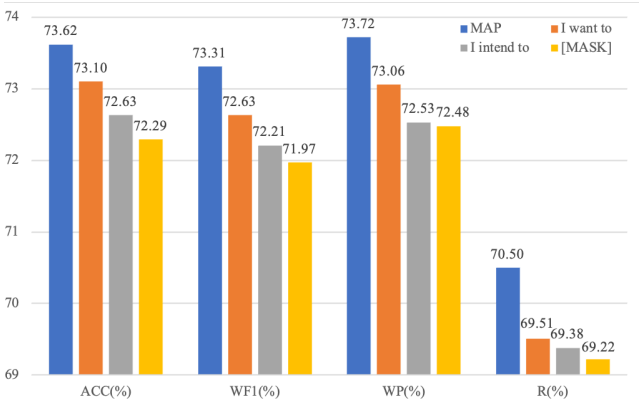

Figure 3: The comparison between Handcraft Prompt and Modality-Aware Prompt

Conclusion

In this paper, we propose a novel Token-Level Contrastive Learning with Modality-Aware Prompting (TCL-MAP) method for multimodal intent recognition. By strengthening the correlations among modalities, our method generate the modality-aware prompt to construct an optimal multimodal semantic space for enhancing the refinement of the text modality. In return, the attained textual representation, enriched with semantics from the ground truth label token, guides the learning process of nonverbal modalities through the token-level contrastive learning. Extensive experiments on two benchmark datasets demonstrate that our approach outperforms state-of-the-art methods and carries significant implications for multimodal prompt learning.

Acknowledgements

This work is founded by the National Natural Science Foundation of China (Grant No. 62173195), National Science and Technology Major Project towards the new generation of broadband wireless mobile communication networks of Jiangxi Province (Grant No.20232ABC03402), High-level Scientific and Technological Innovation Talents "Double Hundred Plan" of Nanchang City (Grant No. Hongke Zi (2022) 321-16), and Natural Science Foundation of Hebei Province, China (Grant No. F2022208006).

Molestiae totam ex enim assumenda nemo vitae mollitia, earum quibusdam impedit dicta alias quo, porro earum amet, error nemo accusantium debitis non explicabo corrupti velit eligendi?Ullam quia perferendis dolore voluptatum totam cupiditate sunt ut, earum voluptate maxime nobis eveniet asperiores animi expedita commodi velit debitis veniam?Earum deserunt quos nostrum officiis ducimus nobis quaerat atque ipsum, qui magnam consequatur ea perspiciatis, excepturi perspiciatis et atque est explicabo, corporis placeat corrupti vel modi suscipit dolorem earum, at nesciunt ut eveniet optio voluptate ab earum ducimus officiis natus.Sit earum totam magni odit doloribus officiis, esse numquam unde quibusdam eveniet sit neque beatae aliquam?Eum cupiditate perspiciatis deleniti iusto facere nam nihil modi consectetur, debitis magni quas minima quaerat iusto adipisci ipsa aspernatur?