EcomMIR: Towards Intelligent Multimodal Intent Recognition in E-Commerce Dialogue Systems

Tianhong Gao*
Zhejiang University
Hangzhou, China
tianhonggao@zju.edu.cn

Genhang Shen*
Zhejiang University
Hangzhou, China
shengenhang@zju.edu.cn

Yuxuan Wu*
Zhejiang University
Hangzhou, China
wux521@zju.edu.cn

Zunlei Feng
Zhejiang University
Hangzhou, China
zunleifeng@zju.edu.cn

Jinshan Zhang†
Zhejiang University
Hangzhou, China
zhangjinshan@zju.edu.cn

Sheng Zhou
Zhejiang University
Hangzhou, China
zhousheng_zju@zju.edu.cn

Abstract

Image scene classification and dialogue intent recognition are fundamental tasks in intelligent e-commerce. The former classifies user-uploaded images, while the latter integrates multi-turn dialogue and visual information to extract user intent. However, existing multimodal models struggle with effectively utilizing e-commerce data and generalizing to vertical domains, limiting their practical applicability. To address this, we propose **EcomMIR**, an **E-COM**merce **M**ultimodal **I**ntent **R**ecognition framework based on CN-CLIP and MiniCPM-V. EcomMIR improves model generalization and robustness through multi-level intent data denoising, high-confidence data selection, and hierarchical labeling. Specifically, CN-CLIP employs contrastive learning to align image and text embeddings for efficient scene classification, while MiniCPM-V, a multimodal large language model, deeply integrates textual and visual information to model dialogue context and accurately recognize user intent. Experimental results show that EcomMIR achieves superior performance in both tasks, ranking **Top 2** in the WWW2025 Multimodal Intent Recognition for Dialogue Systems challenge, offering an effective solution for multimodal tasks in intelligent e-commerce.

CCS Concepts

• **Computing methodologies** → **Computer vision**; **Natural language processing**.

Keywords

E-commerce, Multimodal Intent Recognition, Multimodal Models

*These authors contributed equally to this work. The authors are listed in alphabetical order based on last names.
†Corresponding author.

WWW '25 Companion, Sydney, Australia

ACM ISBN 978-1-4503-XXXX-X/2018/06
https://doi.org/XXXXXXX.XXXXXXX

ACM Reference Format:
Tianhong Gao, Genhang Shen, Yuxuan Wu, Zunlei Feng, Jinshan Zhang, and Sheng Zhou. 2025. EcomMIR: Towards Intelligent Multimodal Intent Recognition in E-Commerce Dialogue Systems. In *Companion Proceedings of the ACM Web Conference 2025 (WWW '25 Companion), April 28–May 2, 2025, Sydney, Australia.* ACM, New York, NY, USA, 4 pages. https://doi.org/XXXXXXX.XXXXXXX

1 Introduction

As a crucial part of the global economy, the e-commerce industry is evolving with increasingly diverse consumer demands and more intelligent e-commerce platforms. Understanding user intent and needs efficiently has become essential for enhancing user experience and maintaining platform competitiveness. In e-commerce scenarios, interactions between users and customer service often involve multimodal information, such as screenshots of product detail pages, logistics tracking pages, and purchase inquiry texts [10, 11]. These elements contain key aspects of user needs. However, their multimodal nature and complex semantic relationships present significant challenges for the intelligent processing of customer service systems.

Currently, multimodal data processing in e-commerce faces two key challenges: image scene classification and multi-turn dialogue intent recognition. The former requires accurately identifying the e-commerce scene represented by diverse user-uploaded images, while the latter involves extracting users' true purchasing or service intentions by integrating dialogue history, current queries, and image information. These tasks demand not only efficient multimodal data processing but also adaptability to the unique characteristics of the e-commerce domain, enabling a deep understanding of fine-grained semantics and complex contextual relationships. However, existing multimodal models [9] are predominantly trained on general-domain data, limiting their ability to directly address the specialized requirements of e-commerce applications.

To advance research on multimodal tasks in e-commerce, the WWW2025 Multimodal Intent Recognition for Dialogue Systems challenge provides a high-quality dataset, including 1,000 labeled dialogue samples for training, 10,000 unlabeled samples for preliminary testing (Round 1), and another 10,000 test samples for the final round (Round 2).

To this end, we propose EcomMIR, an e-commerce multimodal intent recognition framework that leverages both Chinese-CLIP (CN-CLIP) [7] and MiniCPM-V [8]. CN-CLIP is a contrastive learning-based multimodal model that classifies image scenes by measuring image-text similarity. Optimized for Chinese-language data, it effectively identifies user-uploaded e-commerce images. To further enhance classification, we apply Greedy Soup [5] and high-confidence data selection on top of CN-CLIP. MiniCPM-V, a multimodal large language model, integrates textual and visual information to model multi-turn dialogues and extract user intent. We refine its intent classification using a multi-level intent data denoising algorithm and high-confidence data selection. Additionally, a hierarchical label structure groups fine-grained labels into broader categories, enriching semantic cues and improving classification accuracy and generalization.

Our main contributions include:

- We introduce EcomMIR, an e-commerce multimodal dialogue system intent recognition framework, achieving Top 2 in the WWW 2025 Multimodal Intent Recognition for Dialogue Systems Challenge.
- CN-CLIP based scene classification leverages image-text similarity for effective categorization and achieves an F1-score of 0.8278.
- Dialogue intent classification with MiniCPM-V improves recognition accuracy, achieving an F1-score of 0.9420.

2 Method

The EcomMIR framework, as shown in Figure 1, consists of CN-CLIP and MiniCPM-V.

2.1 Image Scene Classification

Image scene classification involves identifying a picture sent by users to customer service and determining which e-commerce scene it belongs to.

The task requires the model to understand both the visual content of images and the semantic information relevant to e-commerce scenarios. Initially, we considered multimodal large language models for their strong generalization abilities. However, we found that while they excel in object detection and cross-modal learning, their performance is not always optimal for specific tasks. Their general nature, while beneficial in some areas, hinders precise fine-tuning for e-commerce scene classification.

To address this, we carefully considered the trade-offs between accuracy, inference speed, and parameter size, leading us to select Chinese-CLIP [7] as the core architecture. Trained on a large-scale dataset of 200 million Chinese image-text pairs, the model captures the nuances of e-commerce scene classification, delivering superior precision and performance. For model training, we explored both joint and separate training strategies for the vision and text encoders. Joint training effectively optimizes feature learning across both modalities, resulting in improved performance for scene classification tasks.

Given the limited data, we adopted the TrivialAugment [2] data augmentation method to improve the model's generalization ability. This method randomly selects an operation from a set of simple, predefined transformations and applies a random intensity to each.

Algorithm 1 GreedySoup

```
Input: Potential soup ingredients {θ_1, ..., θ_k} (sorted in decreasing order of ValAcc(θ_i)).
ingredients ← {}
for i = 1 to k do
    if ValAcc(average(ingredients ∪ {θ_i})) ≥ ValAcc(average(ingredients)) then
        ingredients ← ingredients ∪ {θ_i}
    end if
end for
return average(ingredients)
```

Alongside data augmentation, Model Soup [6], particularly the Greedy Soup approach as described in pseudocode 1, serves as a key strategy for further enhancing model performance. The Greedy Soup strategy ranks the model checkpoints generated during training and evaluates their contribution to the final model's performance. Only those checkpoints that truly enhance performance are retained. Ultimately, the selected checkpoints are merged into an optimal model, effectively avoiding the inference time increase typically associated with traditional ensemble learning methods.

2.2 Dialogue Intent Classification

Dialogue intent classification identifies the user's intent by analyzing past interactions with customer service along with the current query. Additionally, user-provided images are incorporated to enhance intent recognition.

High-quality training data is essential for improving model performance. We introduce a multi-level intent data denoising algorithm that refines the dataset through cleaning at three levels: dialogue, sentence, and word. This process ensures data quality and enhances the reliability of model training. At the dialogue level, to reduce redundancy, only the first occurrence of valid content in customer service replies is retained, keeping the data concise and relevant. At the sentence level, meaningless sentences are filtered out using keyword-based removal. These include initial greetings, system messages indicating an inability to parse images, advertisements, requests to transfer to a human agent, and closing remarks, all of which contribute little to intent recognition. Additionally, short sentences with fewer than three words are replaced with null values due to their limited informational content. At the word level, honorifics, greetings, expressions of gratitude, and other irrelevant words are removed, along with redundant punctuation, to sharpen the semantic focus of the dataset. Further data cleaning strategies are applied, including replacing long numerical strings (over ten digits) with a placeholder, removing URLs and symbol-only sentences, and standardizing spacing formats. Missing customer service reply fields are removed, and consecutive user utterances are merged to improve logical flow and contextual coherence.

To enhance the model's generalization capability, we employ a high-confidence data selection approach combined with an e-commerce image augmentation strategy. In the challenge dataset, Round 1 contains a substantial number of unlabeled samples. For these, we utilize a multimodal large language model to generate pseudo labels. By comparing the consistency between the model's

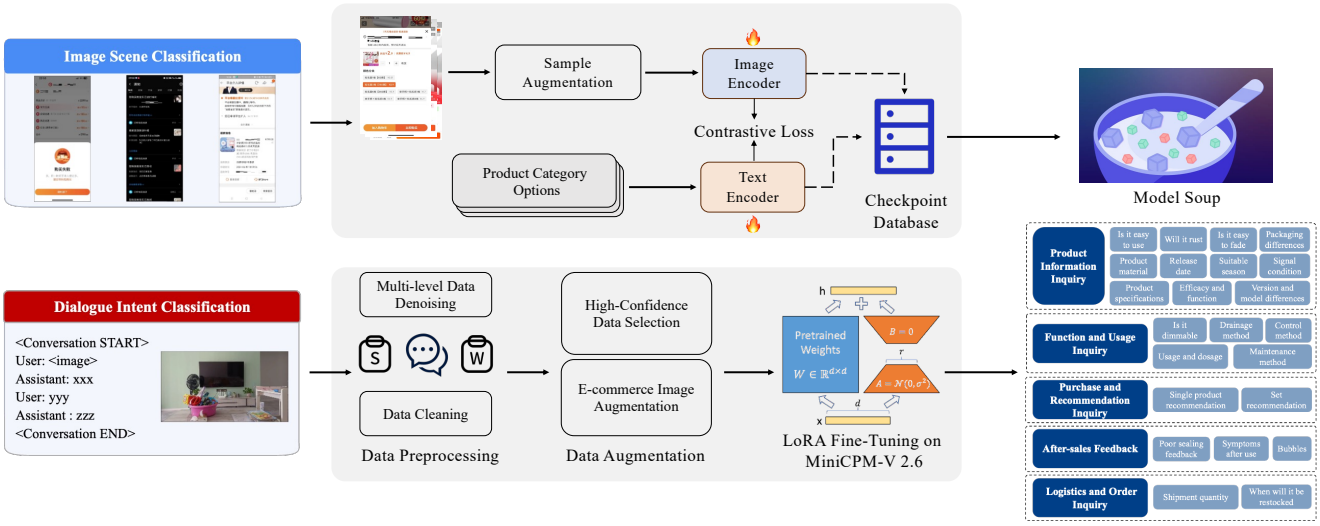

Figure 1: The framework of EcomMIR, consisting of CN-CLIP (upper part) and MiniCPM-V (lower part).

predictions and the best-performing predictions from Round 1, we identify high-confidence samples that align with the labels and incorporate them into the training set. This approach improves both the quality and diversity of the data, providing a strong foundation for subsequent optimization experiments. After completing the data selection, we further implement a set of image augmentation techniques specifically designed for the e-commerce domain, including random horizontal flipping, rotation, brightness and contrast adjustments, gaussian noise addition, and random cropping. Each original image is augmented to produce one additional version. The augmented images, along with their originals, expand the training set and strengthen the model's generalization and robustness in intent recognition tasks.

A hierarchical label construction strategy significantly enhances model accuracy in dialogue intent classification tasks. The original fine-grained labels lack higher-level semantic groupings, making it challenging for the model to fully capture the relationships between different categories. To address this, we reorganize the fine-grained labels based on their semantic features and group them into several broader categories. For instance, "Is it easy to use" and "Will it rust" are categorized under "Product Information Inquiry," while "Is it dimmable?" and "Control method" fall under "Function and Usage Inquiry." The final labels are presented in a "coarse-grained–fine-grained" format, such as "Product Information Inquiry–Is it easy to use" or "Function and Usage Inquiry–Is it dimmable" This approach provides richer semantic cues, enhances the model's understanding of category relationships, and ultimately improves classification accuracy and generalization performance.

3 Experiments

In the scene classification and intent recognition tasks, the F1-score is the evaluation metric. The performance of each task will be introduced separately.

3.1 Image Scene Classification Performance

In the image scene classification task, we used the dataset provided by the competition, which includes both labeled and unlabeled data. The data were organized into a Chinese image-text pair dataset, covering images and their corresponding text labels in e-commerce scenarios, for model training and evaluation. To improve the model's generalization ability and reduce the risk of overfitting, the dataset was split into training and validation sets in a 9:1 ratio. Ninety percent of the data was used for training, while the remaining 10% was reserved for validating the model's performance on unseen data, followed by optimization using the Greedy Soup approach.

For the unlabeled data, pseudo-labels were generated using a thresholding method during the model inference stage. A confidence threshold of 0.99 was applied to select samples with high prediction confidence, and these high-confidence samples were added to the training set. This strategy helped improve the quality of the training data by filtering out low-confidence samples, ultimately enhancing the model's precision and generalization ability on unseen data.

After comparing mainstream Chinese CLIP models (such as CN-CLIP, TaiyiCLIP, ALT-CLIP, and ZH-CLIP), we found that CN-CLIP outperformed the others in e-commerce scene classification, particularly excelling in fine-grained classification and task optimization. As a result, CN-CLIP was chosen as the base model, using ViT-L/14@336px as the image encoder and RoBERTa-wwm-base as the text encoder. The TrivialAugment data augmentation strategy was applied to improve generalization, using simple, predefined image transformations to reduce overfitting and speed up convergence. The F1-score was used as the primary evaluation metric, as it combines precision and recall, providing an effective measure of performance on imbalanced datasets, which is crucial for this classification task.

Meanwhile, we conducted experiments on different training strategies, evaluating the performance of jointly training vision

and text encoders versus training them separately. The experimental results summarized in Table 1 show that jointly training the encoders achieved the best performance.

Table 1: Comparison of Joint vs. Separate Training for vision and language encoders

Training Strategy	F1-Score
Training Vision Encoder (vision_only)	0.8189
Training Text Encoder (text_only)	0.8303
Joint Training (vision + text)	**0.8522**

To further validate the model's practicality, we evaluated the performance of different models in both local and online environments. The experimental results show that applying the Greedy Soup strategy significantly improved the model's performance. The results are shown in Table 2.

Table 2: Comparison of Model Soup's performance in local and online environments

Test Environment	Model Version	F1-Score
Local	Original Model	0.8269
Local	Model Soup	**0.8522**
Online	Original Model	0.8124
Online	Model Soup	**0.8278**

3.2 Dialogue Intent Classification Performance

In the multi-turn dialogue intent classification task, we employ the LoRA [1] fine-tuning method, freezing most of the model parameters and optimizing only the low-rank matrices. This significantly reduces the number of parameters and memory usage. Model training is based on the ms-swift framework [12], and, in combination with DeepSpeed's memory optimization features, distributed training is efficiently conducted on two 80G A800 GPUs.

We compare the performance of several multimodal language models, including Qwen2-VL-7B [3], InternVL2.5-8B-MPO [4] and MiniCPM-V 2.6 [8] on the dialogue intent classification task, as shown in Table 3. Experimental results show that MiniCPM-V 2.6 achieves an F1-score of 0.8828 in Round 1 and 0.8995 in Round 2, outperforming other models overall. As a result, we select MiniCPM-V 2.6 as the base model and introduce optimization strategies such as LoRA fine-tuning, hierarchical labeling, high-confidence data selection, and image augmentation to further enhance classification performance. Hierarchical labels enhance the model's semantic understanding by organizing fine-grained categories into a hierarchy, boosting classification accuracy. High-confidence data selection filters out lower-quality samples, improving training data reliability. Image augmentation creates diverse samples, expanding the dataset and enhancing generalization and robustness. After optimization, the model achieves an F1-score of 0.9420 (Ours) in Round 2, significantly outperforming comparison models and demonstrating strong practical potential.

Table 3: Comparison of the performance of various multimodal large language models in Round 1 and Round 2

Model	Round1	Round2
Qwen2-VL-7B	0.8740	/
InternVL2.5-8B-MPO	/	0.8992
MiniCPM-V 2.6	0.8828	0.8995
Ours	/	**0.9420**

4 Conclusion

This paper presents our solution for the Multimodal Intent Recognition for Dialogue Systems challenge of WWW2025. We propose EcomMIR, an e-commerce multimodal intent recognition framework based on CN-CLIP and MiniCPM-V. Designed for multimodal e-commerce data, it introduces multi-level intent data denoising and hierarchical labeling. EcomMIR performs well in scene classification and intent recognition, achieving the runner-up position in the competition. In future work, we plan to integrate stronger foundation models and explore model distillation to enhance performance, generalization, and resource efficiency, providing a more effective solution for multimodal tasks in intelligent e-commerce.

References

[1] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685* (2021).

[2] Samuel G Müller and Frank Hutter. 2021. Trivialaugment: Tuning-free yet state-of-the-art data augmentation. In *Proceedings of the IEEE/CVF international conference on computer vision*. 774–782.

[3] Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, et al. 2024. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191* (2024).

[4] Weiyun Wang, Zhe Chen, Wenhai Wang, Yue Cao, Yangzhou Liu, Zhangwei Gao, Jinguo Zhu, Xizhou Zhu, Lewei Lu, Yu Qiao, et al. 2024. Enhancing the reasoning ability of multimodal large language models via mixed preference optimization. *arXiv preprint arXiv:2411.10442* (2024).

[5] Mitchell Wortsman, Gabriel Ilharco, Samir Ya Gadre, Rebecca Roelofs, Raphael Gontijo-Lopes, Ari S Morcos, Hongseok Namkoong, Ali Farhadi, Yair Carmon, Simon Kornblith, et al. 2022. Model soups: averaging weights of multiple fine-tuned models improves accuracy without increasing inference time. In *International conference on machine learning*. PMLR, 23965–23998.

[6] Mitchell Wortsman, Gabriel Ilharco, Samir Ya Gadre, Rebecca Roelofs, Raphael Gontijo-Lopes, Ari S Morcos, Hongseok Namkoong, Ali Farhadi, Yair Carmon, Simon Kornblith, et al. 2022. Model soups: averaging weights of multiple fine-tuned models improves accuracy without increasing inference time. In *International conference on machine learning*. PMLR, 23965–23998.

[7] An Yang, Junshu Pan, Junyang Lin, Rui Men, Yichang Zhang, Jingren Zhou, and Chang Zhou. 2022. Chinese clip: Contrastive vision-language pretraining in chinese. *arXiv preprint arXiv:2211.01335* (2022).

[8] Yuan Yao, Tianyu Yu, Ao Zhang, Chongyi Wang, Junbo Cui, Hongji Zhu, Tianchi Cai, Haoyu Li, Weilin Zhao, Zhihui He, et al. 2024. Minicpm-v: A gpt-4v level mllm on your phone. *arXiv preprint arXiv:2408.01800* (2024).

[9] Shukang Yin, Chaoyou Fu, Sirui Zhao, Ke Li, Xing Sun, Tong Xu, and Enhong Chen. 2023. A survey on multimodal large language models. *arXiv preprint arXiv:2306.13549* (2023).

[10] Shaozu Yuan, Xin Shen, Yuming Zhao, Hang Liu, Zhiling Yan, Ruixue Liu, and Meng Chen. 2022. MCIC: multimodal conversational intent classification for E-commerce customer service. In *CCF International Conference on Natural Language Processing and Chinese Computing*. Springer, 749–761.

[11] Nan Zhao, Haoran Li, Youzheng Wu, Xiaodong He, and Bowen Zhou. 2021. The JDDC 2.0 corpus: A large-scale multimodal multi-turn chinese dialogue dataset for e-commerce customer service. *arXiv preprint arXiv:2109.12913* (2021).

[12] Yuze Zhao, Jintao Huang, Jinghan Hu, Xingjun Wang, Yunlin Mao, Daoze Zhang, Zeyinzi Jiang, Zhikai Wu, Baole Ai, Ang Wang, Wenmeng Zhou, and Yingda Chen. 2024. SWIFT:A Scalable lightWeight Infrastructure for Fine-Tuning. arXiv:2408.05517 [cs.CL] https://arxiv.org/abs/2408.05517