A New Formula for Sticker Retrieval: Reply with Stickers in Multi-Modal and Multi-Session Conversation

Bingbing Wang[1,2*], Yiming Du[3*], Bin Liang[3†], Zhixin Bai[1], Min Yang[4], Baojun Wang[5], Kam-Fai Wong[3], Ruifeng Xu[1,2,6†]

[1]Harbin Institute of Technology, Shenzhen, China
[2]Guangdong Provincial Key Laboratory of Novel Security Intelligence Technologies, Shenzhen, China
[3]The Chinese University of Hong Kong, Hong Kong, China
[4]Shenzhen Institutes of Advanced Technology Chinese Academy of Sciences, Shenzhen, China
[5]Huawei Noah's Ark Lab, Shenzhen, China
[6]Peng Cheng Laboratory, Shenzhen, China
{bingbing.wang, baizhixin}@stu.hit.edu.cn, ydu@se.cuhk.edu.hk, bin.liang@cuhk.edu.hk, xuruifeng@hit.edu.cn

Abstract

Stickers are widely used in online chatting, which can vividly express someone's intention, emotion, or attitude. Existing conversation research typically retrieves stickers based on a single session or the previous textual information, which can not adapt to the multi-modal and multi-session nature of the real-world conversation. To this end, we introduce **MultiChat**, a new dataset for sticker retrieval facing the multi-modal and multi-session conversation, comprising 1,542 sessions, featuring 50,192 utterances and 2,182 stickers. Based on the created dataset, we propose a novel Intent-Guided Sticker Retrieval (**IGSR**) framework that retrieves stickers for multi-modal and multi-session conversation history drawing support from intent learning. Specifically, we introduce sticker attributes to better leverage the sticker information in multi-modal conversation, which are incorporated with utterances to construct a memory bank. Further, we extract relevant memories for the current conversation from the memory bank to identify the intent of the current conversation, and then retrieve a sticker to respond guided by the intent. Extensive experiments on our MultiChat dataset reveal the robustness and effectiveness of our IGSR approach in multi-session, multi-modal scenarios.

Introduction

With the advent of instant messaging applications, stickers have become popular in online chats (Zhang et al. 2024). On the one hand, several works on stickers mainly concentrate on sentiment analysis (Liu, Zhang, and Yang 2022; Ge et al. 2022; Zhao et al. 2023). In contrast, stickers present unique advantages in fostering a vibrant and innovative atmosphere within conversations due to their visual nature (Nilasari, Sudipa, and Sukarini 2018; Albar 2018). Therefore, integrating automatic sticker replies based on previous conversations into dialogue systems can make interactions more engaging.

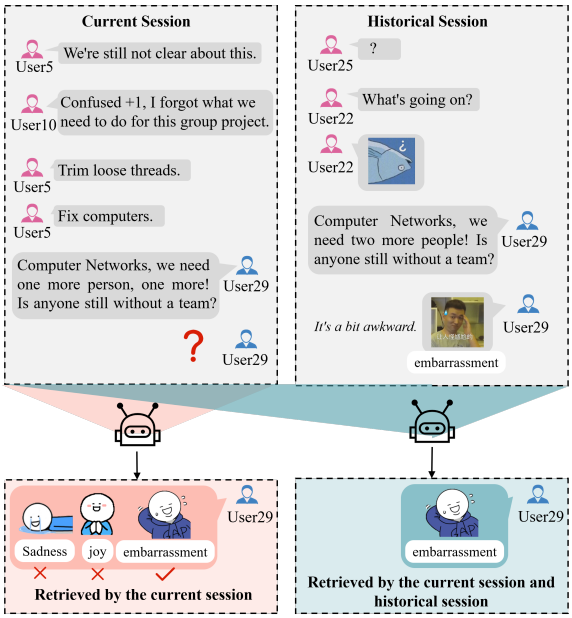

Figure 1: Examples of stickers used in an online multi-session conversation.

Recent research in sticker retrieval has focused on using stickers to respond to text-based contexts or a single conversation session, enhancing the expression of emotions, attitudes, and opinions (Laddha et al. 2020; Gao et al. 2020; Fei et al. 2021a; Zhang et al. 2022, 2024). However, they only focus on the textual content of the current session when retrieving a sticker for a response, ignoring the stickers and previous conversation sessions. Figure 1 shows a conversation with multi-modal and multi-session information in the real-world scenario. This example illustrates two issues not considered in existing sticker retrieval tasks: 1) Combining information from historical sessions is necessary to understand that the current session discusses "course team formation not yet successful"; 2) Learning from the stickers in his-

*These authors contributed equally.
†Corresponding author.

torical sessions is needed to determine a sticker with the expression "embarrassment" to respond to the current session. Furthermore, stickers are utilized to respond to the previous conversation intent, expressing a user's objective for the current utterance within a dialogue session (Shi et al. 2019). As illustrated in Figure 1, User 22 employs a "question mark" sticker to convey the intent of seeking clarification.

To address these above issues, we create a new sticker retrieval dataset with intent labels called **MultiChat**, to encompass more comprehensive real-world scenarios. MultiChat is a multi-modal and multi-session dataset specifically created for open-domain conversation. It contains 1,542 Chinese conversation sessions, featuring 50,192 utterances and 2,182 stickers. To enhance sticker retrieval towards multi-modal and multi-session conversations, we introduce intent information and further propose a novel Intent-Guided Sticker Retrieval framework, to make full use of the multi-modal information in the multi-session conversation for sticker retrieval based on the understanding of stickers' intents, named **IGSR**. To be specific, we first define six attributes, which are combined with the historical conversational utterances and fed into the LLM for constructing a memory bank. Afterward, relevant memories are retrieved from a memory bank based on the current session via OpenAI's LLM-based embedding model (text-embedding-ada-003). Furthermore, a pre-trained Vision Language Model (VLM) (Radford et al. 2021a) plays a central role in our framework, serving as the foundation for both the text and image encoders. The VLM enables the model to concurrently derive user intents and retrieve stickers through a multi-task learning approach. For intent derivation, relevant memories and the current session are fed into the text encoder to obtain memory and context representations. These representations are concatenated to form an intent representation, which is then used by a classifier to predict user intent. For sticker retrieval, sticker representations are generated via the image encoder. These representations, combined with the derived intent representation, are used to retrieve the most suitable sticker. The main contributions of our work can be summarized as follows:

- We create MultiChat, a new dataset designed to facilitate the research of sticker retrieval towards multi-modal and multi-session conversations in social media.
- We propose a novel IGSR framework for sticker retrieval, in which a multi-modal history modeling strategy and a multi-task learning scheme are devised to retrieve the most appropriate stickers for response based on the learning of the sticker's intent.
- Experimental results on our MultiChat dataset demonstrate that our proposed IGSR framework outperforms the baseline models.

Related Work

Sticker Dataset

Stickers have gained substantial attention in recent years (Zhang et al. 2024), particularly within the domain of multi-modal sentiment analysis, where researchers have developed diverse datasets (Liu, Zhang, and Yang 2022; Zhao et al. 2023). Due to their visual nature, stickers offer unique advantages in enhancing conversational dynamics. This insight has prompted researchers to explore context-based sticker retrieval, shifting the focus from simply expressing sentiment to strategically using stickers based on conversational cues. Fei et al. (2021a) presented a meme-incorporated open-domain conversation task with a dataset including 45k Chinese conversations and 606k utterances. Additionally, Ge et al. (2022) introduced a Chinese multi-modal dataset specifically designed for sentiment analysis, comprising 28k text-sticker pairs and 1.5k annotated samples. In these datasets, stickers primarily serve as supplements or responses to the text-based context. However, the conversation context is inherently multi-modal and multi-session in the real world. To address this limitation, this paper creates a comprehensive dataset for sticker retrieval in conversations.

Multi-modal Conversation Method

Several multi-modal studies aim to enhance the efficacy of conversational agents by enriching textual responses with associative vision elements (Huang et al. 2024; Zhang et al. 2024; Maharana et al. 2024). In contemporary social media interactions, using stickers as replies has become commonplace, resulting in a growing body of work focused on sticker retrieval to assist users in selecting the appropriate sticker for responses. Gao et al. (2020) introduced a sticker response selector model that utilizes a convolutional sticker image encoder paired with a self-attention multi-turn dialogue encoder. This model employs a deep interaction network for detailed matching and a fusion network to determine the final matching score. Fei et al. (2021a) presented the Meme Incorporated Open-domain Dialogue (MOD) task, which seamlessly integrates text generation and internet meme prediction into a single sequence generation process. While these methods match conversation contexts with stickers, they fail to model the multi-modal context and the relationships between different sessions. This underscores the need for approaches that better handle the complexities of multi-modal, multi-session conversations.

MultiChat Dataset

Data Preparation

We curated our dataset from the popular social platform WeChat[1], which features a diverse range of conversations and stickers in both individual and group chats. We specifically chose five active chat groups with engaged participants and collected their conversations. These groups engage in open-domain discussions, resulting in a varied and extensive use of stickers.

We established rigorous guidelines and policies for data preprocessing. To safeguard user privacy, all personal information (such as real names, ages, addresses, etc.) is removed, and user IDs are anonymized. Furthermore, any content containing offensive, or insulting language is excluded.

[1]https://weixin.qq.com/

Dataset Statistics	Train	Valid	Test
# sessions	1,120	238	184
# samples	3,447	1,290	1,114
# utterances	30,092	9,711	10,389
# stickers	1,295	637	658
# users	71	60	63
Max. utterances in a session	428	423	436
Avg. utterances in a session	26.87	40.80	56.46
Avg. users in a conversation	5.13	6.56	7.14

Table 1: Statistics of MultiChat. Avg. represents average.

The entire chat content is segmented into distinct conversations to maintain the integrity and independence of each conversation. Following this framework, we systematically examine each sticker in the chat history, capturing its associated context to ensure that each sticker is linked to a corresponding conversation context.

Data Annotation

We recruited five experienced researchers, each with over three years of expertise in multi-modal learning, to serve as annotators. Their tasks were to 1) assess the appropriateness of stickers and 2) categorize each sticker with style and intent tags for every conversation. To enhance the selection of stickers for replies, recognizing sentiment, emotion, and intent is crucial. Therefore, inspired by (Aman and Szpakowicz 2008) and incorporating the labels from GoEmotions (Demszky et al. 2020), we applied these intent tags to capture the diverse and complex nature of conversational expressions.

Quality Assessment

To assess inter-annotator agreement, we use Cohen's Kappa Statistic (Cohen 1960). The average Cohen's Kappa scores for annotator pairs evaluating style and intent in the MultiChat dataset are 0.919 and 0.832, respectively. These substantial Kappa scores demonstrate strong agreement among the annotations, indicating the reliability of the annotations. After the initial annotation, we conducted a post-annotation review process. This involved a detailed examination of randomly selected samples from the dataset by senior researchers to ensure consistency and correctness in annotations, providing another layer of quality.

Dataset Statistics

Table 1 provides a detailed overview of the dataset statistics. In total, there are 1,542 conversation sessions which contain 50,192 utterances, and 2,182 stickers. In this paper, we split each session into multiple samples based on the stickers. For instance, consider a session containing m utterances $C = \{u_1, u_2, \ldots, u_m\}$, where each u can be either text or a sticker. If u_4, u_{10}, and u_{15} are stickers, the three samples are formed by $\{u_1, u_2, \ldots, u_4\}$, $\{u_1, u_2, \ldots, u_{10}\}$, and $\{u_1, u_2, \ldots, u_{15}\}$, respectively. After dividing, there are 4,851 samples. The ratio of the training set, validation set, and test set is approximately 3:1:1. Each conversation session includes 41.38 utterances on average. The average number of users who participate in a conversation is 6.28.

Methodology

In this section, we introduce our novel IGSR framework for multi-modal multi-session sticker retrieval in detail. Each conversation is denoted as $C = \{H_n, D_m, v_m\}$, where $H_n = \{H_1, H_2, ..., H_n\}$ indicates n past sessions. Each session includes multiple utterances or stickers among speakers. $D_m = \{(s_1, u_1), (s_2, u_2), ..., (s_m, u_m)\}$ denotes the context of the current session at m step, where s and u denote the speaker and the utterance/sticker from the corresponding speaker. v_m is the ground truth sticker to D_m with history session H.

To select an appropriate sticker v from the sticker set V or a conversation based on past and current sessions, we propose a pipeline framework, IGSR, to deal with the sticker retrieval task, as shown in Figure 2, which consists of three primary components: 1) ***Multi-modal History Modeling***: This component aggregates multi-modal historical data, including attributes using a Large Language Model (LLM) to create a memory bank. 2) ***Intent Derivation***: This component integrates relevant memory from the memory bank with the current session to predict the user's intent. 3) ***Sticker Retrieval***: This final component selects a sticker based on the derived intent and sticker representations.

Multi-modal History Modeling

History session is crucial in conversation which can help in understanding the context and maintaining the topic. However, existing conversation systems primarily concentrate on the single-modal text from history sessions or even overlook the history sessions entirely. Recognizing that real-world dialogues convey emotions and intentions in a multi-modal manner, we propose to model the multi-modal history.

To better capture the richness of historical interactions, for each sticker in the history session, we design six attributes to represent the key information and reduce unnecessary interference from irrelevant information for the learning of stickers, i.e. *intent* L_I, *style* L_S, *gesture* L_G, *posture* L_P, *facial expression* L_F, and *verbal* L_V. The intent and style attributes are provided from our dataset. For the other four attributes, we use Qwen-VL (Bai et al. 2023a), a Multi-modal Large Language Model (MLLM) to produce attribute-aware sticker descriptions based on the designed prompts $\{A_G, A_P, A_F, A_V\}$:

$$\begin{aligned} &\{L_G, L_P, L_F, L_V\} \\ &= \text{MLLM}(\{A_G, A_P, A_F, A_V\}) \end{aligned} \tag{1}$$

Through several turns of interactions, we use system prompts such as, "This is a sticker used in conversation. Please provide several keywords to describe the gesture, posture, facial expression, and verbal content," to leverage the LLM's ability to generate these descriptive attributes for each sticker. Then, we feed the history sessions, including utterances and the attributes of the stickers, in chronological order into the LLM to generate a summary, which is then stored in the memory bank.

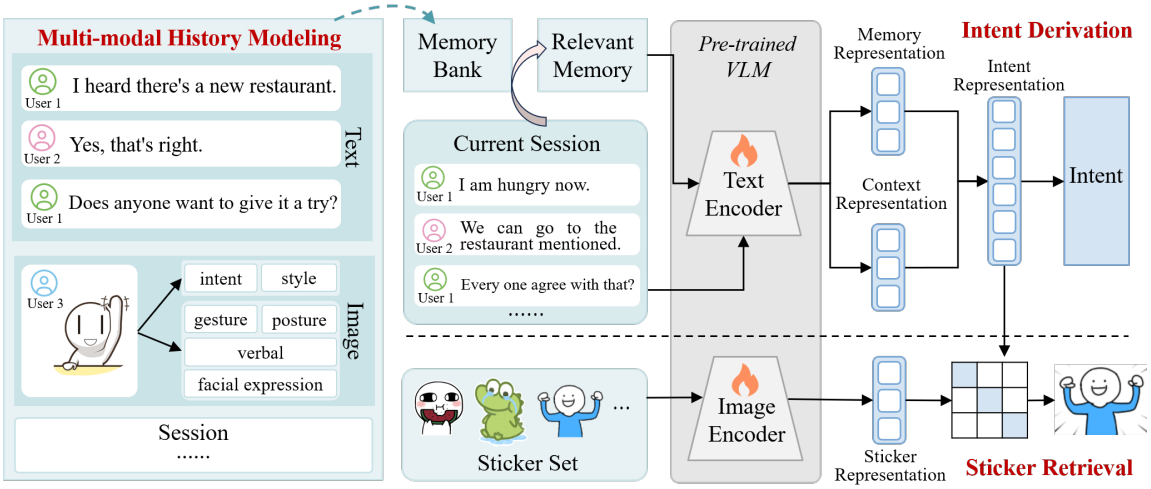

Figure 2: Illustration of our IGSR framework comprising multi-modal history modeling, intent derivation, and sticker retrieval.

$$M_i = \text{LLM}(D_i, P) \tag{2}$$

where M_i indicates multiple sentences that summarise the crucial information based on the current session, and P denotes the prompt of LLM for memory generation: "*Your goal is to summarize the session [D_i]*". This operation is repeated K times until the session ends, resulting in the final memory bank M_K.

Intent Derivation

To capture the precise intent during the conversation, we incorporate as many relevant historical memories as possible to capture intent-related features. We then combine the current session with a designed encoder to predict the intent.

Similar to the process in multi-modal history modeling, we use attributes to represent stickers in the current session. Then, the current session and relevant memories are fed into the text encoder to derive summary and context representations. Using text-davinci-003 embeddings for semantic retrieval, we extract the top-N relevant memories $M_r = \{M_1, ..., M_k\}$ from the memory bank, which stores summaries of historical sessions based on the summary of the current session D_i.

$$\begin{aligned} R_m &= f_{text}(M_r) \\ R_c &= f_{text}(D_i) \end{aligned} \tag{3}$$

where the f_{text} represents the text encoder. The intent representation R_I is obtained by concatenating memory representation R_m and context representation R_c. The combined representation is then fed into a classifier, consisting of a linear layer for dimensionality reduction followed by a softmax function to produce the probability distribution for each intent category. We train the model using the standard gradient descent algorithm, minimizing the cross-entropy loss:

$$\hat{y}_I = \sigma(W_I R_I + b_I), \text{ where } R_I = R_m \oplus R_c \tag{4}$$

$$\mathcal{L}_I = -\sum_{j=1}^{N} y_I^j \log \hat{y}_I^j + \lambda_I ||\Theta_I||^2 \tag{5}$$

where σ denotes the softmax function. $W_I \in \mathbb{R}^{d_r \times d_r}$ is the learnable parameter and b_I is the bias training alone with the model. d_r is the dimension of intent representation. y_I and $\hat{y}_I$ represent the ground truth and predicted label distribution of the intent. Θ_I represents all the learnable parameters of the model, and λ_I denotes the coefficient for L2 regularization.

Sticker Retrieval

To obtain the sticker representation, we apply the pre-trained vision transformer from CLIP model as the image encoder.

$$R_v = f_{image}(v) \tag{6}$$

where R_v is the sticker representation and f_{image} denotes the image encoder.

Loss Function. During the training stage, we follow previous contrastive learning methods (Radford et al. 2021b) and utilize the InfoNCE loss (Oord, Li, and Vinyals 2018) to train our framework. Given a batch of β intent-sticker representation pairs $r_j, v_j, j = 1^{\beta}$ as training data, where $r_j \in R_I$ and $v_j \in R_v$, we calculate the text-to-image contrastive loss $\mathcal{L}_{v2t}$ and the image-to-text contrastive loss $\mathcal{L}_{t2v}$ as follows:

$$\mathcal{L}_{v2t} = -\log \frac{\exp(f_{sim}(v_j, r_j)/\tau)}{\sum_{k=1}^{\beta} \exp(f_{sim}(v_j, r_k)/\tau)} \tag{7}$$

$$\mathcal{L}_{t2v} = -\log \frac{\exp(f_{sim}(r_j, v_j)/\tau)}{\sum_{k=1}^{\beta} \exp(f_{sim}(r_j, v_k)/\tau)} \tag{8}$$

where $f_{sim}(r_j, v_j)$ denotes the cosine similarity and $\tau \in \mathbb{R}^+$ is a temperature factor. β represents the batch size. The total loss of our approach is:

$$\mathcal{L} = \mathcal{L}_{v2t} + \mathcal{L}_{t2v} + \mathcal{L}_I \tag{9}$$

Modality	Model	P@1	P@3	P@5	mAP	GPT-4	Human Evaluation
Text Modality	Baichuan2	7.00	18.22	25.58	13.57	38.60	48.25
	Qwen1.5	8.08	21.36	31.42	15.96	44.74	36.84
	Llama3	8.53	21.36	33.03	16.53	43.86	42.11
	ChatGLM3	14.09	18.85	23.16	16.97	42.98	31.58
Multi-modality	MOD	3.23	4.85	6.46	5.21	21.93	51.75
	SRS	3.77	6.82	9.34	7.09	24.56	52.63
	IRRA	5.57	9.69	14.90	14.55	25.44	55.26
	LGUR	8.26	14.81	17.68	15.23	21.93	58.77
	PCME	9.34	21.72	28.73	16.93	35.09	64.04
	CLIP	9.25	22.62	28.19	16.57	35.97	64.04
	Qwen-VL	9.25	24.69	48.38	14.73	39.47	62.28
	GPT-4	9.96	24.42	34.83	17.85	44.74	57.90
	LLaVA	10.23	28.10	44.08	13.74	30.70	61.40
	IGSR	**14.45***	**34.02***	**54.94***	**19.39***	**64.04***	**67.54***

Table 2: Main results (%) of various methods. Bold indicates that our method surpasses other models. The results with * indicate the significance tests of our IGSR over other baseline models (with p-value < 0.05).

Experiment

Experimental Settings

Implement details. We apply GPT-4 (Achiam et al. 2023) to construct the memory bank in multi-modal history modeling and utilize the Qwen-VL (Bai et al. 2023a) to generate four attributes of stickers. We extract the top 5 relevant memories based on text-davinci-003 embeddings. The CLIP text encoder and image encoder are employed. We use a batch size of 2 and employ the Adam optimizer (Kingma and Ba 2014) for training. The learning rate is set to 1×10^{-4}. All experiments are conducted at Tesla V100s[2].

Evaluation metrics. In our experiments, we utilize four evaluation metrics: P@N, mAP, GPT-4, and human evaluation. P@N measures the precision of the top N predictions, focusing on P@1, P@3, and P@5. A result is correct if the retrieved sticker matches the intention label of the ground truth sticker, acknowledging that multiple stickers can appropriately respond to the same conversation. Mean average precision (mAP) is used as a widely accepted metric for evaluating retrieval accuracy (Lin et al. 2014). We utilize both GPT-4 and human evaluation to ensure a comprehensive assessment of models' performance. Specifically, we randomly sample around 10% of the test cases (114 samples) and ask both GPT-4 and human evaluators to rate the quality on a scale of $\{0, 1\}$, focusing on the background consistency and relevance of the stickers. This dual approach allows us to capture both automated and nuanced human perspectives.

Baselines

To assess the performance of our model, we compare the proposed IGSR against several baseline methods, including existing text-based models and multi-modal models. (1) **Text-based models**: Baichuan2 (Yang et al. 2023), Llama3 (Touvron et al. 2023), ChatGLM3 (Du et al. 2022), Qwen1.5 (Bai et al. 2023b). **Multi-modal models**: IRRA (Jiang and Ye 2023), PCME (Chun et al. 2021), MOD (Fei et al. 2021b), SRS (Gao et al. 2020), LGUR (Shao et al. 2022), CLIP (Radford et al. 2021a), LLaVA (Liu et al. 2024), Qwen-VL (Bai et al. 2023b), GPT-4 (Achiam et al. 2023). Notably, MOD and SRS are two sticker retrieval approaches.

For LLMs including Baichuan2, Qwen1.5 Llama3, ChatGLM3, Qwen-VL, GPT-4, LLaVA, we first design a *text response generation response prompt* that integrates the relevant summary with the current session to generate responses for each session. We then retrieve the appropriate sticker based on the generated response and the sticker attributes using BM25 (Robertson et al. 1995). In text-based LLMs such as Baichuan2, Qwen1.5, Llama3, and ChatGLM3, sticker attributes are derived using Qwen-VL. For multi-modal LLMs like Qwen-VL, GPT-4, and LLaVA, these attributes are directly obtained by the models themselves. We also utilize OpenAI's LLM-based embedding model as a retriever and design a *sticker intent prediction prompt* for LLM to generate sticker intents or descriptions instead of responses for retrieval, but this approach results in decreased performance.

[2]Dataset and code are released in https://github.com/HITSZ-HLT/IGSR.

Model	P@1	P@3	P@5	mAP	GPT-4	Human
IGSR	**14.45**	**34.02**	**54.94**	**19.39**	**64.04**	**67.54**
w/o mem	7.36	25.40	42.46	18.87	29.38	56.01
w/o int	2.69	32.23	53.32	17.24	26.88	52.50

Table 3: Experimental results (%) of ablation study. w/o mem and w/o int mean without memory and intent. "Human" represents the human evaluation.

Main Results

We compare the performance of our IGSR with baselines across various evaluation metrics, as shown in Table 2. IGSR consistently outperforms all baseline models, demonstrating its effectiveness in multi-modal, multi-session sticker retrieval. We observe improved results in Top N precision as N increases, since a larger N allows for a greater number of results, expanding the scope of potential matches and enhancing the likelihood of finding relevant labels.

Model	P@1	P@3	P@5	mAP
	w/wo	w/wo	w/wo	w/wo
Baichuan2	7.0/7.7	18.2/19.6	25.6/29.7	13.6/15.1
Qwen1.5	8.1/8.9	21.4/22.2	31.4/32.9	16.0/16.7
Llama3	8.5/7.9	21.4/20.7	33.0/31.0	16.5/15.6
ChatGLM3	14.1/13.5	18.9/22.0	23.2/27.5	17.0/18.3
Qwen-VL	9.2/9.3	24.7/25.5	48.4/47.2	14.7/14.5
LLaVA	10.2/8.3	28.1/25.4	44.1/42.3	13.7/13.9

Table 4: Experimental results (%) of effect of memory. / splits the results using memory and without using.

Text-based models, predominantly implemented by LLMs, outperform some multi-modal models. This superior performance is due to LLMs' extensive parameterization and sophisticated network architectures, which significantly enhance their ability to understand and generate intricate language and image descriptions. For the metric P@1, ChatGLM3 performs the best, while for P@5, Qwen-VL shows the best performance, reaching 48.384%. All baseline models significantly underperform compared to our IGSR. This further highlights the effectiveness of our approach in intent derivation, underscoring the pivotal role of intention as a key bridging element in the process.

Multi-modal models include both small models (e.g., PCME, MOD, IRRA, LGUR, and CLIP) and large models (e.g., Qwen-VL, LLaVA). Small models primarily focus on capturing semantic relationships between textual and visual content but are not specifically designed for sticker retrieval scenarios, leading to inferior performance compared to our proposed framework. Interestingly, as a sticker retrieval model, MOD exhibits overall inferior performance compared to most multi-modal methods. This disparity can be attributed to MOD's design, which targets retrieving suitable stickers from a limited set of similar candidates, thus emphasizing the distinction of local information among similar sticker expressions. As large models, Qwen-VL and LLaVA perform better than smaller models but still fall short of our approach, which leverages relevant memory, significantly enhancing performance. Across the P@1, P@3, and P@5 metrics, our method achieves minimum improvements of 4.229%, 5.882%, and 6.553%, respectively. In summary, while large multi-modal models outperform smaller ones in sticker retrieval tasks, our IGSR framework surpasses both, demonstrating its exceptional robustness and effectiveness in handling complex multi-modal scenarios.

Furthermore, the results of GPT-4 and human assessments are not entirely consistent. For the text-based method, GPT-4's scores are higher compared to the multi-modal method, whereas human evaluations show the opposite trend. Additionally, human evaluations overall are higher than those of GPT-4. This discrepancy may stem from differing evaluation criteria: GPT-4 focuses on the alignment of image and text features, while human evaluators consider the conversational context and potential emotional nuances.

Char.	Intent	Style	P@1	P@3	P@5	mAP
			7.54	23.79	33.12	12.73
√			9.96	26.30	44.25	19.21
	√		10.95	32.68	52.96	20.58
		√	5.30	23.34	44.43	19.67
√	√		9.61	27.02	43.45	20.41
√		√	7.81	24.42	39.95	18.98
	√	√	9.87	28.28	46.95	18.32
√	√	√	**14.45**	**34.02**	**54.94**	**19.39**

Table 5: Experimental results (%) of different attributes. √ represents the used attribute.

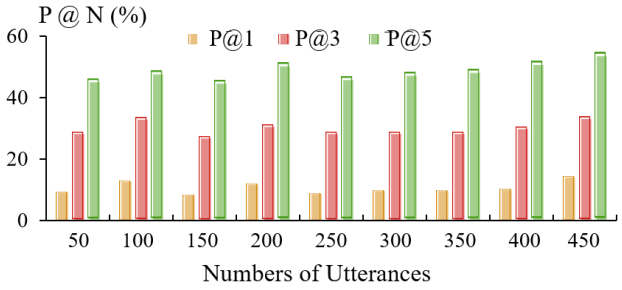

Figure 3: Performance of our approach across all metrics when varying the number of utterances.

Ablation Study

We also perform an ablation study on the use of memory and intent with evaluation results shown in Table 3. All ablation models perform worse than the complete model under all metrics, demonstrating the necessity of each component in our approach. Notably, removing relevant memory ("w/o memory") leads to considerable performance degradation, underscoring the importance of summary in understanding the conversation context. Moreover, the removal of intent ("w/o intent") significantly degrades performance, especially in the metric of P@1, indicating that the intent prediction during the model training improves the learning of sticker representation across different sticker properties.

Effect of Relevant Memory. Based on the results of the ablation study, we further explore the role of memory in sticker retrieval. Specifically, we design a prompt to generate the response without relevant memory. The comparative results are shown in Table 4, where the results with and without memory are separated by a slash. We observe that without using relevant memory, Baichuan2, Qwen1.5, ChatGLM3, and Qwen-VL show an increase in performance. This may be because incorporating relevant memory results in input sequences that are too long, causing LLMs to struggle with processing lengthy inputs effectively, which in turn impacts their performance. Notably, the performance of Llama3 and LLaVA improves when relevant memory is utilized, suggesting that these models have enhanced capabilities for processing long text inputs.

Effect of Attributes. In the process of multi-modal history modeling, six attributes are utilized to represent the key

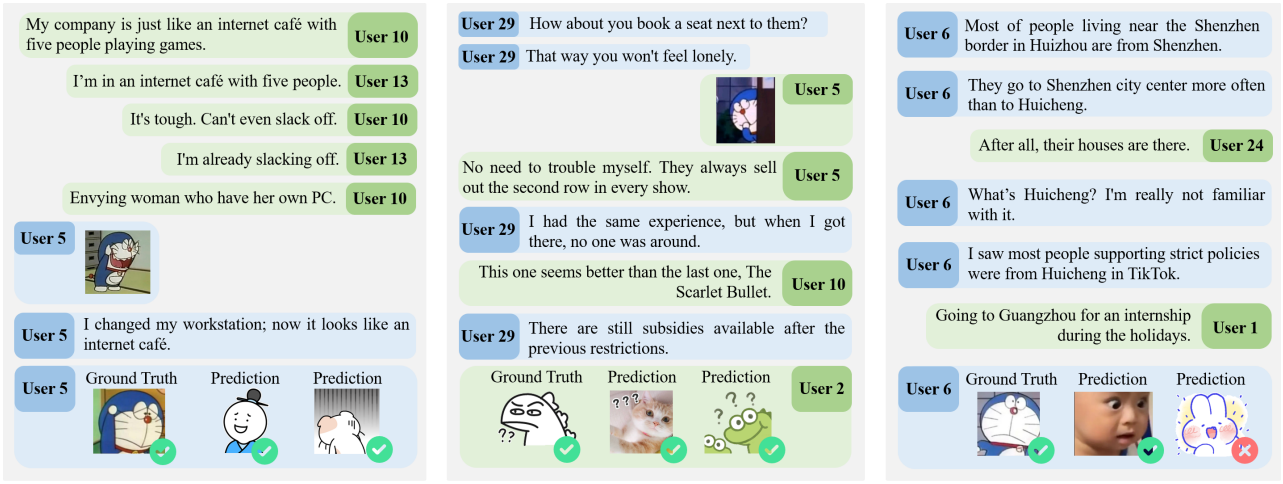

Figure 4: Examples of conversation context and stickers. Prediction represents the sticker retrieved using our method.

information of the sticker. This section examines the effectiveness of different attributes. We consider posture, gesture, and verbal expression as the main sticker characteristics generated by the LLM (referred to as "Char.") and list all the scenarios in Table 5. It can be seen that using all the attributes gains the best performance. Conversely, the exclusion of attributes results in the lowest performance. Using only the intent to represent the sticker achieves performance close to using all attributes, with the lowest difference being just 1.346% at P@3. In contrast, using only the style attribute yields the worst performance. This indicates that the intent provided in our dataset better captures the sticker's function, while the "Character" attribute provides supplementary details. These three types of attributes complement each other, and using all attributes together more comprehensively represents the crucial information of the sticker.

Effect of Max Number of Utterances. In a conversation, the context of the current session can provide valuable information for the response. This section explores and analyzes the effect of varying the maximum number of utterances per session on the performance of our IGSR framework. We conducted experiments with maximum values ranging from 50 to 450, and the results are illustrated in Figure 3. The maximum lengths of utterances in the training, validation, and test phases are 428, 423, and 436, respectively. Therefore, setting the maximum number of utterances to 450 means we do not restrict the utterances in the current session. As shown in Figure 3, we observe that using all utterances yields the best performance, while limiting the number to 50 results in the worst performance. This suggests that insufficient context and data hinder accurate predictions, highlighting the importance of utilizing a more extensive range of utterances for better model performance.

Case Study

Figure 4 showcases various interactive cases retrieved by our approach. These examples highlight the IGSR framework's ability to enhance communication with expressive and contextually relevant stickers. From Figure 4(a), we observe that the expression of sadness can be conveyed through both facial expressions and actions. Our proposed method effectively learns information from various expressions or actions, allowing for appropriate sticker responses. In addition, a significant challenge in sticker retrieval is its diversity in styles. Our model, incorporating intent information, reduces interference from different styles, thereby enabling more precise localization of useful sticker information. As shown in example Figure 4(b), our method retrieves stickers that encompass both realistic and cartoon styles.

Unlike traditional dialogue systems, this study collected open-domain group chat data, which exhibits characteristics of multi-party, multi-modal, multi-turn conversations. Therefore, complex conversational dynamics and interaction patterns may occur during chat dialogues. As seen in Figure 4(c), the content spoken by *User 1* is not strongly related to the previous context, thus affecting the final sticker prediction results. Consequently, in future research, combining user information could be explored to further enhance the performance of sticker retrieval.

Conclusion

We create a new dataset for multi-modal multi-session sticker retrieval, called MultiChat. Unlike previous studies that retrieve stickers based on the current session, our new dataset can cover more realistic scenarios. Based on our created dataset, we propose IGSR, a framework for sticker retrieval in conversation. In which, a multi-modal history modeling strategy is designed for memory bank construction, and a multi-tasking framework is employed to simultaneously derive intents and retrieve stickers. Extensive experiments on our MultiChat dataset highlight the importance of intent and demonstrate that our proposed approach effectively utilizes memory, achieving exceptional performance in multi-modal, multi-session sticker retrieval.

Acknowledgements

This work was partially supported by the National Natural Science Foundation of China 62176076, Natural Science Foundation of GuangDong 2023A1515012922, the Shenzhen Foundational Research Funding JCYJ20220818102415032 and JCYJ20210324115614039, the Major Key Project of PCL2021A06, Guangdong Provincial Key Laboratory of Novel Security Intelligence Technologies 2022B1212010005.

References

Achiam, J.; Adler, S.; Agarwal, S.; Ahmad, L.; Akkaya, I.; Aleman, F. L.; Almeida, D.; Altenschmidt, J.; Altman, S.; Anadkat, S.; et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Albar, D. 2018. Chat Sticker Design as Media Recognition of Character in Instant Messaging Platform. In *International Conference on Business, Economic, Social Science and Humanities (ICOBEST 2018)*, 307–310. Atlantis Press.

Aman, S.; and Szpakowicz, S. 2008. Using roget's thesaurus for fine-grained emotion recognition. In *Proceedings of the Third International Joint Conference on Natural Language Processing: Volume-I*.

Bai, J.; Bai, S.; Yang, S.; Wang, S.; Tan, S.; Wang, P.; Lin, J.; Zhou, C.; and Zhou, J. 2023a. Qwen-vl: A frontier large vision-language model with versatile abilities. *arXiv preprint arXiv:2308.12966*.

Bai, J.; Bai, S.; Yang, S.; Wang, S.; Tan, S.; Wang, P.; Lin, J.; Zhou, C.; and Zhou, J. 2023b. Qwen-vl: A frontier large vision-language model with versatile abilities. *arXiv preprint arXiv:2308.12966*.

Chun, S.; Oh, S. J.; De Rezende, R. S.; Kalantidis, Y.; and Larlus, D. 2021. Probabilistic embeddings for cross-modal retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 8415–8424.

Cohen, J. 1960. A coefficient of agreement for nominal scales. *Educational and psychological measurement*, 20(1): 37–46.

Demszky, D.; Movshovitz-Attias, D.; Ko, J.; Cowen, A.; Nemade, G.; and Ravi, S. 2020. GoEmotions: A Dataset of Fine-Grained Emotions. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 4040–4054.

Du, Z.; Qian, Y.; Liu, X.; Ding, M.; Qiu, J.; Yang, Z.; and Tang, J. 2022. 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)*, 320–335.

Fei, Z.; Li, Z.; Zhang, J.; Feng, Y.; and Zhou, J. 2021a. Towards expressive communication with internet memes: A new multimodal conversation dataset and benchmark. *arXiv preprint arXiv:2109.01839*.

Fei, Z.; Li, Z.; Zhang, J.; Feng, Y.; and Zhou, J. 2021b. Towards expressive communication with internet memes: A new multimodal conversation dataset and benchmark. *arXiv preprint arXiv:2109.01839*.

Gao, S.; Chen, X.; Liu, C.; Liu, L.; Zhao, D.; and Yan, R. 2020. Learning to respond with stickers: A framework of unifying multi-modality in multi-turn dialog. In *Proceedings of the Web Conference 2020*, 1138–1148.

Ge, F.; Li, W.; Ren, H.; and Cai, Y. 2022. Towards Exploiting Sticker for Multimodal Sentiment Analysis in Social Media: A New Dataset and Baseline. In *Proceedings of the 29th International Conference on Computational Linguistics*, 6795–6804.

Huang, M.; Long, Y.; Deng, X.; Chu, R.; Xiong, J.; Liang, X.; Cheng, H.; Lu, Q.; and Liu, W. 2024. DialogGen: Multimodal Interactive Dialogue System for Multi-turn Text-to-Image Generation. *arXiv preprint arXiv:2403.08857*.

Jiang, D.; and Ye, M. 2023. Cross-Modal Implicit Relation Reasoning and Aligning for Text-to-Image Person Retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2787–2797.

Kingma, D. P.; and Ba, J. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Laddha, A.; Hanoosh, M.; Mukherjee, D.; Patwa, P.; and Narang, A. 2020. Understanding chat messages for sticker recommendation in messaging apps. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, 13156–13163.

Lin, T.-Y.; Maire, M.; Belongie, S.; Hays, J.; Perona, P.; Ramanan, D.; Dollár, P.; and Zitnick, C. L. 2014. Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, 740–755. Springer.

Liu, H.; Li, C.; Wu, Q.; and Lee, Y. J. 2024. Visual instruction tuning. *Advances in neural information processing systems*, 36.

Liu, S.; Zhang, X.; and Yang, J. 2022. SER30K: A large-scale dataset for sticker emotion recognition. In *Proceedings of the 30th ACM International Conference on Multimedia*, 33–41.

Maharana, A.; Lee, D.-H.; Tulyakov, S.; Bansal, M.; Barbieri, F.; and Fang, Y. 2024. Evaluating Very Long-Term Conversational Memory of LLM Agents. *arXiv preprint arXiv:2402.17753*.

Nilasari, N. L.; Sudipa, I. N.; and Sukarini, N. W. 2018. Sticker Emoticons Used in LINE Messenger; A Semantic Study. *J. Humanis*, 22: 585–591.

Oord, A. v. d.; Li, Y.; and Vinyals, O. 2018. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. 2021a. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, 8748–8763. PMLR.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. 2021b. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, 8748–8763. PMLR.

Robertson, S. E.; Walker, S.; Jones, S.; Hancock-Beaulieu, M. M.; Gatford, M.; et al. 1995. Okapi at TREC-3. *Nist Special Publication Sp*, 109: 109.

Shao, Z.; Zhang, X.; Fang, M.; Lin, Z.; Wang, J.; and Ding, C. 2022. Learning granularity-unified representations for text-to-image person re-identification. In *Proceedings of the 30th ACM International Conference on Multimedia*, 5566–5574.

Shi, C.; Chen, Q.; Sha, L.; Xue, H.; Li, S.; Zhang, L.; and Wang, H. 2019. We know what you will ask: A dialogue system for multi-intent switch and prediction. In *Natural Language Processing and Chinese Computing: 8th CCF International Conference, NLPCC 2019, Dunhuang, China, October 9–14, 2019, Proceedings, Part I 8*, 93–104. Springer.

Touvron, H.; Lavril, T.; Izacard, G.; Martinet, X.; Lachaux, M.-A.; Lacroix, T.; Rozière, B.; Goyal, N.; Hambro, E.; Azhar, F.; et al. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Yang, A.; Xiao, B.; Wang, B.; Zhang, B.; Bian, C.; Yin, C.; Lv, C.; Pan, D.; Wang, D.; Yan, D.; et al. 2023. Baichuan 2: Open large-scale language models. *arXiv preprint arXiv:2309.10305*.

Zhang, Y.; Kong, F.; Wang, P.; Sun, S.; Wang, L.; Feng, S.; Wang, D.; Zhang, Y.; and Song, K. 2024. StickerConv: Generating Multimodal Empathetic Responses from Scratch. *arXiv preprint arXiv:2402.01679*.

Zhang, Z.; Zhu, Y.; Fei, Z.; Zhang, J.; and Zhou, J. 2022. Selecting Stickers in Open-Domain Dialogue through Multitask Learning. In *Findings of the Association for Computational Linguistics: ACL 2022*, 3053–3060.

Zhao, S.; Ge, Y.; Qi, Z.; Song, L.; Ding, X.; Xie, Z.; and Shan, Y. 2023. Sticker820K: Empowering Interactive Retrieval with Stickers. *arXiv preprint arXiv:2306.06870*.