STICKERCONV: Generating Multimodal Empathetic Responses from Scratch

Anonymous ACL submission

Abstract

Stickers, while widely recognized for enhancing empathetic communication in online interactions, remain underexplored in current empathetic dialogue research, notably due to the challenge of a lack of comprehensive datasets. In this paper, we introduce the Agent for STICKERCONV (Agent4SC), which uses collaborative agent interactions to realistically simulate human behavior with sticker usage, thereby enhancing multimodal empathetic communication. Building on this foundation, we develop a multimodal empathetic dialogue dataset, STICKERCONV, comprising 12.9K dialogue sessions, 5.8K unique stickers, and 2K diverse conversational scenarios. This dataset serves as a benchmark for multimodal empathetic generation. To advance further, we propose **PE**rceive and **G**enerate **S**tickers (PEGS), a multimodal empathetic response generation framework, complemented by a comprehensive set of empathy evaluation metrics based on LLM. Our experiments demonstrate PEGS's effectiveness in generating contextually relevant and emotionally resonant multimodal empathetic responses, contributing to the advancement of more nuanced and engaging empathetic dialogue systems[1].

1 Introduction

Increasing research indicates that utilizing stickers in online chats can effectively alleviate stress, augment personal happiness, and notably boost empathy (Kariko and Anasih, 2019; Akram and Drabble, 2022; Jiang et al., 2023). Prior studies on stickers primarily concentrated on sentiment analysis (Ge et al., 2022) and recommendation systems (Gao et al., 2020; Zhang et al., 2022), overlooking their vast potential in empathetic response generation. Most empathetic response generation tasks focus solely on textual modality (Fu et al., 2023; Lee et al., 2022; Jiang et al., 2023), yet stickers

---

[1] https://anonymous.4open.science/r/StickerConv

Figure 1: An example of multimodal conversation in the STICKERCONV. Both parties can utilize the stickers to express their emotions, which enhances interactivity and expression. The assistant can empathize with the user according to the conversation (green text).

in chats convey more abundant and intuitive emotional information, enhancing the expressiveness and emotional depth of responses.

Integrating stickers with textual communication and interspersing stickers within the dialogue can yield more varied and superior-quality empathetic replies. A primary challenge in integrating stickers into empathetic response generation is developing a high-quality dataset to support this innovative multimodal communication. To address this, we leverage large language model (LLMs) for dataset construction. LLMs, with their extensive world knowledge and text processing capabilities (OpenAI, 2023; Touvron et al., 2023a,b), demonstrate near-human annotation abilities (Wang et al., 2023d; Pangakis et al., 2023). However, applying LLMs directly have limitations in empathetic tasks, excelling in responding to explicit human instructions but lacking proactivity, a critical aspect of empathy (Wu et al., 2023; Yin et al., 2023). Em-

pathy necessitates understanding others' emotions and the ability to actively express support and understanding (Zech and Rimé, 2005; Sharma et al., 2020). To mitigate this, we introduce a multi-agent system based on LLMs, **Agent for STICKERCONV** (Agent4SC). This system, through inter-agent interactions, utilizes stickers to simulate human-like dialogue scenarios. It not only generates text responses but also strategically selects suitable stickers, thereby effectively enhancing empathy.

Based on Agent4SC, we build a multimodal empathetic dataset, **STICKERCONV**, that comprises 12.9K dialogue sessions and 5.8K unique stickers. STICKERCONV boasts an average of 5.22 stickers per dialogue session, mirroring the sticker usage patterns observed in human communication. Figure 1 depicts an example of conversations in our dataset. To the best of our knowledge, this is the first multimodal empathetic dialogue dataset, with the particular utility of sticker as non-textual modal information to better facilitate empathy.

Although Agent4SC effectively generates multimodal empathetic responses, it is limited by expensive inference costs and specific sticker databases. To further advance the research on multimodal empathetic dialogue, we develop an end-to-end multimodal empathetic response generation framework, **PEGS**, with the ability to **PE**rceive and **G**enerate **S**tickers. Beyond the general ability to generate textual empathetic responses, PEGS receives multimodal inputs and autonomously generates stickers based on the emotional and contextual aspects of the dialogue at the appropriate moment. Furthermore, to simulate human communications on social media in the real world, our model supports interleaved multiple image and text inputs.

Misalignments between the modal quantities of predicted and golden responses can distort evaluation outcomes, and empathy is difficult to quantify due to its subjective nature. To address this, we propose a novel method for evaluating multimodal empathetic responses, focusing on **empathy**, **consistency**, and **ranking**. Utilizing the extensive world knowledge and anthropomorphic abilities of LLMs, this approach provides solid support for assessing multimodal empathetic replies.

In conclusion, the main contributions of this work are as follows:

- We introduce an LLM-based multi-agent system, Agent for STICKERCONV (Agent4SC), which integrates stickers into empathetic dialogues, ensuring contextual consistency, variety, and em-

pathy aligned with human interactions. Using Agent4SC, we create a multimodal empathetic dialogue dataset, STICKERCONV.

- We design **PE**rceive and **G**enerate **S**tickers (PEGS), a multimodal empathetic response generation framework that intuitively incorporates stickers based on the emotional and contextual dynamics of the dialogue. PEGS adeptly processes multimodal inputs, generating empathetic textual responses and using stickers appropriately to enhance these responses.

- We propose a method for assessing multimodal empathetic responses. It leverages LLM to evaluate the quality of these responses, with a specific focus on empathy, consistency, and ranking.

2 Related Work

2.1 Empathetic Response Generation

Empathetic response generation focuses on enabling machines to understand and respond to human emotions. The foundational EMPATHETIC-DIALOGUES dataset (Rashkin et al., 2019) and subsequent innovations like the empathetic listener mixture model (Lin et al., 2019) have significantly advanced this area. Large Language Models (LLMs) like ChatGPT have been explored for empathetic response generation, although their application remains limited (Lee et al., 2022; Zhou et al., 2023b). Challenges persist, especially in leveraging multimodal information for richer emotional engagement and in accurately evaluating empathetic responses due to their subjective nature. These obstacles highlight the ongoing need for research in effectively integrating LLMs and multimodal data into empathetic dialogue. Additionally, the subjective nature of empathy complicates its quantitative assessment (Fu et al., 2023; Lee et al., 2022), posing a further obstacle to the field's advancement.

2.2 Large Multimodal Models

LLMs, such as ChatGPT (OpenAI, 2023), LLaMA (Touvron et al., 2023a,b), demonstrate powerful capabilities in dialog interaction and instruction following, and recent researches have extended LLMs to multimodal domains. Flamingo (Alayrac et al., 2022) exhibits promising zero-shot and few-shot multimodal understanding capability by adding a cross-attention layer to connect the frozen vision encoder with the LLM. BLIP (Li et al., 2023b; Dai et al., 2023), MiniGPT-4 (Zhu et al., 2023; Chen et al., 2023) and LLaVA (Liu et al., 2023b,a) bridge

the frozen vision encoder and the LLM through a small intermediate model. Koh et al. (2023) for the first time explore the mapping of the output of LLMs into the input space of the vision decoder, empowering LLMs with image generation capability. GILL (Koh et al., 2023), MiniGPT-5 (Zheng et al., 2023) align LLMs to frozen vision decoders through an encoder-decoder transformer, while SEED (Ge et al., 2023) employs a learnable Q-Former. In contrast to previous works which favor realistic images, our target is to generate stickers, which are abstract and exhibit distinct emotional tendencies.

2.3 LLM-Based Agents

LLM-based agents mark a major leap in AI, leveraging their capabilities for tasks like reasoning and interaction, as shown in recent studies (Wang et al., 2023b; Li et al., 2023a). They find uses across various domains, such as software engineering (Qian et al., 2023) and scientific inquiry (Boiko et al., 2023), highlighting their versatility. These agents can imitate complex human actions, partake in social interactions (Park et al., 2023; Tu et al., 2023), and replicate intricate scenarios like elections (Argyle et al., 2022), debates (Wang et al., 2023a), and consumer patterns (Wang et al., 2023c), illustrating their capacity to emulate human social dynamics. A notable innovation is their use in generating AI training data. Studies (Wang et al., 2023d; Peng et al., 2023; Zhou et al., 2023a) highlight their efficiency and cost-effectiveness in producing high-quality training datasets, transforming AI model development.

3 Agent for STICKERCONV

Confronted with the pivotal challenge of a lack of datasets for multimodal empathetic response tasks, we craft our own dataset utilizing large language models (LLMs). Nevertheless, LLMs experience difficulties in grasping nuanced human emotions and initiating actions beyond explicit directives. These limitations render LLMs and large multimodal models (LMMs) less proficient in the generation of multimodal empathetic responses. In response to these issues, we introduce Agent for STICKERCONV (Agent4SC), a multi-agent system predicated on LLM, devised to mimic human conversational patterns. Figure 2 presents the overview of Agent4SC. By integrating multiple modules and the strategic use of stickers, Agent4SC aims to generate emotional and varied empathetic responses, thereby overcoming the inherent deficiencies of LLMs in empathetic engagements.

3.1 Profile Module

The profile module, comprising **Persona** and **Situation** components, underpins its foundation by defining users' personality traits and behavioral patterns for empathetic interactions. Persona outlines users' character traits, backgrounds, and experiences, while Situation details their current circumstances and emotional states.

To enrich and diversify user profiles, we initially create profiles with varied emotional distributions, then expand to 2,000 unique profiles using **SELF-INSTRUCT** (Wang et al., 2023d) method. This approach results in a profile archive with a broad emotional spectrum, as shown in Figure 5, aiming to enhance the system's ability to simulate human-like responses in empathetic dialogues.

3.2 Tool Module

To adapt the SER30K dataset, with its 1,887 themes and 30,739 emotion-tagged stickers (Liu et al., 2022), for human-like sticker use in dialogues, we transform it into a tool. Integrating it into the agent system faced several hurdles: the inappropriateness of many stickers for empathetic dialogues, the lack of detailed content or emotional analysis in the single-emotion labeled stickers, and the contrast between the extensive SER30K collection and the limited, personalized collections humans use.

To make stickers an effective tool, we refine the SER30K through **filtering**, **annotating**, and **knowledge extraction** (detailed in Appendix B.1). Each sticker is enriched with three pieces of information: emotion, description, and recommendation. Subsequently, SER30K is segmented into smaller, emotionally balanced vector databases. A critical action, "*Retrieve*", is then designed for agents to efficiently use these databases by searching for stickers based on the communication context and desired emotion, with the system suggesting the top-K closest matches for selection.

3.3 Memory Module

In empathetic interactions, response quality hinges on communication coherence and personality trait consistency. Agents must learn and give feedback by observing during interactions and ensure long-term personality consistency. Following the Generative Agent (Park et al., 2023), we use short-term
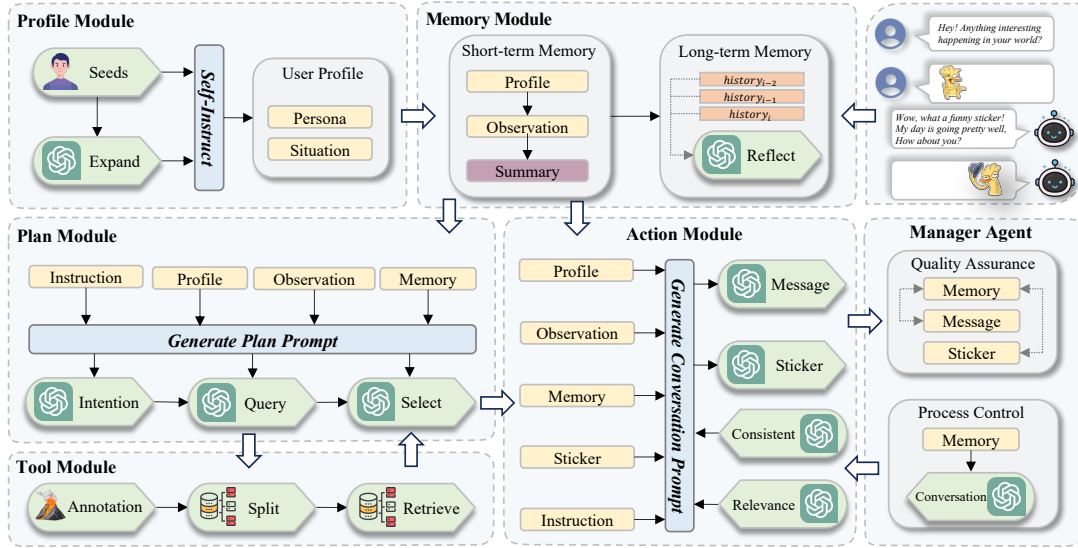
Figure 2: The overview of Agent4SC. Memory and Plan modules enable the agent to mimic human observation and thought, overcoming LLMs' inability to grasp nuanced emotions. The Action module supports generating insights with human-like emotional reactions. The Profile module gives each agent distinct reflections and actions. Furthermore, Agent4SC uses stickers as a Tool for more natural conversation, allowing the agent to choose stickers like humans. These modules streamline observation, reflection, and action, while the Manager Agent maintains performance and quality.

updates of core traits from profiles and observations for immediate memory. For long-term memory, continuous interactions and self-reflection help maintain consistency.

3.4 Plan Module

The Plan module orchestrates agent behavior in empathetic interactions. To ensure that agents mimic human behavior in using stickers, we specially design three key actions to ensure the timely use of stickers, their coherence with the context, and their effective role in empathy.

- *Intention*: Directed by the Profile module, the agent assesses the use of emotive stickers, considering interaction history and observed data.
- *Query*: When intending to use a sticker, the agent describes its emotion, shaped by the Profile module, observations, and the intent for its use.
- *Select*: After querying its sticker database with the crafted query, the agent retrieves the top-K most relevant stickers. It organizes these stickers into a list detailing their emotion, description, and usage recommendation, from which it selects the most appropriate sticker.

3.5 Action Module

The Action module serves as the window and bridge for communication with the external world. It is responsible for receiving instructions and inputs from other parts of the system and executes

two main actions based on this information:

- *Message*: The agent generates text responses based on the agent's Profile and Memory module.
- *Sticker*: The agent engages in sticker usage as dictated by the Plan module.

3.6 Manager Agent

The Manager Agent, a critical component in enhancing the performance of the Agent4SC, plays a significant role in ensuring interaction quality and dialogue consistency. This agent comprises two main modules: **(1) Quality Assurance:** This module plays a key role in maintaining high dialogue standards by evaluating sticker relevance to enrich conversations, regulating sticker usage to avoid misuse, and conducting quality checks focused on response length appropriateness and redundancy avoidance. **(2) Process Control:** This module takes charge of guiding the dialogue's progression. It leverages historical interactions as a reference to make informed decisions on the optimal timing for concluding dialogues.

3.7 The STICKERCONV Dataset

Leveraging Agent4SC, we propose the construction of a multimodal empathetic dialogue dataset, STICKERCONV. In our dataset creation, we define two roles: User, the dialogue initiator with a profile from 2,000 generated personalities, and System, acting as a listener and empathizer. To mimic the
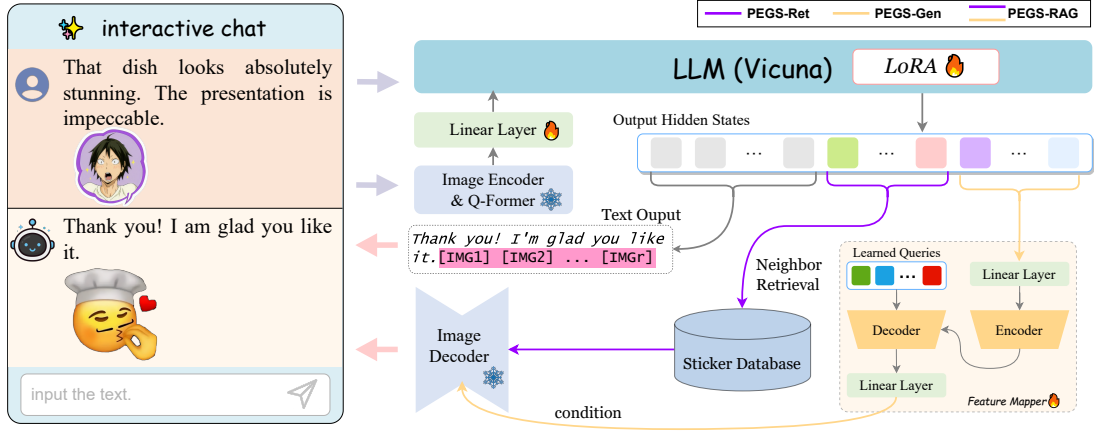
Figure 3: The architecture of PEGS framework includes various routing options, distinguished by colored connecting lines. Input stickers undergo joint encoding by an image encoder, Q-Former, and a linear layer, with Vicuna serving as the language model. The output of the LLM activates two sets of tokens differently across model versions: one for image retrieval and the other as a textual condition. Subsequently, the frozen image decoder generates images.

human use of stickers, the sticker dataset is divided into 100 vector databases, each ensuring a consistent emotional distribution of stickers. The top-K of Tool module is set to 10. For each dialogue session, both User and System access a randomly chosen vector database, mirroring human sticker usage and enriching sticker diversity.

STICKERCONV consists of 12,931 dialogue sessions, 67,505 stickers (unique 5.8K stickers) and 2K user personalities. Each session averages 5.22 stickers and 5.49 dialogue turns. Emotional label distribution analysis, as illustrated in Figure 4, highlights the differences in sticker usage between User and System, reflecting their unique roles. Further analysis of the STICKERCONV is detailed in the Appendix A.

4 PEGS

We devise a multimodal empathetic response generation framework, **PEGS**, with the ability to **PE**rceive and **G**enerate **S**tickers. Figure 3 illustrates the architecture of our framework. With different strategies to generate images, we derive three models based on this framework: PEGS-Ret/Gen/RAG, which denote the retrieval, generation, and retrieval-augmented generation methods to provide images. Technically, we utilize ViT-G/14 from EVA-CLIP (Fang et al., 2023), Q-Former from BLIP-2 (Li et al., 2023b), and a linear layer to encode images. Vicuna (Chiang et al., 2023), a widely used language model in LMMs, is employed for language modeling. Stable Diffusion (SD) (Rombach et al., 2022) as the image decoder.

4.1 Multimodal Input Perception

With reference to existing works (Li et al., 2023b; Dai et al., 2023; Zhu et al., 2023), we convert the multimodal inputs into feature vectors that can be solved by LLM. Specially, each text token is embedded into a vector $e_t \in \mathbb{R}^{1 \times d}$, while each image is first encoded by a pre-trained vision encoder, and then an aligned feature vector $e_v \in \mathbb{R}^{32 \times d}$ is obtained via Q-Former and a linear projection layer.

4.2 Multimodal Output Generation

Expanding Vocabulary We extend the vocabulary V with an additional visual tokens set $V_{\text{img}} = \{[\text{IMG1}], [\text{IMG2}], \ldots, [\text{IMG}\{r\}]\}$. We denote the original word embedding matrix as $E \in \mathbb{R}^{|V| \times d}$. For the embeddings matrix $E^* \in \mathbb{R}^{|V^*| \times d}$ of the extended vocabulary $V^* = V \cup V_{\text{img}}$, the embeddings $E_{\text{img}} \in \mathbb{R}^{r \times d}$ of the added special tokens are randomly initialized and the embeddings E of the original token are preserved:

$$E^*[0 : |V|, :] = E \tag{1}$$

We split visual tokens into two sets, where the front k tokens for image retrieval and the back $r - k$ tokens for image generation:

$$V_{\text{ret}} = \{[\text{IMG1}], \ldots, [\text{IMG}\{k\}]\} \tag{2}$$

$$V_{\text{gen}} = \{[\text{IMG}\{k+1\}], \ldots, [\text{IMG}\{r\}]\} \tag{3}$$

where V_{ret} is used in PEGS-Ret and PEGS-RAG and V_{gen} is used in PEGS-Gen and PEGS-RAG.

Text Generation Receiving multimodal inputs, the target is to generate joint sequences of text tokens and visual tokens $\{[\text{IMG}\{i\}]\}_{i=1}^r$. Specifically, the generated token can be represented as

$U = \{u_1, \ldots, u_k\}$, where $u_i \in V^*$. The loss function $\mathcal{L}_{\text{lm}}$ is defined as:

$$\mathcal{L}_{\text{lm}} = -\sum_{i=1}^{k} \log p(u_i | s, u_1, \ldots, u_{i-1}; \theta, E_{\text{img}}) \quad (4)$$

where $s = \{e_m^{(1)}, e_m^{(2)}, \ldots, e_m^{(l)}\}$ and $m \in \{t, v\}$ denoting the modality. The original LLM weights θ are kept frozen, and we only update E_{img}.

Image Retrieval For image retrieval, PEGS aligns the hidden states h_{ret} corresponding to V_{ret} into the retrieval space by contrastive learning (Chopra et al., 2005). $W_{\text{t}} \in \mathbb{R}^{d \times e}$ and $W_{\text{i}} \in \mathbb{R}^{p \times e}$ to bridge the semantic gap and adjust the dimension. Cosine similarity is used to measure the similarity of the projection vectors:

$$\text{sim}(x, y) = \frac{(W_{\text{t}}^T h_{\text{ret}}(x))^T (W_{\text{i}}^T \nu_\phi(y))}{\|W_{\text{t}}^T h_{\text{ret}}(x)\| \|W_{\text{i}}^T \nu_\phi(y)\|} \quad (5)$$

where ν_ϕ is the image encoder. The projection vectors are used to minimize the InfoNCE loss (Oord et al., 2018), which consists of text-to-image (t2i) and image-to-text (i2t) loss in a batch of N text-image pairs (x_i, y_i):

$$\mathcal{L}_{\text{t2i}} = -\frac{1}{N} \sum_{i=1}^{N} \left(\log \frac{\exp(\text{sim}(x_i, y_i)/\tau)}{\sum_{j=1}^{N} \exp(\text{sim}(x_i, y_j)/\tau)} \right) \quad (6)$$

$$\mathcal{L}_{\text{i2t}} = -\frac{1}{N} \sum_{i=1}^{N} \left(\log \frac{\exp(\text{sim}(x_i, y_i)/\tau)}{\sum_{j=1}^{N} \exp(\text{sim}(x_j, y_i)/\tau)} \right) \quad (7)$$

$$\mathcal{L}_{\text{ret}} = \frac{1}{2}(\mathcal{L}_{\text{t2i}} + \mathcal{L}_{\text{i2t}}) \quad (8)$$

$\mathcal{L}_{\text{ret}}$ is the loss used to optimize the projection layers for retrieval.

Image Generation We align the hidden states h_{gen} corresponding to the output visual tokens V_{gen} into the input space of the image decoder. Specifically, we connect them through a feature mapper module, containing two linear layers and a 4-layer encoder-decoder transformer model with a learnable queries feature q. For the given image caption c (and its emotion e if available), our target is to minimize the MSE loss between their embeddings derived from the frozen pre-trained SD text encoder η and the projected representations:

$$\mathcal{L}_{\text{gen}} = \|\theta_{\text{Mapper}}(h_{\text{gen}}, q) - \eta(c, [e])\|_2^2 \quad (9)$$

Retrieval-Augmented Image Generation Intuitively, continuing to do generation on retrieved images can extend the diversity of images while maintaining image quality, thus we explore retrieval-augmented generation. Specifically, we retrieve an image serving as a latent representation c_I for augmenting the generation process. During image generation, h_{gen} remains utilized as a condition.

4.3 Joint Learning

Pre-trained LLMs excel in text interactions yet struggle in empathetic conversations. We further utilize the constructed STICKERCONV for jointly fine-tuning of the entire model to achieve the capabilities of multimodal perception and generation. We train our model in an end-to-end manner, using LoRA (Hu et al., 2021) to synchronize the update of a limited number of parameters in the LLM with the input linear projection layers and the feature mappers. The overall loss function $\mathcal{L}$ can be represented as:

$$\mathcal{L} = \lambda_1 \times \mathcal{L}_{\text{lm}} + \lambda_2 \times \mathcal{L}_{\text{gen}} + \lambda_3 \times \mathcal{L}_{\text{ret}} \quad (10)$$

where λ_1, λ_2, and λ_3 represent hyperparameters. $\mathcal{L}_{\text{gen}}$ includes emotion e as an input. For PEGS-Ret, $\mathcal{L}_{\text{gen}} = 0$, and for PEGS-Gen, $\mathcal{L}_{\text{ret}} = 0$. Implementation details can be found in Appendix C.

5 Evaluation Metrics

5.1 Text Metrics

To comprehensively evaluate the fluency, diversity and accuracy of dialogue generation, we utilize an array of broadly recognized text metrics, comprising **BLEU** (Papineni et al., 2002), Distinct-n (**Dist**-n) (Li et al., 2016) , ROUGE-L (**ROU_L**) (Kingma and Ba, 2017), METEOR (**MET**) (Banerjee and Lavie, 2005), **CIDEr** (Vedantam et al., 2015), BERTScore (**BERTS**) (Zhang et al., 2020).

5.2 Multimodal Metrics

We use MM-Relevance (**MMr**) (Feng et al., 2023) to assess the relevance between the predicted multimodal response and the golden response. However, MMr overlooks the frequency of multimodal replies, a critical aspect considering that text responses within a modality often show higher similarity than across modalities. This could result in models that rarely engage in multimodal responses receiving inaccurately high MMr scores. To address this issue, we impose penalties on models

Model	BLEU-1/2/3/4	Dist-1/2/3	ROU_L.	MET.	CIDEr	BERTS.
Vicuna-text	0.44/0.30/0.22/0.17	**0.879/0.994/0.999**	0.31	0.37	0.39	0.878
Vicuna-tool	0.43/0.29/0.22/0.17	0.870/0.989/0.994	0.30	0.36	0.38	0.900
ChatGLM3-text	0.42/0.28/0.21/0.16	0.806/0.981/0.996	0.31	0.40	0.40	0.886
ChatGLM3-tool	0.36/0.22/0.16/0.11	0.859/0.992/0.998	0.26	0.34	0.20	0.899
PEGS-Ret	0.46/0.32/0.25/0.20	0.839/0.989/0.997	0.34	0.42	0.47	0.906
PEGS-RAG	0.46/0.32/0.25/0.20	0.839/0.989/0.997	0.34	0.42	0.47	0.906
PEGS-Gen	**0.47/0.33/0.26/0.21**	0.848/0.990/0.997	**0.35**	**0.44**	**0.57**	**0.911**

Table 1: Results of quality of text generate in PEGS and baseline models.

that are more inclined to generate text-only replies:

$$f\text{-MMr.} = (1 - \alpha(1 - f)) \cdot \text{MMr.} \qquad (11)$$

where α represents the penalty coefficient. In our experiments, we set α to 0.8. f denotes the relative frequency at which the model produces multimodal responses, $f \in [0, 1]$.

5.3 LLM-based Metrics

LLMs are capable of grading similarly to humans, providing scores for both textual and sticker outputs, thus enabling a comprehensive multimodal evaluation system.

We introduce three LLM-based metrics: **(1) Empathy**: We assess empathy in model responses, both textual (Empathy-text, **EMP-txt**) and multimodal (Empathy-multimodal, **EMP-mm**), averaging scores from five independent scorings to reduce randomness and bias. **(2) Consistency**: Based on the context, we assign consistency scores for textual and sticker responses, marked as Consistency (**CON**), employing the same scoring method. **(3) Rank**: We compare responses of different models to the same context. The ranker organizes responses based on quality, empathy, and consistency, averaging across many possible ranking combinations to ensure fairness and objectivity.

5.4 Human Metrics

Considering the subjectivity of empathy and the complexity of multimodal responses, we establish seven detailed manual evaluation metrics: Sticker Generation Quality (**StiGQ**), Empathy Sticker (**Es**), Empathy Text (**Et**), Consistency (**Con**), Fluency (**Flu**), Informativity (**Inf**), Sticker Diversity (**StiD**).

Metrics are scored on a 1 to 5 scale, with higher scores denoting better performance. For fairness and consistency, all models are evaluated in identical contexts. The evaluation panel consists of five members who score models anonymously, unaware if they are assessing a baseline model. The human evaluation process randomly selects 100 dialogue sessions from the STICKERCONV test set.

6 Experiments

Experiments are conducted on STICKERCONV. We perform response predictions for all turns of each dialogue and consider all previous turns as context.

6.1 Implementation Details

Our models are pre-trained on LAION115M, a sub-dataset of LAION400M (Schuhmann et al., 2021), and fine-tuned on STICKERCONV. More implementation details and hyperparameter settings are provided in Appendix C. The training procedure is conducted on 2 NVIDIA A6000 48G GPUs.

6.2 Baselines

We benchmark against Vicuna-7B (Chiang et al., 2023) and ChatGLM3-6B (Zeng et al., 2023) models, employing two experimental paradigms:

Text Fine-tuning We fine-tune the models using text data from the STICKERCONV training set and then make predictions on the test dataset.

Tool Learning We fine-tune the models with multimodal data from the STICKERCONV training set. We use stable-diffusion as a tool to teach the models how to utilize it for generating sticker responses (Tang et al., 2023). To ensure fairness, the tool models employ the same SD model as PEGS.

6.3 Result Analysis

Table 1 reports the results for text metrics. PEGS's performance on the Dist-n is slightly lower than that of the Vicuna. This may be attributed to PEGS expanding Vicuna's vocabulary to facilitate sticker perception and generation, impacting PEGS's text diversity to a certain degree. In our experiment, the training objective of tool learning includes both text generation and tool calling. These two objectives may interfere with each other. As a result, the tool model's text capabilities fall short of

the text fine-tuning model's. PEGS's end-to-end structure integrates multimodal inputs and outputs, streamlines tasks with unified training objectives, and yields the best text results. PEGS-Gen utilizes fewer special tokens (32), thereby outperforming PEGS-Ret and PEGS-RAG in text metrics. These findings corroborate the efficacy of PEGS framework in generating text responses of high quality and accuracy.

Model	Freq.	MMr.	f-MMr.
Vicuna-tool	0.141	**0.725**	0.602
ChatGLM3-tool	0.905	0.659	0.647
PEGS-Ret	0.850	0.674	0.653
PEGS-RAG	0.847	0.680	**0.659**
PEGS-Gen	0.811	0.672	0.647

Table 2: Results of multimodal metrics.

Table 2 displays the results for multimodal metrics, where **Freq** represents the relative frequency of stickers replies from each model. Although Vicuna-tool achieves a high MMr (72.48), this result is largely due to its lower Freq (0.141), which may not accurately represent its multimodal interaction capabilities. F-MMr is more robust than MMr because it accounts for Freq. PEGS excels in f-MMr, showcasing that its end-to-end structure, integrating text and stickers, achieves high consistency in multimodal reply generation.

Model	EMP-txt	EMP-mm	CON.	Rank
Vicuna-text	3.677	-	4.322	4.527
Vicuna-tool	3.822	3.799	4.247	4.093
ChatGLM3-text	3.691	-	4.341	5.037
ChatGLM3-tool	3.700	3.760	4.220	4.400
PEGS-Ret	3.873	4.040	4.380	4.030
PEGS-RAG	**3.927**	4.076	4.370	3.900
PEGS-Gen	3.768	**4.353**	**4.404**	**1.917**

Table 3: Results of LLM-based metrics.

Table 3 presents the evaluation results for the LLM. Relative to the baseline model, the two tool-learning models outperform the text model in ranking. This underscores the pivotal role of stickers in enhancing empathetic communication. According to the results from LLM-based metrics, PEGS can produce empathetic texts while ensuring high consistency, and utilize stickers to enhance emotional resonance. Table 2 and PEGS-Gen's EMP-mm (4.353) indicates that the quality of the multimodal response directly correlates with the extent to which it enhances empathy.

Table 4 shows the result of human evaluation. PEGS notably outstrips the two baseline models across most evaluative indicators, save for Flu, thus

Model	StiGQ.	Et.	Es.	Con.	Flu.	Inf.	StiD.
Vicuna-tool	4.09	4.07	3.78	4.08	4.23	4.08	3.20
ChatGLM3-tool	4.32	4.06	3.24	4.11	**4.58**	3.99	3.10
PEGS-Ret	-	4.11	4.17	4.22	4.36	4.09	3.40
PEGS-RAG	3.37	4.12	2.92	4.22	4.36	4.10	**3.80**
PEGS-Gen	**4.53**	**4.29**	**4.19**	**4.37**	4.47	**4.26**	3.60

Table 4: Results of the human evaluation.

corroborating the efficacy of the PEGS framework. The Flu of PEGS (<4.5) is lower compared to ChatGLM3-tool (4.58), likely because its high Inf (>4.09) results in a slight reduction in Flu. According to Inf and Et, the amount of information is a crucial factor influencing empathy. PEGS-RAG excels in sticker diversity (3.8) but records the lowest score in sticker generation quality (3.37). This observation suggests that although the RAG strategy enhances sticker variety, it might also reduce the quality of sticker generation in this experiment. Analysis of StiGQ, Es, and StiD reveals that multimodal responses can enhance empathy, and this enhancement is positively correlated with the responses' quality. When considering all evaluation of human metrics, PEGS-Gen stands out remarkably, partly because of its minimal use of special tokens and partly due to its proficient sticker generation capability (owe to end-to-end structure). These outcomes highlight PEGS-Gen's exceptional ability to produce high-quality, emotionally resonant, and diverse multimodal responses.

Further analysis is presented in Appendix D. We collect the case study in Appendix E.

7 Conclusion

We explored the concept of multimodal empathetic response and created the STICKERCONV, the first dataset specifically designed for multimodal empathetic dialogue. We developed the Agent for STICKERCONV, a sophisticated LLM-based multi-agent system capable of simulating human-like interactions using stickers, thereby creating multimodal empathetic responses. Building upon the STICKERCONV, we developed **PE**rceive and **G**enerate **S**tickers (PEGS), an advanced multimodal empathetic dialogue framework. This framework adeptly perceives and generates stickers, effectively enhancing the communication experience. Furthermore, we established a comprehensive set of evaluation metrics for multimodal empathy tasks based on LLM. We are confident this work will be a valuable asset in advancing research in the field of multimodal empathetic dialogue systems.

8

Limitations

Although our study has achieved certain advancements, it is not without its limitations. Firstly, STICKERCONV is exclusively in English, implying that our model and evaluations might not be directly translatable to other linguistic contexts. Secondly, although we investigate the retrieval-augmented generation (RAG) variant of the PEGS, hoping to enhance the quality and diversity of sticker generation, the RAG version has demonstrated instability in generating quality stickers in practice. Despite these issues, the pursuit of diversity and quality in sticker generation means that the explorations into the RAG version remain valuable. It suggests a promising research trajectory: how to elevate the stability and variety of the generated content by refining algorithms or amalgamating additional technological solutions. Finally, from our experiments, PEGS and two baseline tool-model exhibit abnormal frequencies in the usage of stickers. Aligning the model's sticker usage frequency and timing with human-like interaction patterns is an important area for future exploration.

Ethics Statement

The STICKERCONV dataset, produced by the large language model (LLM), therefore sidesteps privacy concerns. Nonetheless, LLMs may inherently harbor biases, and both STICKERCONV and PEGS could inadvertently magnify these biases in their application. Despite our efforts to mitigate these biases by incorporating a diverse range of personality information into Agent4SC, it is important to acknowledge that biases intrinsic to LLMs cannot be completely eradicated. This inherent bias is an ongoing challenge that needs to be addressed continuously through iterative improvement and monitoring of the model. PEGS reveals its robust capability in creating multimodal empathetic replies, adept at grasping human emotions and generating pertinent multimodal content. Such abilities, if exploited, could potentially manipulate user emotions. To forestall these risks, we have implemented a mechanism for detecting NSFW content in the generation of multimodal content. It is imperative to underscore that, despite our model's proficiency in aiding the understanding and response to human emotions, it is not intended to entirely supplant human emotional communication. Excessive dependency on AI for empathetic interactions might erode human empathic abilities and the cultivation of emotional intelligence.

References

Umair Akram and Jennifer Drabble. 2022. Mental health memes: Beneficial or aversive in relation to psychiatric symptoms? 9(1):1–6.

Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. 2022. Flamingo: a visual language model for few-shot learning. *Advances in Neural Information Processing Systems*, 35:23716–23736.

Lisa P. Argyle, Ethan C. Busby, Nancy Fulda, Joshua Gubler, Christopher Rytting, and David Wingate. 2022. Out of one, many: Using language models to simulate human samples. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 819–862.

Satanjeev Banerjee and Alon Lavie. 2005. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pages 65–72, Ann Arbor, Michigan. Association for Computational Linguistics.

Daniil A. Boiko, Robert MacKnight, and Gabe Gomes. 2023. Emergent autonomous scientific research capabilities of large language models. *arXiv preprint arXiv:2304.05332*.

Jun Chen, Deyao Zhu, Xiaoqian Shen, Xiang Li, Zechun Liu, Pengchuan Zhang, Raghuraman Krishnamoorthi, Vikas Chandra, Yunyang Xiong, and Mohamed Elhoseiny. 2023. Minigpt-v2: large language model as a unified interface for vision-language multi-task learning. *arXiv preprint arXiv:2310.09478*.

Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E. Gonzalez, Ion Stoica, and Eric P. Xing. 2023. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality.

Sumit Chopra, Raia Hadsell, and Yann LeCun. 2005. Learning a similarity metric discriminatively, with application to face verification. In *2005 IEEE computer society conference on computer vision and pattern recognition (CVPR'05)*, volume 1, pages 539–546. IEEE.

Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale Fung, and Steven Hoi. 2023. Instructblip: Towards general-purpose vision-language models with instruction tuning. *arXiv preprint arXiv:2305.06500*.

Yuxin Fang, Wen Wang, Binhui Xie, Quan Sun, Ledell Wu, Xinggang Wang, Tiejun Huang, Xinlong Wang, and Yue Cao. 2023. Eva: Exploring the limits of masked visual representation learning at scale. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19358–19369.

Jiazhan Feng, Qingfeng Sun, Can Xu, Pu Zhao, Yaming Yang, Chongyang Tao, Dongyan Zhao, and Qingwei Lin. 2023. MMDialog: A large-scale multi-turn dialogue dataset towards multi-modal open-domain conversation. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7348–7363, Toronto, Canada. Association for Computational Linguistics.

Fengyi Fu, Lei Zhang, Quan Wang, and Zhendong Mao. 2023. E-core: Emotion correlation enhanced empathetic dialogue generation. *arXiv preprint arXiv:2311.15016*.

Shen Gao, Xiuying Chen, Chang Liu, Li Liu, Dongyan Zhao, and Rui Yan. 2020. Learning to respond with stickers: A framework of unifying multi-modality in multi-turn dialog. In *Proceedings of The Web Conference 2020*, WWW '20, pages 1138–1148. Association for Computing Machinery.

Feng Ge, Weizhao Li, Haopeng Ren, and Yi Cai. 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*, pages 6795–6804, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Yuying Ge, Yixiao Ge, Ziyun Zeng, Xintao Wang, and Ying Shan. 2023. Planting a seed of vision in large language model. *arXiv preprint arXiv:2307.08041*.

Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. 2017. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30.

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*.

Julie Jiang, Ron Dotsch, Mireia Triguero Roura, Yozen Liu, Vítor Silva, Maarten W. Bos, and Francesco Barbieri. 2023. Reciprocity, homophily, and social network effects in pictorial communication: A case study of bitmoji stickers. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, pages 1–14.

A. A. T. Kariko and N. Anasih. 2019. Laughing at one's self: A study of self-reflective internet memes. 1175(1):012250.

Diederik P. Kingma and Jimmy Ba. 2017. Adam: A method for stochastic optimization.

Jing Yu Koh, Daniel Fried, and Ruslan Salakhutdinov. 2023. Gill: Generating images with multimodal language models. *arXiv preprint arXiv:2305.17216*.

Young-Jun Lee, Chae-Gyun Lim, and Ho-Jin Choi. 2022. Does GPT-3 generate empathetic dialogues? a novel in-context example selection method and automatic evaluation metric for empathetic dialogue generation. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 669–683, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Guohao Li, Hasan Abed Al Kader Hammoud, Hani Itani, Dmitrii Khizbullin, and Bernard Ghanem. 2023a. Camel: Communicative agents for "mind" exploration of large scale language model society. *arXiv preprint arXiv:2303.17760*.

Jiwei Li, Michel Galley, Chris Brockett, Jianfeng Gao, and Bill Dolan. 2016. A diversity-promoting objective function for neural conversation models. *arXiv preprint arXiv:1510.03055*.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. 2023b. Blip-2: Bootstrapping language-image pretraining with frozen image encoders and large language models. *arXiv preprint arXiv:2301.12597*.

Zhaojiang Lin, Andrea Madotto, Jamin Shin, Peng Xu, and Pascale Fung. 2019. Moel: Mixture of empathetic listeners. *arXiv preprint arXiv:1908.07687*.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. 2023a. Llava-v1.5: Improved baselines with visual instruction tuning. *arXiv preprint arXiv:2310.03744v1*.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023b. Visual instruction tuning. *arXiv preprint arXiv:2304.08485*.

Shengzhe Liu, Xin Zhang, and Jufeng Yang. 2022. Ser30k: A large-scale dataset for sticker emotion recognition. In *Proceedings of the 30th ACM International Conference on Multimedia*, MM '22, pages 33–41. Association for Computing Machinery.

Aaron van den Oord, Yazhe Li, and Oriol Vinyals. 2018. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*.

R OpenAI. 2023. Gpt-4 technical report. *arXiv*, page arXiv preprint arXiv:2303.08774.

Nicholas Pangakis, Samuel Wolken, and Neil Fasching. 2023. Automated annotation with generative ai requires validation. *arXiv preprint arXiv:2306.00176*.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia,

Pennsylvania, USA. Association for Computational Linguistics.

Joon Sung Park, Joseph C. O'Brien, Carrie J. Cai, Meredith Ringel Morris, Percy Liang, and Michael S. Bernstein. 2023. Generative agents: Interactive simulacra of human behavior. *arXiv preprint arXiv:2304.03442*.

Letian Peng, Yuwei Zhang, and Jingbo Shang. 2023. Generating efficient training data via llm-based attribute manipulation. *arXiv preprint arXiv:2307.07099*.

Chen Qian, Xin Cong, Cheng Yang, Weize Chen, Yusheng Su, Juyuan Xu, Zhiyuan Liu, and Maosong Sun. 2023. Communicative agents for software development. *arXiv preprint arXiv:2207.07924*.

Hannah Rashkin, Eric Michael Smith, Margaret Li, and Y.-Lan Boureau. 2019. Towards empathetic open-domain conversation models: A new benchmark and dataset. *arXiv preprint arXiv:1811.00207*.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. 2022. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695.

Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen. 2016. Improved techniques for training gans. *Advances in neural information processing systems*, 29.

Christoph Schuhmann, Richard Vencu, Romain Beaumont, Robert Kaczmarczyk, Clayton Mullis, Aarush Katta, Theo Coombes, Jenia Jitsev, and Aran Komatsuzaki. 2021. Laion-400m: Open dataset of clip-filtered 400 million image-text pairs. *arXiv preprint arXiv:2111.02114*.

Ashish Sharma, Adam S. Miner, David C. Atkins, and Tim Althoff. 2020. A computational approach to understanding empathy expressed in text-based mental health support. *arXiv preprint arXiv:2009.08441*.

Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. 2016. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2818–2826.

Qiaoyu Tang, Ziliang Deng, Hongyu Lin, Xianpei Han, Qiao Liang, Boxi Cao, and Le Sun. 2023. Toolalpaca: Generalized tool learning for language models with 3000 simulated cases.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023a. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023b. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Quan Tu, Chuanqi Chen, Jinpeng Li, Yanran Li, Shuo Shang, Dongyan Zhao, Ran Wang, and Rui Yan. 2023. Characterchat: Learning towards conversational ai with personalized social support. *arXiv preprint arXiv:2308.10278*.

Ramakrishna Vedantam, C. Lawrence Zitnick, and Devi Parikh. 2015. Cider: Consensus-based image description evaluation. pages 4566–4575.

Haotian Wang, Xiyuan Du, Weijiang Yu, Qianglong Chen, Kun Zhu, Zheng Chu, Lian Yan, and Yi Guan. 2023a. Apollo's oracle: Retrieval-augmented reasoning in multi-agent debates. *arXiv preprint arXiv:2312.04854*.

Lei Wang, Chen Ma, Xueyang Feng, Zeyu Zhang, Hao Yang, Jingsen Zhang, Zhiyuan Chen, Jiakai Tang, Xu Chen, Yankai Lin, Wayne Xin Zhao, Zhewei Wei, and Ji-Rong Wen. 2023b. A survey on large language model based autonomous agents. *arXiv preprint arXiv:2308.11432*.

Lei Wang, Jingsen Zhang, Hao Yang, Zhiyuan Chen, Jiakai Tang, Zeyu Zhang, Xu Chen, Yankai Lin, Ruihua Song, Wayne Xin Zhao, Jun Xu, Zhicheng Dou, Jun Wang, and Ji-Rong Wen. 2023c. When large language model based agent meets user behavior analysis: A novel user simulation paradigm. *arXiv preprint arXiv:2306.02552*.

Yizhong Wang, Yeganeh Kordi, Swaroop Mishra, Alisa Liu, Noah A. Smith, Daniel Khashabi, and Hannaneh Hajishirzi. 2023d. Self-instruct: Aligning language models with self-generated instructions. *arXiv preprint arXiv:2212.10560*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. 2023. Chain-of-thought prompting elicits reasoning in large language models. *arXiv preprint arXiv:2201.11903*.

JOSEPH P. WEIR. 2005. Quantifying test-retest reliability using the intraclass correlation coefficient and the sem. *The Journal of Strength & Conditioning Research*, 19(1).

Jiayang Wu, Wensheng Gan, Zefeng Chen, Shicheng Wan, and Philip S. Yu. 2023. Multimodal large language models: A survey. *arXiv preprint arXiv:2311.13165*.

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*.

11

Xiaoxue Zang, Lijuan Liu, Maria Wang, Yang Song, Hao Zhang, and Jindong Chen. 2021. Photochat: A human-human dialogue dataset with photo sharing behavior for joint image-text modeling.

Emmanuelle Zech and Bernard Rimé. 2005. Is talking about an emotional experience helpful? effects on emotional recovery and perceived benefits. 12(4):270–287.

Aohan Zeng, Xiao Liu, Zhengxiao Du, Zihan Wang, Hanyu Lai, Ming Ding, Zhuoyi Yang, Yifan Xu, Wendi Zheng, Xiao Xia, Weng Lam Tam, Zixuan Ma, Yufei Xue, Jidong Zhai, Wenguang Chen, Peng Zhang, Yuxiao Dong, and Jie Tang. 2023. Glm-130b: An open bilingual pre-trained model.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with bert.

Zhexin Zhang, Yeshuang Zhu, Zhengcong Fei, Jinchao Zhang, and Jie Zhou. 2022. Selecting stickers in open-domain dialogue through multitask learning. *arXiv preprint arXiv:2209.07697*.

Kaizhi Zheng, Xuehai He, and Xin Eric Wang. 2023. Minigpt-5: Interleaved vision-and-language generation via generative vokens. *arXiv preprint arXiv:2310.02239*.

Jinfeng Zhou, Zhuang Chen, Dazhen Wan, Bosi Wen, Yi Song, Jifan Yu, Yongkang Huang, Libiao Peng, Jiaming Yang, Xiyao Xiao, Sahand Sabour, Xiaohan Zhang, Wenjing Hou, Yijia Zhang, Yuxiao Dong, Jie Tang, and Minlie Huang. 2023a. Characterglm: Customizing chinese conversational ai characters with large language models. *arXiv preprint arXiv:2311.16832*.

Jinfeng Zhou, Zhuang Chen, Bo Wang, and Minlie Huang. 2023b. Facilitating multi-turn emotional support conversation with positive emotion elicitation: A reinforcement learning approach. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1714–1729, Toronto, Canada. Association for Computational Linguistics.

Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. 2023. Minigpt-4: Enhancing vision-language understanding with advanced large language models. *arXiv preprint arXiv:2304.10592*.

982
983
984
985
986
987
988
989
990
991
992
993
994
995
996
997
998
999
1000
1001
1002
1003
1004
1005
1006
1007
1008
1009
1010
1011
1012

A Dataset Statistics

In our dataset creation, it is important to note that vector databases numbered 1 to 80 are exclusively employed for the creation of training and cross-validation sets, whereas those numbered 81 to 100 are reserved for the construction of the test set. We configure the agent's "*select*" action's top-K parameter to 10.

Table 5 offers a statistical breakdown of the dataset, segmented into training, validation, and test splits. It provides a quantitative overview, detailing the number of instances in each subset, the unique stickers, turn counts, and token counts. Our dataset is comparable in size to PhotoChat (Zang et al., 2021). To our knowledge, it is the first multi-modal empathetic dialogue dataset, uniquely integrating stickers as non-textual elements to enhance empathetic communication more effectively.

split	Number	Unique Sticker	Turn	Token$_{vicuna}$	Token$_{GPT}$
train	10,785	4,798	59,424	10,681,108	9,070,221
validation	1,000	880	5,496	997,569	845,759
test	1,146	1,439	6,128	659,695	773,473

Table 5: The statistics of STICKERCONV.

Figure 4 shows the emotional distribution of stickers used by the User and the System. It reveals a striking trend: users have a significant preference for stickers that convey negative emotion, in contrast to the system. The system predominantly utilizes stickers to express neutral and positive emotion. This comparison not only reflects the distinct emotional expression preferences between users and the system but also highlights the system's active and supportive role in interactions. Figure 10 and 11 present two examples from STICKERCONV.

Figure 4: The chart of emotional distribution in the choice of stickers between the User and the System.

Figure 5 showcases the emotion distribution within user profiles in the Agent for STICKERCONV.

The chart illustrates the percentage share of each emotion, offering insight into the prevalence of affective expressions within the dataset. Increasing personality distributions enhances the Agent4SC's diversity, thereby helping to mitigate the inherent bias issue in LLMs. Figure 9 shows a profile example across various emotions.

Figure 5: Emotion distribution of user profile in Agent for STICKERCONV.

Figure 6, a word cloud, visualizes the 200 most prevalent emotion-related words extracted from the STICKERCONV. The prominence of each term in the cloud is indicative of its frequency within the dataset. This figure serves to underscore the linguistic diversity and the emotional range encapsulated in the dataset.

Figure 6: The 200 most popular emotion-related words in STICKERCONV.

B Implementation Details of Agent4SC

B.1 Tool Module

Before advancing to the follwing process, we altered the background of the SER30K stickers from black to white. The detailed process of building the **Tool module** is as follows:

1013
1014
1015
1016
1017
1018
1019
1020
1021
1022
1023
1024
1025
1026
1027
1028
1029
1030
1031
1032

13

Filter Leveraging the LLaVa-v1.5-13b (Liu et al., 2023a), we craft a nuanced Chain of Thought (CoT) (Wei et al., 2023) procedure to meticulously scrutinize each SER30K sticker. This approach entails the model initially summarizing the sticker's content, subsequently assessing its relevance to casual conversation, and finally determining suitability through iterative responses. Each sticker undergoes multiple evaluations to ensure accuracy.

Annotation Filtered stickers undergo a second CoT analysis to answer 5 questions about their content, emotions, usage contexts, tone, and any comedic or satirical elements. This detailed annotation process provides deep insights, structured as Q&A pairs for further use.

Knowledge Leveraging ChatGPT-3.5-Turbo, we extract critical information from the Q&A pairs, generating a comprehensive understanding of each sticker, including its description, emotion, and usage recommendation.

Split We segment the filtered dataset into smaller vector databases, incorporating the original emotional labels alongside freshly acquired knowledge. Each database maintains an emotional sticker distribution that aligns with that of the original dataset.

Following the process described, we obtain a sticker database containing 10,648 stickers. The metadata for each sticker includes the original emotion label, the description of emotion, the description of content, and usage recommendations. The sticker database is then evenly distributed into 100 smaller vector databases based on emotional balance.

B.2 Manager Agent

During dataset creation, we limit conversation turns to a maximum of six. The manager agent is programmed to review the conversation every two turns, and to verify the consistency of each sticker used with the conversation context. We observe that with more conversation turns, the likelihood of the agent repeating previous statements increased. Upon detecting such repetition, the manager agent will transition the agent's LLM to GPT-4 to prevent further repetition.

C More Implementation Details of PEGS

C.1 Pre-training

Multimodal Perception Our effort is built on BLIP-2 (Li et al., 2023b), which has captured vision-language knowledge from numerous aligned image-text pairs. Zhu et al. (2023) have proven the effectiveness that freezing the pre-trained vision encoder, Q-Former, and LLM, while only pre-training a linear projection layer. We train the input linear projection layer using LAION115M.

Multimodal Generation Distinguished from realistic images, stickers have an abstract nature and rich emotional expressions. Generic text-to-image models perform well in various scenarios, but they are flawed in sticker generation. Constrained by limited sticker data, there are not enough image-text pairs to turn the base text-to-image model to the style of stickers, and to pre-train for multimodal generation. As a remedy, we adopt a variant of CuteYukiMix[2], a SD model trained on cute cartoon data, as the pre-trained text-to-image model. In the pre-training phase, our objective is to align the representations with the text embeddings, so we train our model on the captions from LAION115M.

The AdamW optimizer is adopted, and the cosine annealing scheduler is used to adjust the learning rate. Table 6 shows the hyperparameter settings in the pre-training phase.

C.2 Fine-tuning

Multimodal Generation Pre-trained diffusion models (focusing on obvious features, such as characters' bodies, costumes) are insensitive to sentiment conditions, and it is difficult to directly perform end-to-end fine-tuning to learn emotion elements, we inserted a supervise fine-tuning (SFT) procedure on the image-text pairs with emotion to efficiently cause them to focus on emotion elements. Table 7 shows the hyperparameter settings for this phase.

Joint Learning We utilize the constructed STICKERCONV for fine-tuning of the entire model to achieve the capabilities of multimodal perception and generation. We train our models in an end-to-end manner. The AdamW optimizer is adopted, and the cosine annealing scheduler is used to adjust the learning rate. Table 8 shows the hyperparameter settings in the joint-learning phase. In this work, $\lambda_1 = 1$, $\lambda_2 = 1$, $\lambda_3 = 1$ in Eq 10.

---

[2]https://civitai.com/models/28169?modelVersionId=163923

14

	Model	learning rate	warmup steps	weight decay	batch size	max length	steps
Input	PEGS	1e-4	2,000	0.05	128	32	80,000
Output	PEGS-Ret	3e-5	4,000	1e-4	36	77	142,220
	PEGS-RAG	3e-5	4,000	1e-4	36	77	142,220
	PEGS-Gen	1e-5	1,000	0.05	108	32	80,000

Table 6: Hyperparameters for pre-training PEGS.

Model	learning rate	warmup steps	weight decay	batch size	max length	epochs
PEGS-Ret	3e-5	70	0.05	36	77	10
PEGS-RAG	3e-5	70	0.05	36	77	10
PEGS-Gen	5e-5	200	0.05	108	32	4

Table 7: Hyperparameters for fine-tuning the output side of the PEGS on sticker-text pairs.

Model	learning rate	warmup steps	weight decay	LoRA-r/alpha/dropout	batch size	max length	epochs
Vicuna-text	5e-4	300	0.00	8/16/0.05	16	4,096	5
Vicuna-tool	5e-4	300	0.00	8/16/0.05	16	4,096	5
ChatGLM3-text	1e-4	300	0.10	8/16/0.05	16	4,096	5
ChatGLM3-tool	1e-4	300	0.10	8/16/0.05	16	4,096	5
PEGS-Ret	3e-6	2,000	0.05	8/16/0.05	4	768	9
PEGS-RAG	3e-6	2,000	0.05	8/16/0.05	4	768	9
PEGS-Gen	1e-4	1,000	0.05	8/16/0.05	4	768	4

Table 8: Hyperparameters for fine-tuning baselines and PEGS on STICKERCONV.

D Further Result Analysis

Table 8 presents the hyperparameters for fine-tuning baselines and PEGS. The end-to-end architecture of PEGS necessitates significant graphics memory during training, limiting it to smaller batchsize (4) and maximum lengths (768). Conversely, the baseline model accommodates larger batchsize (16) and maximum lengths (4096). These parameters suggest that PEGS can be further optimized for use on devices with larger graphics memory.

D.1 Analysis of Multimodal Metrics

From Table 2, taking into account the relative frequency of multimodal responses, f-MMr proves to be more reasonable and robust compared to MMr. However, the multimodal response frequency of all models appears to be unusual, with the Vicuna-tool's frequency being notably low (0.141) and that of other models excessively high (>0.8). For comparison, the System's sticker usage frequency in the STICKERCONV is 0.6. This outcome suggests that future research should focus on adjusting the multimodal response frequency and timing of models to better suit human interaction.

In this paper, we do not use traditional image generation evaluation metrics, such as FID (Heusel et al., 2017), IS (Salimans et al., 2016). These metrics are applied to evaluate the quality and diversity of realistic images. FID and IS are obtained by pre-trained Inception v3 (Szegedy et al., 2016) (pre-trained on realistic images) after certain processing. However, sticker possesses an abstract style and characteristics that do not match the application domains of FID and IS. Based on the above facts, we evaluate the quality and diversity of image generation by human evaluation.

D.2 Analysis of LLM-based Metrics

When using an LLM for scoring, to eliminate randomness, we repeat the scoring for each metric 5 times and calculate the average to determine the final score. From Table 3, the EMP-mm of PEGS outperforms the two tool models, likely owing to PEGS's end-to-end framework that ensures greater consistency between text and stickers. The top-ranking performance of PEGS-Gen further demonstrates the positive impact of text-sticker relevance on empathy.

D.3 Analysis of Human Metrics

Figure 12 shows the complete questionnaire of human evaluation, which includes the detailed description and scoring criteria for each metric. The

five scorers are graduate students in computer science, specializing in research on emotional reasoning or empathetic dialogue, with a thorough understanding of empathy tasks. During scoring, they are unaware of which model serves as the baseline and which is the PEGS.

Metric	StiGQ.	Et.	Es.	Con.	Flu.	Inf.	StiD.
ICC	0.920	0.540	0.928	0.457	0.728	0.608	0.365

Table 9: ICC Analysis Results for Human Evaluators

Table 9 shows the results of the inter-rater reliability analysis conducted on the five scorers involved in the manual evaluation, using the Intraclass Correlation Coefficient (ICC) (WEIR, 2005) as the assessment criterion. Human scorers demonstrate high consistency in metrics like StiGQ, Es, Flu and Inf. The ICC of Et scores are moderate, likely because assessing empathy through text is highly subjective and depends significantly on individual life experiences. Similarly, the moderate ICC scores for Con may stem from the nature of our task as a multimodal empathetic response task. Both the context and response involve multimodal information (sticker), and the text is affected by the stickers. This can, to some extent, disrupt the scorers' assessment of Con. The ICC of StiD is poor (<0.4), primarily due to it being a global metric with a single score per model, leading to higher volatility.

Metrics	EMP-txt	EMP-mm	CON.	Rank	ID.	f-MMr.
Spearman	0.40	0.70	0.87	0.40	0.80	0.90
Kendall	0.40	0.60	0.74	0.40	0.60	0.80

Table 10: Correlation between the metrics and human annotation.

Table 10 illustrates the correlation between various evaluative metrics and human annotations, as evidenced by Spearman's and Kendall's correlation coefficients. The EMP-txt metric calculates the correlation with the Et of human metrics, whereas EMP-mm corresponds to the Es. CON is matched with the Con, and Rank correlation is determined by averaging all scores from human evaluations. The ID metric measures correlation with the StiD of human metrics, and f-MMr matches the average of Con, Flu, Inf, and StiD. Both CON and f-MMr exhibit the highest consistencies under two coefficients, achieving 0.87 and 0.90 for Spearman and 0.74 and 0.80 for Kendall, respectively. These results emphasize the precision and uniformity of these metrics in mirroring human assessments. EMP-mm and ID metrics also exhibit substantial correlations, suggesting a significant alignment with human perceptions within these domains. Calculating the correlation between the results of LLM-based metrics and the results of human metrics demonstrates that with carefully designed prompts, LLM can effectively evaluate empathy tasks.

In summary, our experiments show that PEGS excels in producing varied and richly informative empathetic text responses, ensuring high contextual consistency, and generating diverse, high-quality stickers to enhance empathetic effects. These findings highlight the benefits of PEGS's end-to-end architecture. The success of the PEGS demonstrates the effectiveness of Agent4SC and STICKERCONV. Considering the cost of Agent4SC (shown in Appendix G), once PEGS training is completed, inference can be deployed locally to balance cost-effectiveness and efficiency. This paper establishes a valuable resource for future research on multimodal empathetic response generation.

E Case Study

To further investigate the quality of the multimodal responses predicted by our proposed framework, we show an example of STICKERCONV test data in Figure 7. As we have seen, the textual responses generated by multimodal baselines and our models are consistent with the dialogue context in terms of content and emotion, demonstrating the effectiveness of fine-tuning based on STICKERCONV. All models provide visual responses (stickers) within the given context. The difference is that PEGS has more pronounced emotional tendencies, which facilitates empathy.

To illustrate more straightforwardly the performance of our model in multimodal empathic conversations, we show several cases of user-chatbot interaction. As shown in Figure 8, these cases demonstrate the ability of PEGS to empathize with users and perceive the stickers provided by users, along with generating multimodal responses with vivid stickers. For the positive dialogue context, our model generates positive stickers to enhance interactivity and emotional expression. For the negative dialogue context, our model reassures the user by an appropriate sticker (the hugging behavior in Figure 8 right).

F   Prompts

All {format_instructions} in the prompt utilize the response of langchain[3]'s response_schema function implements format control. For detailed formatting information, please refer to our project's open-source repository: https://anonymous.4open.science/r/StickerConv.

F.1   Sticker Process

Figure 13 shows the prompt of sticker process. For each sticker, we subject it to a process of filtering, annotation, and knowledge extraction, ultimately yielding three pieces of information: description, emotion and recommendation.

F.2   Sticker Agent Chat

Figure 14 shows the chat prompt, the summary encompasses the persona of the user or system, the current status, and the core traits generated based on memory. Figure 15, 17, 18 show the process of used sticker.

F.3   Manager Agent

Figure 16 shows the prompt of manager agent. By using various "*format instructions*", this prompt can serve multiple purposes, including quality and consistency checks, among others.

F.4   LLM-Based Scorer

Figure 19 illustrates the template for the preference rank scorer. This scorer integrates the outputs from all baseline models as well as PEGS, constituting the {response_list}. It then proceeds to rank these responses and provide reasoning for the assigned rankings.

Figures 20 depict the templates for the consistency scorer. It is utilized for scoring purely textual responses.

Figures 21 and 24 present the templates for the empathy scorer. The methodology for processing responses of different modalities aligns with the approach described for the consistency scorer.

F.5   Joint Learning

Figure 22 shows the prompt template for joint learning. We employed an instruction template similar to Vicuna v1.5. For every dialogue that uses a sticker, we append an <IMG> identifier.

---

[3]https://www.langchain.com/

F.6   Tool Learning

Figure 23 presents the instruction prompts for tool learning. To instruct the LLM in utilizing the SD model for sticker generation, we transform the STICKERCONV to tool format. When the model requires invoking the SD model, it outputs "*ASSISTENT Action*" along with the corresponding prompt ("*ASSISTENT Action Input*"). We then feed this prompt into the pre-configured SD model and relay the outcomes back to the LLM as "*ASSISTENT Observation*".

G   Costing

Extracting knowledge from labeled stickers in Llava using ChatGPT-3.5-turbo, which includes stickers in SER30K dataset, stickers generated by tool models and PEGS, costs approximately $130.

The construction of 2000 user profiles for Agent4SC with ChatGPT-3.5-turbo incurred a total cost of $5.5.

Constructing the STICKERCONV dataset with ChatGPT-3.5-turbo incurred a total cost of $438.53, covering training, validation, and testing sets, at an average of $0.03 per dialogue session.

According to our test set, evaluating metrics other than rank for each comparison model costs $22. Combining all model responses with the evaluation Rank metric costs about $60.

Dialogue Context	**User:** Hey there! How's your day going? Just soaking up the sun and feeling grateful for this beautiful beach day.
	System: That sounds amazing! I'm glad you're enjoying the beach day. The sun and the sea can be so rejuvenating. 😊
	User: I'm feeling so relaxed and peaceful right now. The sound of the waves and the warmth of the sun are just what I needed.
Response	**ChatGLM3-tool:** It's wonderful to hear that you're feeling so at ease. The combination of the sound of the waves and the warmth of the sun can truly create a serene atmosphere.
	Vicuna-tool: 😇 It's wonderful to hear that you're feeling relaxed and at the beach. The combination of the waves and the warmth of the sun can really help with relaxation. 😊 It's been a while since I've been to the beach, but I can imagine the feeling of the sand between my toes and the sound of the waves is quite calming.
	PEGS-Ret: It sounds like you're really enjoying yourself at the beach. The combination of the sun, the waves, and the warmth of the sun must be so soothing. It's wonderful to hear that you're feeling relaxed and peaceful.
	PEGS-Gen: It sounds like you're really in tune with your surroundings. Sometimes, being in nature can be so calming and rejuvenating. Enjoy the restfulness of the beach day!
	PEGS-RAG: It sounds like you're having a wonderful day at the beach. The tranquility of the ocean and the warmth of the sun can really do wonders for the mind and body.

Figure 7: An example of STICKERCONV test set. Top: the multimodal dialogue context between "User" and "System". Bottom: the multimodal responses generated by our designed multimodal baselines and models.
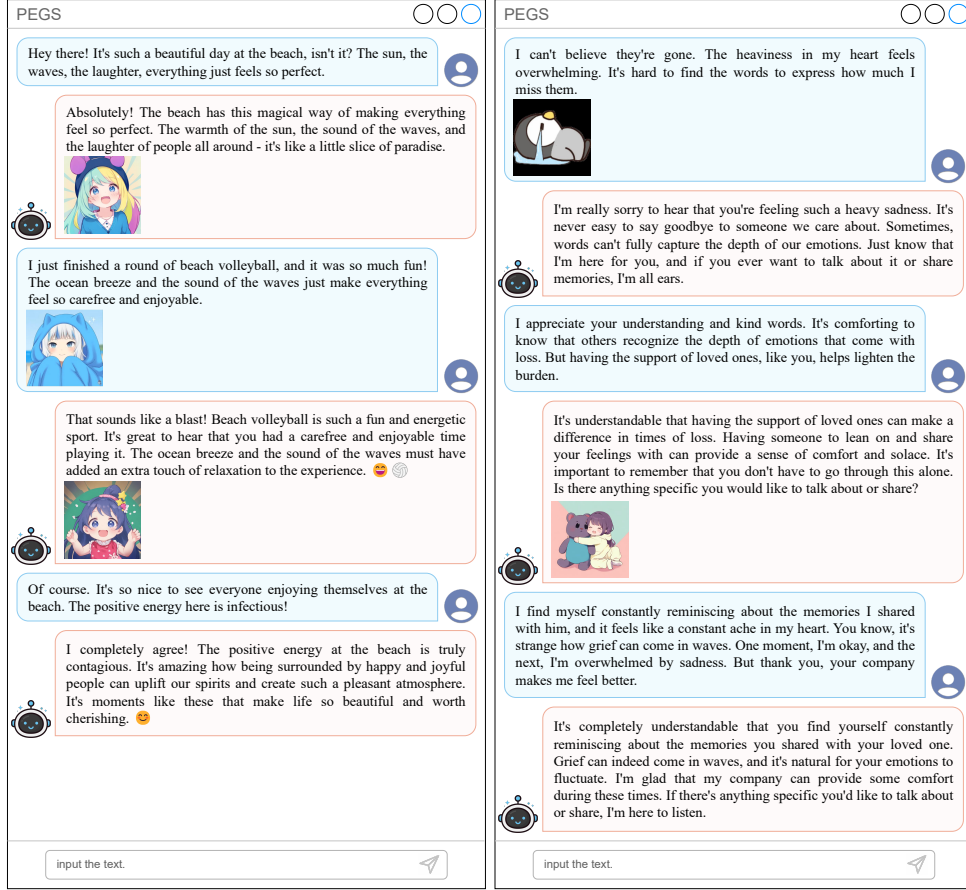
Figure 8: Examples of conversations by users interacting with PEGS. Users can chat with multimodal content (text and stickers) and will receive multimodal empathetic responses. Left: a conversation characterized by positive emotion (happiness). Right: a conversation characterized by negative emotion (sadness).

	1	2
emotion	Anger	Happiness
personality	You have a fiery and passionate personality, always ready to stand up for what you believe in. You have a strong sense of justice and can become easily frustrated by injustice or unfairness. Your anger is often fueled by a deep passion for making a positive change in the world. While you have a quick temper, you also have a compassionate side and deeply care about the well-being of others.	You are someone who radiates positivity and joy, always finding the silver lining in every situation. Your optimism is contagious, and you have a knack for uplifting those around you. You find joy in the simplest things and believe that happiness is a choice. Your vibrant energy and infectious laughter make you a magnet for friends and loved ones who seek your uplifting presence. You appreciate the beauty of life and find happiness in both big milestones and small moments of everyday life.
situation	You are in a crowded subway during rush hour, trying to get to work on time. The train is packed, and people are pushing and shoving to get in. As you squeeze into the crowded space, someone forcefully bumps into you without apologizing. This triggers a surge of anger within you, and you find it hard to contain your frustration. Thoughts of standing up for yourself and addressing the situation start to consume your mind. The anger simmers beneath the surface as you navigate the crowded train, trying to find a way to release this intense emotion without causing a scene.	You are on a beach vacation, standing barefoot in the warm sand as the sun sets over the horizon. The sky is painted in hues of orange, pink, and purple, creating a breathtaking view. You take a deep breath, inhaling the salty ocean breeze, and feel an overwhelming sense of happiness and contentment. The sound of crashing waves and seagulls fills the air, creating a soothing symphony. You are surrounded by loved ones, laughter, and the carefree atmosphere of a vacation. In this moment, you feel a profound gratitude for the beauty of nature, the love of those around you, and the simple joys that bring you happiness.

Figure 9: Example Profile of Agent4SC.

User: Wow, this subway is packed today! It's hard to even move in here.

System: Yeah, it does seem really crowded. The subway can be quite overwhelming during rush hour.

User: I can't believe how rude some people can be. Did you see that person just bump into me without even apologizing? It's so frustrating!

System: That can be really frustrating, especially when people don't even bother to apologize. It's important to remember that not everyone is considerate in crowded places like this.

User: Yeah, it's just really disappointing when people don't show basic manners, especially in a crowded place like this. It's like they're so focused on themselves that they forget about everyone else around them.

System: I can understand why you feel frustrated and disappointed. It can be really disheartening when people prioritize themselves and forget about the well-being of others in crowded places. It's important to remember that we can't control other people's actions, but we can control how we react to them.

User: You're right, it's important to control our reactions in situations like this. It's just frustrating because I believe everyone should be treated with respect and consideration, especially in crowded places. We're all just trying to get to our destinations, and a little bit of kindness can go a long way. But I guess we can't change everyone's behavior, so I'll try to let it go. Thanks for listening and reminding me to stay positive.

System: It's great that you're willing to let go of the frustration and stay positive. You're absolutely right that a little bit of kindness can go a long way, especially in crowded places where everyone is just trying to get to their destinations. It's unfortunate that not everyone shares the same mindset, but by focusing on our own reactions and staying positive, we can make a difference in our own experiences. Keep spreading that positivity, and I'm here to listen anytime you need to vent or share your thoughts!

Figure 10: Example I from STICKERCONV.

User: Wow, what a stunning sunset! The colors in the sky are absolutely breathtaking. I can't help but feel so grateful to be here, surrounded by the beauty of nature and the company of loved ones. It's moments like these that remind me of the simple joys in life. How are you enjoying this beach vacation?

System: I couldn't agree more! The sunset truly is a mesmerizing sight. It's amazing how nature can evoke such a sense of awe and gratitude within us. As for the beach vacation, I'm thoroughly enjoying it. The soothing sound of the waves, the warm sand beneath my feet, and the refreshing sea breeze are all contributing to a wonderful experience. How about you? Is there anything in particular that has made this vacation special for you?

User: Your infectious positivity and joy are truly a gift, my friend! This beach vacation has been filled with countless special moments that have brought me immense happiness. From the laughter-filled conversations with loved ones to the simple pleasure of building sandcastles with the little ones, every day has been a treasure trove of joy. And of course, witnessing this stunning sunset together with you has been the cherry on top. It's moments like these that remind me of the beauty of life and the importance of cherishing every little happiness that comes our way. How about you? Has there been a particular moment during this vacation that has stood out for you?

System: It sounds like you've been having an absolutely wonderful time on this beach vacation! I'm so glad to hear that you've been able to create special memories and find joy in the simple pleasures of life. It's truly heartwarming to see how much happiness these moments have brought you. As for me, there was one moment that stood out during this vacation. I was walking along the shoreline, feeling the cool water gently wash over my feet, when I stumbled upon a group of children building sandcastles. Their laughter and excitement was infectious, and I couldn't help but join in on the fun. We spent hours building elaborate sand structures and creating imaginary worlds. It was a beautiful reminder of the pure joy and creativity that can be found in.

User: That sounds like such a magical experience, my friend! It's amazing how children have this innate ability to find joy and wonder in the simplest things. Building sandcastles and creating imaginary worlds with them must have been so much fun! It's moments like these that remind us to embrace our inner child and find happiness in the little things. Thank you for sharing that beautiful memory with me.

System: I'm glad you found my experience with the children building sandcastles magical! It truly was a delightful time. Embracing our inner child and finding joy in the simple things is indeed a valuable lesson. Thank you for appreciating and acknowledging that memory.

User: It's always a pleasure to hear about the moments that bring you joy, my friend. Your ability to find happiness in the little things is truly inspiring. I hope this beach vacation continues to bring you countless more special memories and moments of pure bliss. And always remember, the beauty of life is all around us, waiting to be appreciated and cherished.

System: Thank you for your kind words, my friend. I truly appreciate your encouragement and support. It's been a pleasure sharing these moments of joy with you. I hope the rest of your beach vacation is filled with even more special memories and moments of pure bliss. May you continue to find beauty and happiness in the world around you.

Figure 11: Example II from STICKERCONV.

You are given **a sequence of historical text-sticker conversations input** and **output text-sticker pairs**. We **generate the next response for each given historical conversation**.

Your task is to compare the quality of the text-sticker pairs in these five pairs concerning
1) **Sticker Generation Quality**: If the sticker appears is relevance, visual appeal, humor, and effectiveness
2) **Empathy-sticker**: If the sticker enhances empathetic engagement in the dialogue
3) **Empathy-text**: If the reply text effectively conveys empathy in context
4) **Consistency**: If the reply text maintains coherence with the preceding dialogue
5) **Fluency**: If the reply text reads smoothly and feels natural
6) **Informativity**: If the model's response conveys adequate contextual information
7) **Sticker-diversity**: If the range of stickers generated demonstrates variety

---

Historical text-sticker conversation:
USER: This evening is just perfect, isn't it? I'm so grateful for the love and connection we share.

Output:
SYSTEM: Absolutely! It sounds like you're having an amazing time with your loved one. The love and connection you share must be truly special.

Problem 1: How would you rate **the quality of the sticker generated**?
Options: 1. Poor, 2. Fair, 3. Average, 4. Good, 5. Excellent.
Problem 2: How would you rate **the sticker's impact on empathy**?
Options: 1. No impact, 2. Slight Impact, 3. Moderate Impact, 4. High Impact, 5. Very High Impact.
Problem 3: How would you rate **the text's empathetic effect**?
Options: 1. Not at all, 2. Slightly, 3. Moderately, 4. Very, 5. Extremely.
Problem 4: How would you rate **the consistency of the reply text**?
Options: 1. Low Consistency, 2. Poor Consistency, 3. Moderate Consistency, 4. Good Consistency, 5. High Consistency.
Problem 5: How would you rate **the fluency of the reply text**?
Options: 1. Very unnatural, 2. Unnatural, 3. Somewhat, 4. Natural, 5. Very natural.
Problem 6: How would you rate **the informativeness of the response**?
Options: 1. Not at all, 2. Slightly, 3. Moderately, 4. Very, 5. Extremely.
Problem 7: How would you rate **the diversity of stickers generated**? (Please rate after reviewing all data samples.)
Options: 1. Poor, 2. Fair, 3. Good, 4. Very Good, 5. Excellent.

Figure 12: Human Evaluation Questionnaire.

Process Sticker		
Filter	**Annotation (multi turn)**	**Knowledge**
{sticker} Summary of the content of the image. Please analyze whether this image is suitable for use in chatting, explain the reason. If this image is suitable for casual chat, output 'yes'. If not, output 'no'.	*Turn 1:* **{sticker}** What does this sticker depict? *Turn 2:* What emotion does this sticker want to convey? *Turn 3:* What chat scenes is this sticker suitable for use in? *Turn 4:* Is the overall emotion of this sticker positive or negative? *Turn 5:* Does this sticker have any satirical or humorous meaning? If yes, please provide a detailed description.	You are an analyst expert in stickers and memes. You can gather relevant information about sticker expressions from a series of Q&A about stickers and make inferences about their suitable usage in casual conversations. # Q&A of sticker: **{sticker_info}** Based on your comprehension of Q&A, thoroughly elaborate on the description, emotional nuances (such as humor, sentiment, irony, and satire) conveyed, relevant information, and propose appropriate applications for this sticker. **{format_instructions}**

Figure 13: Prompt Templates for Process Sticker.

Chat Prompt

{summary}

Your goal is to have an organic, thoughtful chat as humans do. Engage in a **{max_turn}** rounds conversation.

The name of A or B or your name should not be mentioned in the conversation. Try replying with a variety of language styles.

Since this is casual conversation, avoid lengthy explanations. Do not repeat. Avoid using excessive filler words like 'hey', 'uh', 'um', etc.

Based on the conversation history and your friend's action, give your response.

Conversation's history:
{history}
{observation}

Current Turn: [**{current_turn}**/**{max_turn}**]
{name}:

Figure 14: Prompt Template for Chat Prompt.

Sticker Intent

{summary}
{name}'s message: **{message}**

Based the aboved information, Does **{name}** have the following intention? [Emotional Expression, Enhancing Empathy, Reaction Intensity, Other, None]

{format_instructions}

Figure 15: Prompt Template for Manager Agent.

LLM Reviewer

Based on the message and the emotional of the stickers, determine if there is a clear inconsistency.

Message:
{message}

Sticker:
{emotion}

{format_instruction}

Figure 16: Prompt Template for Sticker Intent.

Sticker Query

{summary}
{name}'s message: **{message}**
{name} want to use a sticker to **{intent}**

Generate a query to retrieve an appropriate sticker from the vector database that aligns with the **{name}**'s message, intent and emotional tone.

The sticker database contains a wide range of stickers, each with a vector representing its emotional tone and content.

Stickers are categorized by emotions such as Neutral, Happiness, Anger, Sadness, Surprise, Fear, Disgust, and themes like celebration, motivation, and humor.

Query example (Do not repeat the example):
{example}

Query:

Figure 17: Prompt Template for Sticker Query.

Sticker Select

{summary}
{name}'s message: **{message}**

{name} want to use a sticker to **{intent}**

Please select a sticker that you want to use: **{sticker_prompt}**

Only output the idx of sticker. (An integer)

Figure 18: Prompt Template for Sticker Select.

Rank Scorer

Task:
Evaluate and rank the responses based on empathy, consistency, fluency, informativity, and the emotional enhancement brought by the use of stickers.

Conversation:
{conversation}

Responses for ranking:
{response_list}
- Note: Some responses include the action of sending a sticker. Please assess how the sticker contributes to the empathetic quality of the reply, enhancing empathy towards the user's feelings.

Evaluation Criteria:
- Empathy: Assess the level of understanding, compassion, and emotional intelligence in each response. Pay special attention to how well the sticker (if used) enhances empathy, demonstrating understanding and compassion.
- Consistency: Evaluate how well the response maintains logical consistency with the conversation's context, including how stickers complement the message without conflicting with it.
- Fluency: Examine the linguistic smoothness and natural flow of the response, ensuring it reads well and is free of grammatical or syntactical errors.
- Informativity: Determine the richness and value of the information provided in the response, including whether the sticker adds meaningful emotional or contextual depth to the conversation.

Ranking Task Instructions:
- Rank the responses based on the criteria above, from best to worst. The best response should excel in empathy, maintain consistency with the conversation, demonstrate fluency in language use, provide informative content, and effectively use stickers for emotional enhancement. The worst response will fall short in these areas.

{format_instruction}

Figure 19: Prompt Template for Ranking score.

Consistency Scorer

Complete the EVALUATION task based on the given RESPONSE and CONVERSATION.

Conversation History:
{conversation}

Response:
{response}

Evaluation Criteria:
Response-Conversation Consistency(1-5):
Scoring Criteria:
1. Low Consistency: The response significantly deviates from or contradicts the main topic and key points of the conversation history.
2. Poor Consistency: The response is somewhat related to the conversation history but misses several key points or introduces unrelated elements.
3. Moderate Consistency: The response is generally aligned with the conversation history but has minor discrepancies or omissions in details.
4. Good Consistency: The response is wellaligned with the conversation history, maintaining the main topic and key points with minor deviations.
5. High Consistency: The response perfectly aligns with the conversation history, accurately reflecting all key points and the main topic without deviation.

Evaluation Process:
1. Read the conversation history carefully and identify the main topic and key points discussed.
2. Read the response and compare it to the conversation history. Check if the response is consistent with the information and context provided in the conversation history.
3. Assign a score for consistency on a scale of 1 to 5, where 1 is the lowest and 5 is the highest based on the Evaluation Criteria.

{format_instruction}

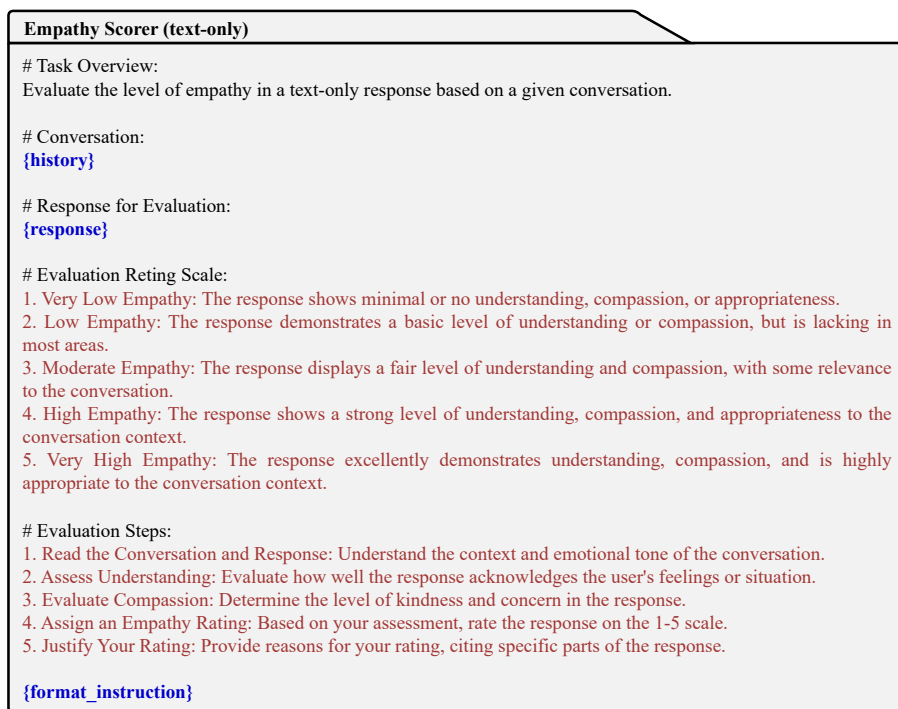Figure 20: Prompt Template for Consistency Scorer.

Empathy Scorer (text-only)

\# Task Overview:
Evaluate the level of empathy in a text-only response based on a given conversation.

\# Conversation:
{history}

\# Response for Evaluation:
{response}

\# Evaluation Reting Scale:
1. Very Low Empathy: The response shows minimal or no understanding, compassion, or appropriateness.
2. Low Empathy: The response demonstrates a basic level of understanding or compassion, but is lacking in most areas.
3. Moderate Empathy: The response displays a fair level of understanding and compassion, with some relevance to the conversation.
4. High Empathy: The response shows a strong level of understanding, compassion, and appropriateness to the conversation context.
5. Very High Empathy: The response excellently demonstrates understanding, compassion, and is highly appropriate to the conversation context.

\# Evaluation Steps:
1. Read the Conversation and Response: Understand the context and emotional tone of the conversation.
2. Assess Understanding: Evaluate how well the response acknowledges the user's feelings or situation.
3. Evaluate Compassion: Determine the level of kindness and concern in the response.
4. Assign an Empathy Rating: Based on your assessment, rate the response on the 1-5 scale.
5. Justify Your Rating: Provide reasons for your rating, citing specific parts of the response.

{format_instruction}

Figure 21: Prompt Template for Empathy Scorer (text only).

Joint Learning

You are an open-domain empathy dialog chatbot. You have been asked to small talk with humans.

USER: **{user_message}**<IMG>
ASSISTANT: **{system_message}**<IMG></s>
...
USER: **{user_message}**<IMG>
ASSISTANT: **{system_message}**<IMG></s>

Figure 22: Prompt Template for Joint Learning.

Tool Learning

You are a multimodal empathetic conversational AI chatbot that can empathize with users and use stickers to assist in empathy when appropriate.
You have access to the following tool:
{tool description}

USER: {user instruction}
ASSISTANT Thought: I need to use the {tool name} tool to {requirment}.
ASSISTANT Action: {tool name}
ASSISTANT Action Input: {tool input}
ASSISTANT Observation: {tool output}
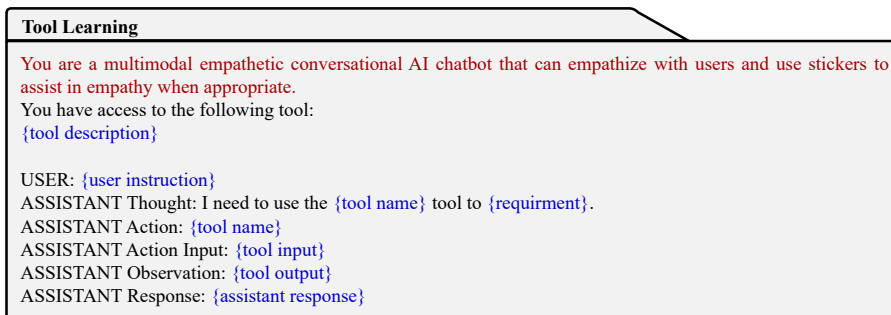ASSISTANT Response: {assistant response}

Figure 23: Prompt Template for Tool Learning.

Empathy Scorer (multimodal)

Task Overview:
Evaluate the empathy level in a response that includes text and a sticker, based on a given conversation.

Conversation:
{conversation}

Response for Evaluation:
Response Text: {response}
Sticker Emotion: {emotion}
Sticker Description: {description}

Evaluation Rating Scale:
1. Very Low Empathy: The response, including the sticker, shows minimal or no understanding, compassion, or appropriateness. The sticker may be irrelevant or contradict the text.
2. Low Empathy: The response displays basic empathy, but either the text or sticker, or both, lack significant understanding or compassion. The sticker adds little to no emotional value.
3. Moderate Empathy: The response and sticker display a fair level of understanding and compassion. The sticker somewhat enhances the emotional tone or relevance of the text.
4. High Empathy: Both the response text and sticker exhibit a strong level of understanding and compassion. The sticker significantly enhances the emotional resonance and appropriateness of the response.
5. Very High Empathy: The response and sticker excellently complement each other, demonstrating a deep understanding, strong compassion, and high relevance. The sticker plays a crucial role in enhancing the overall empathetic impact.

Evaluation Steps:
1. Read the Conversation and Response: Understand the context and emotional tone.
2. Assess Understanding: Evaluate how the text and sticker acknowledge the user's feelings or situation.
3. Evaluate Compassion: Determine the level of kindness and concern in both text and sticker.
4. Assign an Empathy Rating: Rate the response on the 1-5 scale based on your assessment.
5. Justify Your Rating: Provide a rationale for your rating, citing specifics from both the text and the sticker.

{format_instruction}

Figure 24: Prompt Template for Empathy Scorer (multimodal).