Emotion and Intention Guided Multi-Modal Learning for Sticker Response Selection

Yuxuan Hu[1,2*], Jian Chen[3,2*], Yuhao Wang[1], Zixuan Li[2], Jing Xiong[3]
Pengyue Jia[1†], Wei Wang[2†], Chengming Li[2†], Xiangyu Zhao[1†]
[1]City University of Hong Kong, Hong Kong, China
[2]Shenzhen MSU-BIT University, Shen Zhen, China
[3]The University of Hong Kong, Hong Kong, China
yuxuanhu7-c@my.cityu.edu.hk, ccccccj03@connect.hku.hk, licm@smbu.edu.cn, xianzhao@cityu.edu.hk

Abstract

Stickers are widely used in online communication to convey emotions and implicit intentions. The Sticker Response Selection (SRS) task aims to select the most contextually appropriate sticker based on the dialogue. However, existing methods typically rely on semantic matching and model emotional and intentional cues separately, which can lead to mismatches when emotions and intentions are misaligned. To address this issue, we propose **E**motion and **I**ntention **G**uided **M**ulti-Modal **L**earning (**EIGML**). This framework is the first to jointly model emotion and intention, effectively reducing the bias caused by isolated modeling and significantly improving selection accuracy. Specifically, we introduce Dual-Level Contrastive Framework to perform both intra-modality and inter-modality alignment, ensuring consistent representation of emotional and intentional features within and across modalities. In addition, we design an Intention-Emotion Guided Multi-Modal Fusion module that integrates emotional and intentional information progressively through three components: Emotion-Guided Intention Knowledge Selection, Intention-Emotion Guided Attention Fusion, and Similarity-Adjusted Matching Mechanism. This design injects rich, effective information into the model and enables a deeper understanding of the dialogue, ultimately enhancing sticker selection performance. Experimental results on two public datasets show that EIGML outperforms state-of-the-art baselines, achieving higher accuracy and a better understanding of emotional and intentional features.

Code — https://github.com/Applied-Machine-Learning-Lab/EIGML

Introduction

With the rapid proliferation of social media platforms, stickers have attracted significant attention due to their unique ability to express emotions and convey communicative intent (Zhao et al. 2021; Zhang et al. 2024). An appropriately selected sticker can not only effectively highlight the emotions and intentions which users wish to express, but also serve as a substitute for complex textual responses (Chen et al. 2024c; Tang and Hew 2019). Consequently, the task of Sticker Response Selection (SRS) has emerged as an active research area and garnered significant attention.

Figure 1: Examples for potential failures in sticker selection when relying solely on emotional or intentional cues.

Compared to typical image-text matching tasks (Zhang et al. 2022a; Wu et al. 2019; Zhang et al. 2023), SRS poses more complex, highly context-sensitive challenges. Stickers serve not only as visual symbols but also as rich multi-modal carriers of emotional and intentional cues (Wang et al. 2024a; Chen et al. 2025b; Shi and Kong 2024), whose meanings vary greatly across conversational contexts. Thus, models must move beyond surface semantic alignment to effectively capture emotional signals and infer communicative intent for deeper multi-modal understanding (Chee et al. 2025; Ge et al. 2022; Liu, Zhang, and Yang 2022). However, many methods still rely on shallow semantic matching (Gao et al. 2020; Fei et al. 2021), failing to grasp subtle emotional and intentional cues. In real dialogues, stickers convey not only

*These authors contributed equally.
†Corresponding authors.

emotions but also tone, social stance, and complex communicative strategies. Ignoring these leads to semantically plausible yet contextually inappropriate choices.

Although some studies incorporate emotional or intentional information to improve multi-modal alignment (Xia et al. 2024; Wang et al. 2025a), they often treat the two separately, overlooking their intertwined nature in human communication. This separation may cause over-reliance on one aspect and neglect of the other, leading to mismatches in nuanced sticker selection. As shown in Figure 1, models focusing only on emotion may miss subtleties like sarcasm or irony, while those centered on intention may ignore emotional undertones, resulting in contextually inappropriate selections. These cases show that isolating emotion or intention is insufficient and even harm overall performance. A unified emotion-intention framework is thus needed to better capture the contextual communicative functions of stickers.

To address these challenges, we propose the Emotion and Intention Guided Multi-Modal Learning (EIGML) framework to enhance SRS accuracy by jointly modeling emotional and intentional cues. The framework has two core modules targeting key issues. First, to overcome the limitations of semantic matching in capturing subtle emotional variations and implicit intentions, we introduce the Dual-Level Contrastive Framework (DLCF), which boosts sensitivity to emotional and intentional features via intra-modality and inter-modality alignment. Second, to mitigate the bias caused by modeling emotion and intention in isolation, we design the Intention-Emotion Guided Multi-Modal Fusion (IEGMF) mechanism, comprising Emotion-Guided Intention Knowledge Selection (EIKS), Intention-Emotion Guided Attention Fusion (IEGA), and Similarity-Adjusted Matching Mechanism (SAMM). Together, they inject rich, effective information to deepen contextual understanding and improve selection. Extensive experiments on two large-scale datasets, StickerChat and DSTC10-MOD, demonstrate that our method consistently outperforms state-of-the-art models across multiple metrics, proving its effectiveness and robustness in complex scenarios. Our contributions include:

- To our knowledge, this is the first framework to explicitly model both emotion and intention for Sticker Response Selection, reducing mismatches from isolated modeling and improving selection accuracy. It also highlights key challenges in current SRS methods arising from emotional and intentional cues, which remain underexplored.
- We propose a novel EIGML framework, which consists of two core components designed to address the mismatch between emotional and intentional features across visual and textual modalities. Building upon semantic alignment, EIGML introduces cross-modal alignment of emotion and intention representations to enhance overall modality consistency. Moreover, EIGML employs a three-stage progressive fusion strategy, where emotional and intentional cues are gradually injected to guide the integration of visual and textual features, thereby improving the effectiveness of sticker selection.
- Extensive experimental results on two public benchmark datasets validate the effectiveness and efficiency of the proposed EIGML framework.

Related Work

Existing sticker response selection models can be grouped into shallow semantic matching and disjoint emotion–intention modeling.

Shallow Semantic Matching Methods. Early approaches, such as Laddha et al. (2020), use clustering-based methods to predict the next message, which is then replaced with a sticker. Later, a deep interaction network is proposed in Gao et al. (2020), which leverages cross-attention to extract multi-modal features for matching-based sticker retrieval. In Fei et al. (2021), both text and sticker prediction are framed as sequence generation tasks using a unified framework with pretrained GPT-2 (Radford et al. 2019) to jointly encode dialogue context for sticker selection. A multi-task learning framework is introduced by Zhang et al. (2022b) to improve understanding of both sticker and dialogue semantics. While these methods show promising semantic matching results, they often neglect the rich emotional and intentional information in stickers, underscoring the need for explicit modeling of such information.

Separate Emotion and Intention Modeling Methods. In recent years, several studies have attempted to go beyond semantic matching by incorporating emotional or intentional understanding of both stickers and dialogue. CKS (Chen et al. 2024a) integrates commonsense knowledge to better recognize emotional expressions and extract unbiased visual features, thereby improving alignment between stickers and dialogues. PBR (Xia et al. 2024) focuses on emotional features within both stickers and dialogues, enabling emotion-aware sticker selection. Some works (Wang et al. 2025a; Liang et al. 2025) construct datasets with intention labels to incorporate intention into sticker response selection. While these methods introduce emotional or intentional signals, they typically model them separately, failing to capture their intertwined nature. To address this limitation, we propose a unified framework that integrates emotion and intention across modalities, improving both the accuracy and contextual relevance of sticker response selection.

Method

Overview

Problem Definition. Given a multi-turn dialogue history $U = \{u_1, u_2, ..., u_{N_U}\}$ and a set of candidate stickers $S = \{s_1, s_2, ..., s_{N_S}\}$, where N_U and N_S denote the number of utterances and candidate stickers respectively, the SRS task aims to select the most appropriate sticker by understanding various cues in the dialogue. Specifically, we aim to leverage semantic, emotional, and intentional information from the text to train a model that predicts the suitability of each candidate sticker. The training objective minimizes the binary cross-entropy (BCE) loss between the predicted matching scores and ground-truth labels defined as:

$$\min_{\theta} \mathcal{L}_{ma} = -\frac{1}{N} \sum_{i=1}^{N} \text{BCE}\left(f_\theta(U_i, S_i), y_i\right), \qquad (1)$$

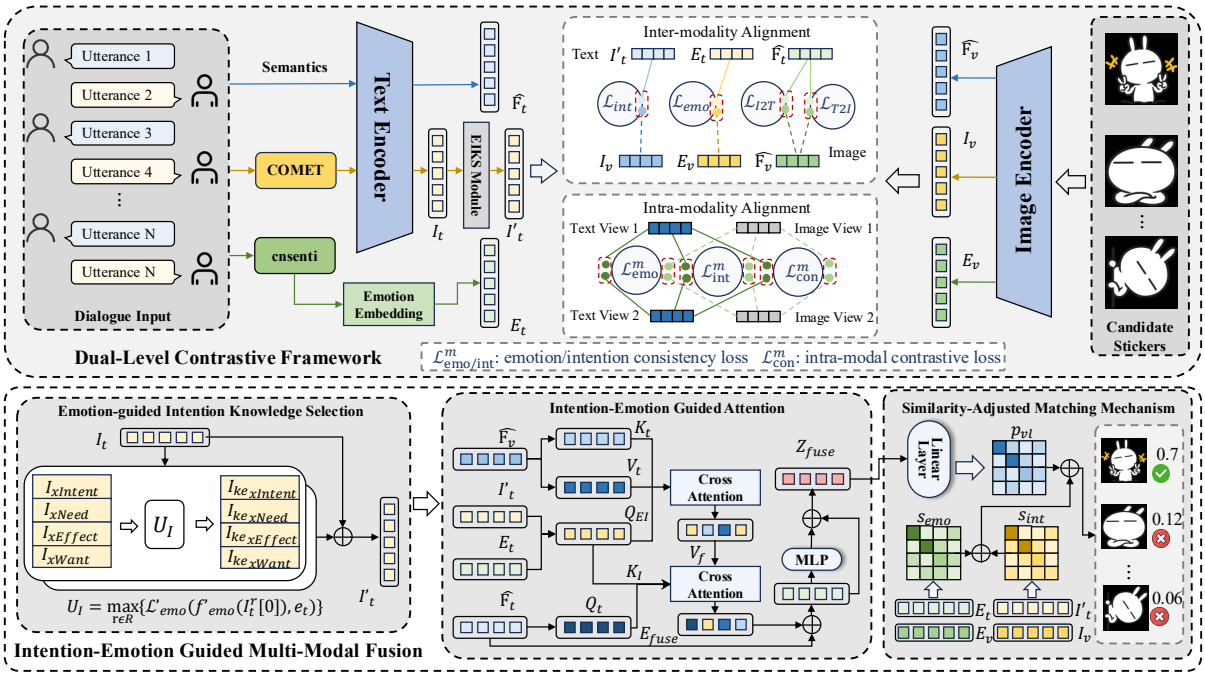


Figure 2: Overview of our proposed approach. Dual-Level Contrastive Framework enhances intra-modality and inter-modality alignment, while Intention-Emotion Guided Multi-Modal Fusion integrates three submodules: (1) Intention-Emotion Attention Fusion, (2) Emotion-Guided Intention Selection, and (3) Similarity-Adjusted Matching Mechanism.

where θ represents the trainable parameters of the model, N is the total number of samples in the dataset, and $y_i = \{y_1^i, y_2^i, ..., y_{N_s}^i\}$ indicates whether the candidate stickers in the i-th sample matches the dialogue information U_i.

Framework Overview. To address the limitations of purely semantic matching in capturing subtle emotions and implicit intentions—and the mismatches caused by modeling them separately—we propose the Emotion and Intention Guided Multi-Modal Learning (EIGML) framework. As shown in Figure 2, EIGML consists of two main components: Dual-Level Contrastive Framework (DLCF) and Intention-Emotion Guided Multi-Modal Fusion (IEGMF). DLCF strengthens emotional and intentional representations and improves multi-modal alignment through intra- and inter-modality strategies. IEGMF integrates Emotion-Guided Intention Knowledge Selection (EIKS), Intention-Emotion Guided Attention Fusion (IEGA), and the Similarity-Adjusted Matching Mechanism (SAMM) to achieve deep visual–textual fusion and capture the complementarity between emotion and intention.

Dual-Level Contrastive Framework

Although emotional and intentional cues are crucial for SRS (Xia et al. 2024; Wang et al. 2025a), existing methods usually model them independently, leading to semantic mismatches. To address this, we propose the first unified framework that jointly models emotion and intention. Its core module, DLCF, performs inter-modality alignment for semantic consistency and intra-modality alignment to strengthen each modality's representations, effectively improving representation quality and reducing mismatches in sticker selection.

Inter-Modality Alignment. Conventional semantic alignment (Gao et al. 2020; Fei et al. 2021) falls short in capturing the subtle emotional–intentional interplay across modalities. To address this, we propose a cross-modal alignment mechanism that explicitly targets emotion and intention. This design enhances the understanding of emotional and intentional cues and semantic consistency between modalities, yielding significantly improved matching performance.

Image and text data are passed through the visual encoder and the textual encoder to get features F_v and F_t, respectively. We adopt the InfoNCE loss (Oord, Li, and Vinyals 2018) for image-text semantic alignment. The `[CLS]` token features of images and texts are projected into a shared embedding space via linear mappings: $\hat{F}_v = F_v^{\text{cls}} W_v$, $\hat{F}_t = F_t^{\text{cls}} W_t$. In image-to-text alignment, a sticker and its paired dialogue form a positive pair, with other dialogues as negatives; in text-to-image alignment, the dialogue and its matched sticker form a positive pair, with other stickers as negatives. The losses are defined as follows:

$$\mathcal{L}_{I2T} = -\frac{1}{K}\sum_{i=1}^{K} \log \frac{\exp(\hat{F}_{v,pos}^i \cdot \hat{F}_t^i / \tau)}{\sum_{j=1}^{K} \exp(\hat{F}_{v,pos}^i \cdot \hat{F}_t^j / \tau)}, \quad (2)$$

$$\mathcal{L}_{T2I} = -\frac{1}{K}\sum_{i=1}^{K} \log \frac{\exp(\hat{F}_t^i \cdot \hat{F}_{v,pos}^i / \tau)}{\sum_{j=1}^{K} \exp(\hat{F}_t^i \cdot \hat{F}_{v,neg}^j / \tau)}, \quad (3)$$

where K is the batch size, $\hat{F}^i_{v,pos}$ and $\hat{F}^i_{v,neg}$ denote the visual embedding of the positive and negative stickers for the i-th dialogue.

For each dialogue, we use the cnsenti (Deng 2019) toolkit to derive an emotion distribution $\mathbf{e}_t$ over seven categories. The most prominent is embedded via the emotion embedding matrix E_{emo} to obtain the textual emotion embedding E_t. For the image, a linear head f_{emo} maps the visual `[CLS]` feature F_i^{cls} to its emotion representation E_v, defined as:

$$E_t = E_{\text{emo}}\left[\arg\max(\mathbf{e}_t)\right], \quad E_v = f_{\text{emo}}(F_v^{\text{cls}}). \tag{4}$$

To achieve cross-modal emotion alignment, we minimize the Kullback–Leibler (KL) divergence between the predicted emotion distribution of the matched image E_v^{pos} and the corresponding text E_t defined as follows:

$$\mathcal{L}_{\text{emo}} = KL(E_v^{\text{pos}}, E_t), \tag{5}$$

where E_v^{pos} means the feature of the correct sticker.

Next, we leverage the generative commonsense transformer model COMET (Bosselut et al. 2019) for commonsense inferences from the input text across four relations $R = \{$*xIntent*, *xNeed*, *xWant*, *xEffect*$\}$. These inferences are concatenated as C_t, defined as:

$$C_r = \text{COMET}(T, r), \quad C_t = \oplus_{r\in R} C_r, \tag{6}$$

where $r \in R$ denotes the relation type, and $\oplus$ is the concatenation operation.

The commonsense inference C_t is encoded via the text encoder to obtain the initial intention embedding $I_t = \mathcal{F}t(C_t)$, which is further refined by the EIKS module into I'_t (detailed in the next section). For the visual modality, a projection head f_{int} maps the sticker's `[CLS]` feature F_i^{cls} to its intention embedding $I_v = f_{\text{int}}(F_i^{\text{cls}})$. The intention alignment loss $\mathcal{L}_{\text{int}}$, based on InfoNCE (Equation (2), (3)), encourages consistency between I_v and I'_t. The total inter-modality loss is defined as:

$$\mathcal{L}_{inter} = \mathcal{L}_{I2T} + \mathcal{L}_{T2I} + w_e \cdot \mathcal{L}_{\text{emo}} + w_i \cdot \mathcal{L}_{\text{int}}. \tag{7}$$

where w_e and w_i are the weights for $\mathcal{L}_{\text{emo}}$ and $\mathcal{L}_{\text{int}}$.

Intra-Modality Alignment. While cross-modal alignment ensures inter-modality consistency, strong intra-modality representations are crucial. Emotional and intentional cues often lie within each modality and may be missed by shallow encoders. To address this, we introduce intra-modality alignment to enhance each modality's expressiveness and discriminability, reinforcing emotion and intention understanding to support improved cross-modal matching.

Specifically, we perform intra-modality alignment by generating two augmented views via independent dropout and separate projection heads (Gao, Yao, and Chen 2021). These views are mapped to emotion and intention spaces using the shared intention head for inter-modality alignment. For each modality $m \in \{v, t\}$, the emotion and intention consistency losses are symmetric KL divergence, which are defined as:

$$\mathcal{L}^m_* = \frac{1}{2}\left[KL(p^m_{1,*} \| p^m_{2,*}) + KL(p^m_{2,*} \| p^m_{1,*})\right], \tag{8}$$

where $p^m_{i,*}$ is the softmax-normalized prediction from the i-th view and $* \in \{\text{emo}, \text{int}\}$. Additionally, we employ InfoNCE-style instance-level contrastive learning between the two views using cosine similarity. The contrastive loss for modality m is defined as:

$$\mathcal{L}^m_{con} = \frac{1}{2}[\mathcal{L}_{NCE}(z^m_1, z^m_2) + \mathcal{L}_{NCE}(z^m_2, z^m_1)], \tag{9}$$

where z_1, z_2 are are normalized features from two augmented views. The final intra-modality loss is the sum of emotion KL losses, intention KL losses, and the instance contrastive loss are defined as:

$$\mathcal{L}_{intra} = \sum_{m\in\{v,t\}} (w^m_e \cdot \mathcal{L}^m_{emo} + w^m_i \cdot \mathcal{L}^m_{int} + \mathcal{L}^m_{con}). \tag{10}$$

where w^m_e and w^m_i are the weights for $\mathcal{L}^m_{emo}$ and $\mathcal{L}^m_{int}$.

Intention-Emotion Guided Multi-Modal Fusion

While emotion and intention are closely linked in real communication, existing approaches (Xia et al. 2024; Wang et al. 2025a) often model them separately, leading to mismatches in SRS. To address this, we propose Intention-Emotion Guided Multi-Modal Fusion (IEGMF), which strengthens cross-modal fusion via three components: (1) EIKS uses emotional cues to retrieve relevant intentional knowledge; (2) IEGA performs fine-grained multi-modal integration under joint emotional–intentional guidance; and (3) SAMM refines the final decision by correcting potential misalignment.

Emotion-Guided Intention Knowledge Selection. To better align intentional knowledge with emotional state and dialogue context, we propose Emotion-Guided Intention Knowledge Selection (EIKS) inspired by Cai et al. (2023). Specifically, a linear classifier f'_{emo} is applied to the text representation F_t^{cls} to predict an emotion distribution via softmax. Cross-entropy loss optimizes both the predicted distribution and learned emotion representation, defined as: $\mathcal{L}'_{emo} = -\sum_{c=1}^{C} e_t \log \sigma(f'_{emo}(F_t^{cls}))$, where C is the number of emotion categories and σ denotes the softmax function. Then, f'_{emo} eliminates irrelevant intentional knowledge by iteratively refining knowledge representations using gradients from the emotion classification task. At each step, the knowledge vector with the highest emotional classification loss is identified as the most irrelevant candidate, defined as:

$$U_I = \max_{r\in R} \{\mathcal{L}'_{emo}(f'_{emo}(I^r_t[0]), e_t)\}. \tag{11}$$

We further employ a nonlinear regression method to model the influence of knowledge exclusion, yielding a guidance matrix $G = \nabla_\theta f$. Next, we utilize this matrix to obtain the adjustment vector δ. This vector is then injected into all knowledge representations to construct knowledge-enhanced representations: $I_{ke} = I_t + \delta$. Following knowledge enhancement, a dynamic gating mechanism f_{gate} is used to adaptively fuse the enhanced knowledge I_{ke} with the knowledge I_t to get I'_t, defined as follows:

$$w_k = \text{Sigmoid}\left(f_{gate}([I_t; I_{ke}])\right), \tag{12}$$

$$I'_t = w_k \odot I_{ke} + (1 - w_k) \odot I_t, \tag{13}$$

where $\odot$ denotes element-wise multiplication.

Intention-Emotion Guided Attention Fusion. To enable fine-grained contextual-level alignment, we design Intention-Emotion Guided Attention Fusion (IEGA) that integrates selected emotional and intentional cues into the multi-modal fusion process, enhancing the joint understanding of dialogue and sticker semantics. Specifically, we first concatenate the intentional knowledge features I'_t with the textual emotion representation E_t as the intention-emotion features E_{IT}. Given image features F_v, textual features F_t, and the Intention-Emotion features E_{IT}, we first compute their linear projections: query vectors Q_T and Q_{EI} from F_t and E_{IT}, and key/value vectors K_V and V_V from F_v. The query from Intention-Emotion features attends to the image keys and values, generating Intention-Emotion knowledge-aware visual value V_{EIV}, defined as:

$$V_{EIV} = \sigma\left(\frac{Q_{EI}K_V^\top}{\sqrt{d_{head}}}\right)V_V. \tag{14}$$

Then, textual queries attend to the intention-emotion knowledge-aware visual features as:

$$E_{fuse} = \sigma\left(\frac{Q_T Q_{EI}^\top}{\sqrt{d_{head}}}\right)V_{EIV}, \tag{15}$$

where d_{head} is the attention head dimension. An MLP f_{fuse} is used for the final output, defined as:

$$Z_{fuse} = f_{fuse}(\texttt{LN}\left(F_t + \texttt{SD}(E_{fuse})\right), \tag{16}$$

where Layer normalization (LN), dropout with stochastic depth (SD) are applied to improve training stability.

Similarity-Adjusted Matching Mechanism. To improve the model's ability to identify the most appropriate sticker from emotional and intentional perspectives, we propose Similarity-Adjusted Matching Mechanism (SAMM), which refines the matching score by integrating semantic similarity in both the emotional and intentional representation spaces. We first use a linear layer to predict the semantic match score p_{vl} based on Z_{fuse}. Specifically, given the textual emotion and intention embeddings E_t, I'_t from the dialogue, and their corresponding sticker-level prototypes E_v, I_v, we compute normalized cosine similarities as:

$$s_{\text{emo}} = \frac{1+\cos(E_t, E_v)}{2}, \quad s_{\text{int}} = \frac{1+\cos(I'_t, I_v)}{2}. \tag{17}$$

These scores are fused into a single similarity score with a learnable parameter α as:

$$s_{\text{EI}} = \alpha \cdot s_{\text{emo}} + (1-\alpha)\cdot s_{\text{int}}, \tag{18}$$

We then combine s_{EI} with the original vision-language matching probability p_{vl} to obtain the final relevance score with a learnable parameter β to balance the contribution of matching confidence and emotion-intention alignment between the vision and text modalities, defined as:

$$p_{\text{final}} = \beta \cdot p_{\text{vl}} + (1-\beta)\cdot s_{\text{EI}}, \tag{19}$$

	Stickers	Dialogue-Sticker Pairs
StickerChat	174,695	340,168
DSTC10-MOD	307	215,117

Table 1: Statistics of StickerChat and DSTC10-MODzhe.

Training & Inference

We train our model using the full combination of losses in an end-to-end way. The whole loss used for training our model can be summarized as:

$$\mathcal{L}_{total} = \mathcal{L}_{\text{itm}} + \mathcal{L}_{\text{inter}} + \mathcal{L}_{\text{intra}} + w'_e \cdot \mathcal{L}'_{\text{emo}}. \tag{20}$$

where w'_e is the weight of loss $\mathcal{L}'_{\text{emo}}$.

During inference, we discard all contrastive components and directly use p_{final} as the predicted score. The sticker with the highest score is regarded as the selected response.

Experiments

Dataset and Metrics

Datasets We conduct experiments on two public datasets: StickerChat (Gao et al. 2020) and the Chinese version of DSTC10-MOD (Fei et al. 2021), with detailed statistics shown in Table 1. StickerChat contains 320,168 dialogue-sticker pairs for training and 10,000 each for validation and testing. DSTC10-MOD includes 45,000 dialogues and 307 stickers; since the test set is unavailable, we follow Zhang et al. (2022b) and evaluate on the validation set. Dialogues are split into multiple samples using the same preprocessing, resulting in 211,575 training and 3,542 test pairs.

Metrics Following prior work (Gao et al. 2020; Xia et al. 2024), we adopt Mean Average Precision (MAP) and Recall at position K among 10 candidates (R_{10}@K) to evaluate sticker selection performance. MAP reflects the overall ranking quality, while R_{10}@K (K = 1, 2, 5) measures the proportion of cases where the correct sticker appears in the top K, indicating retrieval effectiveness at different cutoffs. All results are reported as percentages.

Implementation Details

Following Xia et al. (2024), in StickerChat, each dialogue is paired with one correct sticker and other theme-based distractors, while in DSTC10-MOD, distractors are randomly sampled from the full sticker set. We use the pre-trained bert-base-chinese (Devlin et al. 2019) (max length 512) as the text encoder and ViT-B/16 from ALBEF (Li et al. 2021) as the visual encoder, both producing 768-dimensional features. The model is optimized with AdamW (lr=5e-5) for 5 epochs, with a random seed of 42. Experiments run on 4 NVIDIA A800 GPUs using PyTorch with a batch size of 16 per GPU. Input images are randomly cropped and resized to 128×128. Emotion-related weights (w_e, w'_e, w^t_e, w^v_e) and intention-related weights (w_i, w^t_i, w^v_i) are set to 0.5.

	MAP	R_{10}@1	R_{10}@2	R_{10}@5
PSAC	66.2	53.3	64.1	83.6
LSTUR	68.9	55.8	68.0	87.4
Synergistic	59.3	43.8	56.9	79.8
SRS	70.9	59.0	70.3	87.2
CLIP	70.9	59.1	70.3	86.8
ALBEF	76.8	67.0	75.6	90.0
PBR	79.2	69.3	79.5	93.5
EIGML	**81.2***	**72.3***	**81.3***	**93.9***

Table 2: Evaluation results on the StickerChat dataset, best results in **bold** while second with underline. "*" indicates the statistically significant improvements (i.e., two-sided t-test with $p < 0.05$) over the best baseline.

	MAP	R_{10}@1	R_{10}@2	R_{10}@5
SRS	50.3	30.5	54.2	71.3
MOD-GPT	52.3	31.2	54.8	72.1
CLIP	54.9	38.4	56.5	52.3
MMBERT	57.7	37.1	51.3	85.2
PBR	65.0	47.2	67.1	90.1
EIGML	**67.4***	**50.0***	**70.4***	**92.1***

Table 3: Evaluation results on the DSTC10-MOD dataset, best results in **bold** while second with underline. "*" indicates the statistically significant improvements (i.e., two-sided t-test with $p < 0.05$) over the best baseline.

Comparison with Baselines

Baselines. On the StickerChat dataset, we compare our proposed method, EIGML, with previous methods including PSAC (Li et al. 2019), LSTUR (An et al. 2019), Synergistic (Guo, Xu, and Tao 2019), SRS (Gao et al. 2020), CLIP (Radford et al. 2021), ALBEF (Li et al. 2021), and PBR (Xia et al. 2024). For the DSTC10-MOD dataset, we compare EIGML with SRS, MOD-GPT (Fei et al. 2021), CLIP, MMBERT (Zhang et al. 2022b), and PBR.

Table 2 shows the comparison results on the StickerChat dataset. Compared to visual question answering methods (PSAC and Synergistic), recommendation methods (LSTUR), pre-trained multi-modal retrieval models (CLIP and ALBEF), and even SRS task-specific methods (SRS and PBR), our proposed EIGML achieves the best performance. As shown in Table 2, the best performance of the previous methods is PBR, which achieves 79.2% on MAP and 69.3% on R_{10}@1. By contrast, our EIGML selects stickers that align with both the emotional and intentional requirements of the dialogue while considering the semantics alignment, thus outperforming previous methods with 2% on MAP and 3% on R_{10}@1, which suggests that the joint modeling of emotional and intentional cues is effective in enhancing SRS accuracy and can partially mitigate mismatches that arise when emotions and intentions are misaligned.

To provide a more comprehensive evaluation, we further test our method on the DSTC10-MOD dataset. As shown in Table 3, compared with the strongest baseline (65.0% MAP and 47.2% R_{10}@1), EIGML achieves absolute gains of 2.4% in MAP and 2.8% in R_{10}@1. These results validate the effectiveness and robustness of our approach. By jointly modeling dialogue semantics and integrating emotional and intentional cues, our method improves sticker selection accuracy and demonstrates strong generalization ability.

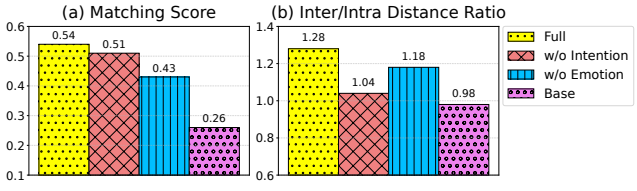


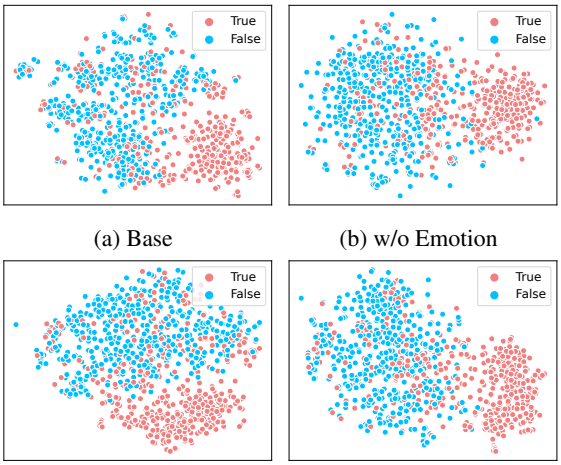
Figure 3: On StickerChat, matching scores and Inter/Intra Distance Ratios. Higher ratios reflect better class separation in the t-SNE space, while higher matching scores indicate stronger alignment between dialogues and correct stickers.

Effect of Emotion and Intention

To carefully evaluate the impact of emotional and intentional information beyond basic semantic alignment, we conduct ablation studies. We remove the intention-related (w/o Intention) and emotion-related (w/o Emotion) modules and measure changes in matching scores for positive sticker–text pairs, showing their distinct roles in improving SRS accuracy. We also compare our full model with a Base model using only text–image semantic alignment and self-attention, highlighting how emotion and intention enhance overall understanding of stickers and dialogue.

The results are shown in Figure 3. We observe that our proposed EIGML model achieves an average matching score of 0.54 for correctly predicted positive sticker-text pairs. When the intention-related and emotion-related modules are removed, the scores drop to 0.51 and 0.43, respectively. In contrast, the base model only achieves a score of 0.26. These results clearly demonstrate the effectiveness of incorporating both intention and emotion modeling. The significant performance drop after removing either component underscores their complementary roles in understanding sticker emotion and intention, which are essential for mitigating mismatches when emotions and intentions are misaligned.

We perform a visualization analysis of sticker representations using t-SNE (Maaten and Hinton 2008) (Figure 4) and compare the Inter/Intra Cluster Distance Ratio of each method (Figure 3) to measure separability and compactness. Results show that models lacking emotional or intentional components produce more scattered distributions, with larger intra-class distances and smaller inter-class separations. In contrast, although EIGML sticker features exhibit some dispersion, they show clearer distinctions between correct and incorrect stickers. This suggests that jointly modeling emotion and intention leads to more effective separation of relevant stickers, highlighting the importance of emotional and intentional cues in understanding context and selecting appropriate stickers.



(a) Base (b) w/o Emotion

(c) w/o Intention (d) EIGML

Figure 4: t-SNE visualization of 500 StickerChat examples using different methods. For each test instance, one ground-truth and one randomly sampled negative image (from nine distractors) are shown.

Inter	Intra	EIKS	IEGA	SAMM	MAP
√	√	√	√		80.48
√	√	√		√	80.58
√	√		√	√	80.64
√		√	√	√	80.99
	√	√	√		80.29
√	√	√	√	√	**81.15**

Table 4: Ablation results on the StickerChat dataset. The meaning of √ is to include a submodule. Since SAMM is built upon Inter-modality emotion and intention alignment, it is also removed when Inter is ablated.

Ablation Study

To better assess the contributions of each key component in EIGML, we perform ablation studies on the StickerChat dataset. EIGML comprises five core components: Inter-Modality Alignment (Inter), Intra-Modality Alignment (Intra), EIKS, IEGA, and SAMM. Since image–text semantic alignment underlies the SRS task, when ablating Inter, we remove only emotion and intention alignment across modalities while keeping basic semantic alignment.

As shown in Table 4, all EIGML components contribute to performance. Removing inter-modality alignment and SAMM causes the largest drop (from 81.31% to 80.29%), highlighting the importance of cross-modal emotion–intention alignment. Excluding intra-modality alignment reduces accuracy to 80.99%, indicating the need for within-modality consistency. Removing EIKS or IEGA further degrades performance (to 80.64% and 80.58%), demonstrating the effectiveness of knowledge supervision and interaction guidance. Disabling similarity-based matching also leads to a noticeable drop (80.48%). Overall, these results confirm the necessity and synergy of all components.

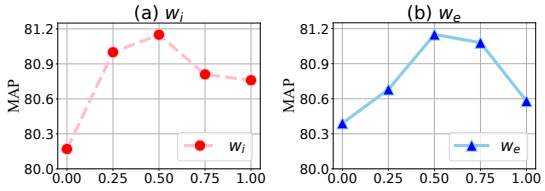


Figure 5: Hyper-parameter analysis on w_e and w_i. When adjusting the w_e, w_i is set as 0.5, and vice versa.

Model	Train Time (it/s)	Params (M)	FLOPs (T)
PBR	0.5732	212.65	3.577
EIGML	**0.5068**	**196.64**	**2.053**

Table 5: Computational cost comparison between EIGML and the strongest baseline PBR under the same setting with batch size 16, and a single A800 GPU. Best results in **bold**.

Hyper-parameter Analysis

We further investigate the hyperparameter used to control the weighting of emotion-related and intention-related losses, denoted as w_e (including w_e, w_e', w_e^t, and w_e^v) and w_i (including w_i, w_i^t, and w_i^v), respectively. As shown in Figure 5, we vary each weight from 0 to 1 in steps of 0.25, while keeping the other fixed at 0.5. The results reveal that performance, measured by MAP, is sensitive to both parameters. For w_e, the MAP peaks at 81.15% when $w_e = 0.5$, suggesting that moderate emphasis on emotion-related supervision yields optimal results. Similarly, the highest MAP for w_i also appears at 0.5, further confirming that balanced incorporation of intention information is most effective. Notably, performance drops when either weight is set to 0 (i.e., the corresponding module is removed), indicating that both emotional and intentional signals are essential for maximizing model performance. These findings validate the importance of jointly modeling both aspects and show that their contributions are best realized when neither dominates.

Computational Cost Analysis

To evaluate computational efficiency, we compare EIGML with the strongest baseline PBR under identical settings. As shown in Table 5, EIGML reduces per-iteration training time, number of parameters, and FLOPS by 11.59%, 7.53%, and 42.61%, respectively, achieving superior performance while substantially lowering computational overhead and validating its efficiency.

Conclusion

In this paper, we propose Emotion and Intention Guided Multi-Modal Learning, a novel framework for sticker response selection that jointly integrates emotional and intentional cues. By modeling emotion and intention together, it enhances contextual understanding and selection accuracy through consistent intra- and inter-modality alignment and progressive contextual knowledge infusion.

Acknowledgments

This research was partially supported by National Natural Science Foundation of China (No.62502404), Hong Kong Research Grants Council (Research Impact Fund No.R1015-23, Collaborative Research Fund No.C1043-24GF, General Research Fund No.11218325), Institute of Digital Medicine of City University of Hong Kong (No.9229503), Huawei (Huawei Innovation Research Program), Tencent (CCF-Tencent Open Fund, Tencent Rhino-Bird Focused Research Program), Alibaba (CCF-Alimama Tech Kangaroo Fund No. 2024002), Didi (CCF-Didi Gaia Scholars Research Fund), Kuaishou, and Bytedance, Innovation Team Project of Guangdong Province of China (No. 2024KCXTD017), Shenzhen Science and Technology Foundation (No. JCYJ20240813145816022).

References

An, M.; Wu, F.; Wu, C.; Zhang, K.; Liu, Z.; and Xie, X. 2019. Neural news recommendation with long-and short-term user representations. In *Proceedings of the 57th annual meeting of the association for computational linguistics*, 336–345.

Bosselut, A.; Rashkin, H.; Sap, M.; Malaviya, C.; Celikyilmaz, A.; and Choi, Y. 2019. COMET: Commonsense transformers for automatic knowledge graph construction. *arXiv preprint arXiv:1906.05317*.

Cai, H.; Shen, X.; Xu, Q.; Shen, W.; Wang, X.; Ge, W.; Zheng, X.; and Xue, X. 2023. Improving empathetic dialogue generation by dynamically infusing commonsense knowledge. *arXiv preprint arXiv:2306.04657*.

Chee, H. E. M.; Wang, J.; Guo, Z.; Ma, W.; and Zhang, M. 2025. PerSRV: Personalized Sticker Retrieval with Vision-Language Model. In *Proceedings of the ACM on Web Conference 2025*, 293–303.

Chen, J.; Cai, Y.; Xu, R.; Wang, J.; Xie, J.; and Li, Q. 2024a. Deconfounded emotion guidance sticker selection with causal inference. In *Proceedings of the 32nd ACM International Conference on Multimedia*, 3084–3093.

Chen, J.; Dong, X.; Wang, W.; Zhou, S.; Yu, L.; and Hu, X. 2025a. DERI: Cross-Modal ECG Representation Learning with Deep ECG-Report Interaction. In *Proceedings of the Thirty-Fourth International Joint Conference on Artificial Intelligence (IJCAI-25)*.

Chen, J.; Hu, Y.; Lai, Q.; Wang, W.; Chen, J.; Liu, H.; Srivastava, G.; Bashir, A. K.; and Hu, X. 2024b. IIFDD: Intra and inter-modal fusion for depression detection with multimodal information from Internet of Medical Things. *Information Fusion*, 102: 102017.

Chen, J.; Hu, Y.; Lu, H.; Wang, W.; Yang, M.; Li, C.; and Hu, X. 2025b. MGHFT: Multi-Granularity Hierarchical Fusion Transformer for Cross-Modal Sticker Emotion Recognition. arXiv:2507.18929.

Chen, J.; Wang, W.; Hu, Y.; Chen, J.; Liu, H.; and Hu, X. 2024c. Tgca-pvt: Topic-guided context-aware pyramid vision transformer for sticker emotion recognition. In *Proceedings of the 32nd ACM International Conference on Multimedia*, 9709–9718.

Deng, X. 2019. cnsenti: An Open-Source Python Library for Chinese Text Sentiment Analysis. https://github.com/hiDaDeng/cnsenti.

Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*, 4171–4186.

Fei, Z.; Li, Z.; Zhang, J.; Feng, Y.; and Zhou, J. 2021. 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.

Gao, S.; Chen, X.; Liu, L.; Zhao, D.; and Yan, R. 2021. Learning to respond with your favorite stickers: A framework of unifying multi-modality and user preference in multi-turn dialog. *ACM Transactions on Information Systems (TOIS)*, 39(2): 1–32.

Gao, T.; Yao, X.; and Chen, D. 2021. Simcse: Simple contrastive learning of sentence embeddings. *arXiv preprint arXiv:2104.08821*.

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.

Guo, D.; Xu, C.; and Tao, D. 2019. Image-question-answer synergistic network for visual dialog. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 10434–10443.

Hu, Y.; Tan, M.; Zhang, C.; Li, Z.; Liang, X.; Yang, M.; Li, C.; and Hu, X. 2024. Aptness: Incorporating appraisal theory and emotion support strategies for empathetic response generation. In *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management*, 900–909.

Jia, P.; Liu, Y.; Li, X.; Zhao, X.; Wang, Y.; Du, Y.; Han, X.; Wei, X.; Wang, S.; and Yin, D. 2024. G3: an effective and adaptive framework for worldwide geolocalization using large multi-modality models. *Advances in Neural Information Processing Systems*, 37: 53198–53221.

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.

Li, J.; Selvaraju, R.; Gotmare, A.; Joty, S.; Xiong, C.; and Hoi, S. C. H. 2021. Align before fuse: Vision and language representation learning with momentum distillation. *Advances in neural information processing systems*, 34: 9694–9705.

Li, X.; Song, J.; Gao, L.; Liu, X.; Huang, W.; He, X.; and Gan, C. 2019. Beyond rnns: Positional self-attention with

co-attention for video question answering. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, 8658–8665.

Li, X.; Zhao, X.; Xu, J.; Zhang, Y.; and Xing, C. 2023. IMF: interactive multimodal fusion model for link prediction. In *Proceedings of the ACM web conference 2023*, 2572–2580.

Liang, B.; Wang, B.; Bai, Z.; Lang, Q.; Sun, M.; Hou, K.; Zhou, L.; Xu, R.; and Wong, K.-F. 2025. Reply with Sticker: New Dataset and Model for Sticker Retrieval. *IEEE Transactions on Audio, Speech and Language Processing*.

Liang, J.; Zhao, X.; Li, M.; Zhang, Z.; Wang, W.; Liu, H.; and Liu, Z. 2023. Mmmlp: Multi-modal multilayer perceptron for sequential recommendations. In *Proceedings of the ACM Web Conference 2023*, 1109–1117.

Liu, Q.; Hu, J.; Xiao, Y.; Zhao, X.; Gao, J.; Wang, W.; Li, Q.; and Tang, J. 2024. Multimodal recommender systems: A survey. *ACM Computing Surveys*, 57(2): 1–17.

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.

Maaten, L. v. d.; and Hinton, G. 2008. Visualizing data using t-SNE. *Journal of machine learning research*, 9(Nov): 2579–2605.

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. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, 8748–8763. PmLR.

Radford, A.; Wu, J.; Child, R.; Luan, D.; Amodei, D.; Sutskever, I.; et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8): 9.

Shi, Y.; and Kong, F. 2024. Integrating Stickers into Multimodal Dialogue Summarization: A Novel Dataset and Approach for Enhancing Social Media Interaction. In *Proceedings of the 32nd ACM International Conference on Multimedia*, 9525–9534.

Tang, Y.; and Hew, K. F. 2019. Emoticon, emoji, and sticker use in computer-mediated communication: A review of theories and research findings. *International journal of communication*, 13: 2457–2483.

Wang, B.; Du, Y.; Liang, B.; Bai, Z.; Yang, M.; Wang, B.; Wong, K.-F.; and Xu, R. 2025a. A new formula for sticker retrieval: Reply with stickers in multi-modal and multi-session conversation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, 25327–25335.

Wang, B.; Liang, B.; Feng, C.-M.; Zuo, W.; Bai, Z.; Huang, S.; Wong, K.-F.; Zeng, X.; and Xu, R. 2024a. Towards Real-World Stickers Use: A New Dataset for Multi-Tag Sticker Recognition. *arXiv preprint arXiv:2403.05428*.

Wang, M.; Xiao, Y.; Wang, B.; Zhang, S.; Ye, S.; Wang, W.; Yin, H.; Guo, R.; and Xu, Z. 2025b. FindRec: Stein-Guided Entropic Flow for Multi-Modal Sequential Recommendation. In *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V. 2*, 3008–3018.

Wang, M.; Zhao, Y.; Liu, J.; Chen, J.; Zhuang, C.; Gu, J.; Guo, R.; and Zhao, X. 2024b. Large multimodal model compression via iterative efficient pruning and distillation. In *Companion Proceedings of the ACM Web Conference 2024*, 235–244.

Wang, Y.; Pan, J.; Li, X.; Wang, M.; Wang, Y.; Liu, Y.; Liu, D.; Jiang, J.; and Zhao, X. 2025c. Empowering Large Language Model for Sequential Recommendation via Multimodal Embeddings and Semantic IDs. In *Proceedings of the 34th ACM International Conference on Information and Knowledge Management*, 3209–3219.

Wu, Y.; Wang, S.; Song, G.; and Huang, Q. 2019. Learning fragment self-attention embeddings for image-text matching. In *Proceedings of the 27th ACM international conference on multimedia*, 2088–2096.

Xia, W.; Liu, S.; Rong, Q.; Jia, G.; Park, E.; and Yang, J. 2024. Perceive before respond: Improving sticker response selection by emotion distillation and hard mining. In *Proceedings of the 32nd ACM International Conference on Multimedia*, 9631–9640.

Xiong, J.; Li, Z.; Zheng, C.; Guo, Z.; Yin, Y.; Xie, E.; Yang, Z.; Cao, Q.; Wang, H.; Han, X.; et al. 2023. Dq-lore: Dual queries with low rank approximation re-ranking for in-context learning. *arXiv preprint arXiv:2310.02954*.

Zhang, C.; Hu, Y.; Yang, M.; Li, C.; and Hu, X. 2023. Skeletal spatial-temporal semantics guided homogeneous-heterogeneous multimodal network for action recognition. In *Proceedings of the 31st ACM International Conference on Multimedia*, 3657–3666.

Zhang, C.; Zhang, H.; Wu, S.; Wu, D.; Xu, T.; Zhao, X.; Gao, Y.; Hu, Y.; and Chen, E. 2025. Notellm-2: Multimodal large representation models for recommendation. In *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V. 1*, 2815–2826.

Zhang, K.; Mao, Z.; Wang, Q.; and Zhang, Y. 2022a. Negative-aware attention framework for image-text matching. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 15661–15670.

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, Y.; and Lu, H. 2018. Deep cross-modal projection learning for image-text matching. In *Proceedings of the European conference on computer vision (ECCV)*, 686–701.

Zhang, Z.; Zhu, Y.; Fei, Z.; Zhang, J.; and Zhou, J. 2022b. Selecting stickers in open-domain dialogue through multi-task learning. *arXiv preprint arXiv:2209.07697*.

Zhao, S.; Yao, X.; Yang, J.; Jia, G.; Ding, G.; Chua, T.-S.; Schuller, B. W.; and Keutzer, K. 2021. Affective image content analysis: Two decades review and new perspectives. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(10): 6729–6751.