

Appendix - Deconfounded Emotion Guidance Sticker Selection with Causal Inference

A More Experimental Details

A.1 Experiment on Hyperparameters

We conduct experiments to explore different hyperparameter settings of the loss function in Eq. 14, including $(\lambda_1, \lambda_2) = \{(0.3, 0.3), (0.5, 0.3), (0.7, 0.3), (0.5, 0.1), (0.5, 0.3), (0.5, 0.5)\}$ and find that the fluctuations in the results are minimal. Therefore, we set hyperparameters λ_1 and λ_2 for the loss function are 0.5 and 0.3, respectively.

A.2 Implementation of Baseline Models

In this subsection, details of baseline models are shown as follows:

- **CLIP** is an encoder-based vision-language model. For the sticker selection task, we fine-tune the pre-trained CLIP model with contrastive loss. The batch size is configured as 64 and the AdamW optimizer in conjunction with a cosine scheduler is for optimization. Specifically, we restrict our textual input to only the last utterance in the dialogue, since incorporating additional utterances leads to a decrease in performance.
- **ViLT** jointly process the textual sequence and image information with a unified transformer encoder. Specifically, we directly fine-tune the ViLT model, which is pre-trained on the image-text retrieval dataset and optimize the model with image-text matching loss.
- **MOD-GPT** is a language model based method, which uses dialogue context as input and adopts a GPT-2 [1] model to match the sticker. Specifically, it directly inputs the embedding of the sticker image into GPT-2. Subsequently, the similarity between the GPT-2 output embedding and features of candidate stickers is computed for sticker selection.
- **MMBERT** is currently the state-of-the-art model in the sticker selection task. It proposes a multitask learning framework with three auxiliary tasks (i.e., emotion classification, sticker semantic prediction and masked context prediction). Specifically, it adopts a multimodal BERT model as the backbone for sticker selection.
- **BLIP-2** is an encoder-decoder based vision-language model. In our implementation, we only fine-tune the image-text matching classifier to align more closely with our sticker selection objectives without significantly altering the parameters of the pre-trained model.
- **FROMAGE** utilizes a pre-trained frozen large language model (LLM) (i.e., OPT-6.7B [2]) for the multimodal dialogue. It is designed for image retrieval and text generation tasks. More importantly, the model has been pre-trained on a large scale of multimodal dialogue data and achieved zero-shot solid performance, we directly apply it to the sticker selection task without additional fine-tunings.

B Formula Derivations

In this paper, we utilize the Normalized Weighted Geometric Mean (NWGM) approximation to compute Eq. 7 of the main paper, i.e., $P(Y | do(V)) = \sum_i P(w_i)P(Y | V, w_i)$. To comprehensively understand the approximation process, we first provide the definition

of Weighted Geometric Mean (WGM) of an exponential function $y(x) = \exp[f(x)]$:

$$\begin{aligned} \text{WGM}(y(x)) &= \prod_x y(x)^{P(x)} \\ &= \prod_x \exp[f(x)]^{P(x)} \\ &= \prod_x \exp[f(x)P(x)] \\ &= \exp \left[\sum_x f(x)P(x) \right] \\ &= \exp \{ \mathbb{E}_x[f(x)] \}, \end{aligned} \quad (1)$$

where \mathbb{E}_x is absorbed into the exponential term, the expected value of $y(x)$ can be approximated by WGM:

$$\begin{aligned} \mathbb{E}_x[y(x)] &= \sum_x y(x)P(x) \\ &\approx \text{WGM}(y(x)) \\ &\approx \exp \{ \mathbb{E}_x[f(x)] \}. \end{aligned} \quad (2)$$

Therefore, the Normalized Weighted Geometric Mean (NWGM) approximation can be represented as:

$$\begin{aligned} \text{NWGM}(y(x)) &= \frac{\prod_x \exp(f(x))^{P(x)}}{\sum_j \prod_x \exp(f(x))^{P(x)}} \\ &= \frac{\exp(\mathbb{E}_x[f(x)])}{\sum_j \exp(\mathbb{E}_x[f(x)])} \\ &= \text{softmax}(\mathbb{E}_x[f(x)]). \end{aligned} \quad (3)$$

In our implementation, we use $\text{softmax}(f_y(V, w_i))$ to emulate the calculation of $P(Y|V, w_i)$, where $f_y(\cdot)$ is neural network layers and $\text{softmax}(f_y(V, w_i)) \propto \exp[f_y(V, w_i)]$. Then, the Eq. 7 of the main paper can be newly formulated as:

$$P(Y | do(V)) = \mathbb{E}_{w_i} [\text{softmax}(f_y(v_p, w_i))]. \quad (4)$$

According to the Eq. 3, the above expectation can be approximated by Normalized Weighted Geometric Mean (NWGM) as:

$$\begin{aligned} P(Y | do(V)) &\approx \text{WGM}(P(Y | v_p, w)) \\ &\approx \text{softmax}(\mathbb{E}_{w_i} [f_y(v_p, w_i)]). \end{aligned} \quad (5)$$

Finally, we obtain the visual features after intervention \bar{v} with a scaled dot-product attention as:

$$\begin{aligned} \bar{v} &= \text{softmax}(\mathbb{E}_{w_i} [f_y(v_p, w_i)]) \\ &= \text{softmax}(\text{FC}((v_p^T W)W)), \end{aligned} \quad (6)$$

where $\bar{v} \in \mathbb{R}^{d^r}$ and $\text{FC}(\cdot)$ is a fully-connected layer.

C Failure Case

Fig. 1 shows a failure case in our model that accurately predicts the speaker's emotion but neglects the individual preference of the speaker during sticker selection. Specifically, in the candidate set, there are two appropriate happy-themed stickers for the dialogue.

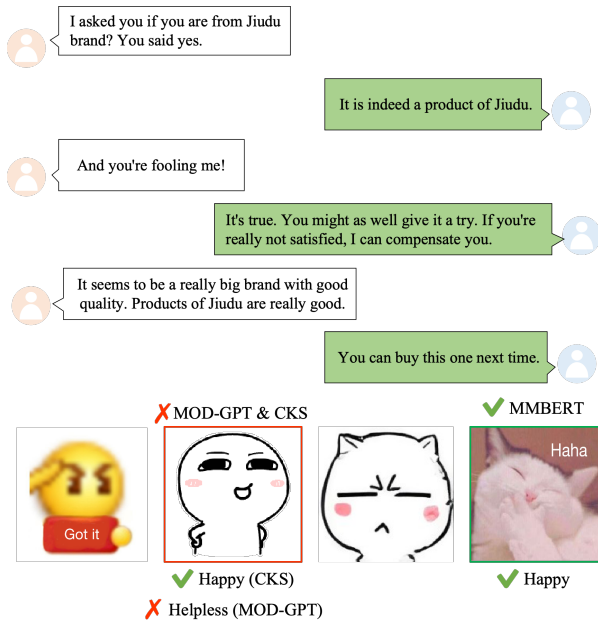


Figure 1: The failure case of the selected sticker and predicted emotion from baselines (i.e., MOD-GPT and MMBERT) and our CKS model.

The speaker typically selects the one that aligns best with the unique preference. Therefore, we can explicitly model user portraits in the future to achieve more accurate sticker selection.

References

- [1] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog* 1, 8 (2019), 9.
- [2] Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuo-hui Chen, Christopher Dewan, Mona T. Diab, Xian Li, Xi Victoria Lin, Todor Mihaylov, Myale Ott, Sam Shleifer, Kurt Shuster, Daniel Simig, Punit Singh Koura, Anjali Sridhar, Tianlu Wang, and Luke Zettlemoyer. 2022. OPT: Open Pre-trained Transformer Language Models. *CoRR* abs/2205.01068 (2022).