FACEPHI: LIGHTWEIGHT MULTIMODAL LARGE LANGUAGE MODEL FOR FACIAL LANDMARK EMOTION RECOGNITION

Hongjin Zhao[1], **Zheyuan Liu**[1], **Jiaxu Liu**[2], **Tom Gedeon**[3]

[1]*Australian National University,* [2]*University of Liverpool,* [3]*Curtin University*

firstname.lastname@anu.edu.au jiaxu.liu@liverpool.ac.uk tom.gedeon@curtin.edu.au

ABSTRACT

We introduce FacePhi, a multimodal large language model (LLM) for emotion recognition through facial landmarks. By focusing on facial landmarks, FacePhi ensures privacy preservation in emotion detection tasks. FacePhi is optimized for computational efficiency by incorporating Phi-2, a LLM with a small number of parameters, as well as utilizing lightweight facial landmark data. This design choice makes FacePhi suitable for deployment in resource-constrained settings. Our investigation highlights the importance of feature alignment during the training phase, indicating its pivotal role in enhancing the model's performance for the challenging task of facial landmark emotion recognition.

1 INTRODUCTION

The introduction of ChatGPT (OpenAI, 2022) has heightened interest in multimodal large language models (LLMs), leading to the release of numerous LLMs, such as LLaMA (Touvron et al., 2023a), LLaMA2 (Touvron et al., 2023b) and Mistral-7B (Jiang et al., 2023). Recently, there is a trend to leverage a single multimodal LLM for diverse application tasks (Xu et al., 2023; Liu et al., 2023). In this work, we investigate the applicability of multimodal LLMs for facial landmark emotion recognition. We aim to empower multimodal LLMs with the capability to interpret human facial expressions, utilizing facial landmark data. This approach is chosen with the dual aim of ensuring user privacy protection and reducing computational cost.

It is non-trivial to implement a multimodal LLM for facial landmark emotion recognition. Initially, we attempt to fine-tune PointLLM (Xu et al., 2023), a multimodal LLM for point cloud understanding, for facial landmark emotion recognition using LoRA (Hu et al., 2021). This approach results in a near-random accuracy in performance. This paper aims to explore training methods for facial landmark emotion recognition with multimodal LLMs. Inspired by PointLLM (Xu et al., 2023), we discover that it is pivotal to pretrain the encoder for feature alignment before finetuning the entire multimodal LLM.

We summarize our contributions as follows:

- We introduce FacePhi, a lightweight multimodal LLM for facial landmark emotion recognition. To the best of our knowledge, we are the first to conduct facial landmark emotion recognition with multimodal LLMs.
- We demonstrate the significance of pertaining the encoder with feature alignment prior to training multimodal LLM for facial landmark emotion recognition.

2 RELATED WORK

2.1 FACIAL LANDMARK EMOTION RECOGNITION

Tautkute et al. (2018) developed an enhanced model named EmotionalDAN, which integrates facial landmarks into the classification loss function. The performance was good, enabling EmotionalDAN to surpass current state-of-the-art sentiment classification methods by a margin of up to 5% on two challenging benchmark datasets. The effectiveness of EmotionalDAN underscores the benefits of utilizing facial landmarks within emotion recognition tasks. Currently, there is a gap in the literature, as there is no work on the use of multimodal Large Language Models for emotion recognition, which we intend to fill.

2.2 LARGE LANGUAGE MODELS AND MULTIMODAL ADAPTATIONS

Large Language Models (LLMs) aim to process and generate natural language responses to given inputs (Touvron et al., 2023a). Traditionally, LLMs such as LLaMA (Touvron et al., 2023a) and LLaMA2 (Touvron et al., 2023b) have focused solely on textual data. Recently, increasing interests have arisen in multimodal LLMs. These models extend the capabilities of LLMs to interpret and analyze diverse data types, including images (Liu et al., 2023; Zhou et al., 2023), point clouds (Xu et al., 2023; Qi et al., 2023), and videos (Lin et al., 2023).

The recent work, Phi-2 (Microsoft, 2023), in contrast to its predecessors' trend for making LLMs larger, contains a mere 2.7 billion parameters. Despite its smaller size, Phi-2 demonstrates superior performance across multiple tasks when compared to larger models, such as LLaMA variants (Touvron et al., 2023a;b) and Mistral-7B (Jiang et al., 2023). In this work, we leverage Phi-2 (Microsoft, 2023) and introduce a lightweight and efficient multimodal LLM designed for facial landmark emotion recognition.

2.3 FEATURE ALIGNMENT

Improving the precision and reliability of large language models (LLMs) in language understanding and generation necessitates the use of text alignment training. One seminal work in this area is CLIP (Radford et al., 2021). CLIP utilizes contrastive learning to discern the relationships between images and their textual descriptions across large datasets, significantly enhancing its performance in various visual tasks.

Similarly, ULIP (Xue et al., 2023) serves as a platform for achieving a cohesive representation of images, text, and 3D point clouds. By leveraging the connections between these different types of data, ULIP trains models to understand and generate language based on a wide range of inputs. This approach promotes the integration of multimodal data into a unified learning model, improving the model's language processing capabilities. Here, we finetune ULIP (Xue et al., 2023) to produce the landmark embeddings that are aligned with the text modality, which is subsequently injected into the LLM for reasoning.

3 METHODS

3.1 FACEPHI ARCHITECTURE

In this section, we elaborate on the details of our FacePhi model, which comprises of two main components, an encoder and a Large Language Model (LLM), as illustrated in Figure 1. We note that in this work, we adopt PointMLP (Ma et al., 2022) finetuned through the ULIP (Xue et al., 2023) pipeline (see Section 3.3) as the encoder, and Phi-2 (Microsoft, 2023) as the LLM. However, our method is agnostic to the choices of the components. A simple MLP-based projection layer connects these two components. Details are as follows.

Encoder. The encoder, denoted M_{Encoder}, is designed to accept a sequential input of facial landmark $w \in \mathbb{R}^{N \times C}$, where N represents the number of points/landmarks and C denotes the number of channels. For the facial landmark dataset adopted in this work, $N = 478$, while $C = 3$,

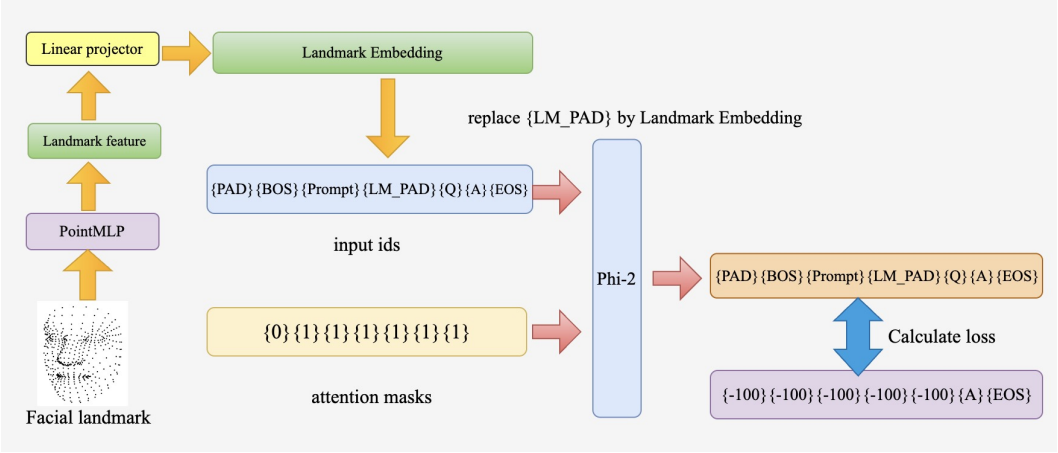

Figure 1: **Model Architecture.** The facial landmark feature will be extracted using PointMLP (Ma et al., 2022), and this feature will be fed into a linear projection layer to ensure that its hidden size matches the input embedding's hidden size. Subsequently, this landmark embedding will be inserted into the Large Language Model (LLM) Phi-2 (Microsoft, 2023) by replacing `[LM PAD]`. We refer readers to Appendix A.2 for details on the tokens noted in this figure. Here, we also demonstrate the attention masks that couple the input tokens, where 0 suggests the corresponding token (in this case, the padding `[PAD]`) is masked and ignored. The output of the model is used for computing the loss, as in Equation 3.

i.e., the three-dimensional coordinates for each point. The encoder processes this input into a one-dimensional feature $e \in \mathbb{R}^d$, where d is the embedding dimension. For PointMLP (Ma et al., 2022), we use the default $d = 256$. The function of M_{Encoder} is then expressed as

$$e = M_{\text{Encoder}}(w). \tag{1}$$

Following the encoder, we insert a simple MLP-based projection layer, whose purpose is to reshape the feature vector $e \in \mathbb{R}^d$ to $e' \in \mathbb{R}^{1\times h}$, where h represents the hidden dimension of the subsequent LLM. To this end, the reshaped embedding is effectively a token embedding for the input of the LLM, whose details are as follows.

Large Language Model (LLM). The M_{LLM} reasons over the facial landmark embedding alongside a given text prompt for an answer in natural language, where the construction of the text prompt is elaborated in Appendix A.1. Incorporating facial landmark embeddings into the input sequence of the LLM is straightforward, as we only need to concatenate e' with the text prompt sequence t. Since we make no significant changes to the architecture of the LLM, we refer readers to its original work (Microsoft, 2023) for details.

3.2 Training and Evaluation

Following previous works (Liu et al., 2023; Xu et al., 2023; Qi et al., 2023), we employ a two-stage training process. Initially, we freeze the parameters of both the LLM and the facial landmark encoder, directing our training focus on the projection layer that serves as a bridge between the encoder and the language model. In the subsequent phase, we enable training for the large language model. The model's performance is evaluated based on its ability to predict the end of a sequence and generate accurate responses, with loss measured through the negative log-likelihood of the correct word predictions.

Loss function. Mirroring Phi-2 (Microsoft, 2023), the loss for a single predicted token can be formulated as:

$$\mathcal{L}_{\text{token}} = -\log P\left(x_T | x_1, x_2, \ldots, x_{T-1}; \theta\right), \tag{2}$$

where $P\left(x_T|x_1, x_2, \ldots, x_{T-1}; \theta\right)$ is the probability assigned by the model parameterized by θ to the correct next token x_T given the previous words in the sequence.

For a batch of N predictions with multiple sequences, the total loss can be expressed as:

$$\mathcal{L}_{\text{batch}} = -\sum_{i=1}^{N} \log P\left(x_T^{(i)}|x_1^{(i)}, x_2^{(i)}, \ldots, x_{T-1}^{(i)}; \theta\right). \tag{3}$$

Evaluation Metric. We formulate the task as a classification task. In essence, the aim is for the model to correctly classify the emotion corresponding to the facial landmark against the pre-defined categories, where we assess the accuracy. To match the natural language output of the LLM against the category labels, we employ a simple keyword-matching scheme. For instance, if the model outputs "This landmark denotes an emotion of happiness." and the corresponding keyword label is 'happiness', the prediction is deemed accurate. Likewise, if the pre-defined keyword labels are not present in the answer, the prediction is deemed inaccurate.

3.3 Feature Alignment

Recall that we utilize PointMLP (Ma et al., 2022) as the facial landmark encoder. Prior to integrating such a module into the LLM pipeline, we first pretrain it following ULIP (Xue et al., 2023), such that its produced embeddings are better aligned with the text modality of the LLM. Note that ULIP (Xue et al., 2023) introduces a tri-modal pre-training strategy that synchronizes text, 3D structural information, and images. However, we do not use images in our setup.

We utilize a handcrafted template to generate text that aligns with the specific emotion of each facial landmark. As an example, the label for a facial landmark indicative of happiness is formatted as "The facial landmark shows emotion of happiness.".

Here, our training objective is the typical contrastive loss as used by CLIP (Radford et al., 2021). The formula for the loss function while training, given a batch of N image-text pairs, is as follows

$$\mathcal{L}_{\text{batch}} = \frac{1}{2N}\sum_{i=1}^{N}\left[-\log\frac{\exp(\kappa(w_i, t_i)/\tau)}{\sum_{j=1}^{N}\exp(\kappa(w_i, t_j)/\tau)} - \log\frac{\exp(\kappa(w_i, t_i)/\tau)}{\sum_{k=1}^{N}\exp(\kappa(w_k, t_i)/\tau)}\right], \tag{4}$$

where $\kappa(\cdot, \cdot)$ is the cosine similarity kernel, implemented as a normalization followed by dot-product. t_i is the text that corresponds to the i-th facial landmark w_i. τ is a learnable temperature parameter that scales the logits before applying the softmax function, as in (Radford et al., 2021).

4 Experiments

Dataset. Considering the scarcity of emotion recognition datasets containing facial landmarks, this study leveraged the mediapipe algorithm (Lugaresi et al., 2019) to extract a total of 478 distinct facial landmarks per image from the AffectNet dataset (Mollahosseini et al., 2017), which contains images categorized into eight basic emotion classes [1]. Every point of the 478 facial landmarks is described by a three-dimensional vector, representing its spatial characteristics, as illustrated in Figure 1 (bottom-left).

Next, we employ ChatGPT4 (OpenAI, 2022) to craft 15 distinct input templates and 10 unique output templates for each category label (see Appendix A.1). Throughout the data generation process, we randomly choose an input template and subsequently generate an output template based on the corresponding label. This ensures that each data entry consists of one input template, one output template, and one facial landmark. Ultimately, after processing, our dataset encompasses approximately 28,000 samples for the training set and 4,000 samples for the validation set. Please refer to the Appendix A.2 for details on data pre-processing.

[1] Eight basic emotion classes are neutral, happiness, sadness, surprise, fear, disgust, anger and contempt.

Table 1: **Emotion Recognition Result on Different LLMs.** The LLMs in the second, third and fifth rows of the table have been fine-tuned, while the models in the other two rows have not. † indicates that ChatGPT4 cannot handle this task.

Multimodal LLMs	**Accuracy (in %)**
ChatGPT4 (OpenAI, 2022)	N/A †
PointLLM (Xu et al., 2023)	11.4
PointLLM (Xu et al., 2023) + LoRA (Hu et al., 2021)	9.3
PointMLP (Ma et al., 2022) + Phi-2 (Microsoft, 2023)	10.2
FacePhi (Ours)	**27.7**

Implementation Details. All experiments were carried out using a single NVIDIA A100 80G, utilizing the float16 datatype for both data and model. The total duration for training amounts to 15 hours.

While fine-tuning the Large Language Model (LLM), for the first stage, the batch size and learning rate are set at 48 and 2.5×10^{-3} respectively. In the second phase, the batch size is decreased to 32 and the learning rate is reduced to 2.5×10^{-5}. We employ AdamW (Loshchilov & Hutter, 2017) as the optimizer for both stages and train each stage for three epochs.

For feature alignment of the PointMLP (Ma et al., 2022) through ULIP (Xue et al., 2023) (Section 3.3), we used AdamW (Loshchilov & Hutter, 2017) as our optimizer and a learning rate of 2×10^{-5}. We train for 20 epochs with a batch size of 64.

4.1 Results and Ablation Studies

Accuracy of FacePhi. As shown in Table 1, the model's accuracy in facial landmark emotion recognition is 28%. We note that this is an exceedingly challenging task. To demonstrate, we also showcase the performance of existing multimodal LLMs. We note that PointLLM (Xu et al., 2023), a recent model pretrained specifically on point cloud data, achieves a much lower accuracy of 11.4%. To bridge the domain gap between conventional 3D point clouds and facial landmarks, we attempt finetuning PointLLM through LoRA (Hu et al., 2021). However, this results in an even lower accuracy of 9.3%, suggesting that closing the domain gap is non-trivial. These outcomes are considered suboptimal for a task involving eight distinct classes, where a theoretical random guess shall yield an average accuracy of 12.5%. It also validates that our method, in comparison, is indeed effective.

We additionally test on ChatGPT4 (OpenAI, 2022). At the time of writing, ChatGPT4 can accept and process `numpy` arrays through Python. To this end, we input the facial landmark data in such a format and test its ability to predict the corresponding emotions. To our disappointment, the model explicitly expresses its inability to perform the emotion recognition task consistently. A typical answer from the model is demonstrated in Appendix A.3, which shows that ChatGPT4 cannot successfully reason over such data.

Necessity of Feature Alignment. Here, we investigated the necessity of feature alignment (Section 3.3) for the facial landmark encoder. Specifically, we train PointMLP (Ma et al., 2022) with a traditional classification task instead of the ULIP pipeline (Xue et al., 2023). After training, the classifier achieves an admirable accuracy of 50%. We then use this as the facial landmark encoder of our FacePhi model and proceed with the training of the Large Language Model. Surprisingly, this configuration only yields an accuracy of 10.2%, as demonstrated in Table 1. We note that this, in comparison to our main result of 27.7% accuracy, effectively constitutes an ablation study that underscores the significance of feature alignment while training multimodal LLMs.

5 Conclusion and Future Work

This paper introduces FacePhi, an multimodal large language model designed to process facial landmarks and recognize emotions without the need for extensive computational resources. FacePhi

demonstrates promising results, achieving a 28% accuracy. Our findings indicate that for multimodal LLMs, fine-tuning the encoder with a text alignment method is essential. Moving forward, we plan to enable large models to accurately recognize facial landmark emotions without decreasing their performance on other tasks.

REFERENCES

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models, 2021.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. Mistral 7b, 2023.

Bin Lin, Yang Ye, Bin Zhu, Jiaxi Cui, Munan Ning, Peng Jin, and Li Yuan. Video-llava: Learning united visual representation by alignment before projection, 2023.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning, 2023.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.

Camillo Lugaresi, Jiuqiang Tang, Hadon Nash, Chris McClanahan, Esha Uboweja, Michael Hays, Fan Zhang, Chuo-Ling Chang, Ming Guang Yong, Juhyun Lee, et al. Mediapipe: A framework for building perception pipelines. *arXiv preprint arXiv:1906.08172*, 2019.

Xu Ma, Can Qin, Haoxuan You, Haoxi Ran, and Yun Fu. Rethinking network design and local geometry in point cloud: A simple residual mlp framework. *arXiv preprint arXiv:2202.07123*, 2022.

Microsoft. Phi-2: The surprising power of small language models. `https://www.microsoft.com/en-us/research/blog/phi-2-the-surprising-power-of-small-language-models/`, 2023. Accessed: February 3, 2024.

Ali Mollahosseini, Behzad Hasani, and Mohammad H Mahoor. Affectnet: A database for facial expression, valence, and arousal computing in the wild. *IEEE Transactions on Affective Computing*, 10(1):18–31, 2017.

OpenAI. Introducing chatgpt. `https://openai.com/blog/chatgpt`, 2022. Accessed: 2024-02-08.

Zhangyang Qi, Ye Fang, Zeyi Sun, Xiaoyang Wu, Tong Wu, Jiaqi Wang, Dahua Lin, and Hengshuang Zhao. Gpt4point: A unified framework for point-language understanding and generation, 2023.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision, 2021.

Ivona Tautkute, Tomasz Trzcinski, and Adam Bielski. I know how you feel: Emotion recognition with facial landmarks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, June 2018.

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

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

Runsen Xu, Xiaolong Wang, Tai Wang, Yilun Chen, Jiangmiao Pang, and Dahua Lin. Pointllm: Empowering large language models to understand point clouds. *arXiv preprint arXiv:2308.16911*, 2023.

Le Xue, Mingfei Gao, Chen Xing, Roberto Martín-Martín, Jiajun Wu, Caiming Xiong, Ran Xu, Juan Carlos Niebles, and Silvio Savarese. Ulip: Learning a unified representation of language, images, and point clouds for 3d understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1179–1189, 2023.

Juexiao Zhou, Xiaonan He, Liyuan Sun, Jiannan Xu, Xiuying Chen, Yuetan Chu, Longxi Zhou, Xingyu Liao, Bin Zhang, and Xin Gao. Skingpt-4: An interactive dermatology diagnostic system with visual large language model, 2023.

A Appendix

A.1 Data Generation

The generation of the AffectLandmarkNet dataset involves a process where instructions from Table 2 are selected randomly, and corresponding answers are also chosen at random based on the labels indicated in Table 3. For instance, if the label for a particular sample is 'Happiness', the input text, inclusive of special tokens, might appear as follows:

"`[PAD] [BOS]` Below is a conversation between a curious user and an AI agent. The assistant gives helpful, detailed, and polite answers to the user's questions, Users: `[LM PAD]` Can you identify the emotion displayed by this 3D facial landmark? AI: It is probable that this 3D facial landmark was created with the intention of embodying the emotion of Happiness. `[EOS]`"

In this context, the text for which the attention mask is set to one would be:

"`[EOS]` Below is a conversation between a curious user and an AI agent. The assistant gives helpful, detailed, and polite answers to the user's questions, Users: `[LM PAD]` Can you identify the emotion displayed by this 3D facial landmark? AI: It is probable that this 3D facial landmark was created with the intention of embodying the emotion of Happiness. `[EOS]`"

Furthermore, the text segment used for loss computation would be:

"It is probable that this 3D facial landmark was created with the intention of embodying the emotion of Happiness. `[EOS]`"

A.2 Data Pre-processing

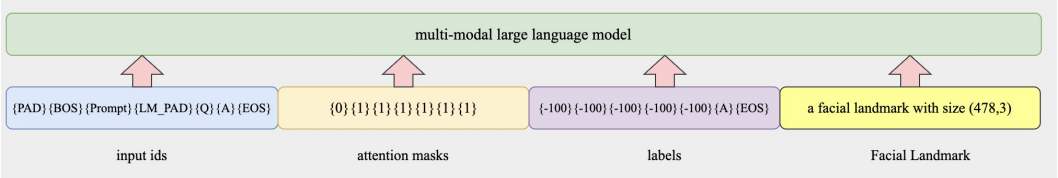

Figure 2: **Input details.** `[PAD]` is the set of pad tokens, `[BOS]` is the begin token, `[prompt]` is the system prompt used for pretraining, `[LM PAD]` is the landmark pad token. `[Q]` and `[A]` are question tokens and answer tokens,respectively. `[EOS]` is the end token.

To enhance the generalization capability of the large language model, facial landmark coordinates are normalized, ensuring their distribution uniformly spans a range between -1 and 1.

Moreover, to facilitate batch processing of inputs with variable text lengths, the integration of padding tokens is imperative to standardize the length of the texts. In this context, we introduced the token `[PAD]` to serve as a padding token for the tokenizer, harmonizing the length of shorter texts with a predefined benchmark. Conversely, texts exceeding this benchmark are truncated to

Table 2: **Instructions Templates** for 3D Facial Landmark Analysis

No.	Instruction
1	Can you identify the emotion displayed by this 3D facial landmark?
2	Look at this 3D facial landmark and judge the emotion it's showing.
3	What emotion is being expressed by this 3D facial landmark?
4	Determine the emotion depicted on this 3D facial landmark.
5	Examine this 3D facial landmark and select the emotion it's exhibiting.
6	Can you spot the emotion on this 3D facial landmark?
7	Identify which emotion is portrayed on this 3D facial landmark.
8	Which emotion is this 3D facial landmark expressing?
9	Identify the emotion on this 3D facial landmark.
10	Judge the emotion that this 3D facial landmark is displaying.
11	Can you tell which emotion this 3D facial landmark is portraying?
12	This 3D facial landmark is expressing an emotion.
13	Determine the emotion is shown by this 3D facial landmark.
14	Evaluate this 3D facial landmark and judge the emotion.
15	Review the expression of this 3D facial landmark and judge its emotion.

conform to the established length. Additionally, `[LM PAD]` is employed to denote the location for the insertion of the landmark embedding within the input embedding.

In the attention mask, positions corresponding to padding tokens are assigned a value of 0, whereas all other tokens retain a value of 1. Regarding the labels, only the response tokens and the end token are preserved, while all remaining tokens are assigned a value of -100. After data pre-processing, each data sample has input ids, attention mask, labels and a facial landmark. Figure 2 depicts the details of the inputs of large language model.

A.3 Experimental Results from ChatGPT4

As shown in Figure 3, we showcase a typical result obtained from ChatGPT4 (OpenAI, 2022), which demonstrates its inability to reason over the facial landmark data.

Table 3: **Instructions Templates** for Happiness Emotion

No.	Instruction
1	It is probable that this 3D facial landmark was created with the intention of embodying the emotion of Happiness.
2	It is likely that the 3D facial landmark was designed to capture the essence of the emotion Happiness.
3	There is a strong likelihood that the 3D facial landmark was developed to represent the emotion of Happiness.
4	There is a good chance that the 3D facial landmark was specifically made to emulate the Happiness emotion.
5	The 3D facial landmark was likely developed with the aim of symbolizing the Happiness emotion.
6	It is quite possible that the 3D facial landmark was crafted to convey the feeling of Happiness.
7	It is reasonable to assume that the 3D facial landmark was designed to embody Happiness.
8	The 3D facial landmark was crafted to convey the feeling of Happiness.
9	The 3D facial landmark was presumably crafted with the goal of reflecting the emotion of Happiness.
10	The 3D facial landmark was created with the express purpose of representing the emotion of Happiness.

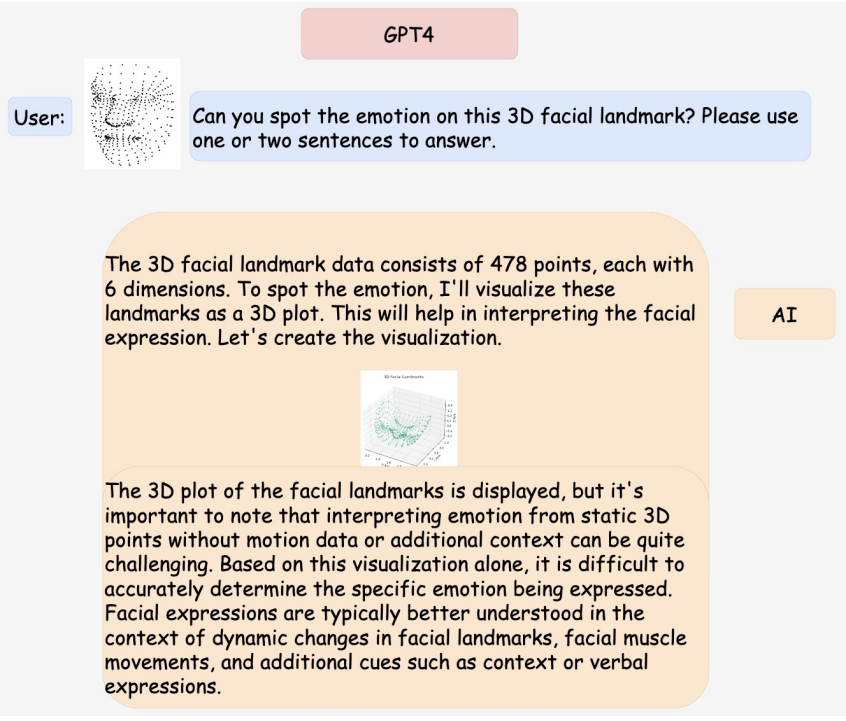

Figure 3: **Results obtained from ChatGPT4.** ChatGPT4 (OpenAI, 2022) engages in conversations with users regarding facial landmark emotion recognition. However, it does not process the ability to reason over such inputs.