Emo3D: Metric and Benchmarking Dataset for 3D Facial Expression Generation from Emotion Description

Mahshid Dehghani[◇,†], **Amirahmad Shafiee**[◇,†]*, **Ali Shafiei**[◇,†]*, **Neda Fallah**[◇,†], **Farahmand Alizadeh**[◇,†], **Mohammad Mehdi Gholinejad**[◇,†], **Hamid Behroozi**[◇], **Jafar Habibi**[◇], **Ehsaneddin Asgari**[§]

[◇] Sharif University of Technology

[†] NLP & DH Lab, Computer Engineering Department, Sharif University of Technology

[§] Qatar Computing Research Institute, Doha, Qatar

easgari@hbku.edu.qa

Abstract

3D facial emotion modeling has important applications in areas such as animation design, virtual reality, and emotional human-computer interaction (HCI). However, existing models are constrained by limited emotion classes and insufficient datasets. To address this, we introduce Emo3D, an extensive "Text-Image-Expression dataset" that spans a wide spectrum of human emotions, each paired with images and 3D blendshapes. Leveraging Large Language Models (LLMs), we generate a diverse array of textual descriptions, enabling the capture of a broad range of emotional expressions. Using this unique dataset, we perform a comprehensive evaluation of fine-tuned language-based models and vision-language models, such as Contrastive Language-Image Pretraining (CLIP), for 3D facial expression synthesis. To better assess conveyed emotions, we introduce Emo3D metric, a new evaluation metric that aligns more closely with human perception than traditional Mean Squared Error (MSE). Unlike MSE, which focuses on numerical differences, Emo3D captures emotional nuances in visual-text alignment and semantic richness. Emo3D dataset and metric hold great potential for advancing applications in animation and virtual reality.

1 Introduction

Automatic translation of character emotions into 3D facial expressions is an important task in digital media, owing to its potential to enhance user experience and realism. Facial Expression Generation (FEG) has a wide range of applications across various industries, including game development, animation, film production, and virtual reality. Previous studies in this domain have primarily focused on generating facial expressions for 2D or 3D characters, often relying on a limited set of predefined classes (Siddiqui, 2022) or driven by audio cues (Karras et al., 2017; Peng et al., 2023). However, there is a growing need for better control in the generation of complex and diverse human facial expressions. Recent studies (Zou et al., 2023; Zhong et al., 2023; Ma et al., 2023) have made notable progress in this area through the use of text prompts, offering a more direct approach to address the challenge of limited control that has been prevalent in earlier works (Siddiqui, 2022; Karras et al., 2017; Peng et al., 2023).

The primary issue with recent works using text prompts is (i) their limited focus on textual descriptions of emotions. Many studies have not deeply explored emotional context. These studies have not offered a comprehensive solution that integrates both textual descriptions and 3D FEG, creating a noticeable gap in the field (Zhong et al., 2023; Zou et al., 2023). Moreover, there is (ii) a scarcity of datasets containing emotional text alongside corresponding 3D facial expressions, impeding the development and training of FEG models for practical applications (Zhong et al., 2023; Zou et al., 2023; Ma et al., 2023). Additionally, (iii) the absence of reliable benchmarks and standardized evaluation metrics in this research area further complicates the assessment of FEG models.

Contributions: This paper tackles key challenges in FEG, focusing on generating expressions from textual emotion descriptions. Our contributions towards addressing the gaps in the field of FEG are as follows:

(i) Emo3D-dataset: We present the Emo3D-dataset, specifically developed to bridge the gap between textual emotion descriptions and 3D FEG. This dataset provides a rich compilation of annotated emotional texts alongside matching 3D expressions for effective training and assessment of FEG models.

*These authors contributed equally to this work

Findings of the Association for Computational Linguistics: NAACL 2025, pages 3158–3172

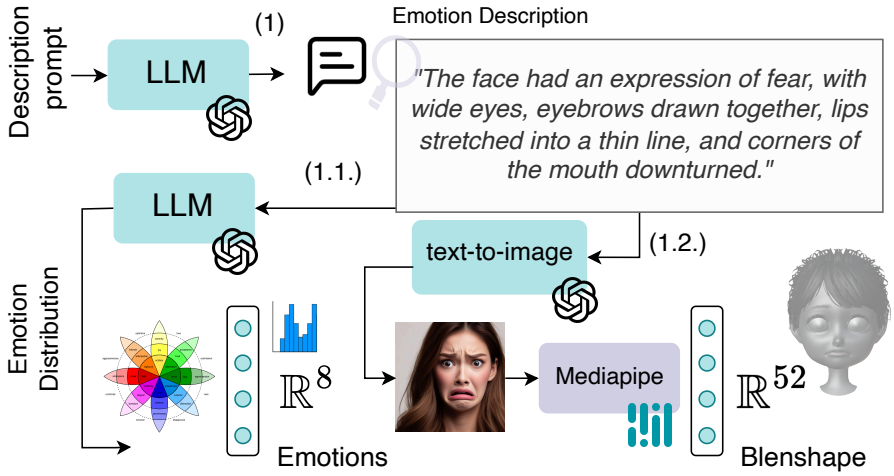


Figure 1: **Emo3D Dataset Creation**: Textual data describing human emotions is initially generated using GPT. We then utilize DALL-E models to synthesize human faces. Each image undergoes face blendshape extraction using MediaPipe. Furthermore, we employ GPT to extract the emotion distribution for each prompt.

(ii) Baseline Models: We propose several baseline models for FEG as benchmarks for future research. These models provide a foundation for evaluating new advancements and measuring progress in translating emotion descriptions into 3D facial expressions. Our baselines include **(i)** fine-tuning pre-trained language models, **(ii)** CLIP-based approaches, and **(iii)** Emotion-XLM, a customized model designed to enhance the functionality of language models for this task.

(iii) Evaluation Metric: To address the lack of standardized evaluation metrics in FEG, we introduce a new metric specifically designed to capture the complexities and nuances of human emotions.

2 Related Work

Audio-based emotion extraction: FEG methods often utilize audio data, leveraging the semantic, tonal, and expressive qualities of voice for 3D FEG. "Audio-driven Facial Animation" (Karras et al., 2017) learns to map audio waveforms to 3D facial coordinates, identifying a latent code for expression variations beyond audio cues. "EmoTalk" (Peng et al., 2023) focuses on creating 3D facial animations driven by speech, aligning expressions with both content and emotion.

CLIP-based baselines: The utility of CLIP's language-and-vision feature space (Radford et al., 2021) in text-to-image generation has been highlighted in several works. MotionCLIP (Tevet et al., 2022) leverages CLIP for a feature space that accommodates dual modalities, enabling out-of-domain actions and motion integration into CLIP's latent space. The 4D Facial Expression Diffusion Model (Zou et al., 2023) introduce a generative framework for creating 3D facial expression sequences, utilizing a Denoising Diffusion Probabilistic Model (DDPM). The framework consists of two tasks: learning a generative model based on 3D landmark sequences and generating 3D mesh sequences from an input facial mesh driven by the generated landmarks. Also, ExpCLIP (Zhong et al., 2023) is an autoencoder designed to establish semantic alignment among text, facial expressions, and facial images. ExpClip introduces a blendshape encoder to map blendshape weights to an embedding, reconstructed by a decoder. Concurrently, a CLIP text encoder (ϵ_{text}) and text projector (P_{text}), along with an image encoder (ϵ_{img}) and an image projector (P_{img}) to map emotion text and images into a joint embedding space.

Additionally, (Li et al., 2023) introduced CLIPER, a unified framework for both static and dynamic facial expression recognition, utilizing CLIP and introducing multiple expression text descriptors (METD) for fine-grained expression

representations, achieving state-of-the-art performance by a two-stage training paradigm which involves learning METD and fine-tuning the image encoder for discriminative features.

Metrics: While a variety of metrics exist for evaluating 2D image generation, the development of effective metrics for 3D FEG remains a challenge. Building upon the approach in (Xu et al., 2017), (Cong et al., 2023) adopted R-precision to measure the alignment between input text and output image. This metric was calculated using a CLIP model fine-tuned on the entire dataset, following the methodology outlined in (Park et al., 2021).

3 Dataset

We introduce the Emo3D-dataset, an assembly of 150,000 instances. Each instance comprises a triad: textual description, corresponding image, and blendshape scores created as follows:

(i) Emotion Descriptions: To generate emotion-specific textual descriptions, we prompted GPT-3.5 (OpenAI, 2023) to focus on eight primary emotions: happiness, anger, surprise, sadness, disgust, contempt, fear, and neutral. Subsequently, we again utilized GPT-3.5 to derive emotion distributions for these textual elements through carefully crafted prompts. This process resulted in eight-dimensional vectors representing distinct emotional profiles, as illustrated in Figure 1.

While concerns may arise regarding the reliability of GPT-3.5 in generating emotion distributions, the human evaluation study in Section 6.1 demonstrates their strong alignment with human perception. That section also provides a comparative analysis of GPT-3.5, GPT-4o-mini, and Gemma-9B, highlighting the reasoning behind our model choice. Additionally, a more in-depth analysis of the linguistic characteristics of the generated data can be found in the supplementary material.

(ii) 2D Image Generation: Subsequently, we utilize DALL-E 3 (Ramesh et al., 2022), an image generation model, to create images that align with the generated textual descriptions. While the reliability of AI-generated images in conveying emotions is a valid concern, previous studies, including the comprehensive evaluation presented in (Lomas et al., 2024), have demonstrated that DALL-E 3 excels in generating images that closely align with human emotional evaluations. These findings reinforce the effectiveness of DALL-E 3 in producing emotionally resonant outputs, providing a strong foundation for our use of this model in emotion-driven image generation.

(iii) Blendshape Scores Estimation: A blendshape is a predefined 3D model deformation used to represent facial movements by blending a neutral face with specific expressions, such as raising eyebrows, smiling, or frowning. These 52 blendshapes, compatible with Apple ARKit, correspond to a wide range of facial expressions, enabling precise control and reconstruction of a 3D face model's emotions or expressions. We utilize the Mediapipe framework (Lugaresi et al., 2019) to generate blendshape scores corresponding to the images synthesized from textual descriptions. Figure 2 presents an overview of the dataset, showcasing three sample data points. Each sample consists of three textual descriptions and their corresponding images, generated using the DALL-E model. As illustrated, the generated images closely match the textual descriptions. Furthermore, for each sample, a vector is provided representing the distribution of eight primary sensory categories associated with the given descriptions.

Primitive Emotion Faces: Additionally, for intrinsic evaluation purposes, we construct a dataset of primitive emotions comprising singular emotion words, each paired with corresponding images that portray males and females exhibiting three distinct intensity levels of emotion. Utilizing Mediapipe (Lugaresi et al., 2019), we subsequently extract blendshape scores for the facial expressions depicted in these images. The emotional distributions associated with these individual words are derived using Emolex (Mohammad, 2018). Figure 3 provides an example of the primitive emotion "surprise" and a set of close words defined using Emolex.

Comparison of Emo3D with Existing Datasets: As shown in Table 1, Emo3D-dataset integrates textual, visual, and blendshape modalities, providing a more holistic representation of emotional expressions compared to single-modal datasets (Saravia et al., 2018; Mollahosseini et al., 2019; Chen et al., 2023). Our dataset comprises 90,000 images and 60,000 texts. It can also be employed for emo-

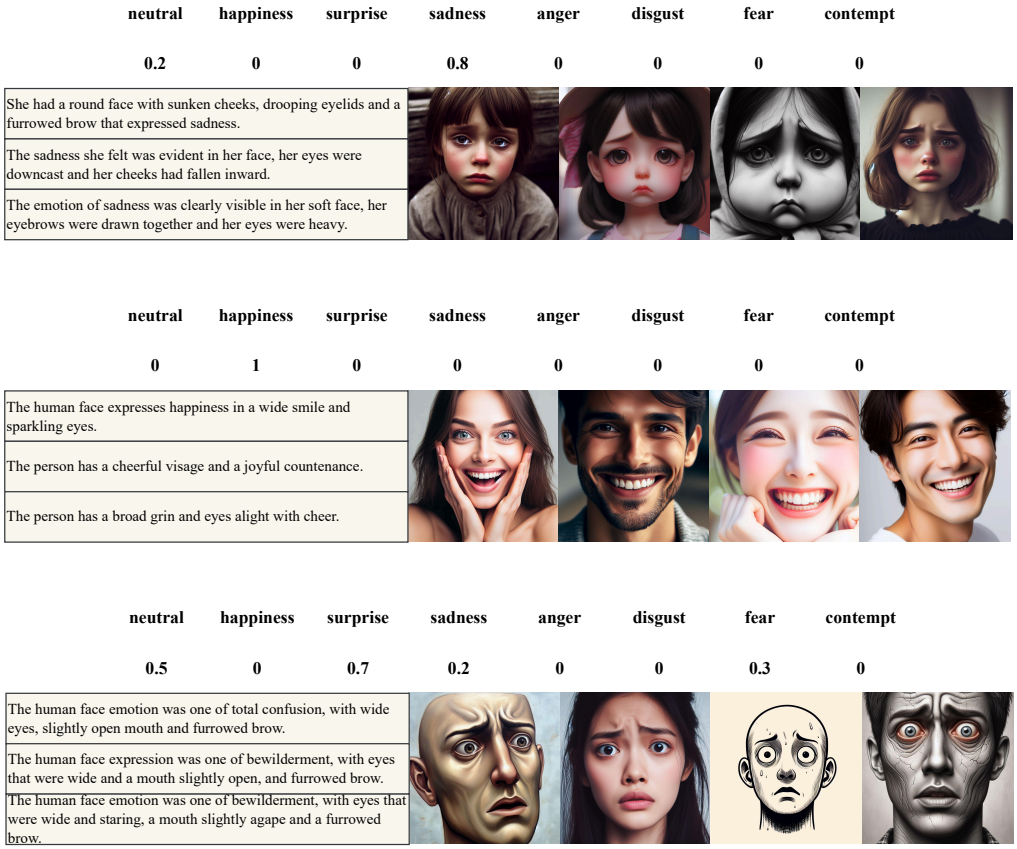


Figure 2: **Samples of Emo3D Dataset**: A glimpse into the rich diversity and complexity of our collected data, paving the way for insightful analysis and discovery.

Figure 3: "Surprise" Emotion Word Cloud: closest words to "surprise" using Emolex based on cosine similarity of emotion distribution.

tion recognition in text and images, thanks to the emotion distributions associated with each sample.

Emo3D-dataset shares similarities with other existing datasets, particularly TEAD(Zhong et al., 2023) and TA-MEAD(Ma et al., 2023), in terms of modality integration and a focus on emotional expressions. TA-MEAD (Ma et al., 2023) dataset, designed for 2D FEG, provides emotion descriptions for videos, along with Action Unit (AU) (Ekman and Friesen) intensity annotations for each video. In contrast, our Emo3D-dataset offers a unique perspective by concentrating on textual emotion expressions, corresponding images, and blendshape scores. TEAD (Zhong et al., 2023) dataset, designed for 3D FEG, features situation descriptions, our Emo3D-dataset distinguishes itself by emphasizing emotion descriptions. Additionally, our dataset includes a distinctive feature with corresponding images for each text, providing a richer

and more comprehensive resource. The Emo3D-dataset, comprising 150,000 samples, stands out significantly in scale when compared to ExpClip, which consists of 50,000 samples.

Dataset	Size	Distribution	Modalities
AffectNet	440,000	Yes	Image
Emo135	700,000	Yes	Image
CARER	417,000	Yes	Text
TEAD	50,000	No	Text Blendshape
TA-MEAD	-	No	Text Video
Emo3D	Text: 60,000 Image: 90,000 Triple: 150,000	Yes	Text Image Blendshape

Table 1: **Comparison of Emo3D with existing datasets:** This table summarizes key attributes of Emo3D alongside established datasets such as AffectNet, Emo135, CARER, TEAD, and TA-MEAD.

4 Method

4.1 Models

In this section, we propose several baseline models for the task of translating emotion descriptions into 3D facial expressions. This includes **(i)** fine-tuning of pre-trained language models, **(ii)** CLIP-based approaches, and **(iii)** Emotion-XLM, a customized XLM model designed to enhance the functionality of language models for this task.

Pretrained LM Baselines: We utilize BERT (Devlin et al., 2019) and Glot500, a highly multilingual variant of XLM-RoBERTa (ImaniGooghari et al., 2023), as the backbones. To map LM outputs into a designated target space, we incorporate a Multi-Layer Perceptron (MLP). The MLP is trained with tuples $T = \{(b, l) \mid b \in \mathbb{R}^{768}, l \in \mathbb{R}^{52}\}$, where b denotes the LM output and l represents the corresponding blendshape scores.

Emotion-XLM: Extending the MLP structure to XLM-RoBERTa, we introduce an emotion-extractor unit. The transformer output is fed into this unit to extract emotion distributions alongside one-hotted vectors. Representing the input space as $B = \{b \mid b \in \mathbb{R}^{768}\}$, the emotion-extractor unit produces output $E = \{(v, o) \mid v, o \in \mathbb{R}^{8}\}$, where v indicates emotion intensities in $V = \{[v_1, \ldots, v_8] \mid v_i \in [0, 1], i = 1, \ldots, 8\}$, and o is the one-hotted vector of v. Pairs of vectors are then passed to the MLP unit, where they are concatenated with the text embedding before being fed to the regression unit, $\mathbb{F}(.) : \mathbb{R}^{784} \rightarrow \mathbb{R}^{52}$. In the training time, 50 % of the time, ground-truth emotion labels are replaced with the emotion-extractor unit's output, to efficiently train both modules, ensuring that the blendshape MLP unit is well-trained while giving enough feedback to the emotion-extractor unit.

$$\mathcal{L} = \lambda_1 \mathcal{L}_{Blendshape} + \lambda_2 \mathcal{L}_{Emotion} \quad (1)$$

Our training methodology employs a combination of MSE losses for blendshapes and extracted emotions, weighted by coefficients to balance their contributions effectively. This model is illustrated in Figure 4.

CLIP Baseline: We employed a Multi-Layer Perceptron (MLP) architecture built upon the CLIP model (Radford et al., 2021). The model consists of three fully connected layers: the first with 256 units, the second with 128 units, and the output layer with 52 units, corresponding to the blendshape scores. All layers use the ReLU activation function, except for the output layer, which uses the sigmoid activation to constrain the predictions between 0 and 1, ensuring the generated blendshape values are valid. The Emo3D dataset provides image-text pairs and their corresponding blendshape scores. By leveraging CLIP's ability to generate embeddings for both image and text, we trained the model to get both modalities as input to predict blendshape scores. This effectively doubles the size of the dataset used for training, distinguishing this approach from Pretrained LM Baselines.

VAE CLIP: We employed a Variational Autoencoder (VAE) to align blendshape scores with their corresponding text and image CLIP (Radford et al., 2021) embeddings, as illustrated in Figure 5. This model consists of an encoder, latent space, and decoder. The encoder processes the input, a 52-dimensional blendshape vector, through two fully connected layers with ReLU activations and dropout regularization, outputting two vectors:

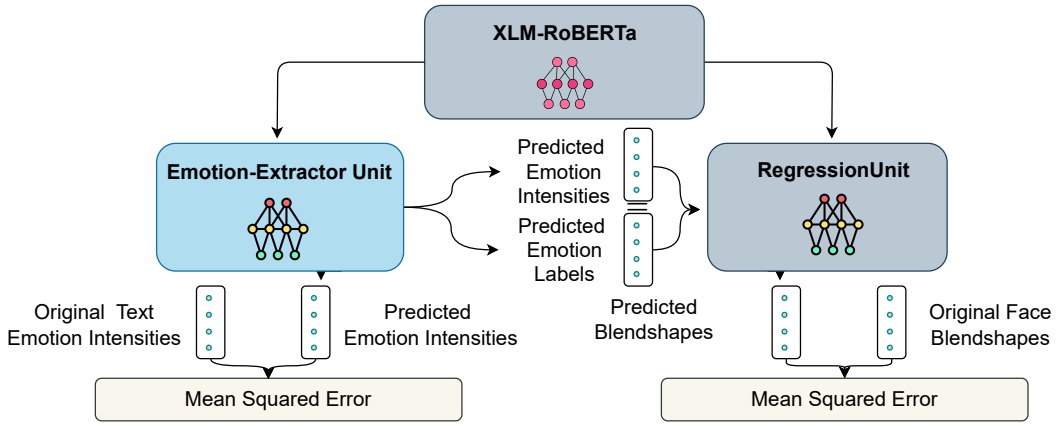


Figure 4: **Architecture and Training Process of the Emotion-XLM Model:** Emotion-XLM uses emotion ground truth to predict facial blendshapes. An Emotion Extractor guides the Regression model with the Teacher-Forcing technique at a 50% ratio. Both units are trained via MSE loss.

mean and log-variance, which define the distribution of the latent space. The latent space is fixed to match CLIP embeddings (corresponding text and image embeddings). The reparameterization trick samples from this latent space to enable backpropagation. The decoder takes the latent vector and reconstructs the original blendshape vector using two more fully connected layers, outputting the reconstructed facial expression. The blendshape scores are generated by the decoder from the CLIP embeddings of text. The model is trained via three distinct losses. Textual-blendshape and Visual-blendshape alignment are addressed using cosine similarity. Moreover, The reconstruction loss is defined by MSE.

$$\mathcal{L}_{text} = 1 - \cos\left(CLIP_{text}, z\right) \quad (2)$$

$$\mathcal{L}_{image} = 1 - \cos\left(CLIP_{image}, z\right) \quad (3)$$

$$\mathcal{L} = \mathcal{L}_{recon} + \lambda_{text}\mathcal{L}_{text} + \lambda_{image}\mathcal{L}_{image} \quad (4)$$

Here, cos(a, b) denotes the cosine similarity between two vectors a and b.

4.2 Metric

We introduce a new 3D FEG metric for evaluating the reconstruction of the original emotion vector from 2D snapshots of the generated 3D faces. We create a test set comprising diverse emotion prompts uniformly selected. To evaluate any proposed FEG model, we generate the corresponding blenshape scores of the input text and project the 3D face model onto a 2D image. Using zero-shot CLIP (Radford et al., 2021), we identify the k-nearest text prompts related to the image. We

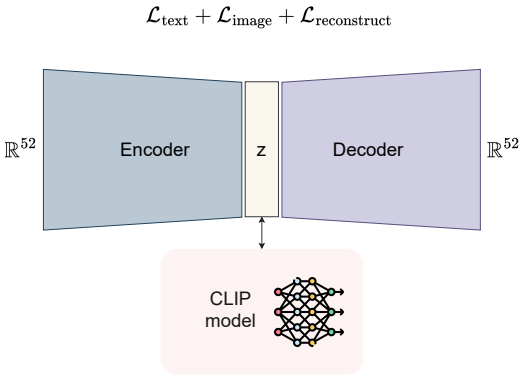


Figure 5: **Architecture and Training Process of the VAE-CLIP Model:** VAE CLIP concurrently reconstructs facial expressions while aligning their latent representation with corresponding text and image representations in the CLIP space.

calculate the emotion distribution for the original prompt and the top-K prompts. This is followed by computing the Kullback-Leibler (KL) divergence between the emotion vector of the original prompt and the average emotion vector of the top-K retrieved prompts. We refer to the normalized KL bounded between 0 and 1 as the "Emo3D metric":

$$D_{KL}(\phi \,||\, \bar{\psi}) = \sum_i \phi(i) \cdot \log\left(\frac{\phi(i)}{\bar{\psi}(i)}\right) \quad (5)$$

$$\text{Emo3D Metric} = \frac{1}{1 + e^{-D_{KL}(\phi \,||\, \bar{\psi})}} \quad (6)$$

where ϕ represents the emotion distribution of the input prompt, and $\bar{\psi}$ represents the mean emotion distribution of the top-k prompts. The steps for

Model	MSE	Emo3D
BERT	0.03	0.796
XLMRoBERTa	0.04	0.789
VAE CLIP	0.002	0.776
Emotion-XLM	0.035	0.756
CLIP Baseline	0.014	0.737

Table 2: Performance comparison of FEG models using MSE vs. Emo3D metrics.

Emo3D calculation are outlined in Figure 6. In our evaluation of the FEG models, we provide both the Emo3D Metric and the MSE scores of the 3D models for comparison purposes. Additionally, to validate the alignment of the Emo3D Metric with human perception, we conducted a human evaluation study. This study assesses the correlation between the metric's rankings and human judgments of emotional alignment. More details on this evaluation can be found in Section 6.2.

5 Results

The FEG model performances are provided in Table 2. It becomes evident that the CLIP With Regression Unit model demonstrates superior performance when evaluated using our Emo3D metric. Our results indicate that the MSE and Emo3D metrics do not consistently align. To better understand this discrepancy, we conducted a human evaluation of the 3D model outputs (details provided in 6.2). The evaluation revealed that samples that performed better according to Emo3D metric also demonstrated a closer visual resemblance to the input prompt, in contrast to samples that showed better performance based on MSE, similar to Figure 7. This can be because in our metric, Emo3D prioritizes visual-text alignment and proximity, tending to capture richer semantic information than distance metrics in 3D space using MSE.

6 Human Evaluation

To ensure the validity of our approach, we conducted a comprehensive human evaluation study. This study serves three main purposes: (1) validating the annotations generated by LLMs, (2) justifying our choice of GPT-3.5, comparing our results with more modern LLMs GPT-4o-mini and Gemma-9B, and (3) demonstrating that our proposed Emo3D metric aligns more closely with human perception compared to MSE.

6.1 Assessing the Quality of Emotion Distributions

To ensure the accuracy and reliability of emotion distributions generated by the models, we conducted a rigorous human evaluation of a dataset consisting of 100 emotionally diverse text scenarios. These scenarios were selected through k-means clustering (k=100) on emotion embeddings obtained from a pre-trained model specific for emotion classification, `roberta-base-go_emotions` (Lowe, 2021). Four independent annotators manually annotated the selected scenarios, resulting in two-way assessments of the emotions conveyed in each text sample. Details of the given instructions are provided in Appendix A.1.

The agreement between annotators was evaluated using Cohen's kappa, yielding a score of 0.79, indicating a substantial agreement level. Discrepancies in emotional assessments were addressed through discussions among the annotators, and when necessary, the judgment of a third annotator was sought to resolve disagreements. The final emotion score for each text scenario was calculated by averaging the agreed-upon assessments. This human judgment served as the reference for comparing model-generated emotion distributions.

Additionally, to evaluate the performance of different language models, we compared GPT-3.5 to Gemma-9B and GPT-4o-mini. GPT-3.5 was initially chosen due to its availability and cost-effectiveness at the time of dataset curation, while Gemma-9B and GPT-4o-mini were selected as recent examples of open-source and commercial LLMs, respectively. For each of the 100 text scenarios, we generated emotion distributions using these models, and the results were compared against the human assessments.

The results of the human evaluation and model performance are summarized in Table 3, which presents the averaged absolute emotion score differences between human judgments and model-generated emotion distributions for each model.

Surprisingly, GPT-3.5 demonstrated better performance based on both the absolute emotion score differences and the averaged cross-entropy over basic emotion distributions. However, statistical analysis using a t-test revealed that neither Gemma-9B nor GPT-4-o-mini performed significantly worse. These findings highlight the suitability and relia-

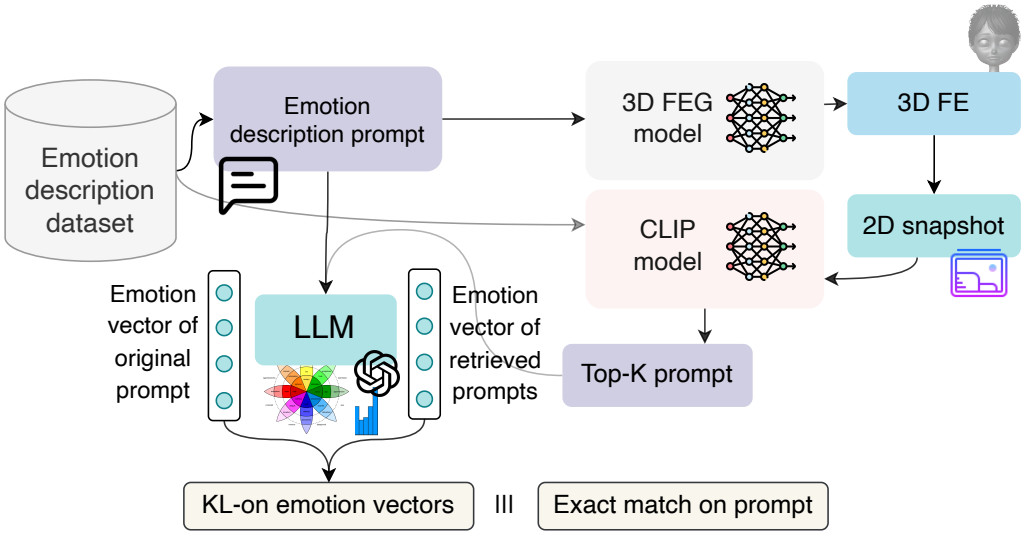


Figure 6: **Overview of the Emo3D Metric Calculation Process:** Our methodology in Emo3D metric entails selecting n prompts with a balanced emotion distribution. For a given input and a generated facial expression by a model, we project the 3D face onto a 2D image and employ zero-shot CLIP to identify the k nearest text prompts. Subsequently, we compute the Kullback-Leibler (KL) divergence between the emotion distribution of the input text and these k prompts.

Figure 7: **Qualitative 3D Face Generation Model Comparison:** For the given text prompt, "The human face exuded joy, with their eyes sparkling with delight and lips curling upwards in a broad beam of happiness", the figure compares the output of the proposed FEG models, i.e., BERT-based, XLMR-based, CLIP-based, Emotion-XLM and VAE Clip models.

bility of GPT-3.5 for this task, with performance comparable to, and sometimes better than, other models. Moreover, the minimal differences from human judgment underline the credibility of the generated emotion distributions.

These findings are noteworthy for several reasons. First, they demonstrate the potential of GPT-3.5 as a highly reliable model for emotion analysis in text, highlighting its cost-effectiveness and suitability for this task. Second, the minimal differences observed across all models emphasize the credibility and reliability of the emotion distributions generated, as they align closely with human judgment despite the models' varying architectures and training data.

In conclusion, the results from both the human evaluation and model performance analysis confirm that GPT-3.5 is a strong candidate for emotion distribution tasks, and its performance is comparable to, if not better than, more advanced models like Gemma-9B and GPT-4o-mini. These findings also validate the overall methodology, suggesting that LLMs are capable of generating emotion distributions that are closely aligned with human perceptions of emotions in text.

6.2 Evaluating the Alignment of Emo3D Metric with Human Perception

To assess the reliability of the Emo3D metric, we selected 100 samples from our test set. For each

Emotion	Gemma-9B	GPT-3.5	GPT-4o-mini
Neutral	0.20 ± 0.07	0.20 ± 0.05	0.18 ± 0.23
Happiness	0.24 ± 0.11	0.23 ± 0.11	0.23 ± 0.32
Surprise	0.22 ± 0.07	0.25 ± 0.09	0.24 ± 0.28
Sadness	0.27 ± 0.09	0.26 ± 0.08	0.28 ± 0.30
Anger	0.19 ± 0.08	0.19 ± 0.07	0.20 ± 0.27
Disgust	0.08 ± 0.03	0.14 ± 0.039	0.11 ± 0.19
Fear	0.22 ± 0.07	0.22 ± 0.07	0.24 ± 0.26
Contempt	0.22 ± 0.13	0.24 ± 0.10	0.19 ± 0.31

Table 3: Averaged absolute emotion score difference between human judgments and model-generated emotion distributions for specific text scenarios.

of these samples, we conducted two-way rankings, where each human ranker scored an image corresponding to an emotion description on a scale of 0–4 based on the front view. These scores were then used to rank the images for each description. The human rankers achieved a Spearman agreement score of 0.62, indicating moderate agreement.

Next, we compared the Emo3D metric against the traditional MSE metric to evaluate its utility. To do this, we measured the correlation between the human rankers' scores and rankings achieved from both Emo3D and MSE using Pearson correlation and Kendall's Tau. These metrics were used to determine how well the Emo3D metric aligns with human evaluations compared to MSE. The results of the comparison between Emo3D, MSE, and human evaluations are presented in Table 4 and Table 5.

Comparison	Human Score	Emo3D	MSE
Human Score	1.00	0.84	0.56
Emo3D	0.84	1.00	0.12
MSE	0.56	0.12	1.00

Table 4: Pearson correlation between Human scores, Emo3D metric, and MSE.

Table 4 and 5 demonstrate that Emo3D aligns significantly better with human rankings compared to MSE. The results of these comparisons support the validity of Emo3D as a more reliable metric than MSE for assessing the emotional alignment of 3D images with textual emotion descriptions. Additionally, the human evaluation shows that the use of a 2D image of a 3D model from the front view is sufficient for assessing emotional alignment. This confirms that the 2D projection captures the necessary features to evaluate facial expressions. The only notable disagreement between Emo3D and human rankings occurs when comparing VAE Clip and Emotion-XLM, where VAE Clip scores higher in human evaluations, while Emotion-XLM scores higher in Emo3D. However, both are close in scores in both settings, suggesting a slight gap. Despite this, the results from the Pearson correlation underline the reliability of Emo3D as a benchmark for emotion assessment in 3D outputs.

Comparison	Human Score	Emo3D	MSE
Human Score	1.00	0.67	0.33
Emo3D	0.67	1.00	0.00
MSE	0.33	0.00	1.00

Table 5: Kendall's Tau correlation between Human Score, Emo3D, and MSE.

7 Conclusion

In this paper, we introduced "Emo3D", a comprehensive "Text-Image-Expression dataset" that covered a wide range of human emotions and their textual descriptions, paired with images and 3D blendshapes. Our use of LLMs to generate prompts captured a variety of emotional expressions and descriptions. To the best of our knowledge, "Emo3D" stood out as the most comprehensive FEG dataset, encompassing sufficiently diverse and complex emotional descriptions. Furthermore, we developed an efficient evaluation metric to provide 3D image synthesis models with a reliable benchmark. Throughout our work, we tested several unimodal and multimodal models as baselines to encourage new entrants to the field. The significance of "Emo3D" lies in its potential to advance 3D facial expression synthesis, holding promising implications for animation, virtual reality, and emotional human-computer interaction.

8 Limitations and Future Work

While our dataset exhibits positive attributes, it is not without errors stemming from the processes involved in its production. Specifically, the use of Mediapipe to obtain blendshape scores introduced

inaccuracies, particularly in the representation of certain emotions and facial expressions. To enhance the dataset in future endeavors, collaboration with skilled animators could be sought to refine and design more accurate blendshape scores.

9 Ethics

This paper leverages GPT-3.5 (OpenAI, 2023) for generating textual emotional descriptions and DALL-E3 (Ramesh et al., 2022) for creating corresponding images. It's vital to recognize the potential biases and privacy concerns inherent in these AI models. Both GPT-3.5 and DALL-E3, like many advanced AI systems, reflect the data on which they were trained, which can include societal biases and inaccuracies. Mitigating and analyzing such biases is beyond the scope of this paper. Studies such as "DALL-EVAL: Probing the Reasoning Skills and Social Biases of Text-to-Image Generation Models" (Cho et al., 2023) and the paper "ChatGPT: A Comprehensive Review on Background, Applications, Key Challenges, Bias, Ethics, Limitations and Future Scope" (Ray, 2023) thoroughly examine biases in the DALL-E and GPT models, respectively. According to these studies, the DALL-E and GPT models are also shown to have certain degrees of biases related to gender, skin tone, professions, and certain attributes and may have privacy or accountability concerns.

References

Keyu Chen, Changjie Fan, Wei Zhang, and Yu Ding. 2023. 135-class emotional facial expression dataset.

Jaemin Cho, Abhay Zala, and Mohit Bansal. 2023. Dall-eval: Probing the reasoning skills and social biases of text-to-image generation models. *Preprint*, arXiv:2202.04053.

Yuren Cong, Martin Renqiang Min, Li Erran Li, Bodo Rosenhahn, and Michael Ying Yang. 2023. Attribute-centric compositional text-to-image generation. *IEEE Trans. Pattern Anal. Mach. Intell.*

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 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)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Paul Ekman and Wallace V Friesen. Facial action coding system. *Environmental Psychology & Nonverbal Behavior*.

Ayyoob ImaniGooghari, Peiqin Lin, Amir Hossein Kargaran, Silvia Severini, Masoud Jalili Sabet, Nora Kassner, Chunlan Ma, Helmut Schmid, André Martins, François Yvon, and Hinrich Schütze. 2023. Glot500: Scaling multilingual corpora and language models to 500 languages. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1082–1117, Toronto, Canada. Association for Computational Linguistics.

Tero Karras, Timo Aila, Samuli Laine, Antti Herva, and Jaakko Lehtinen. 2017. Audio-driven facial animation by joint end-to-end learning of pose and emotion. *ACM Transactions on Graphics (TOG)*.

Sheng Li, Jinpeng Wang, Wei Zhao, Yucong Chen, and Kunpeng Du. 2023. Cliper: A unified vision-language framework for in-the-wild facial expression recognition. *arXiv preprint arXiv:2303.00193*.

James Derek Lomas, Willem van der Maden, Sohhom Bandyopadhyay, Giovanni Lion, Nirmal Patel, Gyanesh Jain, Yanna Litowsky, Haian Xue, and Pieter Desmet. 2024. Improved emotional alignment of ai and humans: Human ratings of emotions expressed by stable diffusion v1, dall-e 2, and dall-e 3. *Preprint*, arXiv:2405.18510.

Sam Lowe. 2021. Roberta-base fine-tuned on goemotions dataset. `https://huggingface.co/SamLowe/roberta-base-go_emotions`.

Camillo Lugaresi, Jiuqiang Tang, Hadon Nash, Chris Mcclanahan, Esha Uboweja, Michael Hays, Fan Zhang, Chuo-Ling Chang, Ming Guang Yong, Juhyun Lee, Wan-Teh Chang, Wei Hua, Manfred Georg, Matthias Grundmann, and Google Research. 2019. Mediapipe: A framework for building perception pipelines. *IEEE Trans. Vis. Comput. Graph.*

Yifeng Ma, Suzhen Wang, Yu Ding, Bowen Ma, Tangjie Lv, Changjie Fan, Zhipeng Hu, Zhidong Deng, and Xin Yu. 2023. Talkclip: Talking head generation with text-guided expressive speaking styles. *IEEE Trans. Circuit Syst. Video Technol.*

Saif M. Mohammad. 2018. Word affect intensities. In *Proceedings of the 11th Edition of the Language Resources and Evaluation Conference (LREC-2018)*, Miyazaki, Japan.

Ali Mollahosseini, Behzad Hasani, and Mohammad H. Mahoor. 2019. Affectnet: A database for facial expression, valence, and arousal computing in the wild. *IEEE Transactions on Affective Computing*, 10(1):18–31.

OpenAI. 2023. Gpt-4 technical report.

Dong Huk Park, Samaneh Azadi, Xihui Liu, Trevor Darrell, and Anna Rohrbach. 2021. Benchmark for compositional text-to-image synthesis. *IEEE Trans. Image Process.*

Ziqiao Peng, Haoyu Wu, Zhenbo Song, Hao Xu, Xiangyu Zhu, Jun He, Hongyan Liu, and Zhaoxin Fan. 2023. Emotalk: Speech-driven emotional disentanglement for 3d face animation. *IEEE Trans. Circuit Syst. Video Technol.*

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. 2021. Learning transferable visual models from natural language supervision. *Proceedings of Machine Learning Research*, 139:8748–8763.

Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. 2022. Hierarchical text-conditional image generation with clip latents. *Preprint*, arXiv:2204.06125.

Partha Pratim Ray. 2023. Chatgpt: A comprehensive review on background, applications, key challenges, bias, ethics, limitations and future scope. *Internet of Things and Cyber-Physical Systems*, 3:121–154.

Elvis Saravia, Hsien-Chi Toby Liu, Yen-Hao Huang, Junlin Wu, and Yi-Shin Chen. 2018. CARER: Contextualized affect representations for emotion recognition. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3687–3697, Brussels, Belgium. Association for Computational Linguistics.

J. Rafid Siddiqui. 2022. Explore the expression: Facial expression generation using auxiliary classifier generative adversarial network. *Preprint*, arXiv:2201.09061.

Guy Tevet, Brian Gordon, Amir Hertz, Amit H. Bermano, and Daniel Cohen-Or. 2022. Motionclip: Exposing human motion generation to clip space. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 13682 LNCS:358–374.

Tao Xu, Pengchuan Zhang, Qiuyuan Huang, Han Zhang, Zhe Gan, Xiaolei Huang, and Xiaodong He. 2017. Attngan: Fine-grained text to image generation with attentional generative adversarial networks. *IEEE Conf. Comput. Vis. Pattern Recog.*

Yicheng Zhong, Huawei Wei, Peiji Yang, and Zhisheng Wang. 2023. Expclip: Bridging text and facial expressions via semantic alignment.

Kaifeng Zou, Sylvain Faisan, Boyang Yu, Sébastien Valette, and Hyewon Seo. 2023. 4d facial expression diffusion model. *IEEE Trans. Pattern Anal. Mach. Intell.*

A Human Evaluation Guidelines

In this section, we provide detailed instructions given to annotators for the two human evaluation experiments conducted in our study. The first experiment assesses the reliability of emotion distributions generated by GPT-3.5, while the second evaluates the alignment of our proposed Emo3D metric with human perception. Below, we outline the annotation guidelines for each experiment.

A.1 Emotion Distribution Annotation

In this experiment, annotators were asked to assign intensity values (ranging from 0 to 1) for eight emotional categories—Neutral, Fear, Anger, Happiness, Contempt, Disgust, Surprise, and Sadness—based on facial expression samples. The goal was to measure how well the GPT-3.5-generated emotion distributions align with human perception. Part of the evaluation form given to the annotators is shown in Figure 8

A.1.1 Objective

Your task is to assign intensity values (on a scale from 0 to 1) for eight emotional categories—Neutral, Fear, Anger, Happiness, Contempt, Disgust, Surprise, and Sadness—for each facial expression sample. The goal is to reflect the emotional profile conveyed by the facial expression as accurately as possible. These intensities will then be normalized into a distribution for further analysis.

A.1.2 General Principles

- **Granular Scale of Emotion Intensity:** Each emotion is rated independently on a scale from 0 to 1, where:
  - 0: The emotion is not present in the expression at all.
  - 1: The emotion is maximally expressed.

  Avoid "forcing" an emotion if it is not discernible in the expression.

- **Independence of Emotions:** Emotions may co-occur; for example, a person may express both *Surprise* and *Happiness*. Score all emotions based on their individual presence in the expression, regardless of overlaps.

- **Facial Action and Emotion Mapping:** Use key visual cues, such as facial action units

(FAUs), to infer emotional intensity. The relationship between facial movements and emotions can help ensure consistency:

  - **Happiness:** Upturned mouth corners (smile), cheek raising.
  - **Anger:** Furrowed brows, clenched jaw, narrowed eyes.
  - **Sadness:** Downturned mouth corners, drooping eyelids, raised inner brows.
  - **Fear:** Wide eyes, raised eyebrows, tense mouth.
  - **Surprise:** Wide-open eyes, raised eyebrows, rounded mouth.
  - **Disgust:** Wrinkled nose, raised upper lip.
  - **Contempt:** Asymmetrical smirk, raised lip on one side.
  - **Neutral:** Lack of pronounced facial muscle activity.

- **Subtle Expressions:**
  - Pay attention to subtle facial cues. Low-intensity emotions should still be scored rather than being ignored if present.
  - When unsure, assign a lower intensity rather than zero unless absolutely no evidence of the emotion exists.

A.1.3 Systematic Process for Assessing Emotional Intensity

1. **Break Down the Expression Into Components (FAUs)**
   - Start by identifying the movements of key facial features (e.g., mouth, eyes, eyebrows) and map them to relevant emotions using the following FAU-to-emotion guide:
     - Neutral: Minimal or no visible activation of facial muscles.
     - Happiness: Upturned mouth corners (AU12: Lip Corner Puller), cheek raising (AU6: Cheek Raiser).
     - Sadness: Downturned mouth corners (AU15: Lip Corner Depressor), raised inner eyebrows (AU1+AU4: Inner Brow Raiser and Brow Lowerer), drooping eyelids.
     - Anger: Furrowed brows (AU4: Brow Lowerer), narrowed eyes (AU7: Lid Tightener), tightened lips or open mouth (AU23: Lip Tightener or AU22: Lip Funneler).
     - Fear: Wide-open eyes (AU5: Upper Lid Raiser), raised eyebrows (AU1+AU2: Inner and Outer Brow Raiser), mouth corners stretched back (AU20: Lip Stretcher).
     - Surprise: Raised eyebrows (AU1+AU2), wide-open eyes (AU5), rounded mouth (AU27: Mouth Stretch).
     - Disgust: Wrinkled nose (AU9: Nose Wrinkler), raised upper lip (AU10: Upper Lip Raiser).
     - Contempt: Asymmetrical lip pull or smirk (AU14: Dimpler).
2. **Assess the Intensity of Each Facial Component**
   - For each FAU, rate its intensity on a scale of 0 to 1:
     - 0: No activation of the feature.
     - 0.5: Moderate activation: facial expressions are noticeable but not pronounced.
     - 1: Strong activation: facial expressions are clear and dominate the expression.
3. **Combine FAU Ratings Into Emotion Scores**
   - Use the intensity ratings of the FAUs to assign a score for each emotion:
     - The most intense and dominant FAUs should correspond to higher emotion scores.
     - Secondary or blended FAUs should result in proportionally lower scores.
     - For ambiguous FAUs, assign a low intensity (e.g., 0.1–0.3) rather than ignoring the emotion entirely.
4. **Evaluate the Expression as a Whole**
   - After scoring individual emotions based on FAUs, assess the overall emotional impression:
     - Does the expression predominantly communicate one emotion Assign it the highest intensity score.

- Are multiple emotions blended or secondary emotions present? Score these proportionally.
- For subtle or fleeting expressions, lower all scores appropriately.

5. **Validate and Adjust Scores**
   - Ensure all eight emotion scores align with your observations:
     - No emotion should be scored non-zero without clear justification from FAU activation or your subjective interpretation.
     - Review for proportionality: The most dominant emotion should generally have the highest score.
     - Adjust scores as needed for consistency with previous samples.

A.1.4 Step-by-Step Workflow

- **Observe the Expression Carefully:** View each facial expression for at least 5 seconds to assess its components (eyes, eyebrows, mouth, etc.).
- **Score the Emotions:** Assign a value from 0 to 1 to each of the eight emotional categories, referencing the FAU mappings and observed intensity.
- **Review and Adjust:** After initial scoring, recheck your annotations to ensure alignment between the FAU-based observations and overall emotional impression.
- **Document Annotations:** Provide intensity values for all eight emotions, ensuring the sum does not need to equal 1, as normalization will be handled in post-processing.

A.1.5 Tips for Consistency and Accuracy

- **Neutral as a Baseline:** Start by assessing whether the face appears mostly neutral and assign scores to other emotions relative to this baseline.
- **Dominant vs. Secondary Emotions:** Identify the primary (most intense) emotion. Assign the highest score to it and lower scores to secondary or blended emotions.
- **Non-Emotional Features:** Ignore non-emotional aspects of the face, such as aesthetic features (e.g., wrinkles, facial asymmetry) or artifacts (e.g., shadows, blurs).
- **Consistency Across Samples:** Use a standardized scoring framework across all samples. If a similar expression appears in multiple samples, ensure your ratings are consistent.

A.2 Metric Evaluation Study

This experiment aimed to assess the effectiveness of our Emo3D metric in capturing human-perceived emotional similarity. Annotators were presented with generated 3D facial expressions and corresponding textual prompts and were asked to evaluate the emotional congruence between them. These human judgments were then compared to the metric scores to analyze its alignment with human perception. A sample of the form provided to the annotators can be seen in Figure 9.

A.2.1 Annotation Guidelines

You will evaluate how well four computational models generate facial expressions based on textual descriptions. Your task is to rank the four output images from 1 (best match) to 4 (least match) for each description. To ensure a systematic and objective evaluation, follow the step-wise procedure outlined below, starting with major considerations and narrowing down to finer details.

Step 1: Understand the Description Read the text carefully to identify:

- **Primary Emotion**: The central emotional state described (e.g., Neutral, Fear, Anger, Happiness, Contempt, Disgust, Surprise, or Sadness).
- **Secondary Nuances**: Additional emotional layers or subtleties, such as mixed emotions (e.g., fearful surprise or sad anger).
- **Intensity**: Determine whether the text describes a subtle, moderate, or intense expression of the emotion.
- **Facial Detail Cues**: Note specific descriptions of facial features (e.g., "raised eyebrows," "tightened lips," "averted gaze") or contextual hints.

Step 2: Evaluate Overall Alignment with the Main Emotion For each image:

- Identify the expressed emotion: Assess which emotion is most clearly communicated by the facial expression in the image.

Sentence	ID	Cluster	neutral	happiness	surprise	sadness	anger	disgust	fear	contempt
The human face was contorted in anger, the tight lips pressing together into a thin line and the eyebrows drawing downwards in a deep frown.	971	46								
The emotion on the face is happiness, with a warm smile indicating cheerfulness.	1305	98								
The human face showed a suggestion of sadness, with dejected eyes and a downturned mouth.	2739	44								
The face expresses a combination of joy and fear, as if overwhelmed by a strong surge of emotion.	84	30								
The human face emotion was a mix of shock and excitement, an expression of surprise that could only be described as awe-inspiring.	6197	86								
The human face was displaying a perplexed emotion, looking as if they could not decide which decision to make.	3373	11								
The face looked exhausted, under immense strain and drained of energy due to their emotions.	1762	12								
The human face conveyed a mixture of happiness and surprise.	3495	13								
The human face emotion shows a contemplative mood with a hint of serenity in its expression.	4711	32								
The face is giving off an emotion of sorrow and gloom, with features showing depression and disappointment.	762	34								

Figure 8: A blank annotation sheet provided to annotators for recording emotion distribution ratings across eight emotional categories.

- Match to the text: Decide how well this primary emotion aligns with the main emotion described in the text.
  - Assign higher preference to images that capture the correct emotional category.
  - If none align perfectly, select the one that aligns most closely.

Step 3: Assess Emotional Intensity

- Compare the intensity of the emotion in each image with the level implied in the text.
- Look for proportionality:
  - Is the smile too exaggerated for a mildly happy description?
  - Does fear appear subdued when the text implies terror?
- Penalize images that fail to match the described intensity.

Step 4: Examine Key Facial Features For each image, evaluate how well the detailed facial components align with the text:

- **Eyes**: Are the eyes widened, narrowed, or averting gaze as described?
- **Mouth**: Is the mouth set, smiling, frowning, or clenched appropriately?
- **Brows & Forehead**: Do they reflect the tension or relaxation described (e.g., furrowed brows for anger)?
- **Other Details**: Consider any additional features like head tilts, cheek tension, or jaw positioning mentioned in the description.

Step 5: Consider Overall Naturalness

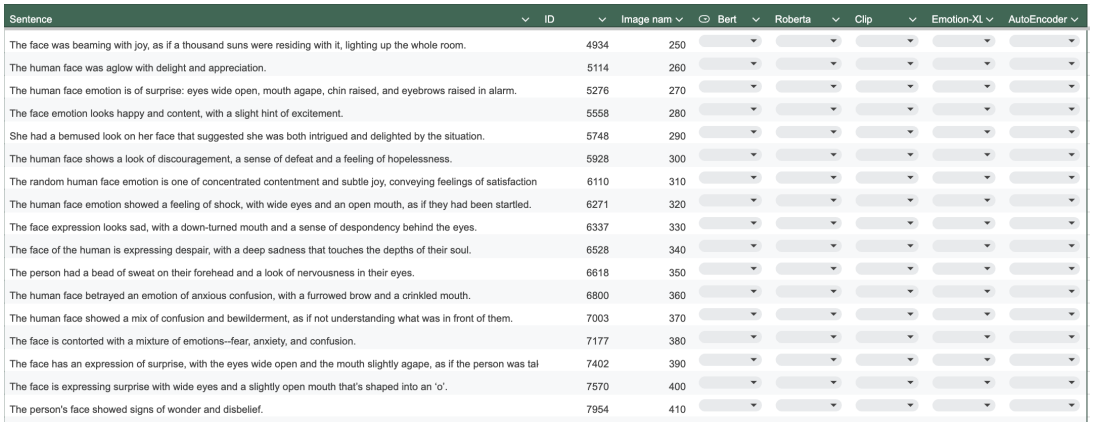
- Does the facial expression look believable and anatomically correct?
- Penalize images with unnatural or distorted features that detract from their realism, even if they align with the text descriptively.

Step 6: Rank the Images

- Compare all four images: Based on the above criteria, rank the images from 1 (best match) to 4 (least match).
- Resolve ties by weighing:
  - Major alignment with the primary emotion.
  - Clarity of facial features.
  - Naturalness and intensity accuracy.

Step 7: Provide Notes (Optional)

- For any ambiguous cases or particularly strong impressions, include brief comments explaining your decisions.
  - Example: Image 3 captures the surprise well, but the mouth is too exaggerated compared to the text.



Sentence	ID	Image nam	Bert	Roberta	Clip	Emotion-XL	AutoEncoder
The face was beaming with joy, as if a thousand suns were residing with it, lighting up the whole room.	4934	250					
The human face was aglow with delight and appreciation.	5114	260					
The human face emotion is of surprise: eyes wide open, mouth agape, chin raised, and eyebrows raised in alarm.	5276	270					
The face emotion looks happy and content, with a slight hint of excitement.	5558	280					
She had a bemused look on her face that suggested she was both intrigued and delighted by the situation.	5748	290					
The human face shows a look of discouragement, a sense of defeat and a feeling of hopelessness.	5928	300					
The random human face emotion is one of concentrated contentment and subtle joy, conveying feelings of satisfaction	6110	310					
The human face emotion showed a feeling of shock, with wide eyes and an open mouth, as if they had been startled.	6271	320					
The face expression looks sad, with a down-turned mouth and a sense of despondency behind the eyes.	6337	330					
The face of the human is expressing despair, with a deep sadness that touches the depths of their soul.	6528	340					
The person had a bead of sweat on their forehead and a look of nervousness in their eyes.	6618	350					
The human face betrayed an emotion of anxious confusion, with a furrowed brow and a crinkled mouth.	6800	360					
The human face showed a mix of confusion and bewilderment, as if not understanding what was in front of them.	7003	370					
The face is contorted with a mixture of emotions--fear, anxiety, and confusion.	7177	380					
The face has an expression of surprise, with the eyes wide open and the mouth slightly agape, as if the person was tal	7402	390					
The face is expressing surprise with wide eyes and a slightly open mouth that's shaped into an 'o'.	7570	400					
The person's face showed signs of wonder and disbelief.	7954	410					

Figure 9: A blank evaluation sheet given to annotators for scoring the quality of generated facial expressions for each model based on textual descriptions.

Final Checklist Before Submission

- Have you ranked all four images distinctly?
- Did you consider all major and minor criteria for each image?
- Are your rankings consistent with the textual descriptions?