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

Anonymous ACL submission

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.

(ii) Baseline Models: We propose several baseline

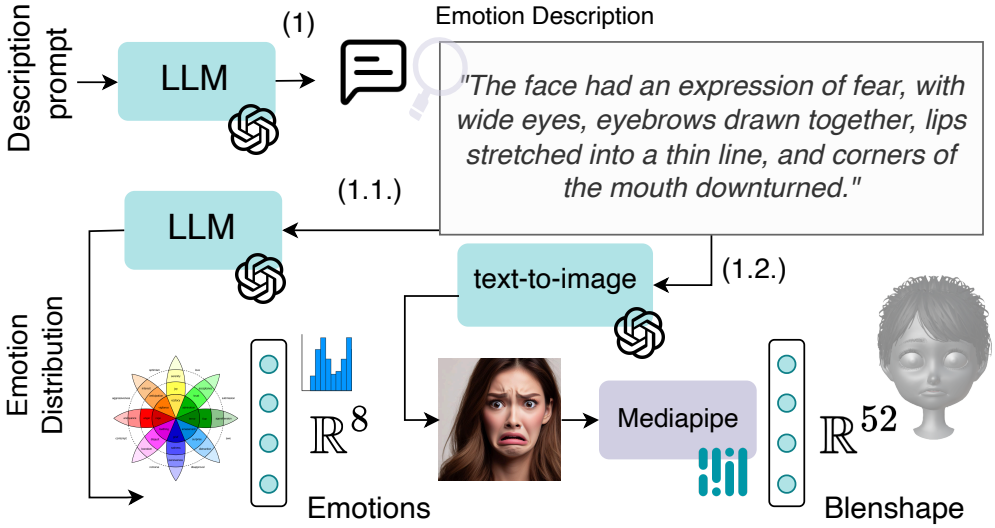

Figure 1: **Emo3D Dataset Creation**: Textual data describing human emotions is initially generated using GPT. We then utilize DALL-E models to synthesize human faces. Each image undergoes face blendshape extraction using MediaPipe. Furthermore, we employ GPT to extract the emotion distribution for each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

models for Facial Emotion Generation (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 XML model designed to enhance the functionality of language models for this task.

(iii) Evaluation Metric: To address the lack of standardized evaluation metrics in Facial Emotion Generation (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 generation. "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

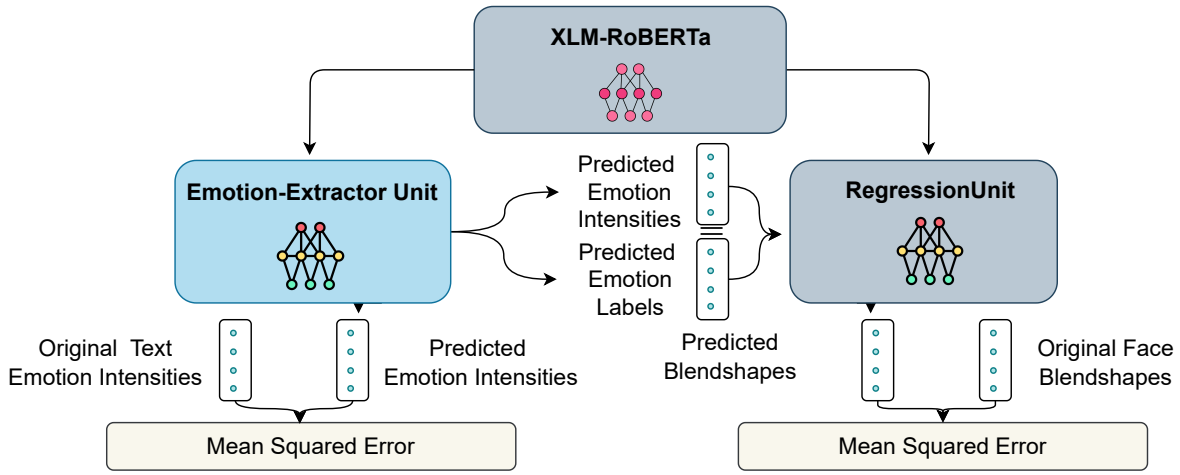

Figure 2: **Architecture and Training Process of the Emotion-XLM Model:** Emotion-XLM uses emotion ground truth to predict facial blendshapes. An Emotion Extractor guides the Regression model with the Teacher-Forcing technique at a 50% ratio. Both units are trained via Mean Squared Error (MSE) loss.

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. 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.

(iii) Blendshape Scores Estimation: We utilize the Mediapipe framework (Lugaresi et al., 2019) to generate blendshape scores corresponding to the images synthesized from textual descriptions. Figure 3 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 4 provides an example of the primitive emotion "surprise" and a set of close words defined using

neutral	happiness	surprise	sadness	anger	disgust	fear	contempt
0.2	0	0	0.8	0	0	0	0

She had a round face with sunken cheeks, drooping eyelids and a furrowed brow that expressed sadness.

The sadness she felt was evident in her face, her eyes were downcast and her cheeks had fallen inward.

The emotion of sadness was clearly visible in her soft face, her eyebrows were drawn together and her eyes were heavy.

neutral	happiness	surprise	sadness	anger	disgust	fear	contempt
0	1	0	0	0	0	0	0

The human face expresses happiness in a wide smile and sparkling eyes

The person has a cheerful visage and a joyful countenance

The person has a broad grin and eyes alight with cheer

neutral	happiness	surprise	sadness	anger	disgust	fear	contempt
0.5	0	0.7	0.2	0	0	0.3	0

The human face emotion was one of total confusion, with wide eyes, slightly open mouth and furrowed brow.

The human face expression was one of bewilderment, with eyes that were wide and a mouth slightly open, and furrowed brow.

The human face emotion was one of bewilderment, with eyes that were wide and staring, a mouth slightly agape and a furrowed brow.

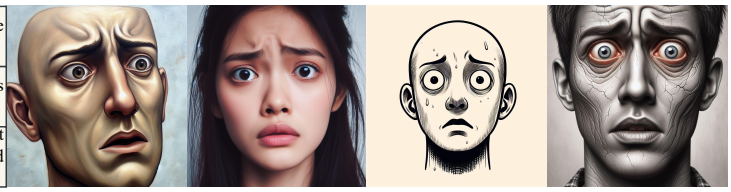

Figure 3: **Samples of Emo3D Dataset**: A glimpse into the rich diversity and complexity of our collected data, paving the way for insightful analysis and discovery.

Emolex.

Human Evaluation: We recognize that there may be inaccuracies in the emotion dataset, particularly in scenarios involving LLM-based generation. Our methodology primarily employs the distribution of emotions found within the text descriptions. Thus, evaluating the emotions in these text descriptions is deemed sufficient. In order to do this we introduced a human evaluation component to verify the accuracy of these text-based emotions. Two independent reviewers assessed the emotional content of 100 text samples, with a Cohen's kappa agreement score of 0.88. We then compared these human evaluations with our system's predictions. The results were encouraging, showing an 81% agreement rate between our system and the human evaluators, which indicates a strong correlation.

Figure 4: "Surprise" Emotion Word Cloud: closest words to "surprise" using Emolex based on cosine similarity of emotion distribution.

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 emotion 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 facial expression generation (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 XML 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 2.

CLIP Baseline: We employed a Multi-Layer Perceptron (MLP) structure on the CLIP (Radford

et al., 2021) backbone. What distinguishes this model from Pretrained LM Baselines is the incorporation of both image and text embeddings during training, effectively doubling the size of our dataset.

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. The encoder maps blendshape scores to their respective text and image representations and the decoder sample from the latent space and reconstructs the blendshape scores. 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 Mean Squared Error (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.

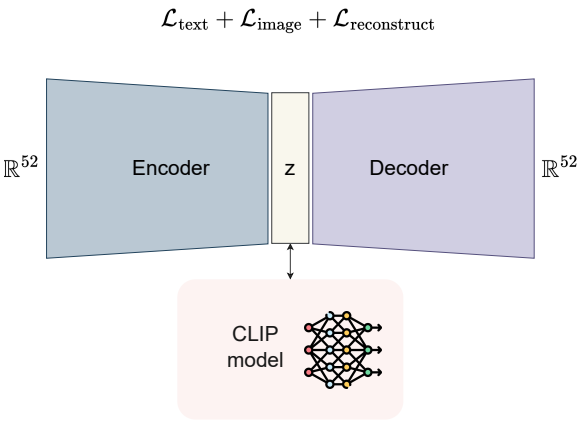

Figure 5: **Architecture and Training Process of the VAE-CLIP Model:** VAE CLIP concurrently reconstructs facial expressions while aligning their latent representation with corresponding text and image representations in the CLIP space.

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 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":

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

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

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

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

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 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.

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. When we examined the 3D model outputs, we observed 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.

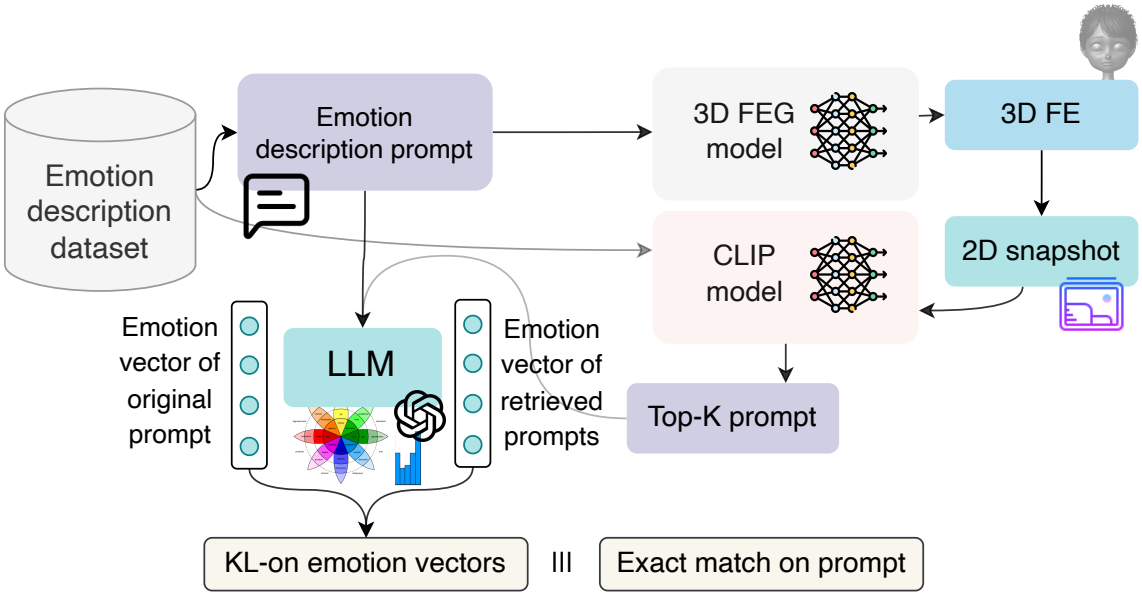

Figure 6: **Overview of the Emo3D Metric Calculation Process:** Our methodology in Emo3D metric entails selecting n prompts with a balanced emotion distribution. For a given input and a generated facial expression by a model, we project the 3D face onto a 2D image and employ zero-shot CLIP to identify 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 AE Clip models.

6 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 Language Models (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.

7 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 sores 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.

8 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*.

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.*