LLM-Generated Dataset for Speech-Driven 3D Facial Animation Models with Text-Controlled Expressivity

Pedro Rodrigues Corrêa, Paula Dornhofer Paro Costa
Dept. of Computer Engineering and Automation (DCA)
Universidade Estadual de Campinas (UNICAMP)
Campinas, Brazil
p243236@dac.unicamp.br, paulad@unicamp.br

***Abstract*—The field of speech-driven 3D facial animation has a fundamental limitation regarding style control by natural language: scarcity of datasets that pair nuanced language with direct, controllable facial parameters. To address this bottleneck, we introduce two core contributions: (1) a large-scale, synthetically generated dataset of over 53,000 items that link emotional transcripts, descriptive tags, and explicit facial descriptions to 51-dimensional blendshape vectors; and (2) a CLIP-based alignment module that learns a shared semantic manifold between this text and the corresponding expressions. Our model employs a Transformer-based blendshape autoencoder and a frozen CLIP text encoder, optimized via a cross-modal objective. We evaluate the structure of this learned manifold using t-SNE projections and latent space traversals, which demonstrate semantically coherent clustering and smooth, plausible transitions in expression. We release our dataset specification, generation methodology, and training recipe to facilitate reproducible research in text-driven facial animation [1].**

I. Introduction

Text-driven style control is a highly desirable modality for facial animation due to its intuitive and editable nature. However, progress in this domain is severely hampered by a critical data gap: the lack of large-scale public datasets that directly connect rich, descriptive text to animation-ready parameters, such as blendshapes. Existing resources typically offer either discrete emotion labels with static images or videos, or require complex processing of audio-visual data, which may not capture the specific nuances conveyed in text [1], [2]. This data scarcity makes it challenging to train models that can translate the semantic richness of language into fine-grained, controllable facial expressions. However, manually annotating tens of thousands of text prompts with corresponding blendshape vectors is prohibitively expensive and requires specialized expertise. A synthetic generation pipeline, while introducing its own biases, provides a scalable and reproducible method for creating a large, structured dataset necessary for training deep representation learning models.

To tackle this foundational problem, we take inspiration from [3] to make two contributions. First, we generate a large-scale synthetic dataset that pairs emotion-rich text with blendshape vectors derived from Facial Action Coding System (FACS) action units (AUs) [4]. Second, we train a CLIP-based alignment architecture designed to learn a shared, continuous manifold from this data [5]. This approach allows text to be projected directly into a space of valid facial expressions. We validate the quality of our learned space through qualitative analysis, including t-SNE visualizations and decoded latent space traversals, which confirm the model's ability to organize expressions semantically. Our primary goal is to provide a foundational dataset and a baseline model to spur further innovation in text-to-expression synthesis.

[1]https://github.com/AI-Unicamp/LLM-Generated-Dataset

II. Related Works

Progress in expressive facial animation has been driven by advances in both data and modeling. Our work builds upon three main areas.

A. Audio- and Text-Driven Animation

Speech-driven facial animation has achieved impressive lip-sync quality and co-speech gesture generation [6]. However, generating emotionally appropriate expressions purely from audio prosody remains challenging [7]–[9]. Text-driven methods offer more explicit control over expressivity, but have often been limited by the datasets discussed next.

B. Facial Expression Datasets

High-quality datasets are the bedrock of facial animation research. Seminal works like AffectNet and MEAD provide valuable pairings of faces with discrete emotion labels or audio [1], [2]. However, they lack the sentence-level, descriptive text needed to train models that can respond to nuanced commands like "*a slight, knowing smile*" versus "*a wide, joyous laugh*". Our work directly addresses this gap.

C. Vision-Language Models in Graphics

The success of contrastive models like CLIP [5] in linking text and images has inspired a wave of innovation in 3D content generation, including text-to-shape and text-to-texture synthesis [10], [11]. We adapt this paradigm by applying it to the specific task of aligning textual descriptions with facial

blendshape parameters, learning a controllable manifold of expressions.

III. Dataset Construction

Each sample in the dataset is a quadruple (t, e, d, b) where t is a short emotional transcript, e is a set of emotion tags, d is a one-sentence description of visible facial changes, and b is a 51-dimensional blendshape vector produced from action units by a fixed AU to blendshape weights mapping. Transcripts t are taken from three corpora for sentiment analysis [12]–[14], and e, d, and the au list are generated to be consistent with t. This structure serves training and testing for the CLIP-based module, which requires paired text and facial targets for semantic alignment.

A. Automatic Annotation Pipeline

We employ a large language model (Llama 3) to create our annotations, guided by carefully designed prompts with few-shot learning. For each input transcript, the pipeline generates a tuple (e, d, au). The prompt explicitly instructs the model to ground the description in observable facial cues (e.g., "eyebrows are drawn together", "lip corners are pulled upwards"), ensuring a strong link between the text and the resulting AUs. The annotation pipeline and a visual example of the dataset are shown in Figure 1.

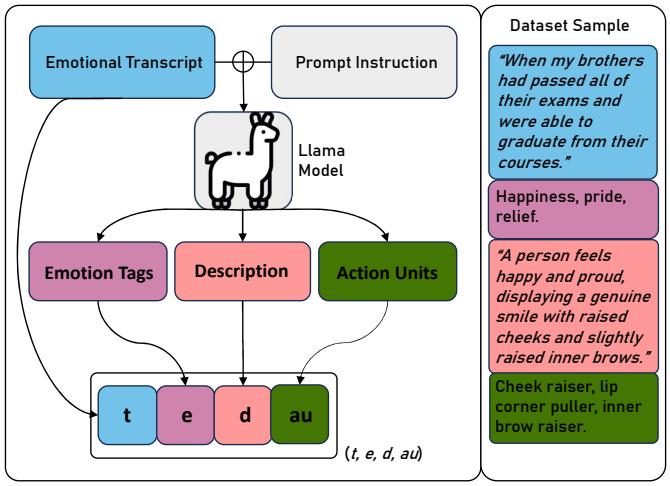

Fig. 1. Synthetic dataset annotation pipeline and example. Left: The Llama model generates emotion tags, a one-sentence description, and action units from an emotional sentence and instruction. AUs are mapped to a 51D blendshape vector by a fixed conversion. Right: an example item.

B. Action Units to Blendshapes

To target a controllable facial rig, we convert AUs to ARKit blendshape weights with a deterministic mapping (see Table I. From the 46 AUs in FACS, we select the 26 that most impact the face and correlate them with the 51 ARKit blendshapes.

TABLE I
Mapping Example of AUs to ARKit Blendshapes

Action Unit	ARKit Blendshape Weight
Inner Brow Raiser	*browInnerUp* (0.7)
Outer Brow Raiser	*browOuterUpLeft* (0.9), *browOuterUpRight* (0.9)
Lip Corner Puller	*mouthSmileLeft* (0.9), *mouthSmileRight* (0.9)

C. Quality control and Split

Recognizing that LLM outputs can be variable, we implemented a quality control step. We manually validated approximately one hundred random samples to check for semantic consistency across the $\{t, e, d, au\}$ tuple. This process helped refine the prompts to fix common issues. The final dataset was split into 80% training and 20% testing sets, stratified by emotion tags.

IV. CLIP-based alignment module

The CLIP-based alignment module structure can be seen in Figure 2, where it shows the Transformer autoencoder architecture combined with the pretrained CLIP Text Encoder and Text Projector.

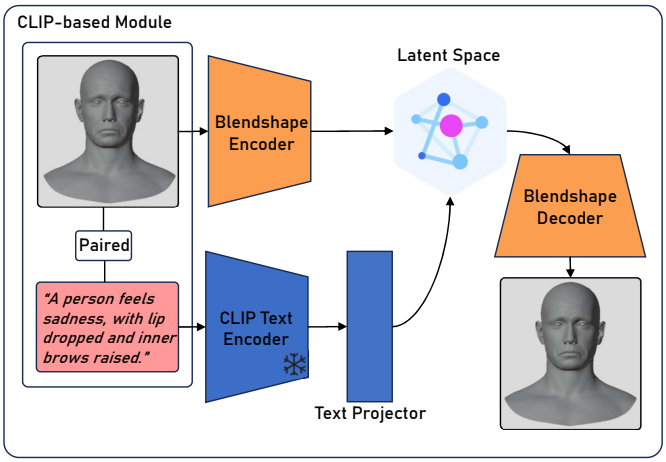

Fig. 2. CLIP-based alignment module. The blendshape autoencoder defines the expression manifold. The frozen CLIP text encoder plus a linear projector aligns text to the same space.

A. Architecture

1) Blendshape Autoencoder: Even though blendshape weight vectors are a non-sequential type of data, they can benefit from self-attention layers by capturing subtle, contextual relationships of certain facial expressions affecting multiple blendshapes simultaneously. Nevertheless, instead of the typical implementation of encoder-decoder blocks, we mirror the blendshape encoder ($\mathcal{E}_b$) to create the blendshape decoder ($\mathcal{D}_b$). This allows us to create an embedding space that stores high-dimensional information of the blendshapes to, subsequently, decode it as 51D vectors again.

2) Pretrained CLIP Text Encoder: In addition to the blendshape autoencoder, we concurrently encode the textual information from the dataset with a CLIP Text Encoder, keeping its parameters frozen during the process. Given the large-scale training pipeline of CLIP on text [5], it has the required robustness necessary to encode a plethora of text formats. Moreover, to correctly match the dimension of the embedded space created by the blendshape autoencoder, we devise a trainable linear text projector.

3) CLIP-base Module Training Pipeline: Inspired by [3], this module is trained via 3 losses represented by (1), where $\mathcal{L}_{ae}$ represents the autoencoder reconstruction loss, $\mathcal{L}_{emb}$ the embedding alignment loss and $\mathcal{L}_{multi}$ the multimodal reconstruction loss. The weights $\{\lambda_1, \lambda_2, \lambda_3\}$ represent hyperparameters that indicate which losses will be more impacted by the optimizer during training.

$$\mathcal{L}_{total} = \lambda_1 \mathcal{L}_{ae} + \lambda_2 \mathcal{L}_{emb} + \lambda_3 \mathcal{L}_{multi} \tag{1}$$

The autoencoder reconstruction (2) loss measures L2 distance between the predicted and input blendshapes.

$$\mathcal{L}_{ae} = \mathbb{E}_{b\sim\tau} \| \mathcal{D}_b(\mathcal{E}_b(b)) - b \| \tag{2}$$

To ensure the alignment of blendshapes and text, we use an embedding alignment loss (3) that implements cosine similarity between these two types of data. The S parameter indicates a randomly sampled text from (*t, e, d*).

$$\mathcal{L}_{emb} = 1 - \cos(\mathcal{P}_{text}(\mathcal{E}_{text}(S)), \mathcal{E}_b(b)) \tag{3}$$

Finally, we implement (4) that also measures the L2 distance between the input blendshape and the decoded version of the text data. This ensures multimodality alignment by making the paired blendshape-text data closer in the latent space.

$$\mathcal{L}_{multi} = \mathbb{E}_{b\sim\tau} \| \mathcal{D}_b(\mathcal{P}_{text}(\mathcal{E}_{text}(S))) - b \| \tag{4}$$

B. Optimization and regularization

In order to improve this module's generalization capabilities, we use data augmentation for textual data and minor random perturbations for the blendshape vectors. Inspired by [15], we apply synonym replacement and word order swap in random samples of (*t, e, d*). Regarding the blendshapes, we create 51D perturbation vectors and proceed to randomly populate them with numbers from 0.0 to 0.2 in each position. Since this method is meant to enhance the model's robustness, it is not desirable to insert larger perturbations due to potential changes in the conveyed emotion of the perturbed blendshapes. The perturbation vectors are added to the blendshape vectors *b* from the dataset.

C. Implementation Details

From [3], we borrow the hyperparameter values and employ weights of $\{\lambda_1, \lambda_2, \lambda_3\} = \{1, 10, 10\}$, a learning rate of 1×10^{-5}, batch size equal to 256 and 200 training epochs in a NVIDIA Quadro RTX 8000.

V. Evaluation in the latent space

A. t-SNE organization of embeddings

We embed the test set into the learned manifold and visualize the distribution with a two-dimensional t-SNE projection. Emotion tags and descriptions form distinct clusters, while transcripts aggregate more centrally and sparsely. This pattern reflects the stronger explicit expressivity in tags and descriptions, and the more heterogeneous or implicit emotion in transcripts. Figure 3 (left) shows the distribution.

B. Path sampling and decoded traversals

To assess local smoothness and coherence, we perform linear interpolation between random points from a straight line in the latent space, and decode the intermediate expressions (Figure 3, right). All the selected paths traversing through the expression manifold yielded graded changes in brow position, cheek activation, and mouth shape. These continuous, interpretable, and artifact-free transitions are highly desirable for controllable animations.

VI. Potential Applications and Risks

Our dataset has broad applications across entertainment, accessibility, and education domains. In media production, content creators can prototype character expressions using natural language descriptions, reducing animation time while enabling more intuitive artistic direction. The system also supports accessibility by providing text-based expression control for users with motor impairments, and enhances virtual reality and social platforms with nuanced avatar control.

The primary risks stem from the synthetic nature of our LLM-generated dataset, which may encode cultural biases and fail to represent diverse emotional expressions accurately across different demographics. While our system generates blendshape parameters rather than photorealistic imagery, integration with rendering systems could enable deepfake applications and emotional manipulation in synthetic media.

VII. Conclusion

We presented a methodology for creating a large-scale synthetic text-to-expression dataset and introduced a CLIP-based module that learns a shared semantic space from this data. Our analysis shows that the learned manifold is well-structured and supports smooth traversals that decode to plausible facial expressions. By releasing our dataset and tools, we aim to provide a valuable resource to the community, facilitating reproducible research and paving the way for more expressive, controllable text-driven style for 3D faces.

Our work has important limitations that open avenues for future research. The dataset is entirely synthetic, generated by an LLM, therefore possibly encoding biases present in the model or our prompts; thus, we recommend human screening before use in sensitive applications and caution that textual descriptions should not be treated as direct evidence of human affect. Moreover, the mapping from AUs to blendshapes is specific to the ARKit blendshape model and may need adaptation for other models. Future work should focus on validating the

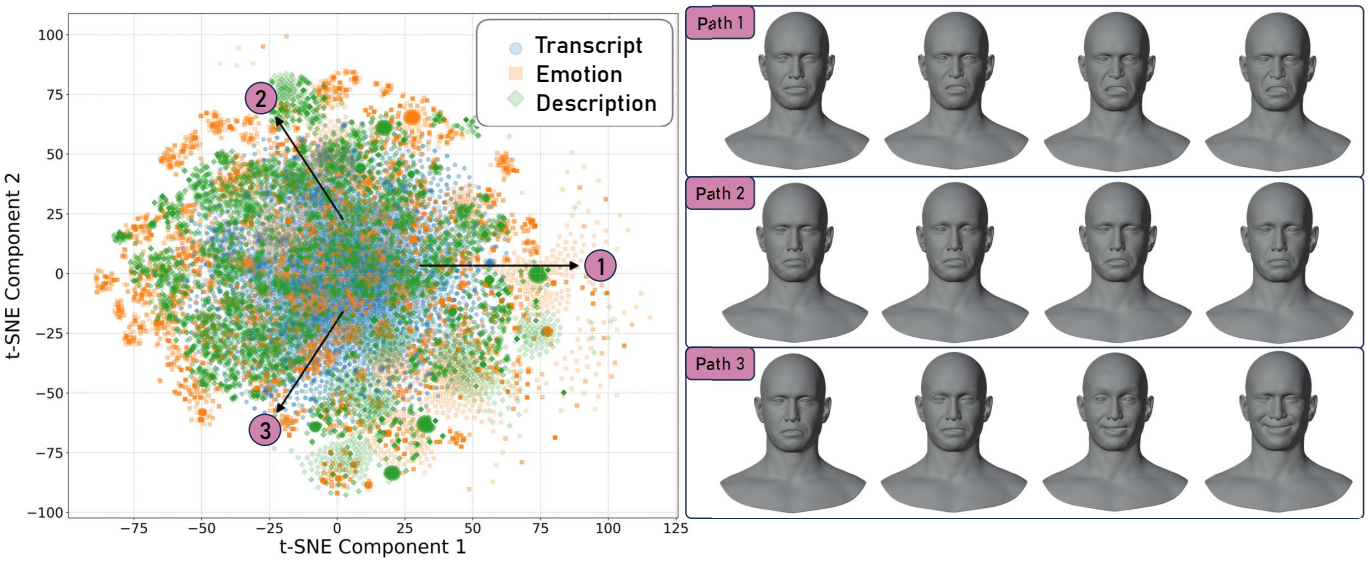

Fig. 3. Latent space analysis. Left: t-SNE of test embeddings showing clusters for tags and descriptions and a central region for transcripts. Right: decoded faces along three latent paths that begin near the center and end in distinct regions.

synthetic dataset with real-world data. We also plan to extend the evaluation method to incorporate quantitative metrics to systematically assess samples from the dataset.

Acknowledgment

This work was partially funded by the Coordenação de Aperfeiçoamento de Pessoal de Nivel Superior – Brasil (CAPES) – Finance Code 001 and by the São Paulo Research Foundation (FAPESP) under grant #2020/09838-0 (BI0S - Brazilian Institute of Data Science). Paula Dornhofer Paro Costa is also supported by the Ministry of Science, Technology, and Innovations, with resources from Law No. 8.248, of October 23, 1991, under the PPI-SOFTEX program, coordinated by Softex and published as Cognitive Architecture (Phase 3), DOU 01245.003479/2024-10. The authors are also affiliated with the Artificial Intelligence Lab, Recod.ai.

References

[1] K. Wang, Q. Wang, H. Peng, Z. Chen, J.-F. Wang, and C. Chen, "MEAD: A large-scale audio-visual dataset for emotional talking-face generation," in *European conference on computer vision*. Springer, 2020, pp. 599–615.

[2] A. Mollahosseini, B. Hasani, and M. H. Mahoor, "Affectnet: A database for facial expression, valence, and arousal computing in the wild," in *2017 12th IEEE international conference on automatic face & gesture recognition (FG 2017)*. IEEE, 2017, pp. 1–8.

[3] Y. Zhong, H. Wei, P. Yang, and Z. Wang, "ExpCLIP: Bridging text and facial expressions via semantic alignment," 2023. [Online]. Available: https://arxiv.org/abs/2308.14448

[4] P. Ekman and W. V. Friesen, "Facial action coding system," *Environmental Psychology & Nonverbal Behavior*, 1978.

[5] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, and I. Sutskever, "Learning transferable visual models from natural language supervision," 2021. [Online]. Available: https://arxiv.org/abs/2103.00020

[6] D. Cudeiro, T. Bolkart, C. Laidlaw, A. Ranjan, and M. J. Black, "Capture, learning, and synthesis of 3d speaking styles," in *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019, pp. 10 093–10 103.

[7] H. Zhou, Y. Liu, Z. Liu, P. Luo, and X. Wang, "Talking face generation by adversarially disentangled audio-visual representation," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 33, no. 01, 2019, pp. 9299–9306.

[8] T. Karras, T. Aila, S. Laine, A. Herva, and J. Lehtinen, "Audio-driven facial animation by joint end-to-end learning of pose and emotion," *ACM Trans. Graph.*, vol. 36, no. 4, 2017. [Online]. Available: https://doi.org/10.1145/3072959.3073658

[9] A. Richard, M. Zollhoefer, Y. Wen, F. de la Torre, and Y. Sheikh, "Meshtalk: 3d face animation from speech using cross-modality disentanglement," 2022. [Online]. Available: https://arxiv.org/abs/2104.08223

[10] A. Jain, B. Mildenhall, J. T. Barron, P. Abbeel, and B. Poole, "Zero-shot text-guided object generation with dream fields," 2022. [Online]. Available: https://arxiv.org/abs/2112.01455

[11] O. Michel, Z. Liu, V. Gkitsas, M. Saffar, D. N. Metaxas, and N. Paragios, "Text2mesh: Text-driven neural stylization for meshes," in *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

[12] K. R. Scherer and H. G. Wallbott, "Evidence for universality and cultural variation of differential emotion response patterning," *Journal of Personality and Social Psychology*, vol. 67, no. 1, p. 55, 1994. [Online]. Available: https://doi.org/10.1037/0022-3514.67.1.55

[13] D. Demszky, D. Movshovitz-Attias, J. Ko, A. Cowen, G. Nemade, and S. Ravi, "GoEmotions: A dataset of fine-grained emotions," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, D. Jurafsky, J. Chai, N. Schluter, and J. Tetreault, Eds. Online: Association for Computational Linguistics, Jul. 2020, pp. 4040–4054. [Online]. Available: https://aclanthology.org/2020.acl-main.372/

[14] S. Mohammad and F. Bravo-Marquez, "Emotion intensities in tweets," in *Proceedings of the 6th Joint Conference on Lexical and Computational Semantics (*SEM 2017)*, N. Ide, A. Herbelot, and L. Màrquez, Eds. Vancouver, Canada: Association for Computational Linguistics, Aug. 2017, pp. 65–77. [Online]. Available: https://aclanthology.org/S17-1007/

[15] J. Wei and K. Zou, "EDA: Easy data augmentation techniques for boosting performance on text classification tasks," in *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, K. Inui, J. Jiang, V. Ng, and X. Wan, Eds. Hong Kong, China: Association for Computational Linguistics, Nov. 2019, pp. 6382–6388. [Online]. Available: https://aclanthology.org/D19-1670/