

Supplementary Material: Audio-Driven Speech Animation with Text-Guided Expression

Sunjin Jung¹ , Sewhan Chun² , Junyong Noh¹ 

¹KAIST, Visual Media Lab

²NAVER Cloud



Figure 1: Examples of arbitrary continuous 15-frame segments

1. Examples of the collected texts

For the paired text data describing facial expressions, we randomly selected five continuous 15-frame segments from video sequences, as shown in Figure 1. The figure shows a part of a video sequence from the MEAD dataset [WWS*20], specifically sentence 025 of actor M003 exhibiting angry emotions at intensity level 3. From these segments, we generated pseudo text data using Large Language Models (LLMs) [GOO23]. The following are the two questions and examples of the sentences collected:

Q1. Describe the emotions portrayed by the person in the images.

- "The man looks like he is enraged."
- "The man looks like he is very angry and frustrated."
- "The man has his eyebrows furrowed, his eyes wide, and his teeth gritted together, conveying anger."
- ...

Q2. Provide a concise description of the person's facial expressions.

- "He wore a face as thunderous as a lightning storm."
- "He scowled with fury as if he had been unfairly wronged."
- "With a thunderous scowl, he appeared as if he wanted to set the world aflame."
- ...

We also collected various sentences describing the expressions using the original emotion labels. For this, we provided a guide sentence including 'he/she' to prevent gender-specific expressions for the speaker. For example, if the original emotion label of the given video was 'sad', the questions and answers are as follows:

Q3. Create five sentences expressing the 'sad' face (e.g., he/she appears deeply sorrowful, as if he/she has lost the entire world).

- "Her lips turned down like an unwatered plant."
- "His downcast eyes betrayed his gloomy mood."
- "Her eyes, usually twinkled with joy, were dull and lifeless, as if a storm had come and extinguished the light within."
- ...

Lastly, we added similar words for each emotion label. For example, if the original emotion label is 'angry', we collected 50 words to express various shades of anger, such as 'furious', 'enraged', 'upset', 'mad', 'displeased', and so on.

2. Implementation details of ExpCLIP

Because the implementation of ExpCLIP [ZWYW24] is not publicly available, we implemented ExpCLIP to the best of our understanding from the paper. We trained ExpCLIP on the same dataset as our method, but instead of predicting the weights of blendshapes as in the original paper, we trained ExpCLIP to predict the vertex positions. Consequently, we fine-tuned the animation generator with text embedding inputs, because training the generator using only facial animation data resulted in bumpy vertex outputs.

References

- [GOO23] GOOGLE: Gemini pro vision. <https://console.cloud.google.com/vertex-ai/publishers/google/model-garden/gemini-pro-vision>, 2023. 1
- [WWS*20] WANG K., WU Q., SONG L., YANG Z., WU W., QIAN C., HE R., QIAO Y., LOY C. C.: Mead: A large-scale audio-visual dataset for emotional talking-face generation. In *European Conference on Computer Vision* (2020), Springer, pp. 700–717. 1
- [ZWYW24] ZHONG Y., WEI H., YANG P., WANG Z.: Expclip: Bridging text and facial expressions via semantic alignment. In *Proceedings of the AAAI Conference on Artificial Intelligence* (2024), vol. 38, pp. 7614–7622. 1