Spinning the Wheel of Emotions: Fine-Grained Facial Editing with VLM-based Annotation

Sihang Chen
Victoria University of Wellington
Wellington, New Zealand
sihang.chen@vuw.ac.nz

Junliang Chen
Victoria University of Wellington
Wellington, New Zealand
junliang.chen@vuw.ac.nz

Fang-Lue Zhang
Victoria University of Wellington
Wellington, New Zealand
fanglue.zhang@vuw.ac.nz

Hedwig Eisenbarth
Victoria University of Wellington
Wellington, New Zealand
hedwig.eisenbarth@vuw.ac.nz

Kevin Romond
Victoria University of Wellington
Wellington, New Zealand
kevin.romond@vuw.ac.nz

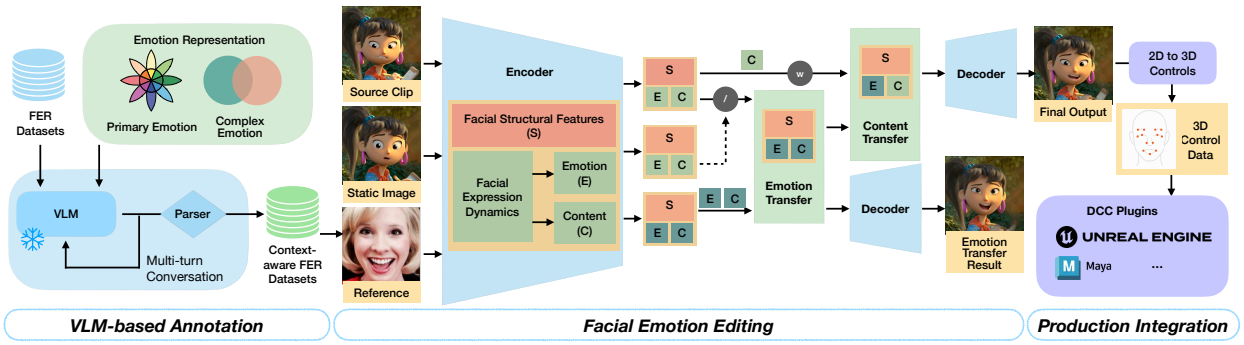

Figure 1: We introduce a novel emotion representation inspired by Plutchik's Wheel of Emotions theory [Plutchik 1980], a VLM-based annotation mechanism, and a facial emotion editing pipeline. Our framework disentangles emotion from performance content, enabling fine-grained, controllable, and production-ready emotion editing. (Images from [Luhn and Hjálmarsson 2021; Mollahosseini et al. 2019])

ABSTRACT

We present a practical framework for fine-grained facial emotion editing that bridges theoretical emotion models and animation workflows. Our approach introduces: (1) A novel emotion representation based on Plutchik's Wheel of Emotions, enabling both primary and complex emotions; (2) A VLM-based annotation pipeline generating context-aware emotion labels with human-level consistency; (3) An emotion-content disentanglement editing method enabling non-destructive emotion transfer and blending, while preserving performance fidelity. The animator-centric GUI provides real-time previews, timeline-based key-able weight controls for both global and regional adjustments, and exports to DCC software via plugins. Evaluations demonstrate superior annotation quality, improved temporal coherence in emotion editing, and production integration. Our framework establishes emotion as a dimension for controllable, practical emotion editing in facial animation.

Conference'17, July 2017, Washington, DC, USA

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM
https://doi.org/10.1145/3757376.3771388

CCS CONCEPTS

• **Computing methodologies** → **Animation**; **Image manipulation**; • **Information systems** → *Multimedia information systems.*

ACM Reference Format:
Sihang Chen, Junliang Chen, Fang-Lue Zhang, Hedwig Eisenbarth, and Kevin Romond. 2025. Spinning the Wheel of Emotions: Fine-Grained Facial Editing with VLM-based Annotation. In . ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/3757376.3771388

1 INTRODUCTION

Facial animation is critical for creating believable and engaging characters. However, current animation techniques struggle to efficiently modify facial emotion once facial animation is established. Animators are hindered by the complexity of facial emotions and the iterative and non-intuitive adjustment, especially when preserving coherence with the original animation. Recent learning-based methods have significantly advanced facial affective computing, leading to breakthroughs in both Facial Emotion Recognition (FER) and the generation of emotionally expressive facial animations. However, these approaches often prioritize automation over creative controls and offer limited operability within animation pipelines.

In this work, we propose frameworks for fine-grained, production-ready facial editing. Our approach introduces a novel emotion representation, inspired by Plutchik's Wheel of Emotions, which

captures a broader spectrum of emotional states while remaining intuitive to interpret. To support this representation, we develop a Vision-Language Model (VLM)-based annotation pipeline that generates context-aware emotion labels. These labels are integrated into a facial editing pipeline that disentangles emotion from performance content, enabling non-destructive transfer and modification. Finally, we provide an animator-centric Graphical User Interface (GUI) featuring real-time previews, timeline-based controls, and export functionality for animation workflows.

Our contributions can be summarized as follows:

- A novel emotion representation inspired by Plutchik's Wheel of Emotions, capturing a broad range of emotion states.
- A VLM-based annotation pipeline that produces context-aware emotion labels, enabling scalable and accurate emotion annotation for new emotion representation across datasets.
- A facial editing pipeline that disentangles emotion from performance content, supports non-destructive manipulation.

2 FACIAL EMOTION REPRESENTATION

We aim to provide an intuitive framework for facial emotion editing. Existing work often relies on categorical models [Ekman 1999] that limit emotion range or dimensional models [Russell 1980] that are less intuitive. Plutchik's Wheel of Emotion [Plutchik 1980] adds nuance to categorical models, but its practical use remains limited. To address these gaps, we propose a novel emotion representation and apply a VLM-based annotation method for re-labeling.

2.1 Emotion Representation

To capture complex emotion blends beyond primary categories, we base on Plutchik's Wheel of Emotion and take the concept of emotion blending by its dyads [Semeraro et al. 2021]. We define primary emotions as E^p and complex emotions as E^c:

$$E^p = [\text{ joy, trust, fear, surprise, sadness, disgust, anger, anticipation }]$$

$$E^c = [\text{ love, submission, alarm, disappointment, remorse, contempt, aggression, optimism, guilt, curiosity, despair, unbelief, envy, cynicism, pride, hope, delight, sentimentality, shame, outrage, pessimism, morbidness, dominance, anxiety, bittersweetness, ambivalence, frozenness, confusion }]$$

where the complex emotions follow the dyad combinations $\oplus$ in Wheel of Emotions (e.g. love = joy $\oplus$ trust) [Semeraro et al. 2021]:

$$E^c_i = E^p_j \oplus E^p_k = E^p_k \oplus E^p_j,$$
$$\text{where } i \in \{0, \ldots, 27\},\ j, k \in \{0, \ldots, 7\},\ j \neq k$$

All emotions are defined as: $\mathbf{E} = [\text{neutral}] \cup E^p \cup E^c$

2.2 VLM-based Facial Emotion Annotation

Recent advances in VLMs have demonstrated strong capabilities in context-aware interpretation of visual data [Deitke et al. 2024]. We use VLM for context-aware facial emotion annotation and labeling using our new emotion representation, as shown in Fig. 2(a).

Table 1: Metrics comparison on the FER+ dataset. We define the Level of Confidence (LoC) as the number of annotators in agreement (out of 10). Accuracy ACC_E for exact match, and ACC_{EM} for either exact or partial matches on primary emotions. Jaccard similarity J measures the similarity between annotation results and ground-truth.

Min GT LoC	ACC_{EM}	ACC_E	J	ACC_{EM}	ACC_E	J
	8 labels - w/ neutral			8 labels - w/o neutral		
0	70.33%	68.66%	0.6947	83.29%	80.87%	0.8203
6	75.65%	75.52%	0.7558	89.01%	88.82%	0.8891
8	83.44%	83.38%	0.8341	93.68%	93.60%	0.9365
10	92.58%	92.58%	0.9258	97.24%	97.24%	0.9726
	37 labels - w/ neutral			37 labels - w/o neutral		
0	60.86%	34.75%	0.4771	84.51%	45.13%	0.6470
6	64.65%	38.65%	0.5165	90.02%	50.09%	0.7007
8	72.45%	45.88%	0.5917	94.08%	55.50%	0.7480
10	85.07%	61.07%	0.7307	97.23%	67.64%	0.8242

We use Molmo-7B-D-0924 [Deitke et al. 2024] as the VLM baseline after testing among pretrained models. To ensure the VLM can better understand emotions and generate context-aware responses, we apply a multi-turn mechanism that allows the VLM to progressively refine its perception of facial emotion in a structured way. The VLM is prompted to describe facial features, infer emotional states, and contextualize the observed expression and emotion. The final annotations are parsed and filtered from VLM's conversations, and each annotation contains: *emotion*, *facial features*, and *scenarios*. In total, we annotated over 350,000 facial images from different FER datasets, samples as Fig. 2(b).

We evaluate our VLM-annotated label on FER+[Barsoum et al. 2016] on its original 8 labels and our new 37-label representation. We perform image feature extraction (CLIP[OpenAI 2025]) and clustering (UMAP), and the results indicate enhanced intra-class compactness and inter-class separability after the VLM annotation, as shown in Fig. 2(c). Qualitative evaluation, as shown in Tab. 1, also indicates VLM has competitive ability as human annotators, and can even detect emotions in some "neutral" faces. Metrics and other FER datasets details are provided in the supplementary.

3 FACIAL EMOTION EDITING

3.1 Methodology

Existing learning-based facial emotion editing techniques have achieved high visual realism. But most of them focus on generating emotional facial animation, failing to preserve critical performance elements such as lip-sync and eye movements, which limits their applications in practical workflows. To address these gaps, our approach disentangles emotion and content from facial animation, enabling controllable and production-ready editing.

3.1.1 Facial Emotion Components. Facial emotion recognition and perception can be primarily analyzed through Facial Action Units (AUs), as described in the widely accepted Facial Action Coding System (FACS) [Ekman and Friesen 1978]. However, it is often observed that even when facial muscles are relaxed, typically defined as a neutral face, emotional expressions may still be perceived [Said et al. 2009]. Therefore, in this paper, we propose that facial emotion can be perceived as comprising two distinct components:

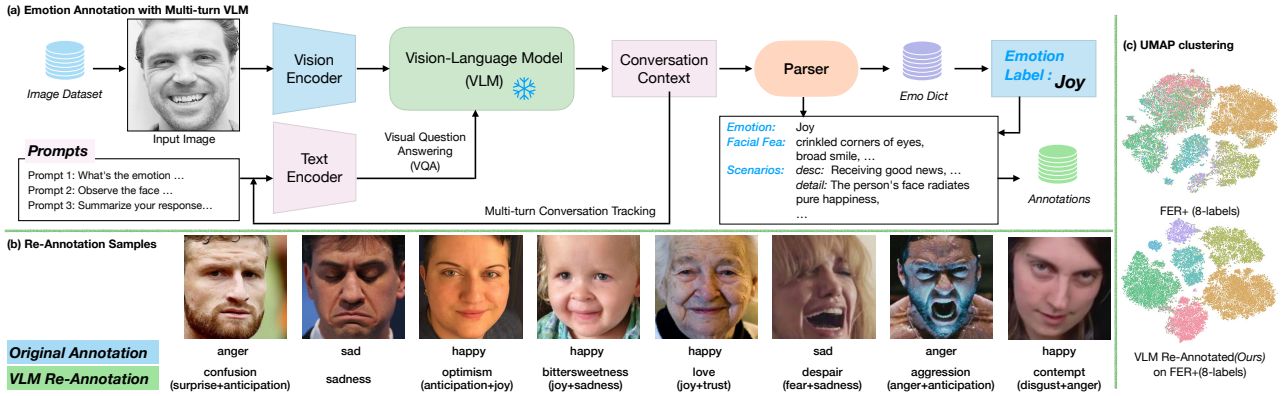

Figure 2: (a) VLM-based multi-turn annotation pipeline. (b) Comparison between original FER+ labels and VLM re-annotations. (c) UMAP clustering results on FER+ dataset, demonstrating improved intra-class compactness and inter-class separability.

- ***Facial Structural Features:*** stable, morphology-related attributes of the face that can influence perceived emotion.
- ***Facial Expression Dynamics:*** the temporal variations in facial muscle contractions that produce expression over time.

3.1.2 Emotion and Content in Facial Expression Dynamics. In practical animation workflows, facial motion is often driven by an actor's performance. In these cases, facial emotion editing is expected to be non-destructive and should preserve indispensable aspects of the facial animation for narrative coherence and production quality. We define these essential elements as content fidelity.

In this paper, we suppose content fidelity as animations around the mouth and eye regions. These regions directly influence key components of character performance, such as lip-sync for dialogue and eye-line positioning for gazing. Loss or distortion in these elements would easily undermine the believability and narrative clarity of facial animation. Based on this, we divide Facial Expression Dynamics into two complementary components:

- ***Emotion Component:*** the transient variations in facial muscle contractions that convey affective states.
- ***Content Component:*** the animation elements that should be preserved for narrative and functional integrity.

3.1.3 Emotion Editing Pipeline. Our emotion editing pipeline begins by decomposing the input animation according to this framework, separating *Facial Structural Features* from *Facial Expression Dynamics*, and disentangling the *Emotion* and *Content* components, enabling emotion editing while preserving content fidelity.

The emotion editing is primarily performed on the *Facial Expression Dynamics*, as *Facial Structural Features* are inherently tied to an individual's facial identity and morphology, so that direct modification would lead to unrealistic or identity-altering changes.

Our proposed pipeline performs two-step transfer, *Emotion Transfer* and *Content Transfer*, on *Facial Expression Dynamics* through weight-based transfer between corresponding components. Specifically, varying weights can be assigned to different components or facial regions, enabling more precise control over blending of emotional cues while preserving functionally critical elements.

Finally, the edited components are decoded as new facial animations with desired emotion adjustments and content integrity.

Figure 3: (a) Emotion Transfer results. (b) Step-by-step results of the pipeline (Emotion Transfer and Content Transfer), which preserve content fidelity such as lip-sync and gaze while modifying emotions.

3.2 Implementation Details

We implement our pipeline based on LivePortrait [Guo et al. 2024].

3.2.1 Input and Feature Extraction. The inputs include source facial images $\mathbf{I}^{source}$, reference emotion images $\mathbf{I}^{ref}$, and an optional static face image I^{static} for stable editing results. All inputs are first cropped and resized to the desired image size with the appropriate facial portion ratio, and then processed by LivePortrait's Appearance and Motion Extractor, obtaining the following facial features: 3D appearance feature volume f, canonical keypoints x_c, expression deformation δ, rotation R, scale s, and translation t. In addition, we define δ_c as the canonical expression deformation, similar to x_c, that follows $\delta = \delta_c R$. The calculated keypoints x, which serve as input to the warping module and decoder, are calculated as:

$$x = s \cdot (x_c R + \delta) + t = s \cdot (x_c + \delta_c) R + t$$

These extracted features align with Facial Emotion Components that: *Facial Structural Features* can be represented by 3D appearance feature volume f and canonical keypoints x_c; while *Facial Expression Dynamics* refer to canonical expression deformation δ_c.

3.2.2 Emotion Transfer. In this step, we calculate the calculated keypoints x_t^{emo} using facial features from $\mathbf{I}^{source}$ and $\mathbf{I}^{ref}$ for each frame time t. The corresponding canonical expression deformation offset $\delta_{c,t}^{ref}$ is interpolated from δ_c^{ref}, and then the keypoints after emotion transfer can be calculated as:

$$x_t^{emo} = s_t^{source} \cdot (x_{c,t}^{source} + \delta_{c,t}^{ref}) R_t^{source} + t_t^{source}.$$

The resulting facial images of x_t^{emo} after warping and decoding are shown in Fig. 3(a). The real-time feedback of emotion transfer results enables intuitive editing of emotions.

3.2.3 Content Transfer. In this step, we apply the animation content from δ_t^{source} back. The goal is to preserve facial motion content while maintaining the emotional characteristics from the previous step. For each frame time t, we perform a weighted calculation:

$$\delta_{c,t}^{final} = W(\delta_{c,t}^{ref}, \delta_{c,t}^{source}, w_t^{global}, w_t^{eyes}, w_t^{mouth})$$

where w_t^{global} is the global weight of content transfer, determining how much δ_t^{final} follows δ_t^{emo} globally for expressiveness; and regional weights w_t^{eyes} and w_t^{mouth} determine how δ_t^{final} should be more aligned with δ_t^{source} on the eyes and mouth regions to preserve content fidelity. The final keypoints are calculated as:

$$x_t^{final} = s_t^{source} \cdot (x_{c,t}^{source} + \delta_{c,t}^{final}) R_t^{source} + t_t^{source}.$$

The step-by-step results are shown in Fig. 3(b).

3.2.4 Decoding and Output. The final x_t^{final} is passed through the warping module and decoder to generate the final facial image. With all frames processed, our pipeline produces sequences with modified emotion expressiveness and performance content fidelity.

3.2.5 Integration to DCC software. We use MediaPipe [Google 2025] to estimate the ARKit face blendshape weights, making the output compatible with DCC software and facial animation workflows through plugins for further editing or real-time applications.

3.3 GUI

We develop a GUI for intuitive and precise control over the emotion editing process, as illustrated in Fig. 4. The *Preview* panel displays the original frame, the intermediate emotion-infused face, and the final output in real-time, enabling intuitive comparison among steps. The *Timeline* and *Keyframe* panels include a current frame indicator slider, emotion reference thumbnails, and keyframe visualization for weights of content transfers, allowing detailed control over content fidelity over time. The *Content Control* panel provides keyable global and regional (eyes, lips) weight sliders. The *Output* panel provides video output and CSV exports for DCC integration.

Overall, the interface is designed for usability and flexibility, offering real-time feedback and precise control through the pipeline.

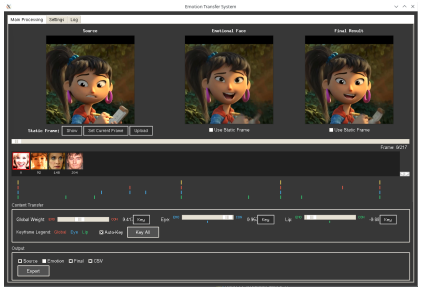

Figure 4: The GUI for facial emotion editing.

4 DISCUSSION AND CONCLUSION

We present a complete framework that bridges the gap between the theoretical emotion model and its practical application in facial animation. By introducing an expanded emotion representation with VLM-based annotation and disentanglement of facial emotion components, our method enables precise emotion editing while preserving content fidelity. The animator-centric GUI addresses industry needs for intuitive, non-destructive workflows.

Limitations include VLM's uncertainty and bias, content-emotion ambiguity in facial animation, and errors accumulated through data conversion. Future work will explore the validation of VLM annotation methods, the disentanglement of facial animation, and more robust and intuitive editing methods. Multi-modal methods also present compelling opportunities for further research.

In conclusion, we establish a framework that integrates emotion modeling with practical facial animation, pointing toward more robust, flexible, and animator-centric editing methods.

REFERENCES

Emad Barsoum, Cha Zhang, Cristian Canton Ferrer, and Zhengyou Zhang. 2016. Training deep networks for facial expression recognition with crowd-sourced label distribution. In *Proceedings of the 18th ACM International Conference on Multimodal Interaction*. ACM, Tokyo Japan, 279–283. https://doi.org/10.1145/2993148.2993165

Matt Deitke, Christopher Clark, Sangho Lee, Rohun Tripathi, Yue Yang, Jae Sung Park, Mohammadreza Salehi, Niklas Muennighoff, Kyle Lo, Luca Soldaini, Jiasen Lu, Taira Anderson, Erin Bransom, Kiana Ehsani, Huong Ngo, YenSung Chen, Ajay Patel, Mark Yatskar, Chris Callison-Burch, Andrew Head, Rose Hendrix, Favyen Bastani, Eli VanderBilt, Nathan Lambert, Yvonne Chou, Arnavi Chheda, Jenna Sparks, Sam Skjonsberg, Michael Schmitz, Aaron Sarnat, Byron Bischoff, Pete Walsh, Chris Newell, Piper Wolters, Tanmay Gupta, Kuo-Hao Zeng, Jon Borchardt, Dirk Groeneveld, Crystal Nam, Sophie Lebrecht, Caitlin Wittlif, Carissa Schoenick, Oscar Michel, Ranjay Krishna, Luca Weihs, Noah A. Smith, Hannaneh Hajishirzi, Ross Girshick, Ali Farhadi, and Aniruddha Kembhavi. 2024. Molmo and PixMo: Open Weights and Open Data for State-of-the-Art Vision-Language Models. https://doi.org/10.48550/arXiv.2409.17146 arXiv: 2409.17146 [cs].

Paul Ekman. 1999. Basic Emotions. In *Handbook of Cognition and Emotion* (1 ed.). Wiley, 45–60. https://doi.org/10.1002/0470013494.ch3

Paul Ekman and Wallace V. Friesen. 1978. Facial Action Coding System. (1978). https://doi.org/10.1037/t27734-000

Google. 2025. Mediapipe Solutions Guide. https://ai.google.dev/edge/mediapipe

Jianzhu Guo, Dingyun Zhang, Xiaoqiang Liu, Zhizhou Zhong, Yuan Zhang, Pengfei Wan, and Di Zhang. 2024. LivePortrait: Efficient Portrait Animation with Stitching and Retargeting Control. *arXiv preprint arXiv:2407.03168* (2024).

Matthew Luhn and Hjalti Hjálmarsson. 2021. Sprite Fright. https://studio.blender.org/projects/sprite-fright/

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 (Jan. 2019), 18–31. https://doi.org/10.1109/TAFFC.2017.2740923

OpenAI. 2025. openai/clip-vit-large-patch14-336 · Hugging Face. https://huggingface.co/openai/clip-vit-large-patch14-336

Robert Plutchik. 1980. A General Psychoevolutionary Theory of Emotion. In *Theories of Emotion*. Elsevier, 3–33. https://doi.org/10.1016/B978-0-12-558701-3.50007-7

James A. Russell. 1980. A Circumplex Model of Affect. *Journal of Personality and Social Psychology* 39, 6 (Dec. 1980), 1161–1178. https://doi.org/10.1037/h0077714

Christopher P. Said, Nicu Sebe, and Alexander Todorov. 2009. Structural resemblance to emotional expressions predicts evaluation of emotionally neutral faces. *Emotion* 9, 2 (2009), 260–264. https://doi.org/10.1037/a0014681

Alfonso Semeraro, Salvatore Vilella, and Giancarlo Ruffo. 2021. PyPlutchik: Visualising and comparing emotion-annotated corpora. *PLOS ONE* 16, 9 (Sept. 2021), e0256503. https://doi.org/10.1371/journal.pone.0256503