Supplementary Material: MReactor: Offline Multiple Appropriate Facial Reaction Generation with Hierarchical Cognitive Disentanglement

A. Overview

This supplementary document contains further details on related work (Sec. B), methodology (Sec. C), experiments (Sec. D), and limitations (Sec. E).

B. Related Work

Related MAFRG Techniques. In this section, we review some representative methodologies in Multiple Appropriate Facial Reaction Generation (MAFRG). Disentangling the underlying driving factors is a crucial component in this task. Prior disentanglement methods either manipulate global latent directions, as in InterfaceGAN [15], or adversarially decouple audio and visual cues for audio driven face synthesis [25, 27]. Latent diffusion models for facial reaction generation, such as ReactDiff [6], EMO [17], and Dreamtalk [10], enhance realism but still leave local (lip, eye, pose) and coarse (head, appearance, expression) factors entangled. Attention-based pipelines have also advanced AFRs generation, including UniFarn with a unified Transformer backbone [7], hierarchical decouple fusion [9], and PerFRDiff for personalised editing [29], along with applications in micro-expression synthesis [3], attention fusion for FER [13], adversarial transfer [24], anonymisation [5], and cross-domain recognition [4]. Although prior approaches enhance robustness and highlight salient regions, they lack explicit control over individual facial components, do not jointly reason over detailed and global expressive patterns, and provide no mechanism for conditioning on subject identity and semantics. In contrast, our MREACTOR performs disentanglement at multiple levels and applies attention-driven motion modelling, enabling coherent, controllable, and subject-aware AFRs generation.

C. Methodology

C.1. Multimodal Perception Behaviour

MPB first utilizes modality-specific encoders (wav2vec [1], RoBERTa [8], and FaceVerse [19]) to generate frame-level representations. Next, building on the work of Zhang et al. [26], a temporal correlation learning module [26] is employed to capture both intra- and inter-modal dependencies, allowing the system to identify frames with prominent behavioral signals. To maintain structural consistency across modalities, MPB applies temporal masking [26], using binary masks to suppress segments with weak behavioral evidence or low cross-modal support. This masking process enforces temporal alignment, and reduces noise for the HCD module.

The resulting masked multimodal representations $\{\hat{\mathbf{A}}_S^{1:T}, \hat{\mathbf{L}}_S^{1:T}, \hat{\mathbf{F}}_S^{1:T}\}$ provide stable, behaviourally aligned cues that support the subsequent coarse- and fine-grained disentanglement in HCD.

C.2. Hierarchical Cognition Disentanglement

Fine-grained Facial Disentanglement. In our fine-grained facial disentanglement, the *lip* $\overline{F}_l^{1:T}$, *eye* $\overline{F}_e^{1:T}$, and *pose* $\overline{F}_p^{1:T}$ embeddings are extracted following the corresponding established methods in prior work [21, 22, 28]. Specifically, the lip embedding is derived based on the articulation-focused representation of Zhou et al. [28], the eye embedding follows the blink and attention heatmap formulation proposed by Yu et al. [22], and the head pose embedding is obtained using the geometric representation adopted in Wood et al. [21]. These prior formulations provide the basis for isolating local facial cues, which we incorporate to construct our fine-grained facial feature subsets.

Coarse-grained Facial Disentanglement. In our coarse-grained facial disentanglement, the global *head* $\tilde{F}_h^{1:T}$, *appearance* $\tilde{F}_a^{1:T}$, and *expression* $\tilde{F}_e^{1:T}$ embeddings are extracted following established methods in prior work [2, 18, 23]. Specifically, the head embedding is obtained using the head orientation and motion representation provided by Baltrušaitis et al. [2], the appearance embedding follows the identity-preserving facial representation introduced in Zhang et al. [23], and the expression embedding is derived based on the affective facial expression estimation framework of Toisoul et al. [18]. These prior formulations provide the foundation for capturing stable and global facial attributes, enabling our model to align coarse-grained speaker cues with linguistic and acoustic information to ensure expressive consistency across modalities.

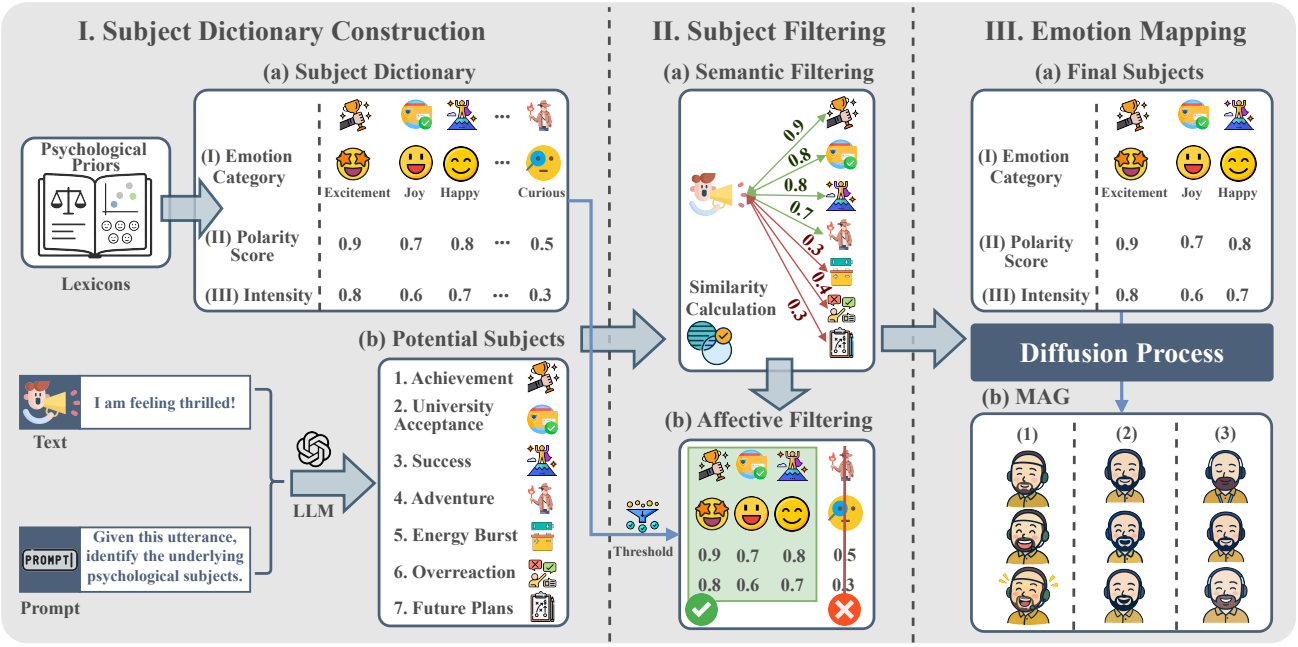

Figure 1. **Three-stage Expressive Recovery Pipeline.** (I) **Subject Dictionary Construction:** Using psychology-informed affective lexicons (e.g., ANEW, VADER, SentiWordNet), we build an emotion dictionary consisting of emotion categories, polarity scores, and intensity values. An LLM receives the utterance (e.g., "I am feeling thrilled!") and generates a set of potential psychological subjects. (II) **Subject Filtering:** Subjects are first filtered by semantic similarity with the utterance and then validated by affective priors to remove emotionally weak or polarity-mismatched entries. (III) **Emotion Mapping:** The remaining subjects are mapped to their corresponding emotion categories, polarity scores, and intensity values, forming psychologically grounded emotion priors. These priors guide the diffusion model to produce multiple facial reactions with coherent emotional alignment.

C.3. Multiple AFR Generation

Expressive Recovery. The expressive recovery component is implemented as a three-stage, psychologically grounded pipeline that enables controlled and diverse emotional response generation. As shown in Fig. 1, the system progressively constructs, filters, and maps subject-related cues into quantitative emotion priors that guide the final generative process. We now describe each stage in detail.

Stage 1: Building the Expressive Subject Dictionary. This stage aims to identify psychologically meaningful subject candidates that may underlie the speaker's utterance. To support this process, we first construct a psychological prior lexicon that links a wide range of subjects to three affective attributes derived from established affective datasets: (1) emotion category, (2) sentiment polarity, and (3) affective intensity. These mappings are obtained from psychology-informed resources such as ANEW, VADER, SentiWordNet, and LIWC [11, 12, 14, 20].

Next, given a speaker's linguistic features $\hat{\mathbf{L}}_S^{1:T}$, we employ a psychologically oriented prompt $\mathcal{P}$ to query a large language model (GPT-4). The prompt asks the model to infer the underlying psychological subjects (e.g., achievement, success, acceptance) that could plausibly evoke an emotional reaction:

$$\mathcal{T} = \text{LLM}(\mathcal{P}, \hat{L}_S^{1:T}), \tag{1}$$

where $\mathcal{T} = \{l_1, l_2, \ldots, l_k\}$ forms the initial expressive subject dictionary. These subjects are semantically tied to the utterance and serve as potential affective triggers for multiple AFRs.

Psychological Prompt: *Given this utterance, identify the underlying psychological subjects.*
Input Sentence: *"I am feeling thrilled!"*
LLM Output:

$$\mathcal{T} = \begin{Bmatrix} \text{Achievement, University Acceptance,} \\ \text{Success, Adventure, Energy Burst,} \\ \text{Overreaction, Future Plans} \end{Bmatrix}$$

These subjects constitute the initial candidate pool, which will subsequently be refined through semantic and affective filtering.

Stage 2: Filtering the Generated Subjects. After obtaining the initial subject dictionary $\mathcal{T}$, the next step is to ensure that each subject is both semantically relevant and affectively meaningful. To achieve this, we apply a two-stage filtering procedure—*semantic filtering* and *affective filtering*—informed by psychologically grounded affective lexicons [11, 12, 14, 20].

Semantic Filtering. Each candidate subject l_i is first evaluated for semantic alignment with the speaker's utterance. We compute the cosine similarity between the embedding of l_i and the embedding of the speaker text $\hat{\mathbf{L}}_S^{1:T}$, retaining only those subjects whose similarity exceeds a threshold:

$$\begin{aligned} s^{\text{sem}}(l_i) &= \cos\Big(e(l_i),\, e(\hat{\mathbf{L}}_S^{1:T})\Big)\,, \\ l_i \in \mathcal{T}_{\text{sem}} \quad &\text{if } s^{\text{sem}}(l_i) \geq \tau_1, \end{aligned} \tag{2}$$

where τ_1 is the semantic similarity threshold. This step removes hallucinated or weakly related subjects, ensuring that the remaining entries are topically consistent with the speaker's intent.

Affective Filtering. Next, each semantically valid subject is evaluated using our psychological prior lexicon, which provides its emotion category, sentiment polarity score, and affective intensity. A subject l_i is retained only if its polarity and intensity are consistent with the overall emotional profile of the utterance (e.g., positive, high-arousal in the case of "thrilled"):

$$l_i \in \mathcal{T}_{\text{aff}} \quad \text{if} \quad |s_i - s_{\text{text}}| \leq \tau_2 \text{ and } \phi_i \geq \tau_3, \tag{3}$$

where τ_2 controls the acceptable polarity deviation between a subject and the utterance, and τ_3 denotes the minimum affective intensity required for a subject to be considered emotionally meaningful. All thresholds are set to 0.5 in our implementation. This step removes emotionally neutral, weak, or contradictory subjects.

Semantic Filtering: Ensures that subjects are *topically relevant* by measuring embedding similarity with the speaker's text.
Affective Filtering: Ensures that subjects are *emotionally valid* by checking whether their polarity and intensity, derived from psychological lexicons (ANEW, VADER, SentiWordNet), match the emotional profile of the utterance.

Why both filters? Semantic filtering guarantees that the subjects "make sense" in context, while affective filtering guarantees that they "feel right" emotionally. Using both ensures that the retained subjects are simultaneously coherent with the speaker's meaning and psychologically consistent with their emotional state.

The filtered set $\mathcal{T}_{\text{filtered}} = \mathcal{T}_{\text{sem}} \cap \mathcal{T}_{\text{aff}}$ forms a refined pool of psychologically grounded subjects, which will be mapped into emotion priors in Stage 3.

Stage 3: Mapping Subjects to Emotion Priors. After filtering, the remaining subjects must be converted into quantitative emotional descriptors that can guide the downstream generative model. Rather than assigning scores arbitrarily, we retrieve each subject's affective attributes from our preconstructed psychological lexicon, which provides three key dimensions for every subject entry: (1) its emotion label (category), (2) sentiment polarity score s_i, and (3) affective intensity $\phi(e_i)$. This lexicon is derived from established affective resources such as ANEW, VADER, SentiWordNet, and LIWC [11, 12, 14, 20].

Specifically, for each filtered subject l_i, we look up its affective attributes in the lexicon:

$$\begin{aligned} \text{EmotionLabel}(l_i) &= \text{Emotion}(l_i), \\ s_i &= \text{Polarity}(l_i), \\ \phi(e_i) &= \text{Intensity}(l_i). \end{aligned} \tag{4}$$

The polarity score s_i reflects the emotional valence (positive, neutral, or negative), whereas the intensity $\phi(e_i)$ captures the arousal strength of the associated emotion.

Emotion Label: Indicates the emotion category associated with a subject (e.g., *Excitement*, *Joy*, *Happiness*).
Sentiment Score s_i: Retrieved from the lexicon to represent the subject's polarity. Higher values indicate positive emotions (e.g., Excitement), while lower values correspond to negative ones (e.g., Anger).
Emotion Intensity $\phi(e_i)$: Represents the subject's affective arousal level, also obtained from the lexicon. High-intensity emotions (e.g., Ecstatic) yield larger values, whereas mild emotions (e.g., Content) receive smaller values.

As an example, the filtered subjects *Achievement*, *University Acceptance*, and *Success* map to the following affective attributes:

Subject	Emotion Label	s_i	$\phi(e_i)$
Achievement	Excitement	0.9	0.8
University Acceptance	Joy	0.7	0.6
Success	Happiness	0.8	0.7

These psychologically grounded emotion priors $(\text{EmotionLabel}(l_i), s_i, \phi(e_i))$ ensure that the model receives precise guidance regarding both the emotional

tone and the emotional strength implied by each subject, enabling the generation of coherent and emotionally appropriate facial reactions.

C.4. Training Strategy

Following the reverse diffusion formulation in [29], the diversity loss generates multiple plausible AFRs by denoising a noisy listener motion sequence $P_{L,d}^{1:T}$ to recover the clean ground-truth listener motion $P_L^{1:T}$. This generation process is supervised by optimizing an MSE loss:

$$\mathcal{L}_g = \mathbb{E}_{P_L^{1:T},\epsilon}\left[\left\| P_L^{1:T} - \hat{P}_L^{1:T}\left(P_{L,d}^{1:T},\, c,\, v,\, d\right)\right\|^2\right], \quad (5)$$

where $\hat{P}_L^{1:T}(P_{L,d}^{1:T},\, c,\, v,\, d)$ denotes the denoised listener AFR prediction obtained from the noisy motion input in the reverse diffusion process. The conditioning set $c = \{\mathbf{C},\, \mathbf{S}\}$ incorporates speaker behaviour cues and motion–smoothness control, while v provides stochastic variation that enables the model to generate diverse AFRs. This diversity-driven objective ensures that the model produces multiple consistent yet distinct listener reactions that align with the multimodal context.

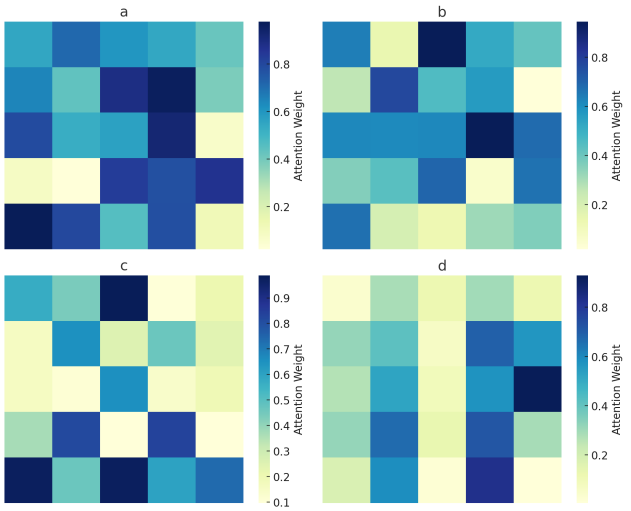

Figure 2. Visualization of different attention mechanisms on *MREACTOR*: (a) *self-attention* in [29], (b) *cross-attention* in [29], (c) combined *self- and cross-attention*, and (d) our MAG module's *relative and subject attention*.

D. Experiments

D.1. Attention Analysis

As shown in Figure. 3, the self-attention (a), cross-attention (b), and their combination (c) fail to highlight the behaviourally salient cues, producing diffuse attention patterns that overlook the core temporal and affective dependencies. In contrast, our proposed relative and subject attention (d) learn substantially more structured and meaningful weights: the relative attention incorporates the decay term $\epsilon(x-y)$ to suppress distant irrelevant frames and emphasise contextually relevant recent cues, while the subject attention modulates the value space using the expressive subject prior $\mathbf{S}$, enabling the model to capture sentiment strength and emotion intensity. Together, these mechanisms allow MREACTOR to focus on the truly informative behavioural signals and generate more coherent and semantically grounded AFRs.

In contrast, MREACTOR goes beyond surface-level mimicry by producing personalised and semantically coherent reactions that remain temporally stable while preserving expressive diversity. As illustrated in Figure 3, our model conditions on the full multi-modal speaker behaviour (audio, transcript, and face) and generates two distinct yet appropriate reaction trajectories for the same speaker behaviour. Compared with existing baselines [16, 29] , which often under-react or collapse into similar neutral patterns, MREACTOR is able to generate multiple appropriate AFRs for the same speaker behaviour. The two sampled AFRs display nuanced and psychologically plausible variations in gaze, eyebrow motion, and mouth dynamics, reflecting different yet contextually coherent interpretations of the unfolding narrative. This demonstrates that MREACTOR not only preserves identity and affective consistency but also captures the inherent one-to-many nature of human listener responses.

D.2. The Number of Employed Attention Layers

Finally, we analyze how the depth of the proposed *relative and subject attention* influences AFR performance. The system remains relatively stable across different attention depths, yet a clear performance optimum appears. As the number of layers increases from 1 to 9, the variations in *FRCorr* and *FRVar* stay within moderate ranges (below 0.44 and 0.13, respectively), while *FRDiv* changes within 0.26 (see Figure 2). For fidelity measures, *FRDist* and *FRSyn* vary less than 15.74 and 1.61, corresponding to relative fluctuations of only 8.1% and 3.3%. These observations indicate that the attention mechanisms remain robust to stacking depth. Notably, six attention layers provide the most favorable performance, achieving the highest *FRCorr*, *FRDiv*, and *FRVar* together with the lowest *FRDist* and *FRSyn*. Beyond six layers, performance gradually weakens across all metrics, suggesting that excessive depth introduces redundancy and reduces temporal discrimination. Overall, six layers strike the most effective balance between contextual modeling, expressive richness, and temporal fidelity.

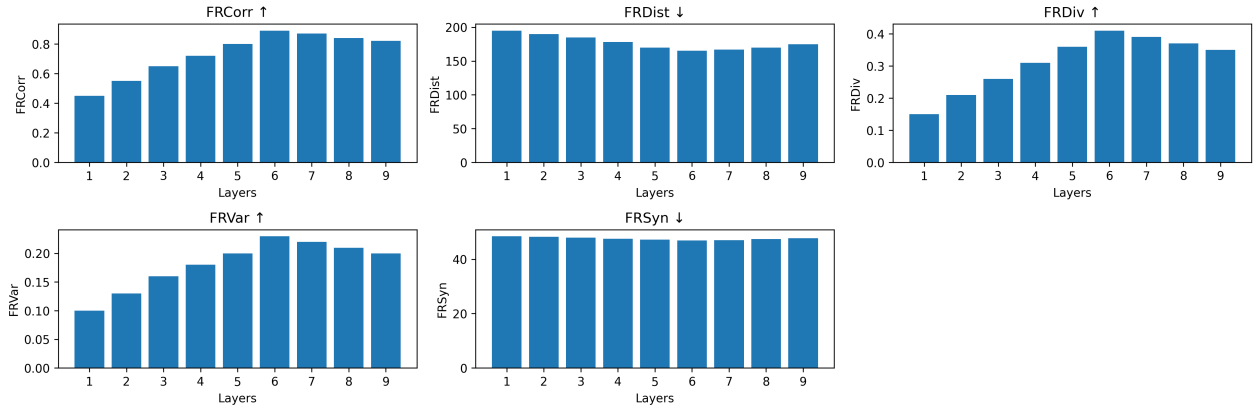

Figure 3. Impacts of different number of attention layers on the facial reaction generation performances.

E. Limitations and Future Work

Limitations. Despite the performance gains achieved by MREACTOR, several limitations remain. First, the current framework operates in an offline manner, requiring access to complete input sequences before generating facial reactions. While this design facilitates globally coherent and context-aware responses, it restricts applicability in real-time interactive scenarios, such as live dialogue systems, where low-latency generation is essential. Second, MREACTOR is primarily trained and evaluated on the REACT2025/MARS dataset. Although this dataset provides high-quality annotations, it may not fully represent the diversity of real-world interactions across cultures, languages, and social contexts, which could limit generalization and introduce dataset-specific biases.

Ethical Considerations. The generation of facial reactions raises ethical considerations related to misuse and social impact. In particular, generative facial technologies may be exploited to create misleading or manipulative content, including deceptive expressions or emotionally persuasive media. To address these concerns, future research and deployment should emphasize transparency regarding AI-generated content, systematic evaluation of demographic and cultural biases, and responsible data practices that respect consent and privacy. We advocate that MREACTOR be applied primarily in beneficial contexts, such as assistive communication systems and socially aware virtual agents, and that its deployment be accompanied by clear usage guidelines and ongoing ethical review.

Future Work. Future work will explore extending MREACTOR to online and real-time settings by incorporating causal attention mechanisms and incremental inference strategies, enabling low-latency facial reaction generation. To improve robustness and fairness, we plan to evaluate and adapt the model on more diverse datasets, including multi-party, cross-cultural, and cross-linguistic interactions. In addition, integrating longer-term psychological factors, such as personality traits and emotional history, may allow for more personalized and consistent facial behaviors. Finally, modeling additional non-verbal signals, including gaze dynamics and upper-body gestures, represents a promising direction for enhancing expressiveness and realism in human-centered interactive systems.

References

[1] Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. wav2vec 2.0: A framework for self-supervised learning of speech representations. *In NeurIPS*, 2020. 1

[2] Tadas Baltrušaitis, Amir Zadeh, Yao Chong Lim, and Louis-Philippe Morency. Openface 2.0: Facial behavior analysis toolkit. In *FG*, 2018. 1

[3] Xinqi Fan, Ali Raza Shahid, and Hong Yan. Edge-aware motion based facial micro-expression generation with attention mechanism. *Pattern Recognition Letters*, 2022. 1

[4] Yanli Ji, Yuhan Hu, Yang Yang, and Heng Tao Shen. Region attention enhanced unsupervised cross-domain facial emotion recognition. *IEEE Trans. Knowl. Data Eng.*, 2021. 1

[5] Zhenzhong Kuang, Xiaochen Yang, Yingjie Shen, Chao Hu, and Jun Yu. Facial identity anonymization via intrinsic and extrinsic attention distraction. In *CVPR*, 2024. 1

[6] Jiaming Li, Sheng Wang, Xin Wang, Yitao Zhu, Honglin Xiong, Zixu Zhuang, and Qian Wang. Reactdiff: Latent diffusion for facial reaction generation. *Neural Netw.*, 2025. 1

[7] Cong Liang, Jiahe Wang, Haofan Zhang, Bing Tang, Junshan Huang, Shangfei Wang, and Xiaoping Chen. Unifarn: Unified transformer for facial reaction generation. In *ACM MM*, 2023. 1

[8] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized bert pretraining approach. *arXiv*, 2019. 1

[9] Qincheng Lv, Xiaofeng Liu, Jie Li, Rongrong Ni, Pujun Xue, and Siyang Song. Hierarchical multimodal decoupling-fusion framework for offline multiple appropriate facial reaction generation. In *ICASSP*, 2025. 1

[10] Yifeng Ma, Shiwei Zhang, Jiayu Wang, Xiang Wang, Yingya Zhang, and Zhidong Deng. Dreamtalk: When expressive talking head generation meets diffusion probabilistic models. *arXiv*, 2023. 1

[11] Saif M. Mohammad. Word affect intensities. In *ACL*, 2018. 2, 3

[12] Saif M Mohammad. Practical and ethical considerations in the effective use of emotion and sentiment lexicons. *arXiv*, 2020. 2, 3

[13] Xuan-Bach Nguyen, Hoang-Thien Nguyen, Thanh-Huy Nguyen, Nhu-Tai Do, and Quang Vinh Dinh. Emotic masked autoencoder on dual-views with attention fusion for facial expression recognition. In *CVPR*, 2024. 1

[14] James W Pennebaker. *Linguistic inquiry and word count: LIWC 2001*. Lawrence Erlbaum Associates, 2001. 2, 3

[15] Yujun Shen, Jinjin Gu, Xiaoou Tang, and Bolei Zhou. Interpreting the latent space of gans for semantic face editing. In *CVPR*, 2020. 1

[16] Siyang Song, Micol Spitale, Cheng Luo, German Barquero, Cristina Palmero, Sergio Escalera, Michel Valstar, Tobias Baur, Fabien Ringeval, Elisabeth Andre, et al. React2023: The first multiple appropriate facial reaction generation challenge. In *ACM MM*, 2023. 4

[17] Linrui Tian, Qi Wang, Bang Zhang, and Liefeng Bo. Emo: Emote portrait alive generating expressive portrait videos with audio2video diffusion model under weak conditions. In *ECCV*, 2024. 1

[18] Antoine Toisoul, Jean Kossaifi, Adrian Bulat, Georgios Tzimiropoulos, and Maja Pantic. Estimation of continuous valence and arousal levels from faces in naturalistic conditions. *Nat. Mach. Intell.*, 2021. 1

[19] Lizhen Wang, Zhiyuan Chen, Tao Yu, Chenguang Ma, Liang Li, and Yebin Liu. Faceverse: a fine-grained and detail-controllable 3d face morphable model from a hybrid dataset. In *CVPR*, 2022. 1

[20] Amy Beth Warriner, Victor Kuperman, and Marc Brysbaert. Norms of valence, arousal, and dominance for 13,915 english lemmas. *Behavior research methods*, 2013. 2, 3

[21] Erroll Wood, Tadas Baltrušaitis, Charlie Hewitt, Sebastian Dziadzio, Thomas J Cashman, and Jamie Shotton. Fake it till you make it: face analysis in the wild using synthetic data alone. In *ECCV*, 2021. 1

[22] Baosheng Yu and Dacheng Tao. Heatmap regression via randomized rounding. *IEEE Trans. Pattern Anal. Mach. Intell.*, 2021. 1

[23] Dan Zeng, Zhiyuan Lin, Xiao Yan, Yuting Liu, Fei Wang, and Bo Tang. Face2exp: Combating data biases for facial expression recognition. In *CVPR*, 2022. 1

[24] Feifei Zhang, Tianzhu Zhang, Qirong Mao, Lingyu Duan, and Changsheng Xu. Facial expression recognition in the wild: A cycle-consistent adversarial attention transfer approach. In *ACM MM*, 2018. 1

[25] Wenxuan Zhang, Xiaodong Cun, Xuan Wang, Yong Zhang, Xi Shen, Yu Guo, Ying Shan, and Fei Wang. Sadtalker: Learning realistic 3d motion coefficients for stylized audio-driven single image talking face animation. In *CVPR*, 2023. 1

[26] Zhicheng Zhang, Lijuan Wang, and Jufeng Yang. Weakly supervised video emotion detection and prediction via cross-modal temporal erasing network. In *CVPR*, 2023. 1

[27] Hang Zhou, Yu Liu, Ziwei Liu, Ping Luo, and Xiaogang Wang. Talking face generation by adversarially disentangled audio-visual representation. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pages 9299–9306, 2019. 1

[28] Hang Zhou, Yasheng Sun, Wayne Wu, Chen Change Loy, Xiaogang Wang, and Ziwei Liu. Pose-controllable talking face generation by implicitly modularized audio-visual representation. In *CVPR*, 2021. 1

[29] Hengde Zhu, Xiangyu Kong, Weicheng Xie, Xin Huang, Linlin Shen, Lu Liu, Hatice Gunes, and Siyang Song. Perfrdiff: Personalised weight editing for multiple appropriate facial reaction generation. In *ACM MM*, 2024. 1, 4