How Do You Perceive My Face? Recognizing Facial Expressions in Multi-Modal Context by Modeling Mental Representations

Florian Blume, Runfeng Qu, Pia Bideau, Martin Maier, Rasha Abdel Rahman, Olaf Hellwich

How Do You Perceive My Face? Recognizing Facial Expressions in Multi-Modal Context by Modeling Mental Representations

Florian Blume$^{\star 1,4[0000-0002-7557-1508]}$✉, Runfeng Qu$^{\star 1,4[0009-0008-7885-8812]}$, Pia Bideau$^{2[0000-0001-8145-1732]}$, Martin Maier$^{3,4[0000-0003-4564-9834]}$, Rasha Abdel Rahman$^{3,4[0000-0002-8438-1570]}$, and Olaf Hellwich$^{1,4[0000-0002-2871-9266]}$

[1] Technische Universität Berlin
[2] Univ. Grenoble Alpes, Inria, CNRS, Grenoble INP, LJK
[3] Humboldt-Universität zu Berlin
[4] Science of Intelligence, Research Cluster of Excellence, Berlin
✉ `florian.blume@tu-berlin.de, runfeng.qu@tu-berlin.de`

Abstract. Facial expression perception in humans inherently relies on prior knowledge and contextual cues, contributing to efficient and flexible processing. For instance, multi-modal emotional context (such as voice color, affective text, body pose, etc.) can prompt people to perceive emotional expressions in objectively neutral faces. Drawing inspiration from this, we introduce a novel approach for facial expression classification that goes beyond simple classification tasks. Our model accurately classifies a perceived face and synthesizes the corresponding mental representation perceived by a human when observing a face in context. With this, our model offers visual insights into its internal decision-making process. We achieve this by learning two independent representations of content and context using a VAE-GAN architecture. Subsequently, we propose a novel attention mechanism for context-dependent feature adaptation. The adapted representation is used for classification and to generate a context-augmented expression. We evaluate synthesized expressions in a human study, showing that our model effectively produces approximations of human mental representations. We achieve State-of-the-Art classification accuracies of 81.01% on the RAVDESS dataset and 79.34% on the MEAD dataset. We make our code publicly available[5].

1 Introduction

Integrating multi-modal contextual information is crucial for generating adaptive behavior and enabling an agent to respond appropriately to its environment. More specifically, contextual information encompasses the multi-modal information that enhances the agent's perception and thus is a prerequisite for adaptive behavior. Latest work in cognitive psychology has shown that the human brain leverages contextual cues and prior knowledge to *dynamically* adjust

[⋆] equal contribution

[5] https://github.com/tub-cv-group/recognizing-by-modeling

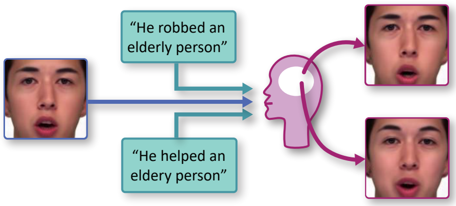


Fig. 1: Visualization of the influence of context on **human perception** of facial expressions. The mental representation shifts congruently with context.

future predictions about incoming sensory input [6, 34]. Concurrent sound, textual cues, or prior knowledge offer additional information that shapes social perception [26, 45]. For example, in the interpretation of facial expressions, the individual's voice plays a significant role in understanding their overall expression. As illustrated in Fig. 1, the same neutral face is perceived as displaying a more positive or negative expression when presented in the context of the respective prior beliefs about the person. In this work, we refer to the perceived facial expression as the *synthesized mental representation*.

Previous works in computer vision typically either only perform facial expression recognition (FER) [4, 11, 17, 20, 27, 28] or generate expressions [3, 9, 19, 47, 51]. Only few approaches exist that perform both tasks jointly [36, 42, 48, 51]. Addressing recognition and generation jointly, however, poses an essential element in modelling human social interaction and creating effective communication between humans and artificial agents, enabling agents to mimic the expressions of conversational partners in a context-sensitive manner and in direct alignment with the recognized expression.

We propose a novel mechanism of fusing expressions and multi-modal context information encoded in the latent space of a variational-autoencoder (VAE) [19]. In particular, using an attention mechanism, our model dynamically adapts previously learned representations of facial expressions using context. Operating on lower-dimensional representations of facial expressions enables us to simultaneously produce well-aligned classifications and an approximation of the perceived expression. We verify the validity of these approximations in a rating study with 160 human observers and show SOTA classification accuracy on the RAVDESS and MEAD datasets. Our model performs a task that is similar to the studies in [2, 30, 41], where human participants were presented with individual photographs paired with affective-semantic contexts. Our contributions are threefold: **(1)** We present a model that for the first time simultaneously classifies expressions and produces approximations of their mental representations, which are inherently well aligned with the predicted class. **(2)** We evaluate these approximations in a human study. They capture the fine-grained effects of emotional context on human perception. **(3)** Our model is explainable due to its ability to visualize the adapted features by generating a context-augmented expression.

2 Related Works

We discuss three types of works: (1) context-sensitive FER-only, (2) synthesizing expressions, and (3) performing both tasks jointly.

Multi-modal context-sensitive facial expression recognition. Multi-modal context-sensitivity in FER ranges from incorporating visual surroundings as additional information cues [17, 18, 20, 28], to drawing on audio [5, 7, 12, 24, 29, 52, 53], text [50, 53], body pose [28] or combining multiple context sources [4, 11, 21]. Transformers have successfully been incorporated into FER in multiple approaches [4, 5, 21, 24]. Contrastive learning schemes have shown to extract general features that ensure good classification performance on unseen data [4, 11, 50]. Franceschini et al. [11] outperform state-of-the-art (SOTA) methods on RAVDESS using an unsupervised contrastive learning scheme on four modalities. Luna-Jiménez et al. [24] use the transformer architecture and action units to predict expressions on RAVDESS. In contrast to our work, generating context-augmented versions of the input face is not straightforward in these approaches.

Generating facial expressions in context. Generative adversarial networks (GANs) have been used for generating realistic looking face images [3, 19, 47]. Larsen et al. [19] combined them with a VAE to allow smooth transitions between representations. Fang et al. [10] employ a GAN to generate a talking face from audio. Peng et al. [32] generate a 3D talking head based on the audio of the RAVDESS dataset, which is capable of producing facial expressions. Using artificial characters in a rating study results in a different perceptual experience for humans [8], which makes them inapplicable to our goal. Xu et al. [46] train their network in a CLIP-like [33] fashion to generate sequences of talking faces. Stypułkowski et al. [40] employ latent diffusion [35] for the same task. None of these works target joint generation of mental representations and classification.

Joint Facial Expression Generation and Recognition. Few works exist that perform the task of simultaneously generating facial expressions while also predicting expression classes. Sun et al. [42] train two GANs cooperatively to recognize facial expressions under large view angle changes. Yan et al. [48] employ a GAN [14] to overcome the lack of labeled training data in FER by jointly training it together with an expression recognition network. Context sensitivity is not part of their work. Zhang et al. [51] also draw on a GAN-based architecture and argue that by generating expressions, they help solve the issue of appearance variance in FER and lack of training data. Their network processes only input images and does not take additional modalities into account. None of the disucssed works model mental representation through synthesized expressions.

3 Method

In this Section, we describe our multi-modal approach that adapts an expression image using affective audio. Our proposed adaption module allows us to *classify* a facial expression in light of context and at the same time *synthesize* a novel facial expression as it would have been perceived by a human. We employ a

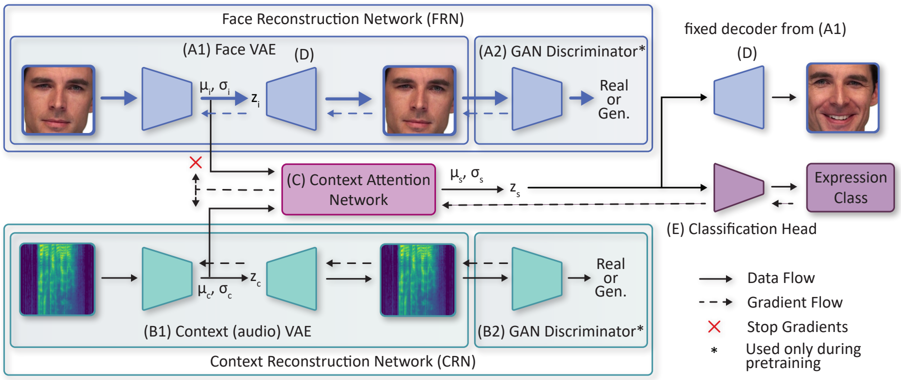


Fig. 2: Overview of our full network architecture. The face and context reconstruction networks FRN and CRN are a VAE-GAN combination. The mean and variance of a facial input image and audio context (Mel spectrogram) are adapted by the CAN, which shifts the face representation using the context representation. The classification head (E) classifies the shifted features and we use the fixed decoder of (A1) to visualize the expression.

two-stream encoder-decoder backbone, based on a VAE-GAN combination [19], and an attention mechanism to combine their latent spaces in a context-sensitive way. Fig. 2 depicts an overview of this design: The representations learned by the face reconstruction net (FRN) and the context reconstruction net (CRN) are adapted in the context attention net (CAN), by shifting the facial features using the context features. The shifted representation is visualized using the fixed decoder (D) of the FRN and classified with the classification head (E). Our model operates and is trained on individual frames together with audio context. To evaluate our model on videos, we perform majority voting over classes of the frames.

3.1 Face and Context Reconstruction Network

The face and context reconstruction networks (called FRN and CRN respectively) both consist of a VAE and GAN discriminator. We follow [19] and add a GAN discriminator for training to increase image quality.

VAE Module. Similarly to [15], we add skip connections to the en- and decoder of the VAEs to speed up the training process and allow processing of larger image resolutions. Let Enc_I be the expression encoder of (A1) and $\boldsymbol{x}_I \in \mathbb{R}^{m \times n \times 3}$ an input expression image:

$$(\boldsymbol{\mu}_I, \boldsymbol{\sigma}_I) = Enc_I(\boldsymbol{x}_I) \tag{1}$$

where $\boldsymbol{\mu}_I \in \mathbb{R}^d$ and $\boldsymbol{\sigma}_I \in \mathbb{R}^d$ denote the mean and variance of a Gaussian distribution, respectively. Mean $\boldsymbol{\mu}_C \in \mathbb{R}^d$ and $\boldsymbol{\sigma}_C \in \mathbb{R}^d$ variance of the context

Mel spectrogram $\boldsymbol{x}_C \in \mathbb{R}^{u \times v}$ are computed using the encoder of (B1) and the following formulas are applied analogously.

The **prior loss** term keeps the latent distribution close to a Gaussian:

$$\mathcal{L}_{prior} = D_{KL}(q(\boldsymbol{z}_I|\boldsymbol{x}_I)||p(\boldsymbol{z}_I)) \tag{2}$$

D_{KL} is the Kullback–Leibler divergence (KL-Divergence), $q(\boldsymbol{z}_I|\boldsymbol{x}_I)$ is the posterior of the latent vector $\boldsymbol{z}_I \in \mathbb{R}^d$ under input $\boldsymbol{x}_I$ and $p(\boldsymbol{z}_I)$ is the Gaussian prior over the latent vector.

The **reconstruction loss** term penalizes the feature map of the discriminator (A2 and B2 in Fig. 2) at a certain level, as proposed in [19], using MSE:

$$\mathcal{L}_{reconst} = MSE(Dis_I^l(Dec_I(Enc_I(\boldsymbol{x}_I))), Dis_I^l(\boldsymbol{x}_I)) \tag{3}$$

where $\boldsymbol{x}_I$ is the input image and Dis_I^l the l-th feature map of the discriminator.

GAN Module. The GAN discriminator's task is to distinguish between input images $\boldsymbol{x}_I$ from the dataset and reconstructions $\hat{\boldsymbol{x}}_I$. In addition, it is tasked to identify reconstructions from random noise $\boldsymbol{z}_p \sim \mathcal{N}(0,1)$ to enforce generation capabilities in the VAE. The overall loss is the following:

$$\begin{aligned}\mathcal{L}_{GAN} = &\log(Dis_I(\boldsymbol{x}_I)) + \log(1 - Dis_I(Dec_I(Enc_I(\boldsymbol{x}_I)))) \\ &+ \log(1 - Dis_I(Dec_I(\boldsymbol{z}_p)))\end{aligned} \tag{4}$$

Joint Training. In the pretraining phase, FRN and CRN are trained unsupervised for input reconstruction and are fixed in subsequent training. We follow the training algorithm of [19] and compute the joint update as follows:

$$\boldsymbol{\theta}_{Enc} \xleftarrow{+} -\nabla_{\boldsymbol{\theta}_{Enc}} (\beta\mathcal{L}_{prior} + \mathcal{L}_{reconst}) \tag{5}$$

$$\boldsymbol{\theta}_{Dec_I} \xleftarrow{+} -\nabla_{\boldsymbol{\theta}_{Dec_I}} (\mathcal{L}_{reconst} - \mathcal{L}_{GAN}) \tag{6}$$

$$\boldsymbol{\theta}_{Dis_I} \xleftarrow{+} -\nabla_{\boldsymbol{\theta}_{Dis_I}} \mathcal{L}_{GAN} \tag{7}$$

where $\beta \in \mathbb{R}$ is a hyperparameter to weigh the prior loss.

3.2 Context-Attention Network (CAN) and Classification Head

Our proposed attention mechanism shifts the facial expression distribution of the FRN based on the context distribution of the CRN. Fig. 3 illustrates this fusion technique.

Context-Attention Net. The CAN computes attention maps based on mean and variance of the distributions of the context and the facial expression input. We use these maps to compute offsets $\boldsymbol{o}_\mu \in \mathbb{R}^d$ and $\boldsymbol{o}_\sigma \in \mathbb{R}^d$ to shift the face mean and variance context-dependently. We compute the following parameters for the attention mechanism:

$$\boldsymbol{q}_\mu = \boldsymbol{W}_q^\mu \boldsymbol{\mu}_C \tag{8}$$

$$\boldsymbol{k}_\mu = \boldsymbol{W}_k^\mu \boldsymbol{\mu}_I \tag{9}$$

$$\boldsymbol{v}_\mu = \boldsymbol{W}_v^\mu \boldsymbol{\mu}_I \tag{10}$$

$\boldsymbol{W}_q^\mu, \boldsymbol{W}_k^\mu, \boldsymbol{W}_v^\mu \in \mathbb{R}^{d \times d}$ are the trainable parameters of linear layers (bias omitted for simplicity). We compute the attention map $\boldsymbol{A} \in \mathbb{R}^{d \times d}$ as

$$\boldsymbol{A} = \mathrm{softmax}(\boldsymbol{q}_\mu \boldsymbol{k}_\mu^T) \tag{11}$$

where the softmax function is applied row-wise. Note that we reverse the attention mechanism from [43] - we do not compute the dot products of the query with all keys, instead, we compute the dot product of a key to all queries. We do this to get attention on the facial mean based on the context mean. The offset $\boldsymbol{o}_\mu$ and the resulting shifted mean $\boldsymbol{\mu}_S$ are then computed as

$$\boldsymbol{o}_\mu = \boldsymbol{A}\boldsymbol{v}_\mu \tag{12}$$

$$\boldsymbol{\mu}_S = \boldsymbol{o}_\mu + \boldsymbol{\mu}_I \tag{13}$$

Fig. 3 provides a visualization of these relationships. During inference we can vary the strength of the context influence by multiplying the offset with a weight, to allow smooth modulation of the offset:

$$\boldsymbol{\mu}_S = m \cdot \boldsymbol{o}_\mu + \boldsymbol{\mu}_I \tag{14}$$

We compute the new (shifted) variance $\boldsymbol{\sigma}_S$ analogously, the only difference being that we operate in log scale.

Joint Context Attention Network and Classification Head Training. To train the CAN and classification head jointly, we first initialize the latter by training it directly on the facial features using the expression classes and cross-entropy loss. Next, we train the network together using the following loss:

$$\mathcal{L} = \mathrm{CE}(\mathrm{E}(\boldsymbol{\mu}_S), y) + \alpha D_{KL}(p(\boldsymbol{z}_I|\boldsymbol{x}_I)||q(\boldsymbol{z}_s|\boldsymbol{x}_I, \boldsymbol{x}_C)) \tag{15}$$

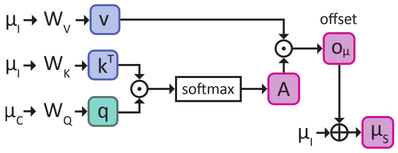


Fig. 3: Detailed view of the CAN from Fig. 2 for adapting the means. $\odot$ is element-wise multiplication, $\oplus$ addition. We left out index μ on the weights for simplicity. The shifted variance $\boldsymbol{\sigma}_S$ is computed analogously.

where CE is the cross-entropy, E is the classification head and $\alpha \in \mathbb{R}$ is a hyperparameter to regularize the shift. Furthermore, during training, we propose a novel data augmentation approach for multi-modal settings, where we swap contexts for a specific actor within its expression class.[6]

4 Experiments

We evaluate our approach on publicly available datasets [23, 44]. We discuss the results in Section 4.3 and 4.4. An ablation study is shown to support the understanding of our proposed approach for FER in multi-modal context.

4.1 Datasets

CelebA (Pretraining). We pretrain the FRN unsupervisedly for face reconstruction on the large-scale dataset CelebA [22]. CelebA is a prevalent dataset for face attribute recognition and consists of roughly 200k images showing 10k different identities. The CRN is trained using the respective downstream datasets.
RAVDESS (Downstream). RAVDESS [23] consists of videos of 24 identities. Each video is labeled with one of the seven expression classes calm, happy, sad, angry, fearful, surprise, and disgust, with an additional binary label for the intensity. We extract 16 frames from each video at regular intervals and use the video's label for each. Following [5] the neutral class is omitted to reduce noise.
MEAD (Downstream). MEAD [44] is a large-scale dataset targeting talking-face generation, which also features FER labels. Similar to [23], MEAD contains videos of 60 actors speaking with different emotions at different intensity levels. We use the frontal view recordings as proposed by [36] and apply the same frame extraction approach as for RAVDESS.

4.2 Implementation Details

We set the batch size to 256, β (Eq. (5)) and α (Eq. (15)) to 0.00001 and $m = 1.0$ (Eq. (14)). The dimension of the latent space of the FRN and CRN is $d = 512$. We use MTCNN [49] to detect faces. We then resize them to 128×128 pixels, which is also the size of the reconstructed and generated images. We apply random horizontal flipping as data augmentation. For generating the Mel spectrograms, we chose 128 Mel bins, a sample rate of 22050, a window and FFT length of 1310, and a hop length of 755. We use the Adam optimizer [16] with a learning rate of 0.00003 and a weight decay of 0.01.
Reconstruction Pretraining. We pretrain the FRN unsupervised for facial image reconstruction on CelebA and the CRN for context reconstruction on the Mel spectrograms of the downstream datasets. Pretraining is run for 400 epochs. The learning rate is decreased by factor 10 after 150 and 300 epochs.

[6] Note, this data augmentation is only possible for multi-modal datasets, where content comes with different context variations within their respective expression class.

Table 1: Classification accuracy of SOTA methods on **RAVDESS**. V = videos, A = audio, K = facial keypoints. Gen and Class are generative and classification approaches, respectively. Bold are best results in the respective group.

Model	Modalities	Year	Gen	Class	Acc
Classification-Only Approaches					
Franceschini et al. [11]	A + V + K	2022		✓	78.54
Fu et al. [12]	A + V	2021		✓	75.76
Ghaleb et al. [13]	A + V	2020		✓	76.30
Chumachenko et al. [5]	A + V	2022		✓	81.58
Dahmouni et al. [7]	A + V	2023		✓	**85.76**
Joint Classification-Generation Approaches					
Sadok et al. [36]	A + V	2024	✓	✓	68.8
Ours	A + V	2024	✓	✓	**81.01**

Table 2: Classification accuracy of SOTA methods on **MEAD**.

Model	Modalities	Year	Gen	Class	Acc
wav2vec [37]	A	2019		✓	68.4
Sadok et al. [36]	A + V	2024	✓	✓	73.2
Ours	A + V	2024	✓	✓	**79.0**

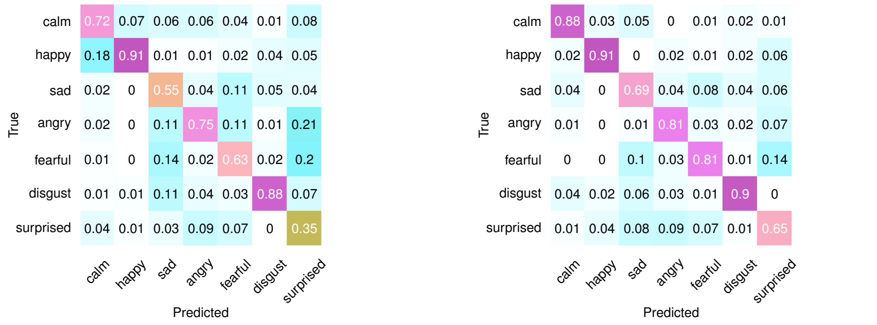
Downstream Classification Training. During downstream training, the FRN is fixed and the last two layers of the CRN are fine-tuned. We first initialize the one-layer classification head by training it directly on the facial features of the FRN to obtain a suitable initialization for its weights. Next, we train the CAN and the single-layer classification head jointly together using the loss from Eq. (15). Note that the decoder from (A1), which we use to visualize the shifted expression, is fixed and not trained in this step.

For both RAVDESS and MEAD, we performed k-fold cross validation with $k = 10$, similarly to other works [11, 29], splitting the folds along the identities (*i.e.* one identity can only occur in test, validation or train set).

4.3 Facial Expression Recognition Performance

We provide classification results for RAVDESS in Tab. 1, and for MEAD in Tab. 2. All compared methods use a dataset split by identities, ensuring no identity seen during training appears during testing. For our method, the final prediction for a test video is obtained by majority voting across frames. Our model achieves an accuracy of 81.01% on RAVDESS and 79.34% on MEAD, matching SOTA performance on classical FER. We largely outperform methods - *i.e.* by 17.75% on RAVDESS and 8.47% on MEAD, that tackle the dual problem of classification and the synthesize of the corresponding percept.

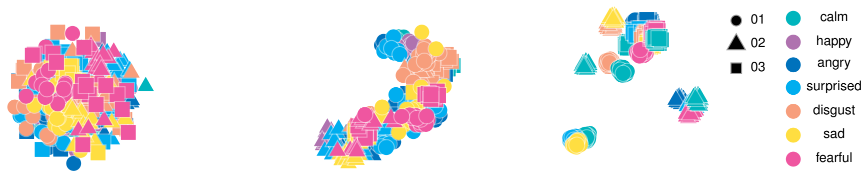
The influence of context on the final FER accuracy is highlighted by visualizing the per-class accuracy in confusion matrices for two conditions: in Fig. 4a, the



(a) Confusion matrix of CAN, face features shifted with face features.

(b) Confusion matrix of CAN, face features shifted with context features.

Fig. 4: Confusion matrices for our model. (a) Uni-modal setting with simulated missing context. (b) Multi-modal setting showing the clear diagonal.



(a) Face-only latent features obtained from the FRN without taking context into account.

(b) Features obtained from the CAN by providing the face features again as context.

(c) **Proposed approach**: Features obtained from the CAN by shifting using the context.

Fig. 5: Comparison of t-SNE visualizations of all samples (*i.e.* frames plus audio) from the RAVDESS identities 01, 02 and 03.

CAN computed the offset for the facial features based on the facial (instead of the context) features to simulate missing context. In Fig. 4b, it received the face together with the context features, as intended by our approach. The latter exhibits higher (or equal) probabilities on the diagonal for every class. This proves that the reported accuracy cannot be attributed solely to the computational capabilities of the CAN but is a result of the meaningful adaption procedure.

The t-SNE plots in Fig. 5 visualize the structure of the latent space in different conditions: (1) The face-only features as we receive them from the FRN do not cluster in any particular way (Fig. 5a). (2) Employing our CAN but providing the face features twice instead of combining with the context leads to a more structured latent space (Fig. 5b) and (3) taking the audio context into account leads to a clustered latent space that makes classification easier (Fig. 5c).

Ablation Study. Tab. 3 lists classification performance of our CAN and its components. We assess the quality of face and context features independently by classifying learned features directly using a single-layer classifier. Experi-

Table 3: Performance ablation study on RAVDESS of a VGG16 network, a single linear layer on the features of the FRN and CRN, and the CAN.

Model	Modality	Acc. % ↓
Baseline VGG16 [38]	Video	64.40
Single-Layer	Video	68.06
CAN	Video	68.99
Single-Layer	Audio	48.66
CAN	Audio	50.05
CAN (strict audio)	Video + Audio	79.64
CAN (Ours)	Video + Audio	**81.01**

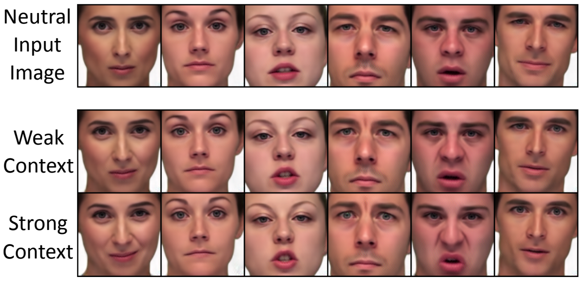


Fig. 6: **Neutral** faces generated with intensity variations as provided by RAVDESS. Left to right: happy, sad, angry, fearful, disgusted, surprised.

ments for the CAN in a simulated uni-modal mode are run by providing face or context twice. If no context is provided, the CAN performs similarly to the single layer classifier using learned representations. However, when learned features are adapted to context, the final classification performance improves by 15.44% (strict audio). Additionally, our new data augmentation technique for multi-modal settings further enhances performance by 1.72%.

4.4 Mental Representations: Context-Augmented Expressions

We show qualitative results of our approach on RAVDESS in different conditions (additional generations for RAVDESS and MEAD in the suppl. material). To visualize the adapted features obtained from the CAN, we use the decoder (D) from Fig. 2. Fig. 6 depicts the effect of the two kinds of intensities provided in RAVDESS on neural input faces. The stronger the intensity, the stronger the facial expression in the generated image. The strength of the effect is subtle but resembles human perception [26], which is our goal.

To our knowledge, the only publications showing generations on RAVDESS are Sinha et al. [39], Ma et al. [25] and Fang et al. [10]. They aim at rendering an input face that strongly resembles the emotion in the accompanying audio

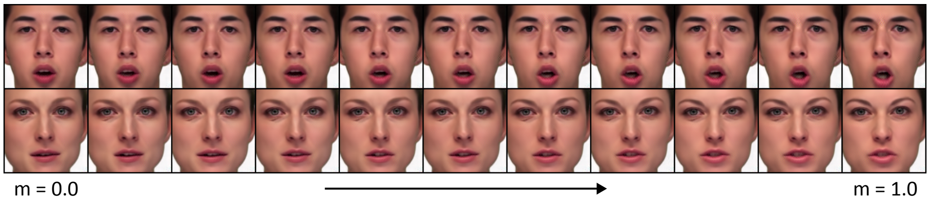


Fig. 7: Exploration of modulating the strength of the context offset by increasing m (see Eq. (14)) from 0 to 1 from left to right in steps of 0.1.

file. This defies our goal of capturing the subtle changes in appearance a human observer would have, nevertheless we provide a comparison in the suppl. material.

How influential is the context in which humans perceive facial expressions? To empirically evaluate the strength of contextual influence that best captures context effects in humans, we aimed to manually control the computed offset using m of Eq. (14). As shown in Fig. 7, this allowed us to generate facial expressions with varying degrees of contextual influence, which makes them useful as experimental stimuli. The context weight m is increased from 0 to 1 in 0.1 steps from left to right. Note that this weight is independent of the two intensities provided by RAVDESS (Fig. 6). The smooth transition demonstrates that our model creates continuous representations of the input data.

4.5 Human Study: Verifying Synthesised Mental Representations

To assess our model's ability to replicate the contextual impact of emotional speech on facial expression perception in human observers, we conducted two experiments with a total of 160 participants. In **the first experiment**, 80 participants evaluated neutral facial expressions from the RAVDESS dataset while listening to the depicted actor's speech with either happy or angry prosody. We restricted the study to these two classes because they are of opposite valence. This allowed us to obtain more fine-grained ratings on a continuous Likert scale. We aimed to measure the impact of the audio's emotion on perceived facial expressions. In **the second experiment**, a different group of 80 participants rated facial expressions synthesized by our model under happy or angry contextual influences. Each face was presented with five different context weights m. To test whether the model's generations can approximate human-like perception, we compared ratings with those obtained in Experiment 1. This comparison was done for each context weight, determining the parameter that best approximates human responses to contextual influences.

Results. Mean facial expression ratings for different emotional context conditions across both experiments are depicted in Fig. 8. In **experiment 1**, a linear mixed effects model was employed, with the factor emotional context (happy vs. angry audio). A significant effect of emotional context on facial expression ratings was observed ($b = 1.00$, $p < .001$), indicating that identical neutral faces were perceived as more negative when accompanied by angry speech compared

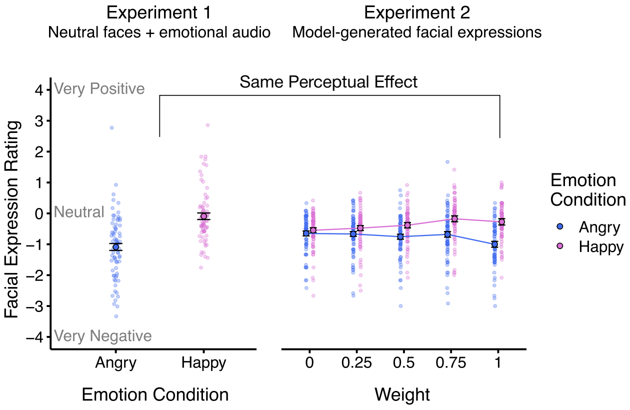


Fig. 8: Rating study results depicting mean facial expression ratings in Experiment 1 (left), in which participants rated neutral faces presented in the context of emotional audio, and Experiment 2 (right), in which participants rated faces generated by our model with happy vs. angry audio context (with 5 different values for the context weight m of Eq. (14)). At $m = 1$, ratings are equal across experiments, our model captures the effect of emotional context in human observers. Figure with more statistical details in supplementary material.

to happy speech. For **experiment 2**, a linear mixed effects model was run with the factors emotional context (happy vs. angry generated expression) and weight ($m = 0$, 0.25, 0.5, 0.75, and 1). To compare rating differences within each weight, the emotional context factor was nested within the weight factor. As shown in Fig. 8 (Experiment 2), expression ratings for faces generated in the context of angry vs. happy prosody did not differ at weights 0 and 0.25 ($bs \leq 0.19$, $ps > .269$), but showed a significant and increasing impact of emotional context at weights 0.5 ($b = 0.37$, $p = .033$), 0.75 ($b = 0.51$, $p = .004$), and 1 ($b = 0.73$, $p < .001$). **Discussion.** The results of our rating study reveal two key findings: 1) Our model effectively captures the impact of context, exemplified by emotional prosody, on human facial expression perception. Our generations reflect how a neutral face would subjectively appear to a human observer when associated with a positive or negative context. 2) The model's efficacy in shifting facial appearance towards contextual emotions does not require further adjustment post-training, as evidenced by the optimal context weight being $m = 1$.

5 Conclusion

In this work, we introduced a novel approach that simultaneously enhances expression class predictions by taking affective context into account, and provides the means to generate an approximation of the expression a human would perceive. Our model achieves SOTA accuracy on RAVDESS and MEAD, and out-performs joint competitor methods. The implications of our rating study showcase that our model not only accurately quantifies human context-sensitive perception but also successfully mirrors the altered subjective experience back to human observers. This has significant potential, particularly for social artificial agents, that could leverage contextual information to adapt to human mental and emotional states, ensuring successful communication. Our model also addresses the dual nature of context-sensitivity of human perception: on the one hand, leveraging context enhances perceptual efficiency and flexibility [26, 31], while on the other hand, it bears the potential for adversely biased perception, e.g., when contextual information originates from untrustworthy sources [2].

Acknowledgements

Funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Germany's Excellence Strategy – EXC 2002/1 "Science of Intelligence" – project number 390523135. In addition, this work has been partially supported by MIAI@Grenoble Alpes, (ANR-19-P3IA-0003)

Disclosure of Interests

The authors have no competing interests to declare that are relevant to the content of this article.

Bibliography

1. Bates, D., Mächler, M., Bolker, B., Walker, S.: Fitting linear mixed-effects models using lme4. Journal of Statistical Software **67**(1), 1–48, 2015, https://doi.org/10.18637/jss.v067.i01
2. Baum, J., Rabovsky, M., Rose, S.B., Abdel Rahman, R.: Clear judgments based on unclear evidence: Person evaluation is strongly influenced by untrustworthy gossip. Emotion **20**(2), 248–260, 2020, https://doi.org/10.1037/emo0000545
3. Bouzid, H., Ballihi, L.: Facial expression video generation based-on spatio-temporal convolutional GAN: FEV-GAN. Intelligent Systems with Applications **16**, 200139, 2022, https://doi.org/10.1016/j.iswa.2022.200139
4. Chudasama, V., Kar, P., Gudmalwar, A., Shah, N., Wasnik, P., Onoe, N.: M2FNet: Multi-modal Fusion Network for Emotion Recognition in Conversation. In: 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), 2022, https://doi.org/10.1109/CVPRW56347.2022.00511
5. Chumachenko, K., Iosifidis, A., Gabbouj, M.: Self-attention fusion for audiovisual emotion recognition with incomplete data. In: 2022 26th International Conference on Pattern Recognition, ICPR 2022, 2022, https://doi.org/10.1109/ICPR56361.2022.9956592
6. Clark, A.: Whatever next? Predictive brains, situated agents, and the future of cognitive science. Behavioral and Brain Sciences **36**(3), 181–204, 2013, https://doi.org/10.1017/S0140525X12000477
7. Dahmouni, A., Rossamy, R., Hamdani, M., Guelzim, I., Ait Abdelouahad, A.: Bimodal Emotional Recognition based on Long Term Recurrent Convolutional Network. In: Proceedings of the 6th International Conference on Networking, Intelligent Systems & Security, NISS '23, 2023, https://doi.org/10.1145/3607720.3607740
8. Eiserbeck, A., Maier, M., Baum, J., Abdel Rahman, R.: Deepfake smiles matter less—the psychological and neural impact of presumed AI-generated faces. Scientific Reports **13**(1), 16111, 2023, https://doi.org/10.1038/s41598-023-42802-x
9. Eskimez, S.E., Zhang, Y., Duan, Z.: Speech Driven Talking Face Generation From a Single Image and an Emotion Condition. IEEE Transactions on Multimedia **24**, 3480–3490, 2022, https://doi.org/10.1109/TMM.2021.3099900
10. Fang, Z., Liu, Z., Liu, T., Hung, C.C., Xiao, J., Feng, G.: Facial expression GAN for voice-driven face generation. The Visual Computer **38**(3), 1151–1164, 2022, https://doi.org/10.1007/s00371-021-02074-w
11. Franceschini, R., Fini, E., Beyan, C., Conti, A., Arrigoni, F., Ricci, E.: Multimodal Emotion Recognition with Modality-Pairwise Unsupervised Contrastive Loss. 2022 26th International Conference on Pattern Recognition (ICPR) pp. 2589–2596, 2022, https://doi.org/10.1109/ICPR56361.2022.9956589

12. Fu, Z., Liu, F., Wang, H., Qi, J., Fu, X., Zhou, A., Li, Z.: A cross-modal fusion network based on self-attention and residual structure for multimodal emotion recognition. ArXiv 2021
13. Ghaleb, E., Niehues, J., Asteriadis, S.: Multimodal Attention-Mechanism For Temporal Emotion Recognition. In: 2020 IEEE International Conference on Image Processing (ICIP), 2020, https://doi.org/10.1109/ICIP40778.2020.9191019
14. Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y.: Generative adversarial nets. In: Advances in Neural Information Processing Systems, 2014
15. Huang, H., Li, Z., He, R., Sun, Z., Tan, T.: IntroVAE: Introspective Variational Autoencoders for Photographic Image Synthesis. In: Neural Information Processing Systems, 2018
16. Kingma, D.P., Ba, J.: Adam: A Method for Stochastic Optimization. https://doi.org/10.48550/arXiv.1412.6980
17. Kosti, R., Alvarez, J., Recasens, A., Lapedriza, A.: Context Based Emotion Recognition using EMOTIC Dataset. IEEE Transactions on Pattern Analysis and Machine Intelligence pp. 1–1, 2019, https://doi.org/10.1109/TPAMI.2019.2916866
18. Kosti, R., Alvarez, J.M., Recasens, A., Lapedriza, A.: Emotion Recognition in Context. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, https://doi.org/10.1109/CVPR.2017.212
19. Larsen, A.B.L., Sønderby, S.K., Larochelle, H., Winther, O.: Autoencoding beyond pixels using a learned similarity metric. In: Proceedings of The 33rd International Conference on Machine Learning, 2016
20. Lee, J., Kim, S., Kim, S., Park, J., Sohn, K.: Context-Aware Emotion Recognition Networks. In: 2019 IEEE/CVF International Conference on Computer Vision (ICCV), 2019, https://doi.org/10.1109/ICCV.2019.01024
21. Li, Y., Wang, Y., Cui, Z.: Decoupled Multimodal Distilling for Emotion Recognition. In: 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, https://doi.org/10.1109/CVPR52729.2023.00641
22. Liu, Z., Luo, P., Wang, X., Tang, X.: Deep Learning Face Attributes in the Wild. In: 2015 IEEE International Conference on Computer Vision (ICCV), 2015, https://doi.org/10.1109/ICCV.2015.425
23. Livingstone, S.R., Russo, F.A.: The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS): A dynamic, multimodal set of facial and vocal expressions in North American English. PLoS ONE **13**(5), e0196391, 2018, https://doi.org/10.1371/journal.pone.0196391
24. Luna-Jiménez, C., Kleinlein, R., Griol, D., Callejas, Z., Montero, J.M., Fernández-Martínez, F.: A Proposal for Multimodal Emotion Recognition Using Aural Transformers and Action Units on RAVDESS Dataset. Applied Sciences **12**(1), 327, 2021, https://doi.org/10.3390/app12010327
25. Ma, Y., Zhang, S., Wang, J., Wang, X., Zhang, Y., Deng, Z.: DreamTalk: When Expressive Talking Head Generation Meets Diffusion Probabilistic Models. https://doi.org/10.48550/arXiv.2312.09767

26. Maier, M., Blume, F., Bideau, P., Hellwich, O., Abdel Rahman, R.: Knowledge-augmented face perception: Prospects for the Bayesian brain-framework to align AI and human vision. Consciousness and Cognition **101**, 103301, 2022, https://doi.org/10.1016/j.concog.2022.103301
27. Mittal, T., Bhattacharya, U., Chandra, R., Bera, A., Manocha, D.: M3ER: Multiplicative Multimodal Emotion Recognition using Facial, Textual, and Speech Cues. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 34, 2020, https://doi.org/10.1609/aaai.v34i02.5492
28. Mittal, T., Guhan, P., Bhattacharya, U., Chandra, R., Bera, A., Manocha, D.: EmotiCon: Context-Aware Multimodal Emotion Recognition Using Frege's Principle. 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) pp. 14222–14231, 2020, https://doi.org/10.1109/CVPR42600.2020.01424
29. Mocanu, B., Tapu, R., Zaharia, T.: Multimodal emotion recognition using cross modal audio-video fusion with attention and deep metric learning. Image and Vision Computing **133**, 104676, 2023, https://doi.org/10.1016/j.imavis.2023.104676
30. Müller, V.I., Habel, U., Derntl, B., Schneider, F., Zilles, K., Turetsky, B.I., Eickhoff, S.B.: Incongruence effects in cross-modal emotional integration. NeuroImage **54**(3), 2257–2266, 2011, https://doi.org/10.1016/j.neuroimage.2010.10.047
31. Otten, M., Seth, A.K., Pinto, Y.: A Social Bayesian Brain: How Social Knowledge Can Shape Visual Perception. Brain and Cognition **112**, 69–77, 2017, https://doi.org/10.1016/j.bandc.2016.05.002
32. Peng, Z., Wu, H., Song, Z., Xu, H., Zhu, X., He, J., Liu, H., Fan, Z.: EmoTalk: Speech-Driven Emotional Disentanglement for 3D Face Animation. In: 2023 IEEE/CVF International Conference on Computer Vision (ICCV), 2023, https://doi.org/10.1109/ICCV51070.2023.01891
33. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., Sutskever, I.: Learning Transferable Visual Models From Natural Language Supervision. In: International Conference on Machine Learning, 2021
34. Rao, R.P.N., Ballard, D.H.: Predictive coding in the visual cortex: A functional interpretation of some extra-classical receptive-field effects. Nature Neuroscience **2**(1), 79–87, 1999, https://doi.org/10.1038/4580
35. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-Resolution Image Synthesis with Latent Diffusion Models. In: 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, https://doi.org/10.1109/CVPR52688.2022.01042
36. Sadok, S., Leglaive, S., Girin, L., Alameda-Pineda, X., Séguier, R.: A multimodal dynamical variational autoencoder for audiovisual speech representation learning **172**, 106120, https://doi.org/10.1016/j.neunet.2024.106120
37. Schneider, S., Baevski, A., Collobert, R., Auli, M.: Wav2vec: Unsupervised pre-training for speech recognition. In: Proc. Interspeech 2019, https://doi.org/10.21437/Interspeech.2019-1873

38. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. 3rd International Conference on Learning Representations (ICLR 2015) 2015
39. Sinha, S., Biswas, S., Yadav, R., Bhowmick, B.: Emotion-Controllable Generalized Talking Face Generation. In: Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence, 2022, https://doi.org/10.24963/ijcai.2022/184
40. Stypułkowski, M., Vougioukas, K., He, S., Zięba, M., Petridis, S., Pantic, M.: Diffused heads: Diffusion models beat GANs on talking-face generation. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024
41. Suess, F., Rabovsky, M., Abdel Rahman, R.: Perceiving emotions in neutral faces: Expression processing is biased by affective person knowledge. Social Cognitive and Affective Neuroscience **10**(4), 531–536, 2015, https://doi.org/10.1093/scan/nsu088
42. Sun, N., Lu, Q., Zheng, W., Liu, J., Han, G.: Unsupervised Cross-View Facial Expression Image Generation and Recognition. IEEE Transactions on Affective Computing **14**(1), 718–731, 2023, https://doi.org/10.1109/TAFFC.2020.3029531
43. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. In: Guyon, I., Luxburg, U.V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., Garnett, R. (eds.) Advances in Neural Information Processing Systems, vol. 30, 2017
44. 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: Vedaldi, A., Bischof, H., Brox, T., Frahm, J.M. (eds.) Computer Vision – ECCV 2020, vol. 12366, https://doi.org/10.1007/978-3-030-58589-1_42
45. Wieser, M.J., Brosch, T.: Faces in Context: A Review and Systematization of Contextual Influences on Affective Face Processing. Frontiers in Psychology **3**, 2012, https://doi.org/10.3389/fpsyg.2012.00471
46. Xu, C., Zhu, J., Zhang, J., Han, Y., Chu, W., Tai, Y., Wang, C., Xie, Z., Liu, Y.: High-Fidelity Generalized Emotional Talking Face Generation with Multi-Modal Emotion Space Learning. In: 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, https://doi.org/10.1109/CVPR52729.2023.00639
47. Xu, Y., Deng, B., Wang, J., Jing, Y., Pan, J., He, S.: High-resolution Face Swapping via Latent Semantics Disentanglement. In: 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, https://doi.org/10.1109/CVPR52688.2022.00749
48. Yan, Y., Huang, Y., Chen, S., Shen, C., Wang, H.: Joint Deep Learning of Facial Expression Synthesis and Recognition. IEEE Transactions on Multimedia **22**(11), 2792–2807, 2020, https://doi.org/10.1109/TMM.2019.2962317
49. Zhang, K., Zhang, Z., Li, Z., Qiao, Y.: Joint face detection and alignment using multitask cascaded convolutional networks **23**(10), 1499–1503, https://doi.org/10.1109/LSP.2016.2603342

50. Zhang, S., Pan, Y., Wang, J.Z.: Learning Emotion Representations from Verbal and Nonverbal Communication. In: 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, https://doi.org/10.1109/CVPR52729.2023.01821
51. Zhang, X., Zhang, F., Xu, C.: Joint Expression Synthesis and Representation Learning for Facial Expression Recognition. IEEE Transactions on Circuits and Systems for Video Technology **32**(3), 1681–1695, 2022, https://doi.org/10.1109/TCSVT.2021.3056098
52. Zhang, Z., Wang, L., Yang, J.: Weakly Supervised Video Emotion Detection and Prediction via Cross-Modal Temporal Erasing Network. In: 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, https://doi.org/10.1109/CVPR52729.2023.01811
53. Zheng, W., Yan, L., Wang, F.Y.: Two Birds With One Stone: Knowledge-Embedded Temporal Convolutional Transformer for Depression Detection and Emotion Recognition. IEEE Transactions on Affective Computing **14**(4), 2595–2613, 2023, https://doi.org/10.1109/TAFFC.2023.3282704

How Do You Perceive My Face? Recognizing Facial Expressions in Multi-Modal Context by Modeling Mental Representations

Supplementary Material

Abstract. In our work, we presented a model that encodes facial images and audio context using VAEs. We developed a context fusion network called context attention net (CAN) which shifts the latent facial distribution, to allow more accurate classifications by the classification head, as well as generate an approximation of the facial expression a human would perceive. Here, we provide additional details on the joint training of the CAN and classification head and provide a derivation of their loss function. We also list more detailed parameters of our rating study. Lastly, we provide more facial expression generations that show the capabilities of our model to purposefully fuse a facial image with affective audio context. The generations are based on neutral face images paired with different audio contexts to mimic the setting in our rating study. In addition, we release the code of our work.

6 Architecture Details

The number of layers of the components of our model are given in Tab. 4.

Component	# Layers
VAE Encoder	16
VAE Decoder	16
CAN	3
Classification Head	1

Table 4: Number of layers in each component of our model.

7 Rating Study

In this section we provide additional information about the procedure used for our rating study with human participants. The study was conducted according to the principles expressed in the Declaration of Helsinki and was approved by the ethics committee of Department of Psychology at Humboldt-Universität zu Berlin. All participants gave their informed consent.

7.1 Materials

In **Experiment 1**, faces of 24 actors from the RAVDESS database with a neutral expression and audio files in which the actors said the sentence “Kids are playing by the door” either with angry or happy emotional prosody served as stimuli. The images were neutral as generated by our network with a weight parameter of 0. We chose a generated face instead of the original frame from the RAVDESS database to eliminate potential effects of low-level visual differences between generated images and images from the database.

In **Experiment 2**, neutral faces of the same 24 actors were shifted towards the model’s representation of the face in the context of either a happy audio or an angry audio, with five different weights: 0 (i.e. the neutral expression also presented in Experiment 1), 0.25, 0.5, 0.75, and 1 (i.e. the model’s originally trained weight parameter).

In both experiments, we used counterbalancing across participants, such that one participant saw each actor either in the happy or in the angry emotion condition and each face was shown equally as often in each emotion condition.

Participants The study adhered to the principles of the Declaration of Helsinki, approved by the ethics committee of Department of Psychology at Humboldt-Universität zu Berlin. Participants were recruited from Prolific.com and received monetary compensation. For each experiment, 80 participants were recruited. The final samples included 72 English native speakers aged 18–35 years ($M = 29.01$) in Experiment 1 and 77 English native speakers aged 18–35 years ($M = 28.58$) in Experiment 2. The samples were balanced, with 50% of participants identifying as female and 50% identifying as male. We used the following criteria for inclusion in the final samples: Participants who reported not being highly distracted during the experiment, participants who reported not giving random ratings, participants who reported being able to hear all the audio files (Experiment 1), and participants who did not report rating only the audio files (Experiment 1). Participants were pre-screened for the following criteria based on their data provided to Prolific.com

Age: 18–35; Prison: No; Approval Rate: 90–100%; Units of alcohol per week: 0, 1-4, or 5-9; Neurodiversity: No; Dyslexia: No; Vision: Yes; Hearing difficulties: No; Cochlear implant: No; Colourblindness: No; Head Injury—Knock out history: No; Head Injury: No; Mental health/illness/condition - ongoing: No; Medication use: No; Mild cognitive impairment/Dementia: No; Autism Spectrum Disorder: No; Depression: No; Mental illness daily impact: No; Anxiety: No; ADD/ADHD: No; Anxiety Severity: No; Mental Health Diagnosis: No; Mental Health Treatment: None; First Language: English.

7.2 Procedure

Both experiments followed a similar procedure: After providing informed consent, participants rated the facial expressions of stimuli described above in random order on a 9-point Likert scale ranging from “very negative” to “very pos-

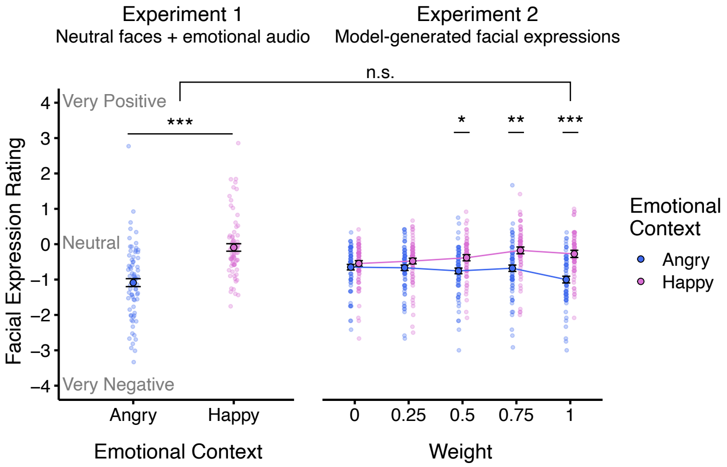


Fig. 9: Rating study results **with additional statistical information**, depicting mean facial expression ratings in Experiment 1 (left), in which participants rated neutral faces presented in the context of emotional audio, and Experiment 2 (right), in which participants rated faces generated by our model with happy vs. angry audio context (with 5 different values for the context weight m of Eq. (14)). At $m = 1$, ratings in Experiment 2 are equal to those of Experiment 1, demonstrating our model's capability to successfully capture the effect of emotional context in human observers. Small dots represent mean ratings per participant, large dots denote grand means across participants, and error bars indicate 95% confidence intervals. Statistical significance levels are denoted by asterisks: * ($p < .05$), ** ($p < .01$), and *** ($p < .001$); "n.s." are non-significant differences.

itive" with "neutral" in the middle. In Experiment 1, ratings pertained to a neutral face presented with happy or angry audio, while in Experiment 2, faces modified by our model to depict contextual influence were rated. Post-rating, participants answered questions about their task experience (e.g., whether they were distracted, whether they gave their ratings randomly, and their potential awareness of the hypothesis tested in the study), were debriefed on the study's purpose, and directed back to Prolific.com.

7.3 Statistical Analyses

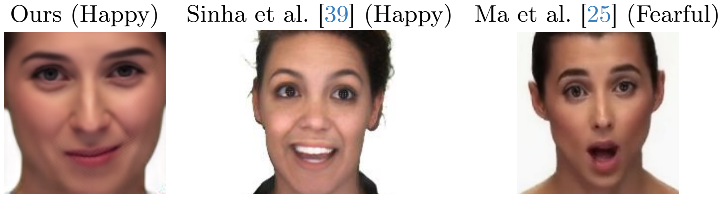
For Experiment 1, we ran a linear mixed effects model with the factor emotion (angry vs. happy audio) coded as a sliding difference contrast (meaning that the estimated effect reflects the predicted mean difference between faces seen with an angry audio vs. a happy audio). For Experiment 2, we ran a linear mixed effects model with the factor emotion (generated face shifted towards an angry vs. a happy expression) and the factor weight (with five levels, 0, 0.25, 0.5, 0.75, and 1). To compare the differences in expression ratings between the happy and the angry condition within each of the weights, the factor emotion was nested within the factor weight. For both experiments, we modelled random intercepts for participants and items (i.e. depicted actors), as well as random slopes for the effect of emotion over participants and items [1]. The significance of fixed effects coefficients ($p < 0.05$) was tested by Satterthwaite approximation.

7.4 Results

Separate models were run for each weight level, including the factors emotional context, experiment (2 vs. 1), and their interaction. With a weight of 1, no significant interaction between experiment and emotional context was found ($b = -0.22$, $p = .084$), whereas this interaction was significant for all other weights ($bs < -0.42$, $ps < .001$). Fig. 9 shows the results including interaction significance.

8 RAVDESS Comparison to SOTA Generations

In the main paper we omitted a comparison to SOTA generation methods [10, 25, 39] because they pursue a different purpose—making a neutral input images resemble strongly the accompanying audio context. Due to our objective of synthesizing mental representations, such heavy shifts in the facial expression would be too strong for a human observer. Changes in subjective appearance are rather subtle [41]. For the sake of completeness, we provide a comparison to SOTA generations in Fig. 10. The two SOTA competitors use a target ground-truth sequence they want to model, whereas we compute our adaption only based on the expression class and the model implicilty learns a sensible shift.



Ours (Happy) Sinha et al. [39] (Happy) Ma et al. [25] (Fearful)

(a) Comparison of expression generation with SOTA methods.

Fig. 10: Comparison of our approach to the SOTA generation methods presented in [25, 39]. All methods use a neutral input image and generate the result using affective audio context. The two SOTA competitor methods aim at generating a video stream that strongly resembles the emotion in the audio. We aim at synthesizing the subjective facial expression a human observer would perceive.

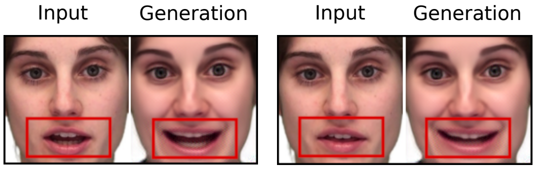


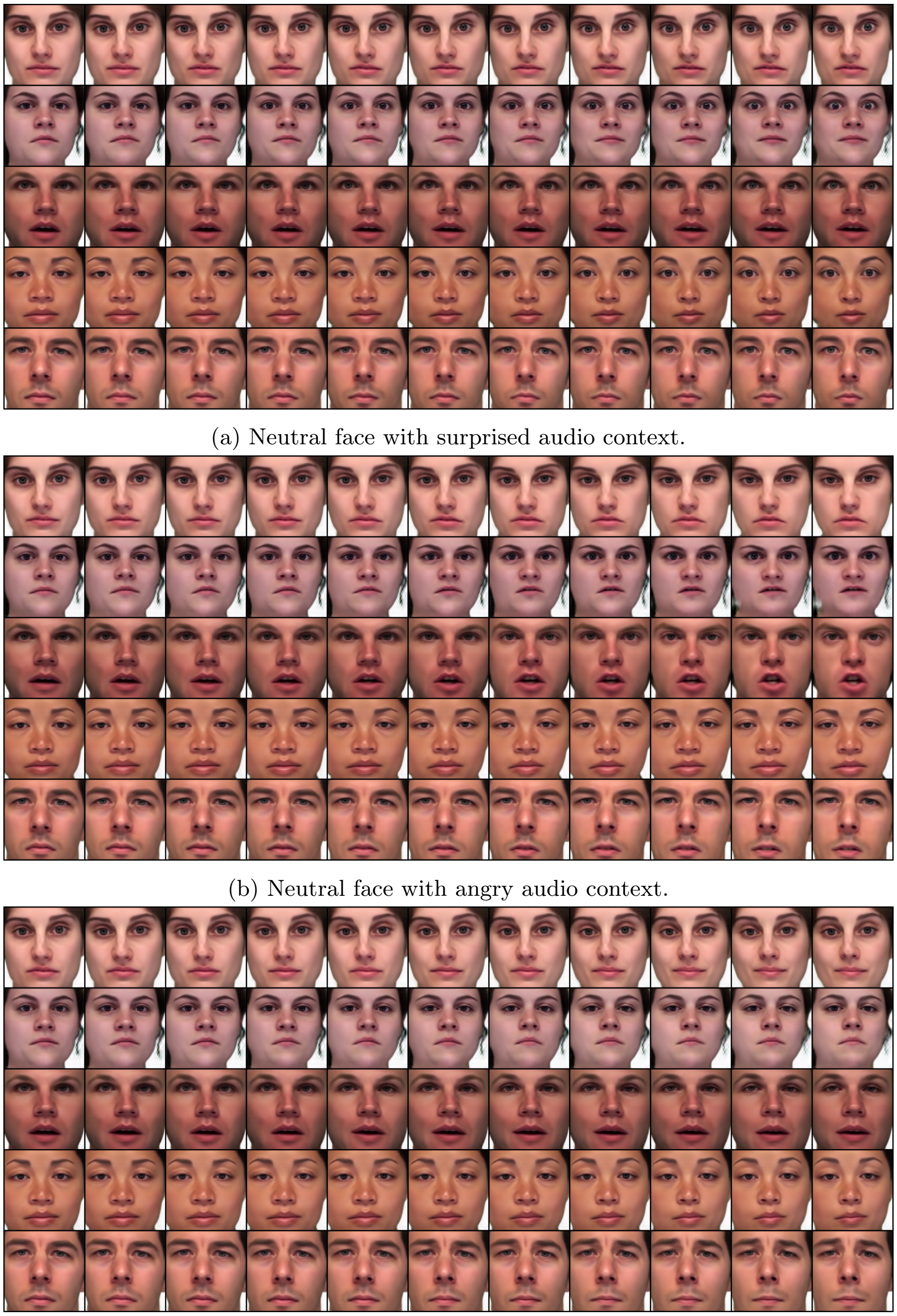
Fig. 11: Generations of neutral images paired with happy audio context. The opening of the mouth influences the opening of the mouth generated by our model. Best viewed zoomed in.

9 Additional Synthesized Facial Expressions

We provide additional generations of facial expressions on the RAVDESS dataset using our proposed model. Fig. 11 shows that our model picks up the subtle differences in the mouth opening in the generations. Figs. 12 to 14 show generations for a neutral input image (left column) paired with different contexts. The context weight m is increased from 0 (second column) to 1 (rightmost column) in 0.1 steps. Fig. 15 shows additional generations for variations of neutral input images with different contexts, to showcase that our model adapts the style of the input image. The leftmost column is the neutral input image, from left to right follows the reconstruction without context, calm audio context, happy audio context, sad audio context, fearful audio context, disgusted audio context, and surprised audio context. Fig. 16 shows generations of a neutral input image (top row) paired with normal (middle row) and strong context intensity (bottom row). The RAVDESS dataset provides these two intensities as a binary label for every sample. The context expression classes are from left to right: happy, sad, angry, fearful, disgusted, surprised.

10 MEAD Generations

Reconstructions on the MEAD dataset are shown in Fig. 17.

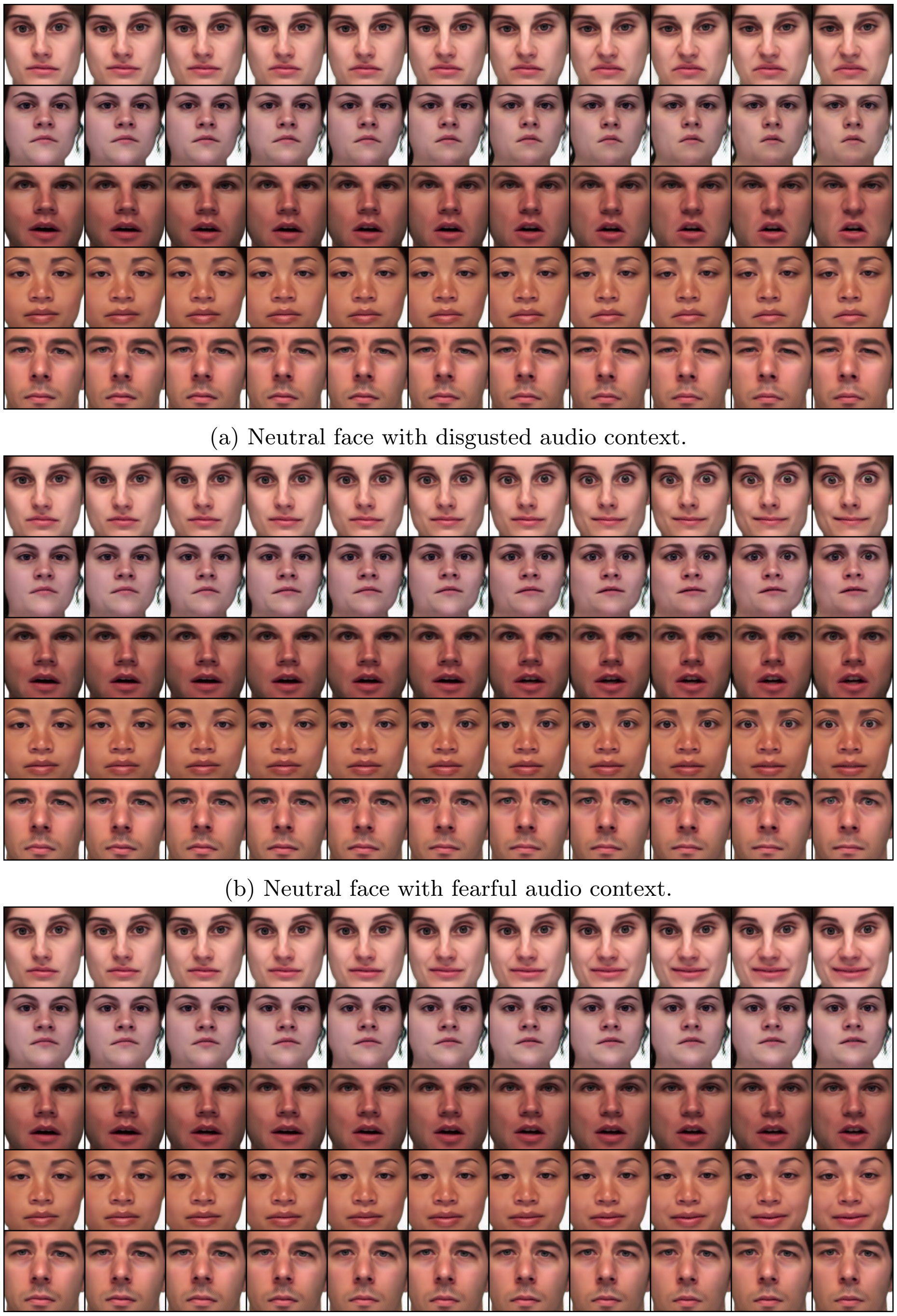


(a) Neutral face with surprised audio context.

(b) Neutral face with angry audio context.

(c) Neutral face with calm audio context.

Fig. 12: Facial expression generations with varying context weight m from 0 to 1 in 0.1 steps.

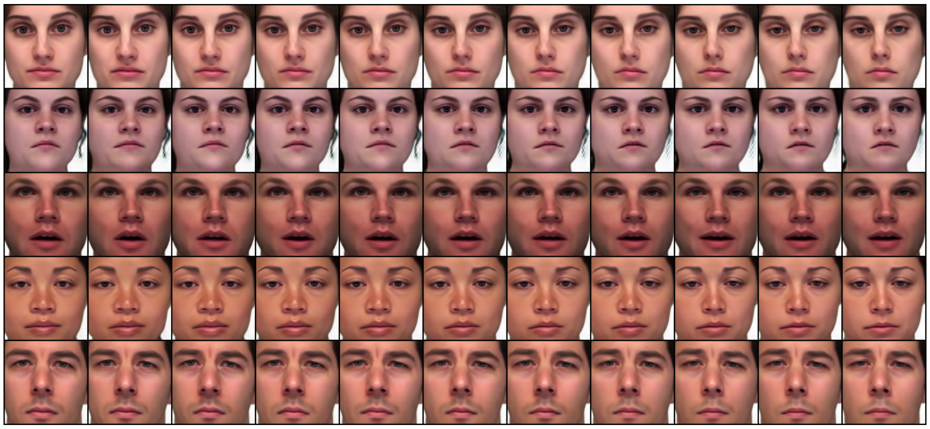


(a) Neutral face with disgusted audio context.

(b) Neutral face with fearful audio context.

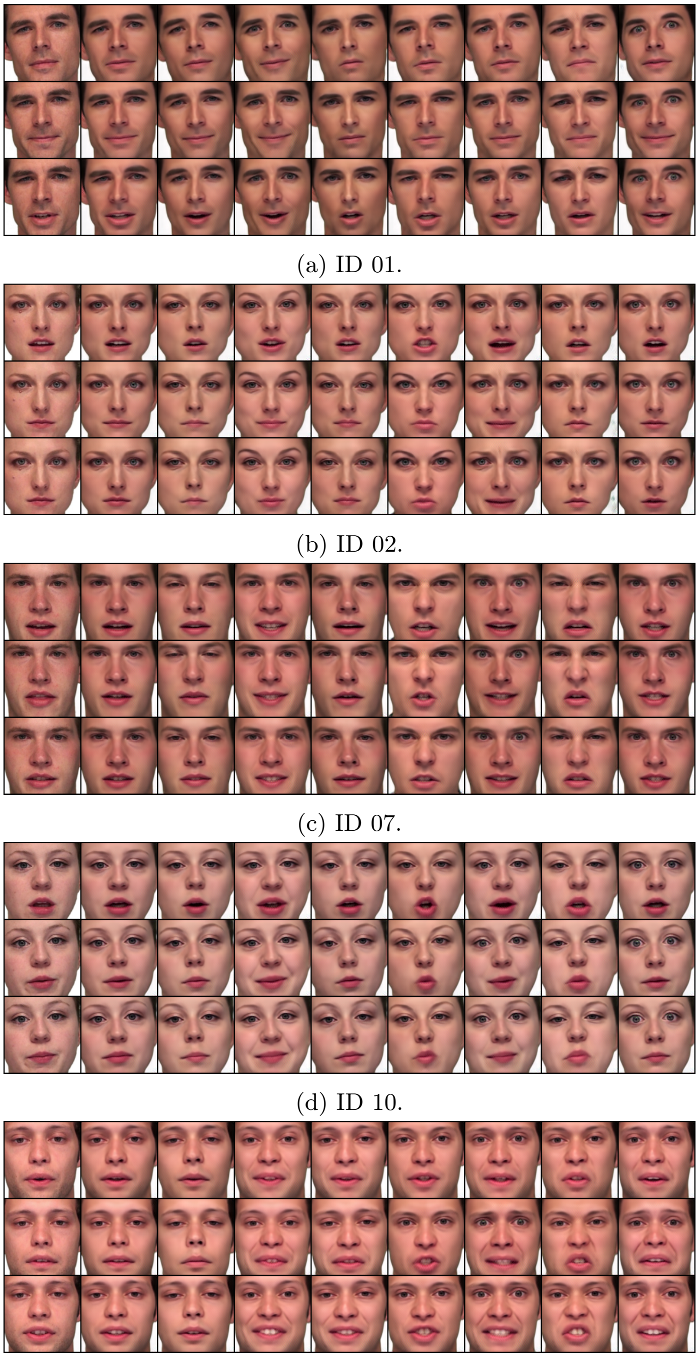
(c) Neutral face with happy audio context.

Fig. 13: Facial expression generations with varying context weight m from 0 to 1 in 0.1 steps (continued).



(a) Neutral face with happy audio context.

Fig. 14: Facial expression generations with varying context weight m from 0 to 1 in 0.1 steps (continued).



(a) ID 01.

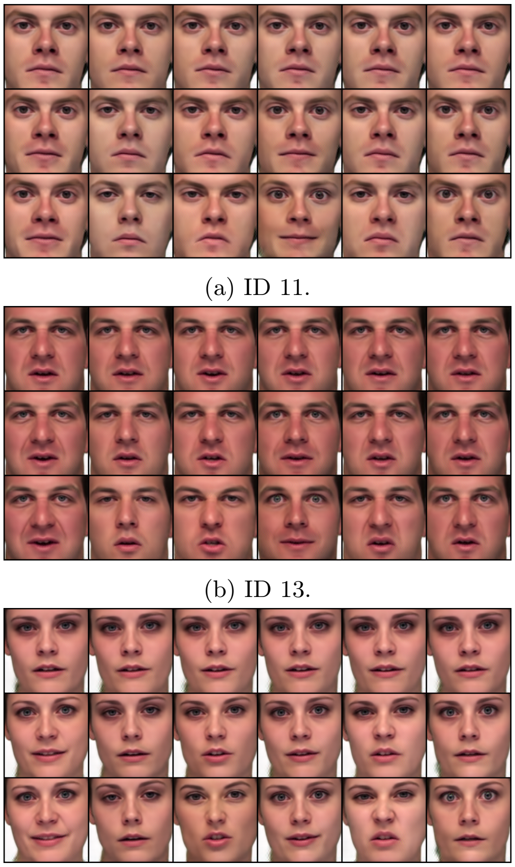
(b) ID 02.

(c) ID 07.

(d) ID 10.

(e) ID 15.

Fig. 15: Generations of neutral faces paired with contexts of all 7 expressions for a selection of IDs from the RAVDESS dataset. From left to right: Input image, reconstruction without context, calm, happy, sad, angry, fearful, disgusted, surprised. We chose five IDs from RAVDESS.



(a) ID 11.

(b) ID 13.

(c) ID 14.

Fig. 16: Generations of neutral faces paired with contexts of all 7 expressions **with normal and strong intensity** for a selection of IDs from the RAVDESS dataset. Top row: input image, middle row: normal intensity, bottom row: strong intensity. From left to right: happy, sad, angry, fearful, disgusted, surprised. We chose three IDs from RAVDESS. The RAVDESS comes with *normal* and *strong* binary labels for the expressions.

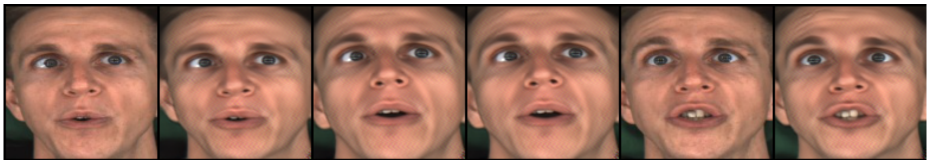


Fig. 17: Reconstructions of our method on the MEAD dataset. Pairs of input image (left) and reconstruction (right).