FERGI: Automatic Scoring of User Preferences for Text-to-Image Generation from Spontaneous Facial Expression Reaction

Shuangquan Feng*   Junhua Ma*   Virginia R. de Sa
University of California San Diego, La Jolla, CA 92093
{s1feng,jum002,desa}@ucsd.edu

Abstract

Researchers have proposed to use data of human preference feedback to fine-tune text-to-image generative models. However, the scalability of human feedback collection has been limited by its reliance on manual annotation. Therefore, we develop and test a method to automatically score user preferences from their spontaneous facial expression reaction to the generated images. We collect a dataset of Facial Expression Reaction to Generated Images (FERGI) and show that the activations of multiple facial action units (AUs) are highly correlated with user evaluations of the generated images. We develop an FAU-Net (Facial Action Units Neural Network), which receives inputs from an AU estimation model, to automatically score user preferences for text-to-image generation based on their facial expression reactions, which is complementary to the pre-trained scoring models based on the input text prompts and generated images. Integrating our FAU-Net valence score with the pre-trained scoring models improves their consistency with human preferences. This method of automatic annotation with facial expression analysis can be potentially generalized to other generation tasks. The code is available at https://github.com/ShuangquanFeng/FERGI, and the dataset is also available at the same link for research purposes.

1 Introduction

The rapid recent advancements in text-to-image generative models have enabled the generation of high-fidelity images aligned with text prompts [54, 58, 60–63]. To better improve the fidelity of the generated images and their alignment with the text prompts, researchers collected large-scale datasets of human preference feedback on images generated by text-to-image models [35, 42, 77–79, 86] and proposed methods of fine-tuning the models with data of human preference feedback [24, 41, 79].

*Equal Contribution

However, the scalability of human feedback collection has still been limited by its reliance on manual annotation. Researchers have used the CLIP score [59], Aesthetic score [64], BLIP score [43], ImageReward score [79], PickScore [35], and HPS v2 score [77] to automatically score human preferences of text-to-image generation based on the input text prompts and the generated images. In this study, we develop and test a system to automatically score user preferences from their spontaneous facial expression reaction to the generated images, which is complementary to the models based only on the texts and images. This automatic scoring requires zero additional effort from the real users of text-to-image generators.

We present the Facial Expression Reaction to Generated Images (FERGI) dataset, which comprises video recordings of 33 participants' facial expression reaction to 2827 images generated by Stable Diffusion (SD) v1.4 [62] based on 576 different self-drafted text prompts, along with their feedback on the generated images from manual input. We estimate the activation of facial action units (AUs), as defined in the Facial Action Coding System (FACS, a comprehensive system breaking down facial expressions into individual components of AUs of muscle movements [21]), in the facial expression reaction videos with a model we trained on external datasets and show that the activations of multiple AUs are highly correlated with user evaluations of the generated images. We further propose a method of automatically scoring user preferences by detecting the activations of the AUs in the facial expression reaction to the generated images and feeding them into a neural network.

2 Related Work

Text-to-Image Generation and Evaluation. Various models have been developed for text-to-image generation, including Generative Adversarial Networks (GANs) [5, 15, 16, 25, 31, 32], Variational autoencoders (VAEs) [10, 34, 68], flow-based models [17, 18, 33], autoregressive models (ARMs) [9, 11, 71, 72, 84], and diffusion models (DMs) [28, 66, 67]. More recent advancements in DMs

[14, 54, 58, 61–63] have achieved great success in generating high-quality images and attracted widespread public attention. As traditional evaluation metrics like Inception Score (IS) [3], Fréchet Inception Distance (FID) [27], and CLIP score [59] failed to comprehensively capture human preferences for text-to-image generation, researchers have proposed to specifically train human preference scoring models based on large-scale human feedback datasets, such as Human Preference Score (HPS) [78], ImageReward [79], PickScore [35], and Human Preference Score v2 (HPS v2) [77], and demonstrated the effectiveness of using them to improve the text-to-image generative models [24, 35, 77, 79].

Human Feedback Datasets. There are multiple datasets of human feedback on text-to-image generation, including AGIQA-1K [86], Human Preference Dataset (HPD) [78], ImageReward [79], Pick-a-Pic [35], AGIQA-3K [42], and Human Preference Dataset v2 (HPD v2) [77]. To the best of our knowledge, our FERGI dataset is the first to include both manual annotations of generated images and the associated facial expression reaction videos.

Facial Expression Recognition and Applications. Automatic facial expression recognition has advanced rapidly in recent years [44]. As emotional responses to image generation can be more complicated than single categories of the most widely-researched seven basic emotions [19, 20, 50], we decided to analyze the facial expression reactions directly in terms of muscle movements, as defined by AUs in FACS [21]. AU detection and estimation has attracted increasing interest [12, 29, 36, 37, 45, 46, 49, 65, 69, 70, 73, 87–89] and been used for facial emotion recognition [80–83]. For example, researchers have shown that the activation of the corrugator (activated in AU4) and zygomaticus (activated in AU12) are associated with negative and positive emotions respectively and are the most investigated muscle activations in studying affect and emotion [76]. Also, the activity of the corrugator supercilii (activated in AU4) is positively associated with amygdala and negatively associated with ventromedial prefrontal cortex activity [26] and increases with negative stimuli, such as negative images [6, 7, 39, 40]. For applications, [4, 75] showed the effectiveness of using facial expression to infer/analyze human preferences.

3 FERGI Dataset

3.1 Participants

39 participants were recruited from the SONA system of University of California San Diego (UCSD) and completed the study asynchronously on personal computers. 6 participants were removed from the dataset for various reasons, including failure to participate, failure in video recording,

and not permitting sharing of their video recordings for research. This resulted in a dataset of 33 participants that will be made available for research. The collection and usage of human participant data in the dataset were approved by the Institutional Review Board (IRB) of UCSD.

3.2 Data Collection Procedure

Each participant completed multiple sessions of data collection. In each session, the participant drafted one text prompt and viewed 5 images generated from the input text prompt. A flow chart of the procedure of each session is shown in Fig. 1. Details are explained below.

3.2.1 Prompt Creation

There are two types of sessions, structured input and free-form input. In both types, the participant can freely draft any text prompt as long as it is not NSFW. The only difference between the two types lies in how the input text prompt is created. In a structured input session, the participant fills out a form to separately specify different elements desired in the images to generate: animate objects, inanimate objects, interactional relations (between two objects), positional relations (between two objects), location, style (of the image), and keywords. Based on the contents of the form, a text prompt is then automatically generated by the Large Language Model (LLM) OpenAI gpt3.5-turbo. The participant can freely adjust the generated prompt as the final input text prompt. In a free-form input, the participant directly enters the entire final input text prompt.

3.2.2 Recording

After the input text prompt is finalized, the participant is directed to a webcam preview to confirm that the webcam captures their face appropriately. The webcam is then used to record throughout the rest of the session. The configuration of the raw videos varied based on the participant's device and browser. Though, recorded videos of most participants have a resolution of 640×480, and all recorded videos were standardized to 30 FPS afterwards.

3.2.3 Image Presentation, Annotation, and Ranking

In each session, a total of 5 images generated by SD v1.4 [62] from the input text prompt are presented and annotated sequentially and ranked afterwards (present image 1 → annotate image 1 → ... → present image 5 → annotate image 5 → rank 5 images). The image presentation and annotation followed the webcam preview. The detailed procedures for the presentation and annotation of each generated image are as follows:
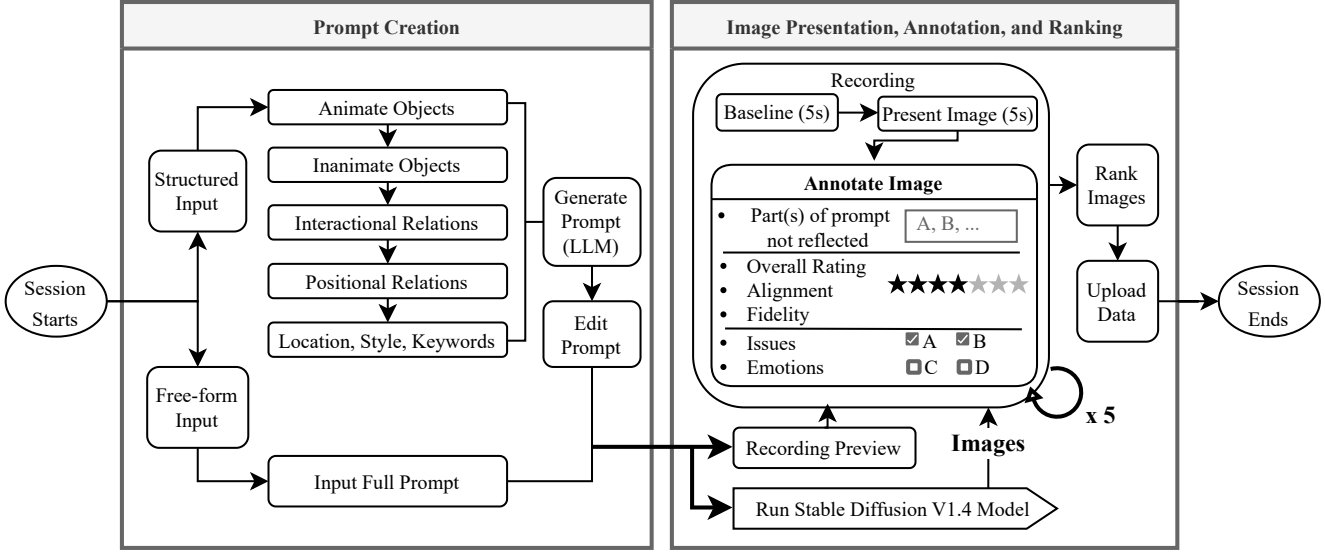
Figure 1. Procedure of one session in data collection.

- **Baseline** (5 seconds): The website displays the "preparing image" status text at the center of the page along with the input text prompt. No specific facial-expression-eliciting event is expected to happen during this period.
- **Present** (5 seconds): The generated image is presented at the center of the webpage along with the input text prompt. The participant's facial expression reaction to the generated image is collected within this interval.
- **Annotate** (no time restriction): The website displays the prompt, the generated image (moved to the right side), and the annotation survey. The participant annotates the image manually by filling out the survey during this interval. The survey was adapted from the survey designed for ImageReward [79] and includes the following questions: important elements not reflected in the image (free-response question), overall rating (on a scale of 1 (worst) to 7 (best)), image-text alignment rating (on a scale of 1 to 7), fidelity rating (on a scale of 1 to 7), issues of the image[1] (multi-choice question), and emotions felt when seeing the image (multi-choice question).

After the annotation of the last image, the participant ranks the 5 generated images from best to worst, and the data of the session is uploaded to the web server ending the session.

3.3 Data Preprocessing

For each image, two clips are extracted from the recorded video: a 5-second "baseline clip" of the participant's facial

---

[1]For the first 7 out of 33 participants (or 8 out of the total 39 participants), there was a bug (later fixed) in checkbox selection for answering the questions regarding "issues of the image" that made the recorded answers unreliable, so the data regarding their answers for this question should not be used in analysis.

expression during the baseline period and a 5-second "reaction clip" of the participant's facial expression during the image presentation period.

SD v1.4 [62] outputs an all-black image when it detects the original output image might be inappropriate. A total of 53 such images out of 2880 total generated were generated in the dataset. The data associated with these images were excluded from further analysis, resulting in fewer than 5 valid images for some input text prompts.

4 AU Model Training

We analyze the facial expression reaction of the users by estimating the activation of their AUs. We trained our AU estimation model on the DISFA [52, 53] and DISFA+ [51] datasets. They are among the largest, most popular, and most cited AU datasets with intensity annotations. In these datasets, each frame is manually annotated by a human expert with intensities on a scale of 0 to 5, for activation of AU1 (inner brow raiser), AU2 (outer brow raiser), AU4 (brow lowerer), AU5 (upper lid raiser), AU6 (cheek raiser), AU9 (nose wrinkler), AU12 (lip corner puller), AU15 (lip corner depressor), AU17 (chin raiser), AU20 (lip stretcher), AU25 (lips part), and AU26 (jaw drop). (See Fig. S10 for a visual reference guide for the analyzed AUs.) Video frames are first preprocessed with face detection [48], face alignment [8], and a combination of histogram equalization and linear mapping [38] before being fed into the model for training, testing, or inference. For model training, we used the neural network IR-50 [13] pre-trained on Glink360k [1] and fine-tuned it on the DISFA and DISFA+ datasets [51, 53]. A single network learned the estimation of all la-

beled AUs with both regression and ordinal classification [55]. More details of the training of the AU model is elaborated in the Appendix / Supplementary Materials.

The comparison between our AU estimation model's performance on the DISFA dataset and other state-of-the-art models, including OpenFace 2.0 [2], CCNN-IT [73], 2DC [47], SCC-Heatmap [23], and iARL [65] is shown in Tab. 1. Our model's Pearson's r (higher is better) is close to that of OpenFace 2.0; our model's ICC(3,1) (higher is better) is the best among all models. Although our model's MAE (lower is better) is less competitive among all models, we believe that ICC(3,1) is a better metric for evaluating the performance of AU estimation models on the DISFA dataset considering the high imbalance of the dataset.

5 Facial Feature Extraction

In this section, we elaborate the procedure of extracting facial features from the video clips associated with each generated image based on the trained AU recognition model. Similar procedures were used to extract features from baseline clips and reaction clips. Further details are in the Appendix / Supplementary Materials.

5.1 Data Filtering

Robustness against face occlusion and faces with non-frontal poses is an ongoing challenge for facial expression recognition models [30, 85]. Our AU model was trained on datasets consisting predominantly of frontal facial images without occlusion and thus for each 5-second video clip, we exclude frames with a low face detection confidence score (given by MediaPipe [48]) and frames with an off-angle pitch or yaw (estimated based on detected facial landmarks [8]). If a clip has more than 20% of its total frames excluded, the whole clip, and its associated data (survey and ranking results) are excluded from further analysis.

5.2 AU Activation Value

We compute an AU activation value α_i for each trained AUi in each video clip. The AU activation value is computed as follows:

- Firstly, the intensity of each AU for each frame of the video clip is estimated with the AU model: $\hat{y}_{i,\mathrm{reg}}^{(1)}, \hat{y}_{i,\mathrm{reg}}^{(2)}, \ldots$, where the superscript indicates the index of the frame.
- Secondly, a moving window mean of the estimated AU intensities for every 0.1 seconds (3 frames for 30 FPS) is computed:

$$\bar{\hat{y}}_{i,\mathrm{reg}}^{(k)} = \frac{1}{3}\sum_{k'=k}^{k+2}\hat{y}_{i,\mathrm{reg}}^{(k')} \tag{1}$$

for all k with defined $\hat{y}_{i,\mathrm{reg}}^{(k+1)}$ and $\hat{y}_{i,\mathrm{reg}}^{(k+2)}$ (as they may be undefined for frames at the end of the video clip or frames adjacent to excluded frames).

- Finally, the AU activation of the clip as the difference between the maximum AU intensity within the video clip and the AU intensity of the first 0.1 seconds is computed:

$$\alpha_i = \max_k(\bar{\hat{y}}_{i,\mathrm{reg}}^{(k)}) - \bar{\hat{y}}_{i,\mathrm{reg}}^{(1)}. \tag{2}$$

6 Experiments

6.1 Statistical Analysis

We started with analyzing statistical relationships between the AU activation values and answers in the annotation survey for each image. The primary results are as follows:

- **Overall Ratings.** The participant's overall ratings (on a scale of 1 to 7) of the generated images have a significant positive correlation with the activation values of AU2 and AU12 and a significant negative correlation with the activation values of AU4 and AU9 as shown in Figs. 2a, 2b, 2f and 2g (Spearman correlation: $\rho \approx 0.09$ and $p < 5 \times 10^{-5}$ for AU2, $\rho \approx -0.18$ and $p < 5 \times 10^{-17}$ for AU4, $\rho \approx -0.09$ and $p < 5 \times 10^{-5}$ for AU9, $\rho \approx 0.08$ and $p < 5 \times 10^{-4}$ for AU12).
- **Extremity of Overall Ratings.** The extremities of the participant's overall ratings (on a scale of 1 to 7) of the generated images (computed as the magnitude of deviation from the midpoint rating (|overall rating $- 4$|) have a significant positive correlation with the activation values of AU1, AU2, AU5, AU12, and AU20 as shown in Figs. 2c to 2g (Spearman correlation: $\rho \approx 0.10$ and $p < 5 \times 10^{-6}$ for AU1, $\rho \approx 0.11$ and $p < 5 \times 10^{-7}$ for AU2, $\rho \approx 0.08$ and $p < 5 \times 10^{-4}$ for AU5, $\rho \approx 0.14$ and $p < 5 \times 10^{-11}$ for AU12, $\rho \approx 0.08$ and $p < 1 \times 10^{-4}$ for AU20).
- **Emotions.** For the question "Did you feel any of the following emotions when you saw the image?", the participant can independently select 0 to 6 answers from the six options: surprised, disgusted, amused, scared, disappointed, and satisfied. Each of the reported emotions was found to be significantly correlated with the activation values of multiple AUs.
  - **Surprised.** "Surprising" images are associated with higher activation values of AU9, AU12, AU25, and AU26 as shown in Fig. 3a (Wilcoxon rank-sum tests: $z \approx 3.78$ and $p < 5 \times 10^{-4}$ for AU9, $z \approx 4.29$ and $p < 5 \times 10^{-5}$ for AU12, $z \approx 4.01$ and $p < 1 \times 10^{-4}$ for AU25, $z \approx 4.45$ and $p < 1 \times 10^{-5}$ for AU26).
  - **Disgusted.** "Disgusting" images are associated with higher activation values of AU4, AU5, and AU9 as shown in Fig. 3b (Wilcoxon rank-sum tests: $z \approx 8.43$ and $p < 1 \times 10^{-16}$ for AU4, $z \approx 3.94$ and $p < 1 \times 10^{-4}$

Table 1. Comparison of our AU estimation model's performance on the DISFA dataset with other models

| AU	1	2	4	5	6	9	12	15	17	20	25	26	Avg.
Pearson's r (higher is better)													
OpenFace 2.0	**.64**	**.50**	.70	**.67**	**.59**	.54	**.85**	.39	**.49**	**.22**	.85	**.67**	**.59**
Our Model	.53	.46	**.74**	.57	.56	**.60**	**.85**	**.42**	.41	.21	**.94**	.66	.58
ICC(3,1) (higher is better)													
CCNN-IT	.18	.15	.61	.07	.65	.55	.82	.44	.37	**.28**	.77	.54	.45
2DC	.70	**.55**	.69	.05	.59	**.57**	**.88**	.32	.10	.08	.90	.50	.50
SCC-Heatmap	**.73**	.44	**.74**	.06	.27	.51	.71	.04	.37	.04	.94	**.78**	.47
iARL	.13	.36	.68	.22	**.56**	.36	.86	**.52**	.37	.12	**.96**	.60	.48
Our Model	.53	.46	**.74**	**.56**	.55	**.57**	.85	.41	**.40**	.20	.93	.66	**.57**
MAE (lower is better)													
CCNN-IT	87	.63	.86	.26	.73	.57	.55	.38	.57	.45	.81	.64	.61
SCC-Heatmap	**.16**	**.16**	**.27**	**.03**	**.25**	**.13**	.32	.15	**.20**	.09	.30	.32	**.20**
iARL	.30	.31	.52	.04	.36	.30	**.31**	**.05**	.33	**.08**	**.29**	**.26**	.26
Our Model	.37	.39	.50	.14	.37	.21	.33	.24	.47	.23	.32	.42	.33
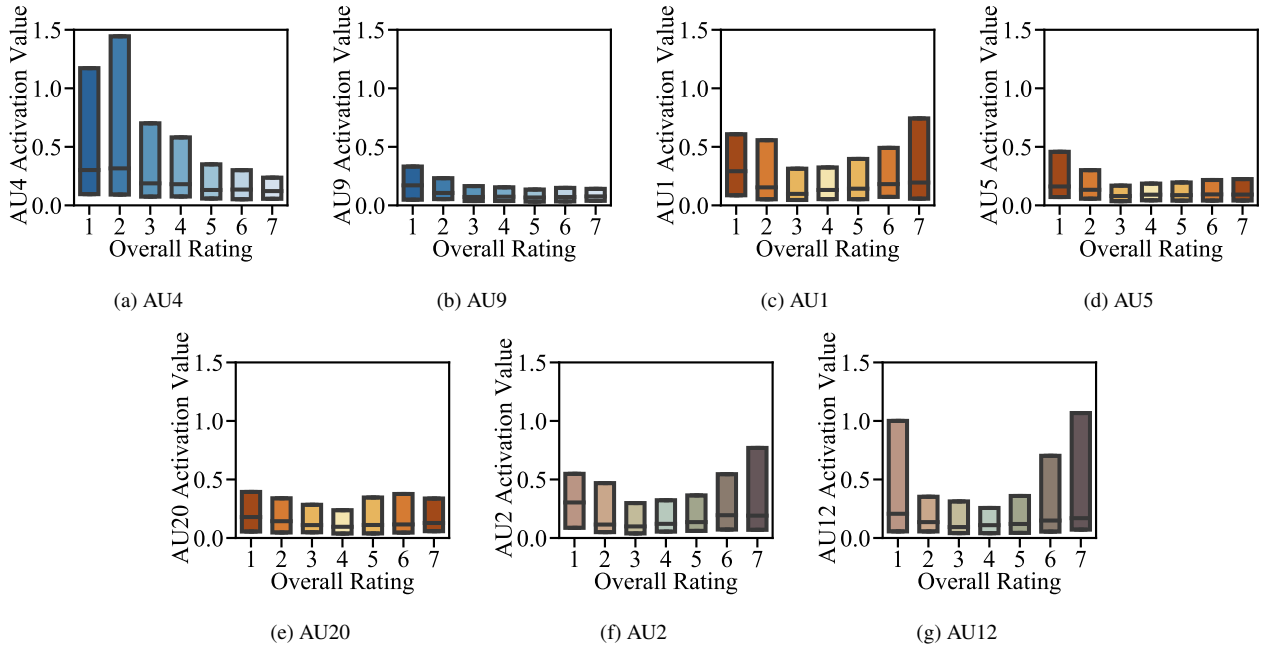
Figure 2. **Overall ratings highly correlated with AU activation values.** The distributions of the activation values of multiple AUs for images of different ratings. Each subfigure shows the results for a different AU (indicated in the captions) and contains 7 boxplots for AU activation values (on the y-axis) for images with 7 different ratings along the x-axis. In each boxplot, the bottom/top of the box represents the first/third quartile (25th/75th percentile) of the AU activation values, and the line in the middle of the box represents the median. Blue bars are used for AUs that are significantly correlated with the ratings while brown bars are used for AUs that are significantly correlated with the extremity of the ratings (computed as |overall rating -4|). AU2 and AU12 are shown in blue and brown reflecting significance (after multiple comparisons) for both ratings and extremity of ratings. The darker blue and darker brown colors show the ratings associated with higher AU activations. The numbers of ratings from 1 to 7 are 104, 185, 258, 456, 581, 403, and 216 respectively.

for AU5, $z \approx 4.87$ and $p < 1 \times 5^{-6}$ for AU9).

– **Amused.** "Amusing" images are associated with higher activation values of AU6, AU9, AU12, AU20, and AU25 as shown in Fig. 3c (Wilcoxon rank-sum tests: $z \approx 7.00$ and $p < 1 \times 10^{-11}$ for AU6, $z \approx 5.01$ and $p < 1 \times 10^{-6}$ for AU9, $z \approx 9.15$ and $p < 1 \times 10^{-19}$ for AU12, $z \approx 6.12$ and $p < 1 \times 10^{-9}$ for AU20, $z \approx 3.85$ and $p < 1 \times 10^{-4}$ for AU25).

– **Scared.** "Scary" images are associated with higher activation values of AU4 and AU5 as shown in Fig. 3d
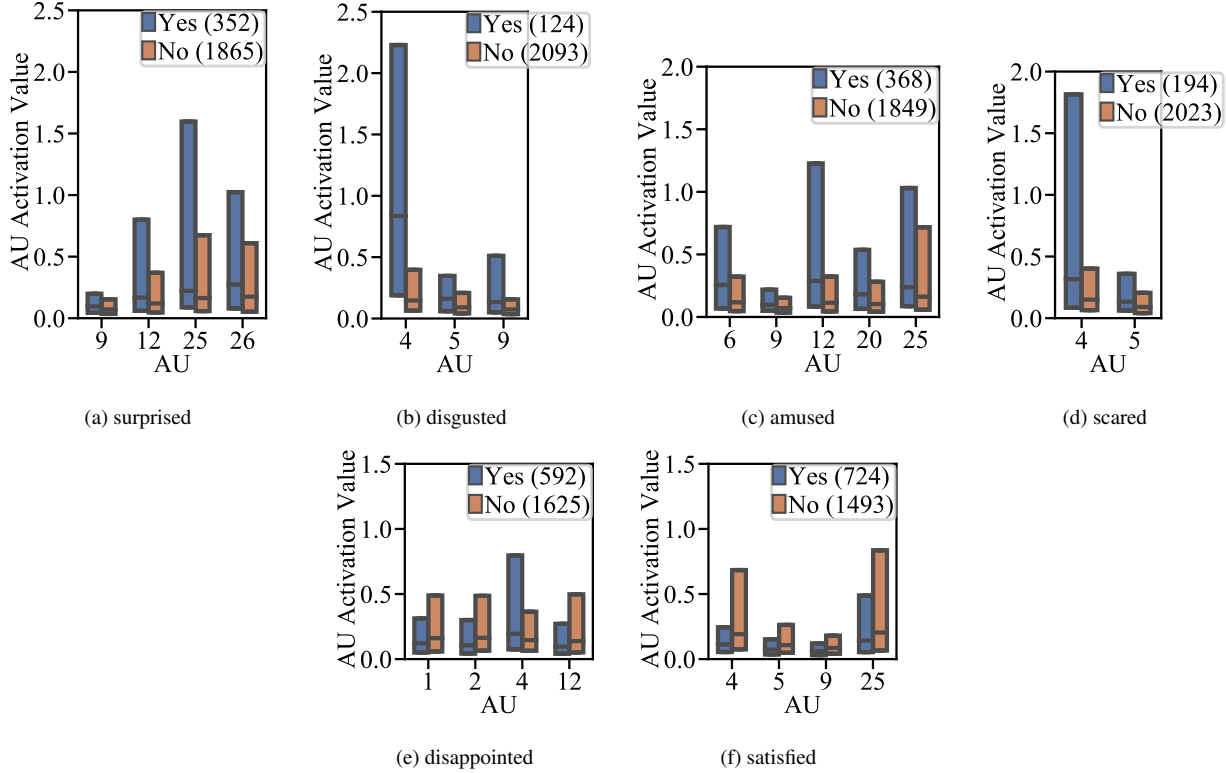
Figure 3. **Reported emotions highly correlated with AU activation values.** The distributions of the activation values of multiple AUs for images eliciting different emotions of the participants as reported in answers to the question "Did you feel any of the following emotions when you saw the image?". Each subfigure shows results for a different emotion (indicated in the captions) and contains multiple pairs of boxplots representing the results for different AUs. The indices of the AUs are along the x-axis, while the y-axis represents the AU activation values in the reaction clip of each image. The parenthesized numbers in the legends give the number of responses of that type. In each pair of boxplots, the boxplot on the left/right side represents the AU activation values for images that did/didn't elicit the corresponding emotion. In each boxplot, the bottom/top of the box represents the first/third quartile (25th/75th percentile) of the AU activation values, and the line in the middle of the box represents the median.

(Wilcoxon rank-sum tests: $z \approx 5.60$ and $p < 5 \times 10^{-8}$ for AU4, $z \approx 4.39$ and $p < 5 \times 10^{-5}$ for AU5).

- **Disappointed.** "Disappointing" images are associated with lower activation values of AU1, AU2, and AU12 and higher activation values of AU4 as shown in Fig. 3e (Wilcoxon rank-sum tests: $z \approx -4.25$ and $p < 5 \times 10^{-5}$ for AU1, $z \approx -5.82$ and $p < 1 \times 10^{-8}$ for AU2, $z \approx 4.82$ and $p < 5 \times 10^{-6}$ for AU4, $z \approx -4.86$ and $p < 5 \times 10^{-6}$ for AU12).
- **Satisfied.** "Satisfying" images are associated with lower activation values of AU4, AU5, AU9, and AU25 as shown in Fig. 3f (Wilcoxon rank-sum tests: $z \approx -9.07$ and $p < 5 \times 10^{-19}$ for AU4, $z \approx -6.62$ and $p < 5 \times 10^{-11}$ for AU5, $z \approx -6.93$ and $p < 5 \times 10^{-12}$ for AU9, $z \approx -3.79$ and $p < 5 \times 10^{-4}$ for AU25).

Note that the results presented above selectively include the AUs whose associated p-values (as reported before corrections) are below the Bonferroni-corrected significance

threshold of $0.05/[12 \times (2 + 6)]$ (12 investigated AUs, 2 tests for overall ratings (1 for the raw rating and 1 for the extremity of the rating), and 6 emotions).

6.2 Automatic Annotation of User Preferences between Image Pairs

6.2.1 Pre-trained Scoring Models

Researchers have used the CLIP score [59], Aesthetic score [64], BLIP score [43], ImageReward score [79], PickScore [35], and HPS v2 score [77] to estimate human preferences of text-to-image generation based on the input text prompts and the generated images.

Among these baseline scoring models, ImageReward, PickScore, and HPS v2 were all trained specifically on datasets of human preferences of text-to-image generation and are supposed to outperform the other models [35, 77, 79]. To introduce a potentially more competitive

6

baseline score, we further combine them to form an ensemble baseline score

$$s_{\text{ens}} = w_{\text{IR}}s_{\text{IR}} + w_{\text{Pick}}s_{\text{Pick}} + w_{\text{HPSv2}}s_{\text{HPSv2}} \quad (3)$$
$$\text{subject to} \quad w_{\text{IR}} + w_{\text{Pick}} + w_{\text{HPSv2}} = 1, \quad (4)$$

where s_{ens} represents the ensemble baseline score, s_{IR}, s_{Pick}, and s_{HPSv2} represent the ImageReward score, PickScore, and HPS v2 score respectively, and w_{IR}, w_{Pick}, and w_{HPSv2} represent their respective weights.

A leave-one-participant-out (LOPO) procedure was applied in the fitting of the three weights w_{IR}, w_{Pick}, and w_{HPSv2}; we used grid search with a granularity of 0.1 to optimize the accuracy of image preference prediction. Since the scales of the scores are different, they were each z-scored using their mean and standard deviation on the training set. We obtained the same results of $w_{\text{IR}} = 0.1$, $w_{\text{Pick}} = 0.6$, and $w_{\text{HPSv2}} = 0.3$ for all LOPO training sets.

6.2.2 FAU-Net Valence Score

We aim to use the estimated AU activation values to automatically score user preferences between images generated from the same input text prompt and propose that this scoring more directly reflects the user's reaction to the generated image and is complementary to the pre-trained scoring models. To achieve this goal, we train an FAU-Net (Facial Action Units Neural Network) to do automatic scoring. In the FAU-Net, the estimated activation values of the 12 investigated AUs are used as the inputs; each activation value is first preprocessed with a linear transformation with trained parameters and sigmoid function ($\sigma(a \cdot \alpha + b)$ with the initializations of $a = 1$ and $b = -1$ for encouraging transformations of positive correlation for better interpretability); then the preprocessed AU activation values are used as inputs to a hidden layer with 16 neurons which feeds into a single ouput neuron representing the AU evaluation score of the image; this output unit receives inputs from both the 16 neurons of the hidden layer and the 12 preprocessed AU activation values. The sigmoid function is used as the activation function for all neurons. The network is trained with a RankNet loss function using the reaction clips and ranking order given by the participant for image pairs generated from the same text prompts and an L1 regularization with a coefficient of 10^{-4}. For model training, we use the Adam optimizer with an initial learning rate of 10^{-3} and a single batch for all data to train the neural network for 300 epochs using a single NVIDIA RTX A6000 48GB GDDR6 GPU for about 15 minutes.

Our FAU-Net is trained with an LOPO procedure and outputs an FAU-Net valence score s_{FAU} reflecting the evaluation of the generated image derived from the facial expression reaction. Note that the AU estimation model is frozen and not trained with the FERGI data.

6.2.3 Integrate FAU-Net Valence Score with Pre-trained Scoring Models

We integrate our FAU-Net valence score and the pre-trained scoring model with a simple linear combination:

$$s_m + a_m s_{\text{FAU}}, \quad (5)$$

where m represents the specific pre-trained scoring model, s_m the score from model m, and a_m the weight for integrating m and the FAU-Net valence score.

We separately investigated the effect of integrating the FAU-Net valence score with each of the three scoring models trained on large human preference datasets (ImageReward, PickScore, and HPS v2), and their ensemble. Here, we only fit the weights a_{IR}, a_{Pick}, a_{HPSv2}, and a_{ens} individually with an LOPO procedure using grid search (from 0.0 to 10.0 with a granularity of 0.1); the FAU-Net weights and the weightings w_{IR}, w_{Pick}, w_{HPSv2} computed in Eq. (3) are not modified at this stage.

Table 2. Accuracy of image preference prediction for a total of 3734 image pairs. The second column shows the accuracy of the models given in the first column when they make predictions independently while the third column shows the accuracy of the best few when they are integrated with the FAU-Net valence score.

Model	Ind. Acc.	Acc. w/ AUs
CLIP Score	57.74	/
Aesthetic Score	50.32	/
BLIP Score	50.05	/
FAU-Net Valence Score	59.35	/
ImageReward Score	60.47	65.21
PickScore	63.28	66.28
HPS V2 Score	62.00	66.87
Ensemble Baseline Score	65.80	**68.64**

The results are shown in Tab. 2. Integration with FAU-Net valence score improves the performance of all four baseline models. Specifically, by further integrating all three pre-trained human preference scoring models together with the FAU-Net valence score, we achieve the highest accuracy of 68.64%, outperforming the ensemble of the three models by a margin of 2.84% ($p < 0.01$ with a two-proportion z-test).

6.2.4 Complementarity between FAU-Net and Pre-trained Scoring Models

The complementarity between FAU-Net and pre-trained scoring models is demonstrated in Fig. 4. The figure shows how annotation accuracy changes if only a subset of the data are selected to annotate by the different scoring models (image pairs are selected based on the largest absolute
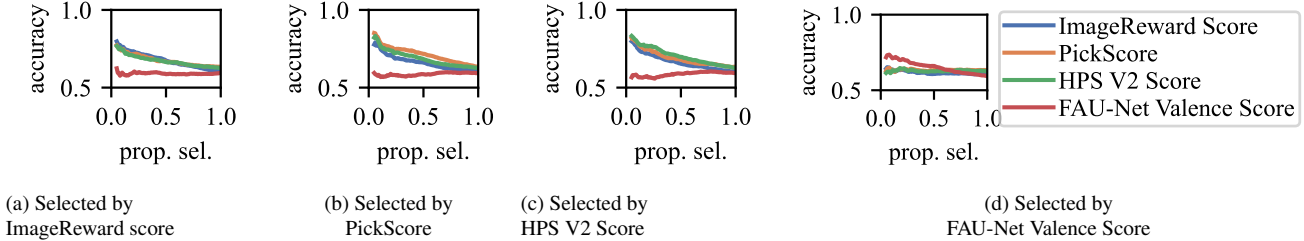
Figure 4. **FAU-Net valence score is complementary to the other pre-trained scoring models.** Each subfigure shows how annotation accuracy changes if only a subset of the data are selected to annotate by the different scoring models (given in the subcaption label), with the x-axis representing the proportion of data selected (image pairs are selected based on the largest absolute difference of scores for the two images) and the y-axis representing the annotation accuracy within the selected subset. Each model is good at estimating the image pairs that it will perform well on as evidenced by increased accuracy for a smaller selected proportion. Note the blue, green, and orange curves are almost on top of each other.

difference scores for the two images given by the associated scoring model). We can clearly see that the three pre-trained scoring models are highly correlated showing similar performance curves (almost overlapping). In contrast, our FAU-Net valence score selects different image pairs, explaining why the FAU-Net valence scores combine synergistically with the other scores.

6.2.5 FAU-Net Analysis

We analyze the weights of the trained FAU-Net to better understand the model. Figure 5 shows the weights from the input preprocessed AU activation values to the nodes in the hidden layer (given that the preprocessings of activation values are still all increasing functions after model training) with the subcaptions illustrating the weight from the corresponding hidden node to the output node. Generally, for hidden nodes positively affecting the output score, they receive positive weights from the activations of AU2, AU6, AU12, AU17, AU20, and AU26 and negative weights from the activations of AU4, AU5, AU9, AU17, and AU25, and for hidden nodes negatively affecting the output score, they receive weights of the opposite signs from the AUs. Similar trends are observed for the weights from the input preprocessed AU activation values and the output node as shown in Fig. 6.

7 Discussion and Conclusion

We propose that user facial expression reactions to text-to-image generation can be used to score user preferences for the generated images, which is complementary to existing scoring models which estimate human preferences based only on the input text prompts and generated images. We present the FERGI dataset, consisting of video recordings of facial expression reaction to text-to-image generation,

and show that multiple AUs are correlated with the participant's evaluation of and emotional reaction to the generated image.

Specifically, we developed an FAU-Net that receives the detected AU activations in the facial expression reaction as inputs and outputs an FAU-Net valence score as the estimation of the user preference. Integrating the FAU-Net valence score with pre-trained scoring models improves their accuracy in image preference prediction, which can be potentially helpful when the scoring models are used to label human preferences for fine-tuning text-to-image generative models.

Importantly, previous work has shown that even if the accuracy of the predictions is not very high, it can still be effective for fine-tuning of text-to-image models: Xu et al. [79] used the reward model ImageReward (trained from manually labeled human preferences), which only has an accuracy of 65.14% on their own dataset and 60.47% on our FERGI dataset, to fine-tune SD v1.4 and obtained a new model whose generated images were preferred by humans 58.79% of the time over those generated by the original SD v1.4. (see Tab. 4 and Fig. 6 of [79]).

We also demonstrated the complementarity between our FAU-Net and the pre-trained scoring models, explaining why integrating them effectively improves the annotation accuracy. As they are respectively better at annotating different sets of image pairs, it is also potentially feasible to use them to fine-tune each other on data where one model gives larger absolute difference of scores for the two images.

In conclusion, we have demonstrated the feasibility of automatically scoring user preferences for image generation from facial expression reaction. The application of this method is not limited to text-to-image generation but may also be potentially applicable to other image generation tasks, such as image-to-image translation [57], image inpainting [22], and super-resolution [74].
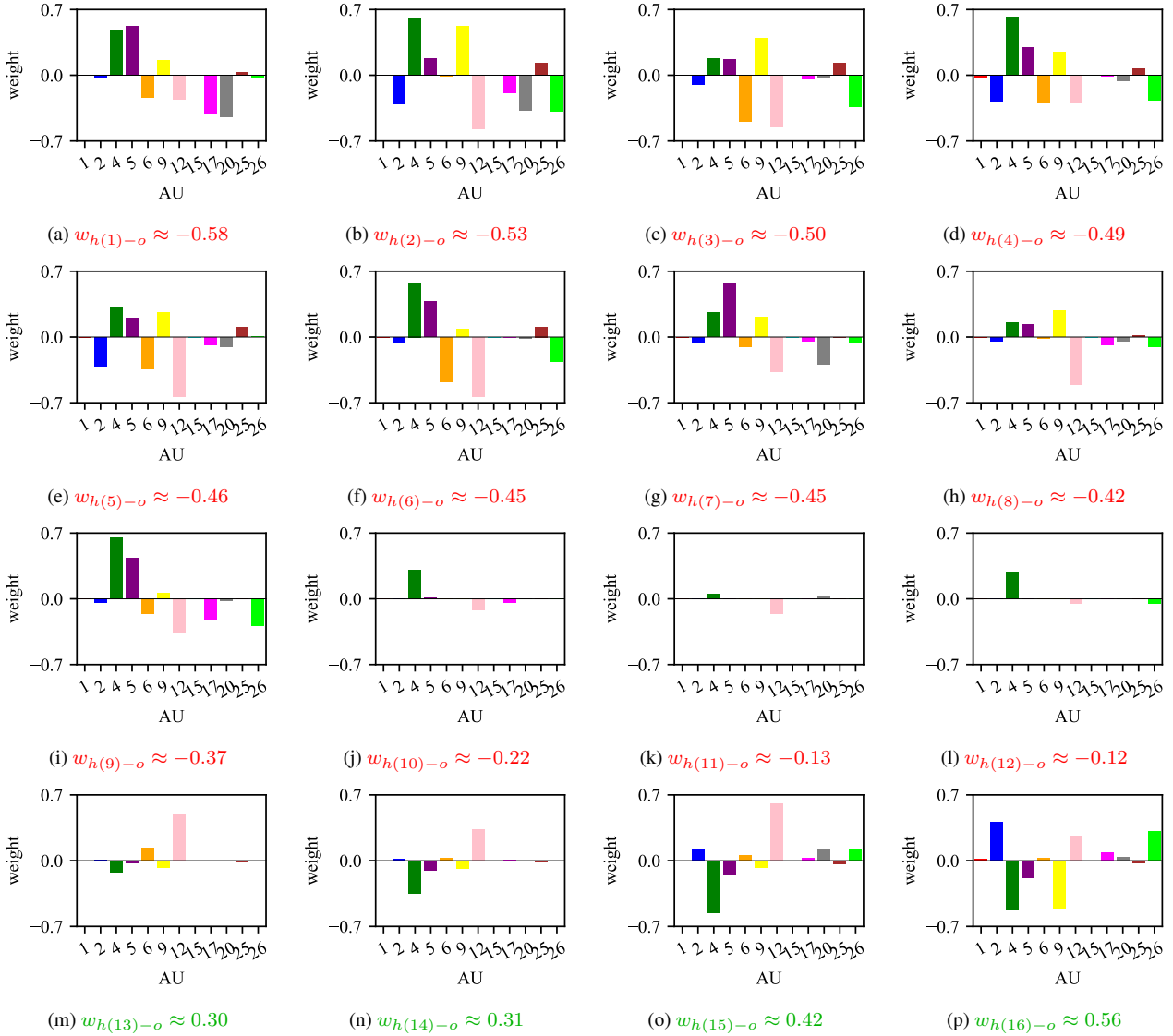
(a) $w_{h(1)-o} \approx -0.58$ (b) $w_{h(2)-o} \approx -0.53$ (c) $w_{h(3)-o} \approx -0.50$ (d) $w_{h(4)-o} \approx -0.49$

(e) $w_{h(5)-o} \approx -0.46$ (f) $w_{h(6)-o} \approx -0.45$ (g) $w_{h(7)-o} \approx -0.45$ (h) $w_{h(8)-o} \approx -0.42$

(i) $w_{h(9)-o} \approx -0.37$ (j) $w_{h(10)-o} \approx -0.22$ (k) $w_{h(11)-o} \approx -0.13$ (l) $w_{h(12)-o} \approx -0.12$

(m) $w_{h(13)-o} \approx 0.30$ (n) $w_{h(14)-o} \approx 0.31$ (o) $w_{h(15)-o} \approx 0.42$ (p) $w_{h(16)-o} \approx 0.56$

Figure 5. **Weights of hidden nodes.** Each subfigure shows the weights from the input preprocessed activation values of the 12 AUs to a hidden node in the FAU-Net, with each subcaption illustrating the weight from the hidden node to the output node (red/green for negative/positive weights). The hidden nodes are ordered by the weights from them to the output node in this figure.
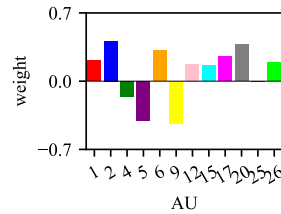

Figure 6. **Weights of the output node.** The weights from the input preprocessed activation values of the 12 AUs to the output node in the FAU-Net.

8 Acknowledgments

9

References

[1] Xiang An, Jiankang Deng, Jia Guo, Ziyong Feng, XuHan Zhu, Jing Yang, and Tongliang Liu. Killing two birds with one stone: Efficient and robust training of face recognition CNNs by partial FC. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4042–4051, 2022. 3, 4

[2] Tadas Baltrušaitis, Peter Robinson, and Louis-Philippe Morency. Openface: an open source facial behavior analysis toolkit. In *2016 IEEE winter conference on applications of computer vision (WACV)*, pages 1–10. IEEE, 2016. 4

[3] Shane Barratt and Rishi Sharma. A note on the inception score. *arXiv preprint arXiv:1801.01973*, 2018. 2

[4] Federico Becattini, Xuemeng Song, Claudio Baecchi, Shi-Ting Fang, Claudio Ferrari, Liqiang Nie, and Alberto Del Bimbo. PLM-IPE: A pixel-landmark mutual enhanced framework for implicit preference estimation. In *Proceedings of the 3rd ACM International Conference on Multimedia in Asia*, pages 1–5, 2021. 2

[5] Andrew Brock, Jeff Donahue, and Karen Simonyan. Large scale gan training for high fidelity natural image synthesis. *arXiv preprint arXiv:1809.11096*, 2018. 1

[6] John T Cacioppo and Louis G Tassinary. Inferring psychological significance from physiological signals. *American psychologist*, 45(1):16, 1990. 2

[7] John T Cacioppo, Richard E Petty, Mary E Losch, and Hai Sook Kim. Electromyographic activity over facial muscle regions can differentiate the valence and intensity of affective reactions. *Journal of personality and social psychology*, 50(2):260, 1986. 2

[8] Cunjian Chen. PyTorch Face Landmark: A fast and accurate facial landmark detector, 2021. Open-source software available at https://github.com/cunjian/pytorch_face_landmark. 3, 4, 5

[9] Mark Chen, Alec Radford, Rewon Child, Jeffrey Wu, Heewoo Jun, David Luan, and Ilya Sutskever. Generative pretraining from pixels. In *International conference on machine learning*, pages 1691–1703. PMLR, 2020. 1

[10] Rewon Child. Very deep vaes generalize autoregressive models and can outperform them on images. *arXiv preprint arXiv:2011.10650*, 2020. 1

[11] Rewon Child, Scott Gray, Alec Radford, and Ilya Sutskever. Generating long sequences with sparse transformers. *arXiv preprint arXiv:1904.10509*, 2019. 1

[12] Wen-Sheng Chu, Fernando De la Torre, and Jeffrey F Cohn. Learning spatial and temporal cues for multi-label facial action unit detection. In *2017 12th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2017)*, pages 25–32. IEEE, 2017. 2

[13] Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. Arcface: Additive angular margin loss for deep face recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4690–4699, 2019. 3, 4

[14] Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. *Advances in neural information processing systems*, 34:8780–8794, 2021. 2

[15] Ming Ding, Zhuoyi Yang, Wenyi Hong, Wendi Zheng, Chang Zhou, Da Yin, Junyang Lin, Xu Zou, Zhou Shao, Hongxia Yang, et al. Cogview: Mastering text-to-image generation via transformers. *Advances in Neural Information Processing Systems*, 34:19822–19835, 2021. 1

[16] Ming Ding, Wendi Zheng, Wenyi Hong, and Jie Tang. Cogview2: Faster and better text-to-image generation via hierarchical transformers. *Advances in Neural Information Processing Systems*, 35:16890–16902, 2022. 1

[17] Laurent Dinh, David Krueger, and Yoshua Bengio. Nice: Non-linear independent components estimation. *arXiv preprint arXiv:1410.8516*, 2014. 1

[18] Laurent Dinh, Jascha Sohl-Dickstein, and Samy Bengio. Density estimation using real nvp. *arXiv preprint arXiv:1605.08803*, 2016. 1

[19] Paul Ekman. Strong evidence for universals in facial expressions: a reply to Russell's mistaken critique. *Psychological Bulletin*, 115(2):268–287, 1994. 2

[20] Paul Ekman and Wallace V Friesen. Constants across cultures in the face and emotion. *Journal of personality and social psychology*, 17(2):124, 1971. 2

[21] Paul Ekman and Wallace V Friesen. Facial action coding system. *Environmental Psychology & Nonverbal Behavior*, 1978. 1, 2

[22] Omar Elharrouss, Noor Almaadeed, Somaya Al-Maadeed, and Younes Akbari. Image inpainting: A review. *Neural Processing Letters*, 51:2007–2028, 2020. 8

[23] Yingruo Fan, Jacqueline Lam, and Victor Li. Facial action unit intensity estimation via semantic correspondence learning with dynamic graph convolution. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 12701–12708, 2020. 4

[24] Ying Fan, Olivia Watkins, Yuqing Du, Hao Liu, Moonkyung Ryu, Craig Boutilier, Pieter Abbeel, Mohammad Ghavamzadeh, Kangwook Lee, and Kimin Lee. Dpok: Reinforcement learning for fine-tuning text-to-image diffusion models. *arXiv preprint arXiv:2305.16381*, 2023. 1, 2

[25] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks. *Communications of the ACM*, 63(11):139–144, 2020. 1

[26] Aaron S Heller, Regina C Lapate, Kaitlyn E Mayer, and Richard J Davidson. The face of negative affect: trial-by-trial corrugator responses to negative pictures are positively associated with amygdala and negatively associated with ventromedial prefrontal cortex activity. *Journal of cognitive neuroscience*, 26(9):2102–2110, 2014. 2

[27] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. GANs trained by

a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30, 2017. 2

[28] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020. 1

[29] Geethu Miriam Jacob and Bjorn Stenger. Facial action unit detection with transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7680–7689, 2021. 2

[30] Mahdi Jampour and Malihe Javidi. Multiview facial expression recognition, a survey. *IEEE Transactions on Affective Computing*, 13(4):2086–2105, 2022. 4

[31] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4401–4410, 2019. 1

[32] Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Analyzing and improving the image quality of StyleGAN. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8110–8119, 2020. 1

[33] Durk P Kingma and Prafulla Dhariwal. Glow: Generative flow with invertible 1x1 convolutions. *Advances in neural information processing systems*, 31, 2018. 1

[34] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013. 1

[35] Yuval Kirstain, Adam Polyak, Uriel Singer, Shahbuland Matiana, Joe Penna, and Omer Levy. Pick-a-pic: An open dataset of user preferences for text-to-image generation. *arXiv preprint arXiv:2305.01569*, 2023. 1, 2, 6

[36] Dimitrios Kollias. ABAW: valence-arousal estimation, expression recognition, action unit detection & multi-task learning challenges. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2328–2336, 2022. 2

[37] Dimitrios Kollias, Panagiotis Tzirakis, Alice Baird, Alan Cowen, and Stefanos Zafeiriou. ABAW: valence-arousal estimation, expression recognition, action unit detection & emotional reaction intensity estimation challenges. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5888–5897, 2023. 2

[38] Chieh-Ming Kuo, Shang-Hong Lai, and Michel Sarkis. A compact deep learning model for robust facial expression recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pages 2121–2129, 2018. 3, 4

[39] Peter J Lang, Mark K Greenwald, Margaret M Bradley, and Alfons O Hamm. Looking at pictures: Affective, facial, visceral, and behavioral reactions. *Psychophysiology*, 30(3):261–273, 1993. 2

[40] Jeff T Larsen, Catherine J Norris, and John T Cacioppo. Effects of positive and negative affect on electromyographic activity over zygomaticus major and corrugator supercilii. *Psychophysiology*, 40(5):776–785, 2003. 2

[41] Kimin Lee, Hao Liu, Moonkyung Ryu, Olivia Watkins, Yuqing Du, Craig Boutilier, Pieter Abbeel, Mohammad Ghavamzadeh, and Shixiang Shane Gu. Aligning text-to-image models using human feedback. *arXiv preprint arXiv:2302.12192*, 2023. 1

[42] Chunyi Li, Zicheng Zhang, Haoning Wu, Wei Sun, Xiongkuo Min, Xiaohong Liu, Guangtao Zhai, and Weisi Lin. AGIQA-3K: an open database for AI-generated image quality assessment. *arXiv preprint arXiv:2306.04717*, 2023. 1, 2

[43] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. BLIP: bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International Conference on Machine Learning*, pages 12888–12900. PMLR, 2022. 1, 6

[44] Shan Li and Weihong Deng. Deep facial expression recognition: A survey. *IEEE transactions on affective computing*, 13(3):1195–1215, 2020. 2

[45] Wei Li, Farnaz Abtahi, and Zhigang Zhu. Action unit detection with region adaptation, multi-labeling learning and optimal temporal fusing. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1841–1850, 2017. 2

[46] Wei Li, Farnaz Abtahi, Zhigang Zhu, and Lijun Yin. Eacnet: Deep nets with enhancing and cropping for facial action unit detection. *IEEE transactions on pattern analysis and machine intelligence*, 40(11):2583–2596, 2018. 2

[47] Dieu Linh Tran, Robert Walecki, Stefanos Eleftheriadis, Bjorn Schuller, Maja Pantic, et al. Deepcoder: Semi-parametric variational autoencoders for automatic facial action coding. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3190–3199, 2017. 4

[48] Camillo Lugaresi, Jiuqiang Tang, Hadon Nash, Chris McClanahan, Esha Uboweja, Michael Hays, Fan Zhang, Chuo-Ling Chang, Ming Guang Yong, Juhyun Lee, et al. Mediapipe: A framework for building perception pipelines. *arXiv preprint arXiv:1906.08172*, 2019. 3, 4

[49] Brais Martinez, Michel F Valstar, Bihan Jiang, and Maja Pantic. Automatic analysis of facial actions: A survey. *IEEE transactions on affective computing*, 10(3):325–347, 2017. 2

[50] David Matsumoto. More evidence for the universality of a contempt expression. *Motivation and Emotion*, 16:363–368, 1992. 2

[51] Mohammad Mavadati, Peyten Sanger, and Mohammad H Mahoor. Extended DISFA dataset: Investigating posed and spontaneous facial expressions. In *proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pages 1–8, 2016. 3, 4

[52] S Mohammad Mavadati, Mohammad H Mahoor, Kevin Bartlett, and Philip Trinh. Automatic detection of non-posed facial action units. In *2012 19th IEEE International Conference on Image Processing*, pages 1817–1820. IEEE, 2012. 3

[53] S Mohammad Mavadati, Mohammad H Mahoor, Kevin Bartlett, Philip Trinh, and Jeffrey F Cohn. DISFA: a spontaneous facial action intensity database. *IEEE Transactions on Affective Computing*, 4(2):151–160, 2013. 3, 4

[54] Alex Nichol, Prafulla Dhariwal, Aditya Ramesh, Pranav Shyam, Pamela Mishkin, Bob McGrew, Ilya Sutskever, and

Mark Chen. GLIDE: towards photorealistic image generation and editing with text-guided diffusion models. *arXiv preprint arXiv:2112.10741*, 2021. 1, 2

[55] Zhenxing Niu, Mo Zhou, Le Wang, Xinbo Gao, and Gang Hua. Ordinal regression with multiple output CNN for age estimation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4920–4928, 2016. 4

[56] Melinda Ozel. FACS LITE – STILL & FAST, 2024. A visual reference guide for FACS at https://melindaozel.com/facs-lite-still-fast (last accessed on 03/13/2024). 11

[57] Yingxue Pang, Jianxin Lin, Tao Qin, and Zhibo Chen. Image-to-image translation: Methods and applications. *IEEE Transactions on Multimedia*, 24:3859–3881, 2021. 8

[58] Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. SDXL: improving latent diffusion models for high-resolution image synthesis. *arXiv preprint arXiv:2307.01952*, 2023. 1, 2

[59] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021. 1, 2, 6

[60] Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. Zero-shot text-to-image generation. In *International Conference on Machine Learning*, pages 8821–8831. PMLR, 2021. 1

[61] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with CLIP latents. *arXiv preprint arXiv:2204.06125*, 1 (2):3, 2022. 2

[62] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022. 1, 2, 3

[63] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, et al. Photorealistic text-to-image diffusion models with deep language understanding. *Advances in Neural Information Processing Systems*, 35:36479–36494, 2022. 1, 2

[64] Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, et al. LAION-5B: an open large-scale dataset for training next generation image-text models. *Advances in Neural Information Processing Systems*, 35:25278–25294, 2022. 1, 6

[65] Zhiwen Shao, Zhilei Liu, Jianfei Cai, Yunsheng Wu, and Lizhuang Ma. Facial action unit detection using attention and relation learning. *IEEE transactions on affective computing*, 13(3):1274–1289, 2019. 2, 4

[66] Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *International conference on machine learning*, pages 2256–2265. PMLR, 2015. 1

[67] Yang Song and Stefano Ermon. Generative modeling by estimating gradients of the data distribution. *Advances in neural information processing systems*, 32, 2019. 1

[68] Arash Vahdat and Jan Kautz. Nvae: A deep hierarchical variational autoencoder. *Advances in neural information processing systems*, 33:19667–19679, 2020. 1

[69] Michel F Valstar, Timur Almaev, Jeffrey M Girard, Gary McKeown, Marc Mehu, Lijun Yin, Maja Pantic, and Jeffrey F Cohn. FERA 2015-second facial expression recognition and analysis challenge. In *2015 11th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG)*, pages 1–8. IEEE, 2015. 2

[70] Michel F Valstar, Enrique Sánchez-Lozano, Jeffrey F Cohn, László A Jeni, Jeffrey M Girard, Zheng Zhang, Lijun Yin, and Maja Pantic. FERA 2017-addressing head pose in the third facial expression recognition and analysis challenge. In *2017 12th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2017)*, pages 839–847. IEEE, 2017. 2

[71] Aaron Van den Oord, Nal Kalchbrenner, Lasse Espeholt, Oriol Vinyals, Alex Graves, et al. Conditional image generation with PixelCNN decoders. *Advances in neural information processing systems*, 29, 2016. 1

[72] Aäron Van Den Oord, Nal Kalchbrenner, and Koray Kavukcuoglu. Pixel recurrent neural networks. In *International conference on machine learning*, pages 1747–1756. PMLR, 2016. 1

[73] Robert Walecki, Vladimir Pavlovic, Björn Schuller, Maja Pantic, et al. Deep structured learning for facial action unit intensity estimation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3405–3414, 2017. 2, 4

[74] Zhihao Wang, Jian Chen, and Steven CH Hoi. Deep learning for image super-resolution: A survey. *IEEE transactions on pattern analysis and machine intelligence*, 43(10):3365–3387, 2020. 8

[75] Karin Wendin, Bodil H Allesen-Holm, and Wender LP Bredie. Do facial reactions add new dimensions to measuring sensory responses to basic tastes? *Food quality and preference*, 22(4):346–354, 2011. 2

[76] T.S.H Wingenbach. Facial EMG – investigating the interplay of facial muscles and emotions. In *Social and Affective Neuroscience of Everyday Human Interaction*. Springer, Cham., 2023. 2

[77] Xiaoshi Wu, Yiming Hao, Keqiang Sun, Yixiong Chen, Feng Zhu, Rui Zhao, and Hongsheng Li. Human preference score v2: A solid benchmark for evaluating human preferences of text-to-image synthesis. *arXiv preprint arXiv:2306.09341*, 2023. 1, 2, 6

[78] Xiaoshi Wu, Keqiang Sun, Feng Zhu, Rui Zhao, and Hongsheng Li. Human preference score: Better aligning text-to-image models with human preference. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2096–2105, 2023. 2

[79] Jiazheng Xu, Xiao Liu, Yuchen Wu, Yuxuan Tong, Qinkai Li, Ming Ding, Jie Tang, and Yuxiao Dong. ImageReward: learning and evaluating human preferences for text-to-image generation. *arXiv preprint arXiv:2304.05977*, 2023. 1, 2, 3, 6, 8

[80] Xiaojing Xu and Virginia R. de Sa. Exploring multidimensional measurements for pain evaluation using facial action units. In *15th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2020)*, pages 559–565, 2020. 2

[81] Xiaojing Xu, Kenneth D. Craig, Damaris Diaz, Matthew S. Goodwin, Murat Akcakaya, Büşra Tuğçe Susam, Jeannie S. Huang, and Virginia R. de Sa. Automated pain detection in facial videos of children using human-assisted transfer learning. In *Lecture Notes in Artificial Intelligence 11326 Artificial Intelligence in Health Revised Selected Papers from the First International Workshop, AIH 2018*, pages 162–180, 2018.

[82] Xiaoxing Xu, J.S. Huang, and V.R. de Sa. Pain evaluation in video using extended multitask learning from multidimensional measurements. In *Proceedings of Machine Learning Research, (Machine Learning for Health ML4H at NeurIPS 2019)*, 2019.

[83] Jiannan Yang, Fan Zhang, Bike Chen, and Samee U Khan. Facial expression recognition based on facial action unit. In *2019 Tenth International Green and Sustainable Computing Conference (IGSC)*, pages 1–6. IEEE, 2019. 2

[84] Jiahui Yu, Yuanzhong Xu, Jing Yu Koh, Thang Luong, Gunjan Baid, Zirui Wang, Vijay Vasudevan, Alexander Ku, Yinfei Yang, Burcu Karagol Ayan, et al. Scaling autoregressive models for content-rich text-to-image generation. *arXiv preprint arXiv:2206.10789*, 2(3):5, 2022. 1

[85] Ligang Zhang, Brijesh Verma, Dian Tjondronegoro, and Vinod Chandran. Facial expression analysis under partial occlusion: A survey. *ACM Computing Surveys (CSUR)*, 51 (2):1–49, 2018. 4

[86] Zicheng Zhang, Chunyi Li, Wei Sun, Xiaohong Liu, Xiongkuo Min, and Guangtao Zhai. A perceptual quality assessment exploration for aigc images. *arXiv preprint arXiv:2303.12618*, 2023. 1, 2

[87] Kaili Zhao, Wen-Sheng Chu, Fernando De la Torre, Jeffrey F Cohn, and Honggang Zhang. Joint patch and multi-label learning for facial action unit detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2207–2216, 2015. 2

[88] Kaili Zhao, Wen-Sheng Chu, and Honggang Zhang. Deep region and multi-label learning for facial action unit detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3391–3399, 2016.

[89] Ruicong Zhi, Mengyi Liu, and Dezheng Zhang. A comprehensive survey on automatic facial action unit analysis. *The Visual Computer*, 36:1067–1093, 2020. 2

FERGI: Automatic Scoring of User Preferences for Text-to-Image Generation from Spontaneous Facial Expression Reaction

Supplementary Material

A Supplementary Visuals

To better illustrate the idea of our paper, we present Fig. S1 showing an example of a strong AU4 activation in response to a low-quality image generation and Fig. S2 an example of a strong AU12 activation in response to a high-quality image generation.

B Data Collection

This section provides more details of the collection of the FERGI dataset not provided in Sec. 3. Figs. S3 to S9 display screenshots of an example of one session in data collection.

B.1 Participants

39 participants[1] were recruited from the SONA system of University of California San Diego (UCSD). and received academic credits as compensation. They underwent the informed consent process and received instruction for the study from the researcher via a videoconference and completed the study asynchronously afterwards on their personal computers. The videos of 3 participants failed to record and displayed only a logo of OBS Studio (the reason has not yet been fully understood by the researchers); 1 participant dropped out of study under the instruction of the researcher because the only available personal device was a tablet, which is not compatible with some procedures of the study; 1 participant did not complete any study sessions asynchronously at all after receiving the instruction in the videoconference. Among the other 34 participants with valid data and videos recorded, only 1 participant chose not to permit video recordings being shared with other researchers for research purposes; therefore, we decided to also exclude the data of this participant in our analysis for enhancing reproducibility of our study within the research community. We ended up with a dataset of 33 participants that will be made available to researchers for research purposes.

B.2 Prompt Creation

There are two forms of sessions, structured input (SI) sessions and free-form input (FFI) sessions. Most participants completed the same number of structured input sessions and free-form input sessions. The number of sessions completed by each participant are shown in Tab. S1. By default, each participant is expected to complete 9 structured input sessions and 9 free-form input sessions (for the first 13 out of the 33 participants chronologically) or to complete 10 structured input sessions and 10 free-form input sessions (for the last 20 out of the 33 participants chronologically). However, the time consumption highly varies for each participant because of numerous reasons (e.g. internet connection), and various technical issues also occur occasionally, so the total number and categories of completed sessions also vary.

B.2.1 Structured Input

The purpose of having structured input sessions is to encourage creation of more diverse text prompts featuring different types of objects, interactions, settings, and styles. Structured input contains 5 sections:

- **Animate Objects** (Fig. S3a): Any object that can do things, including humans, animals, robots, fantasy creatures, etc. Participants are required to enter the name and quantity of the object, with activity and a list of characteristics as optional inputs.
- **Inanimate Objects** (Fig. S3b): Any object that cannot do things, including plants, vehicles, furniture, etc. Participants are required to enter the name and quantity of the object, with a list of characteristics as optional inputs.
- **Interactional Relations** (Fig. S3c): Relationship between an animated object and any other object (something an animate object does to another object). An animate object can only be in a relation if its "activity" field is empty. All objects can only be involved in at most one relation (either interactional or positional).
- **Positional Relations** (Fig. S3d): Relationship between any two objects. An animate object can only be in a relation if its "activity" field is empty. All objects can only be involved in at most one relation (either interactional or positional).
- **Other Inputs** (Fig. S4): The location of image as background or general environment, the style of the image, and a list of keywords to append at the end of the prompt.

After participants finish with the input sections and click "generate prompt", a prompt will be generated based on all of the inputs with LLM (OpenAI gpt3.5-turbo model). Participants can freely adjust the generated prompt as the final prompt. Screenshots of an example of the inputs and

---

[1]The user IDs are up to 040 as shown in Tab. S1 because the same participant used both 027 and 028 due to some technical issues. Though, no data was collected under the user ID 027.
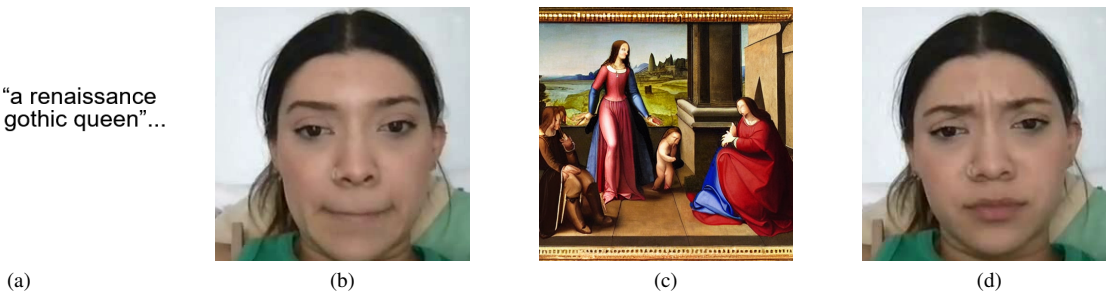
Figure S1. **An example of AU4 activation in response to a low-quality image generation.** (a) The input text prompt. (b) The facial expression of the participant before seeing the generated image. (c) The generated image. (d) The facial expression of the participant after seeing the generated image (the frame with the highest estimated AU4 intensity). In the annotation survey, the participant gave an overall rating of 2, an image-text alignment rating of 2, and a fidelity rating of 1; she identified the image with the issues "output contains unwanted content that was not mentioned in the text prompt" and "existence of body problem"; she reported that she felt "disappointed" when seeing the image; she ranked the image as the worst among the five images generated from the same prompt.
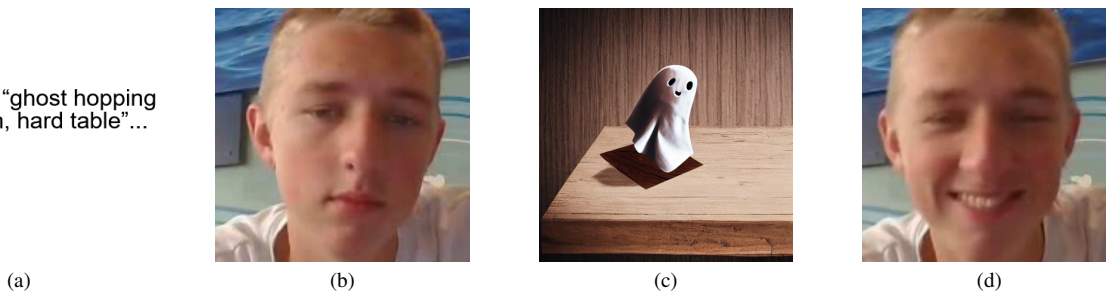

Figure S2. **An example of AU12 activation in response to a high-quality image generation.** (a) The input text prompt. (b) The facial expression of the participant before seeing the generated image. (c) The generated image. (d) The facial expression of the participant after seeing the generated image (the frame with the highest estimated AU12 intensity). In the annotation survey, the participant gave ratings of 7 for each of image-text alignment, fidelity, and overall rating; he did not identify any issues with the image; he reported feeling "satisfied", "surprised", and "amused" when seeing the image; he ranked the image as the best among the five images generated from the same prompt.

prompt generation for a structured input session are shown in Figs. S3 and S4.

B.2.2 Free-Form Input

The purpose of free-form input is to provide freedom and flexibility for participants to experiment with any prompts to complement the less flexible input style of structured input. The participants are provided with an example prompt and a text box to enter the entire prompt as the final prompt. A screenshot of an example of the prompt creation page for a free-form input session is shown in Fig. S5.

B.3 Image Presentation

Screenshots of an example of the pages for image presentation and its baseline are shown in Figs. S6 and S7.

B.3.1 Image Annotation

The annotation survey was adapted from the survey designed for ImageReward [79] and has the following components:

- Participants can optionally provide a typed, comma-separated response indicating any phrase or parts of the prompt that are not reflected by the generated image.
- Participants provide star ratings from 1 to 7 for image-text alignment, fidelity, and overall rating for the generated image.
- Participants can optionally select from a list of common issues to report the issues the generated image has.
- Participants can optionally select from a list of emotions to report their feelings when seeing the image.

A screenshot of an example of the image annotation page (including the details of the questions) is shown in Fig. S8.

User ID	Number of SI Sessions	Number of FFI Sessions	Number of Valid Images
001	10	10	100
002	5	5	50
004	9	9	87
005	9	9	85
006	9	9	90
007	9	9	88
008	9	9	85
009	9	9	88
010	9	8	84
011	9	9	87
012	8	8	80
015	9	9	89
016	10	10	100
017	3	0	15
018	11	4	75
020	10	10	92
021	9	9	87
022	9	6	75
023	10	10	99
024	10	2	58
025	3	10	63
026	10	10	100
028	0	20	99
030	10	10	99
031	10	10	100
032	10	10	99
033	10	10	99
034	10	10	98
035	10	10	99
037	10	10	100
038	10	10	96
039	10	10	96
040	9	4	65
Total	288	288	2827

Table S1. Information of the FERGI dataset. Numbers of structured input (SI) sessions, numbers of free-form input (FFI) sessions, and numbers of valid images for each participant are shown.

B.3.2 Image Ranking

After the participants finish annotating all 5 generated images, they will be asked to rank them from best to worst. All 5 images are displayed side by side, with the left side labeled "Best" and right side labeled "Worst", and the participants provide ranking by dragging the images to reorder them. A screenshot of an example of the image ranking page is shown in Fig. S9.

C   AU Model Training

This section provides more details of the training of the AU model not provided in Sec. 4.

C.1   Datasets

The DISFA dataset [53] contains facial video recordings of 27 participants' spontaneous facial expression while viewing video clips with approximately $130,000$ frames in total. Each frame is manually annotated by a human expert with intensities of AU1 (inner brow raiser), AU2 (outer brow raiser), AU4 (brow lowerer), AU5 (upper lid raiser),

AU6 (cheek raiser), AU9 (nose wrinkler), AU12 (lip corner puller), AU15 (lip corner depressor), AU17 (chin raiser), AU20 (lip stretcher), AU25 (lips part), and AU26 (jaw drop) on a scale of 0 to 5. See Fig. S10 for a visual reference guide for the analyzed AUs.

The DISFA+ dataset [51] is an extension of the DISFA dataset [53]. It contains facial video recordings of 9 participants' posed and spontaneous facial expression with each frame being annotated with the same 12 AUs on a scale of 0 to 5.

C.2 Video Preprocessing

For each frame of the video, we first detect and crop the largest face in the image using MediaPipe [48] and then align the face with Pytorch Face Landmark Detection [8]. Then, we further employ a combination of histogram equalization and linear mapping [38] to increase the global contrast of the facial image. The facial images are resized to 112×112 pixels before being fed into the neural network. Similar preprocessing is also applied to the videos we analyze in the FERGI dataset.

C.3 Model Training

We use the neural network IR-50 [13] pre-trained on Glink360k [1] and fine-tune it on the DISFA and DISFA+ datasets [51, 53]. The last layer of the network is modified so that it outputs the estimation of the AUs in two formats: for estimating the intensity of the ith AU y_i, it outputs 1 value $\hat{y}_{i,\text{reg}}$ representing the numerical estimation of the AU intensity (in the format of regression) and 5 values $\hat{y}_{i,\text{class}(1)}, \hat{y}_{i,\text{class}(2)}, \hat{y}_{i,\text{class}(3)}, \hat{y}_{i,\text{class}(4)},$ and $\hat{y}_{i,\text{class}(5)}$ respectively representing the estimated probability of the AU intensity being higher than or equal to 1, 2, 3, 4, and 5 [55] (in the format of binary classifications). The loss function consists of three parts:

$$E = E_{\text{reg,MSE}} + E_{\text{reg,cos}} + E_{\text{class}}, \quad (6)$$

where $E_{\text{reg,MSE}}$, $E_{\text{reg,cos}}$, and E_{class} respectively represent a mean squared error (MSE) loss for the numerical estimations

$$E_{\text{reg,MSE}} = \Sigma_{i=1}^n w_{i,y_i}(y_i - \hat{y}_{i,\text{reg}})^2, \quad (7)$$

a cosine similarity loss for the numerical estimations

$$E_{\text{reg,cos}} = 1 - \frac{\Sigma_{i=1}^n y_i \hat{y}_{i,\text{reg}}}{(\Sigma_{i=1}^n y_i^2)(\Sigma_{i=1}^n \hat{y}_{i,\text{reg}}^2)}, \quad (8)$$

and a cross entropy loss for the binary classification estimations

$$E_{\text{class}} = \Sigma_{i=1}^n \Sigma_{j=1}^5 w_{i,j,\chi_{y_i \geq j}} CE(\chi_{y_i \geq j}, \sigma(\hat{y}_{i,\text{class}(j)})), \quad (9)$$

with the cross entropy function being

$$CE(y, \hat{y}) = -[y_i \log(\hat{y}_i) + (1 - y_i) \log(1 - \hat{y}_i)]. \quad (10)$$

The weights for the MSE loss and those for the cross entropy loss are both inverse-frequency weighted (for the MSE loss these are based on the merged groups of $\{0, 1\}$ and $\{2, 3, 4, 5\}$). Specifically, the weights for the MSE loss are defined as

$$w_{i,j} = \begin{cases} \dfrac{2 \cdot \frac{1}{\Sigma_{j'=0}^1 n_{i,j'}}}{2 \cdot \frac{1}{\Sigma_{j'=0}^1 n_{i,j'}} + 4 \cdot \frac{1}{\Sigma_{j'=2}^5 n_{i,j'}}}, & \text{for } j = 0, 1 \\[4ex] \dfrac{4 \cdot \frac{1}{\Sigma_{j'=2}^5 n_{i,j'}}}{2 \cdot \frac{1}{\Sigma_{j'=0}^1 n_{i,j'}} + 4 \cdot \frac{1}{\Sigma_{j'=2}^5 n_{i,j'}}}, & \text{for } j = 2, 3, 4, 5 \end{cases} \quad (11)$$

while the weights for the cross entropy loss are defined as

$$\begin{cases} w_{i,j,1} = \dfrac{\frac{1}{\Sigma_{j'=j}^5 n_{i,j'}}}{\Sigma_{j''=1}^5 (\frac{1}{\Sigma_{j'=0}^{j''-1} n_{i,j'}} + \frac{1}{\Sigma_{j'=j''}^5 n_{i,j'}})} \\[4ex] w_{i,j,0} = \dfrac{\frac{1}{\Sigma_{j'=0}^{j-1} n_{i,j'}}}{\Sigma_{j''=1}^5 (\frac{1}{\Sigma_{j'=0}^{j''-1} n_{i,j'}} + \frac{1}{\Sigma_{j'=j''}^5 n_{i,j'}})}, \end{cases} \quad (12)$$

where $n_{i,j}$ represents the number occurrences of the ith AU with an intensity of j.

Notably, although we train the neural network to learn both numerical estimations and binary classification estimations of the AU intensities, only the numerical estimations are used in model inference.

For model training, we employ the Adam optimizer with an initial learning rate of 10^{-4} for parameters of the last layer and an 10^{-5} for other parameters, a weight decay of 5×10^{-4}, and a batch size of 64. We train the model on all data from the DISFA and DISFA+ datasets [51, 53] for a total of 3 epochs using a single NVIDIA GeForce GTX 1080Ti 11G GPU for about 1 hour.

D Data Filtering

This section provides the details of frame exclusion not provided in Sec. 5.1.

We aim to exclude frames with occlusions and frames with off-angle pitch or yaw. To achieve the goal, we follow the following three exclusion rules:
- Frames with a face detection confidence score (FDCS) lower than 0.9 are excluded. The FDCS is given by MediaPipe [48]. Low FDCS is likely caused by occlusions.
- Frames with a yaw indicative ratio (YIR) out of the range of $[0.3, 0.7]$. The YIR is computed as

$$\text{YIR} = \frac{d_{\text{eye-edge,left}}}{d_{\text{eye-edge,left}} + d_{\text{eye-edge,right}}}, \quad (13)$$

where $d_{\text{eye-edge,left}}$ represents the horizontal distance between the left eye and the left edge of the face while $d_{\text{eye-edge,right}}$ represents the horizontal distance between the right eye and the right edge of the face, both of which are computed based on the facial landmarks detected with [8].

- Frames with a pitch indicative ratio (PIR) out of the range of $[0.55, 0.85]$. The PIR is computed as

$$\text{PIR} = \frac{d_{\text{nostrils-eyes}}}{d_{\text{eyes}}}, \qquad (14)$$

where $d_{\text{nostrils-eyes}}$ represents the vertical distance between the center of the nostrils and the center of the eyes while d_{eyes} represents the horizontal distance between the two eyes, both of which are computed based on the facial landmarks detected with [8].

(a) Animate objects

(b) Inanimate objects

(c) Interactional relations

(d) Positional relations

Figure S3. An example of the basic inputs for the prompt creation of structured input sessions.

Other Inputs

Location
classroom at sunset

Style
anime cartoon

Keywords (separated by comma)
vivid colors, high resolution

Generate Prompt (2 attempts left)

An anime cartoon–style image depicts a classroom at sunset, with a red bird looking at a flower, a furry orange cat sleeping, and two red big books on a table.

Continue

Figure S4. An example of other inputs and prompt generation for the prompt creation of structured input sessions.

Enter Prompt

Prompts can be followed by a list of comma separated keywords for additional clarifications.

e.g. A painting of a dog in the style of Van Gogh. highly detailed, night, dreamy

Figure S5. An example of the prompt creation page for the free-form input sessions.

A realistic photo of a furry Calico Cat sleeping with a flower.

Preparing image... ↻

Figure S6. An example of a baseline page.

A realistic photo of a furry Calico Cat sleeping with a flower.

Figure S7. An example of an image presentation page.

A realistic photo of a furry Calico Cat sleeping with a flower.

Please enter phrases from the text that you think are important but not reflected in the generated image (separated by commas)

Overall Rating (1=worst, 7=best) ⓘ

★★★★★★★

Image–Text Alignment ⓘ

★★★★★★★

Fidelity (Image quality) ⓘ

★★★★★★★

Does the image has any of the following issues?
- Output contains unwanted content that was not mentioned in the text prompt
- Obvious "repeated generation" resulting in unreality
- Existence of body problem
- Too blurry to see objects
- Causes psychological discomfort
- Output contains sexual content
- Output contains violent content
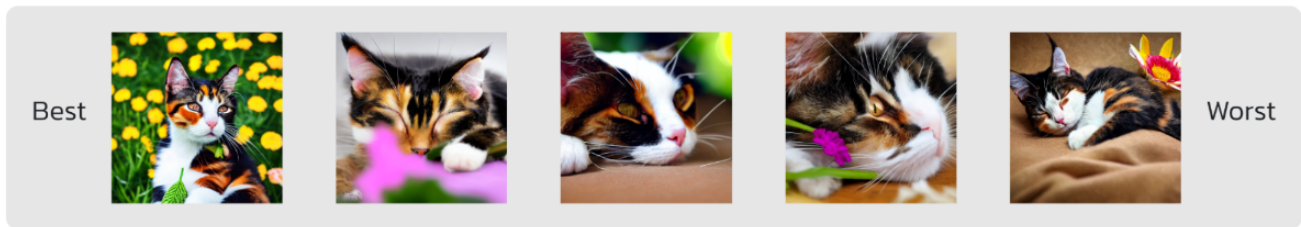- Output contains content that defames certain groups

Did you feel any of the following emotions when you saw the image?
- disappointed
- satisfied
- surprised
- disgusted
- amused
- scared

Figure S8. An example of an image annotation survey page.

Rank Images

Figure S9. An example of an image ranking page.

neutral — **AU1 - inner brow raiser** | **neutral** — **AU2 - outer brow raiser**

neutral — **AU4 - brow lowerer** | **neutral** — **AU5 - upper lid raiser**

neutral — **AU6 - cheek raiser** | **neutral** — **AU9 - nose wrinkler**

neutral — **AU12 - lip corner puller** | **neutral** — **AU15 - lip corner depressor**

neutral — **AU17 - chin raiser** | **neutral** — **AU20 - lip stretcher**

neutral — **AU25 - lips part** | **neutral** — **AU26 - jaw drop**

Figure S10. A visual reference guide for the analyzed AUs extracted from [56].