

# Synthesizing Images on Perceptual Boundaries of ANNs for Uncovering Human Perceptual Variability on Facial Expressions

Haotian Deng<sup>1,\*</sup>, Chi Zhang<sup>1,\*</sup>, Chen Wei<sup>1,2,†</sup>, Quanying Liu<sup>1,†</sup>

<sup>1</sup>*Department of Biomedical Engineering, Southern University of Science and Technology, Shenzhen, China*

<sup>2</sup>*University of Birmingham, Birmingham, United Kingdom*

{12313204, 12210315, 12150103}@mail.sustech.edu.cn; liuqy@sustech.edu.cn

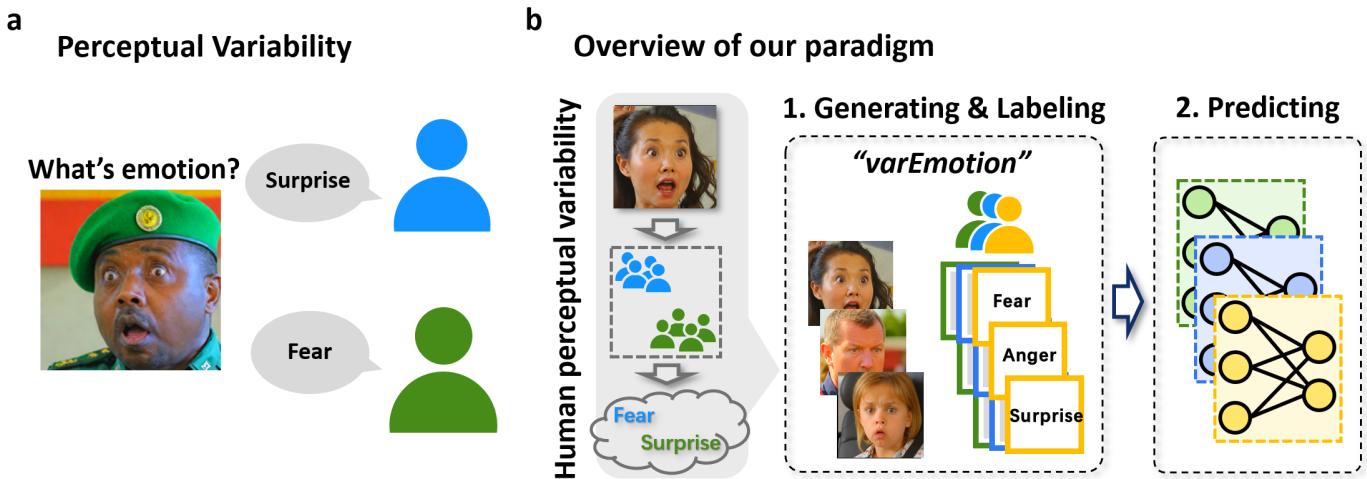


Fig. 1: **Overview of our paradigm.** (a) **Motivation:** An example of perceptual variability. (b) Our approach consists of two main components: **1. Generating & labeling:** Sampling images from ANN decision boundaries and using them in human behavioral experiments to construct the high-variability dataset varEmotion; **2. Predicting:** Finetuning models with human behavioral data to align them with human perceptual variability at the group and individual levels, enhancing behavior prediction accuracy.

**Abstract**—A fundamental challenge in affective cognitive science is to develop models that accurately capture the relationship between external emotional stimuli and human internal experiences. While ANNs have demonstrated remarkable accuracy in facial expression recognition, their ability to model inter-individual differences in human perception remains underexplored. This study investigates the phenomenon of high perceptual variability—where individuals exhibit significant differences in emotion categorization even when viewing the same stimulus. Inspired by the similarity between ANNs and human perception, we hypothesize that facial expression samples that are ambiguous for ANN classifiers also elicit divergent perceptual judgments among human observers. To examine this hypothesis, we introduce a novel perceptual boundary sampling method to generate facial expression stimuli that lie along ANN decision boundaries. These ambiguous samples form the basis of the varEmotion dataset, constructed through large-scale human behavioral experiments. Our analysis reveals that these ANN-confusing stimuli also provoke heightened perceptual uncertainty in human participants, highlighting shared computational

principles in emotion perception. Finally, by fine-tuning ANN representations using behavioral data, we achieve alignment between ANN predictions and both group-level and individual-level human perceptual patterns. Our findings establish a systematic link between ANN decision boundaries and human perceptual variability, offering new insights into personalized modeling of emotional interpretation.

**Index Terms**—Perceptual Variability, Facial Expression Recognition, Emotion Perception, Human-AI Alignment

## I. INTRODUCTION

A core goal of affective cognitive science is to develop models that accurately capture the relationship between external emotional stimuli and human internal experiences. The advancement of artificial neural networks (ANNs) has significantly contributed to this goal, particularly as their latent representations have been shown to strongly correlate with human psychological representations [1]–[7]. This study focuses on a critical phenomenon: even when exposed to the same emotional stimuli, individuals may exhibit significant differences

\* Equal contribution.

† Corresponding author.

in their internal perceptual experiences. While such perceptual variability has been widely studied in complex cognitive tasks (e.g., aesthetic or moral judgments), individual differences in simpler visual decision tasks, such as facial expression recognition, have often been overlooked. As illustrated in Figure 1(a), when different individuals observe the same stimulus, they may categorize it as different emotions (e.g., “anger” vs. “fear”). However, this *high perceptual variability* remains inadequately explored in this field, despite modern neural networks achieving remarkable accuracy in facial expression recognition [8]. Inspired by the similarity between ANNs and human perception, we hypothesize that facial expression samples that are ambiguous for ANN classifiers are also difficult for human participants to recognize. These stimuli serve as key examples that elicit divergent perceptual judgments across individuals, highlighting systematic differences in emotional interpretation.

Emerging methodologies using ANNs as perceptual probes offer promising research avenues. The discovery that imperceptible image perturbations alter both machine and human judgments [10] suggests shared computational principles in visual processing. Recent work by [11] further establishes that minimal stimulus modifications can induce perceptual conflicts across biological and artificial systems. Building on the conceptual framework of model metamers [12]—stimuli equivalent for ANNs but distinguishable by humans—we develop a novel paradigm for facial expression analysis. [13], [14] introduced *controversial stimuli*, designed to elicit divergent judgments across models, further highlighting their misalignment with human perception. Our method directly links ANN decision boundaries to human perceptual variability through facial expression stimulus generation.

Our methodological framework is composed of three components: **1. Generation of high emotional variability stimuli**: Making use of the perceptual boundary sampling approach (as described in Sec. III), we create a set of facial expressions along the decision boundaries of ANNs for six fundamental emotion categories. These stimuli retain their photorealistic authenticity due to the generative uncertainty constraints. **2. Behavioral Validation:** We select images from the perceptual boundaries of ANNs and build the *varEmotion* dataset via human behavioral experiments. This enables us to systematically record the inter-individual differences in emotional perception. **3. Individual Alignment:** We achieve alignment of the models for perceptual variability at both the group and individual levels by fine - tuning ANN models with the utilization of human behavioral data.

Our key contributions are as follows:

(1) We engineered a sophisticated algorithm tailored to sample precisely on the classification boundaries of Artificial Neural Networks (ANNs) employed in facial expression recognition. By leveraging the unique characteristics of these boundaries, this algorithm generates samples that present formidable challenges to ANNs, causing them to struggle in reaching definitive decisions. These samples are of great value as they push the ANNs to their decision - making limits,

enabling a deeper exploration of the network’s performance and robustness in facial expression recognition scenarios.

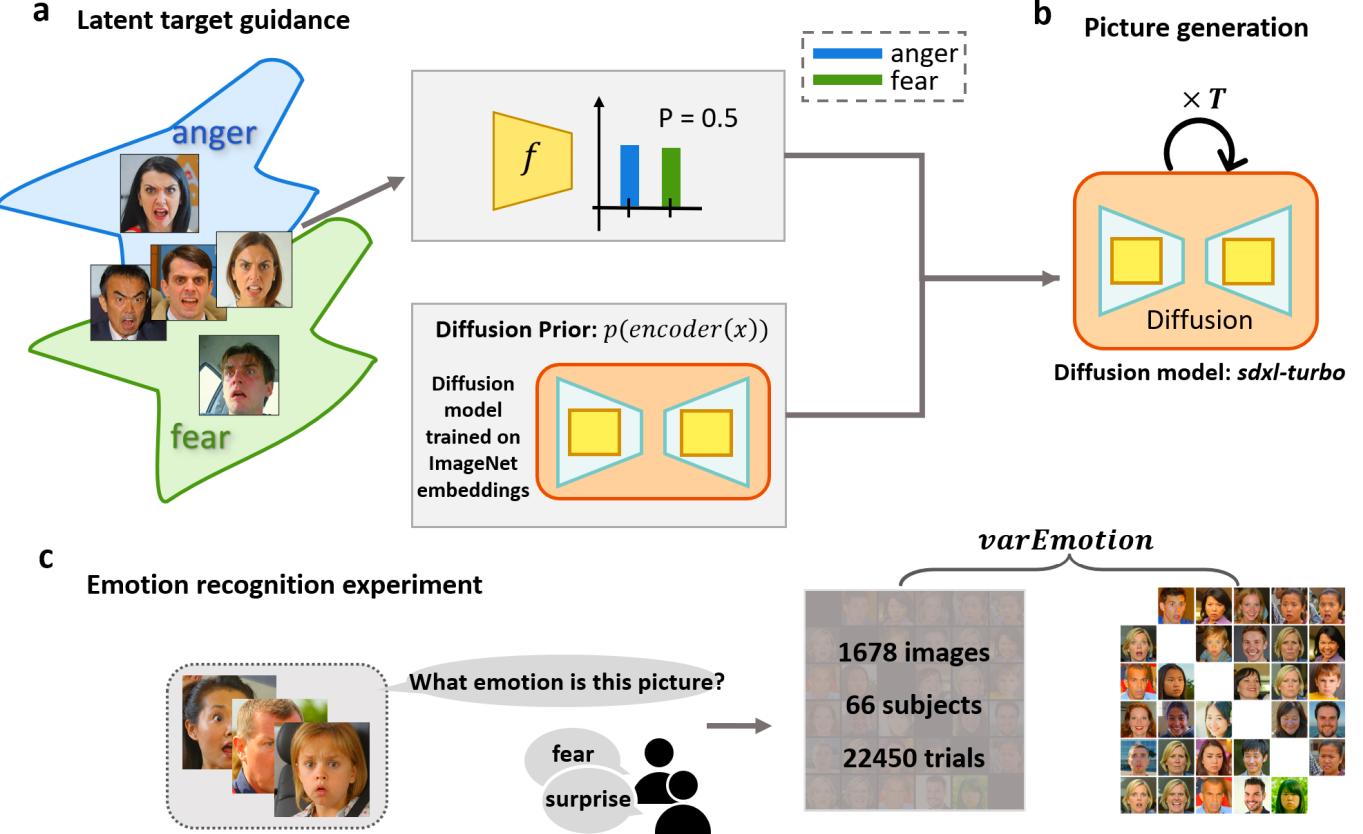
(2) Through extensive large - scale behavioral experiments, we sampled data from the classification boundaries of ANNs and utilized the results to construct the *varEmotion* dataset. A comprehensive quantitative analysis was then carried out on this dataset. The findings clearly indicate that the samples which confound ANNs also substantially heighten the decision - making uncertainty among human subjects. This connection between ANN and human decision - making difficulties provides new insights into the shared cognitive processes underlying facial expression perception.

(3) We achieved a successful alignment of ANNs with human subjects at both the group and individual levels. Our in - depth analysis reveals that individuals display distinct and significant preferences when performing facial expression recognition tasks. Remarkably, these individual - specific preferences can be effectively learned and modeled using a relatively limited number of experimental samples. This discovery paves the way for the development of more personalized and accurate facial expression recognition systems.

## II. RELATED WORKS

Researchers have widely employed ANN - generated synthetic images to explore human perceptual space. They’ve found disparities between model and human perception and refined generation techniques for greater influence on human cognition. [13], [14] used controversial stimuli to show classification differences in neural networks. [10] showed that adversarial perturbations can affect both ANN classifications and human perceptual choices, indicating shared sensitivities. But [11] noted that standard ANN perturbations don’t impact human perception, while robustified ANN models can generate low - norm perturbations that disrupt human percepts. Some studies took different approaches. Feather, Nanda et. [12], [15]–[17] studied *model metamers*, revealing mismatches between model activations and human recognition. [18] introduced DreamSim, a metric using synthetic and human experimental data to better reflect human similarity judgments and fix flaws in traditional metrics. Recent work, like [19], [20], aimed to align vision models with human perceptual representations by adding human - like concepts, improving alignment and performance. For studying human perceptual variability, generated images must strongly influence human cognition. Since samples from ANN perceptual boundaries are often noisy, better methods are needed for natural - looking images. Machine - learning studies on adversarial examples and counterfactual explanations, such as [21], [2], [22], [23], [24], and [25], use diffusion models with training - free guidance [26]–[28] as regularizers. This helps introduce prior distributions, enhancing image naturalness and their impact on human perception.

In the field of psychology, researchers have been searching for the most basic facial expressions. Many studies ([29], [30], [31], [32], [33]) suggest that there are six universal



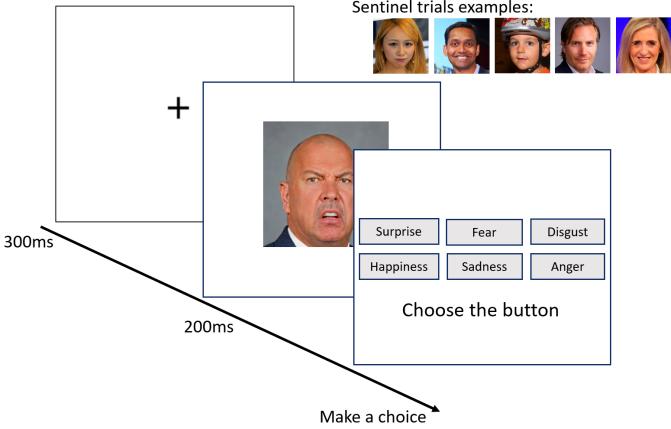
**Fig. 2: Generating images to elicit human perceptual variability.** (a) This example demonstrates how to generate embeddings by sampling from the perceptual boundaries of the expressions ‘anger’ and ‘fear’ in an ANN using the uncertainty guidance method. The goal of uncertainty guidance is to focus the ANN’s prediction of the embeddings on ‘anger’ and ‘fear’. The diffusion model follows the first stage generation process from CoCoG [1], taking the original image embeddings as a prior and guiding the denoising process toward the desired target direction. The ANN model used for guidance is an MLP pre-trained on image embeddings from the RAF-DB dataset, with an output of six channels representing the emotions [‘surprise’, ‘fear’, ‘disgust’, ‘happiness’, ‘sadness’, ‘anger’]. (b) Two-stage image generation. We reference the two-stage generation method from DALL-E 2 [9], using the sdxl-turbo model without a classifier to generate images from the embeddings generated in the first stage. Only prompt guidance is applied during the second-stage generation. (c) Using the methods described above, we generated a series of images capable of inducing uncertainty in the ANN and used these images in a human experiment. In the experiment, human participants are shown the generated images and asked to choose the emotion of the face depicted in each image. A total of 1678 images were used, with 22450 trials conducted across 66 participants, resulting in the high perceptual variability dataset *varEmotion*.

basic expressions, namely ‘surprise’, ‘fear’, ‘disgust’, ‘happiness’, ‘sadness’, ‘anger’. However, Snoek [34], Cordaro [30], and others have pointed out that individuals differ in their perception and recognition of these expressions, and that these differences are related to cultural factors. Therefore, we propose that different individuals have different perceptual boundaries when recognizing facial expressions, and that ANNs also have specific perceptual boundaries for facial expression recognition. In the next phase of the study, we will sample the perceptual boundaries of expressions in ANNs and generate a series of facial images with uncertain expressions using a diffusion model. These images will be used in human experiments, where participants will identify the emotions

corresponding to these images. Based on the feedback data from the participants, we will fine-tune the ANN to enable it to more accurately fit the human perceptual boundaries.

### III. METHOD

In this section, we introduce a method for generating pictures of human facial expression with high perceptual variability. We also conduct human experiment online and constructed the *varEmotion* dataset: a facial dataset with high perceptual variability. Our goal is to generate images that evoke significant human perceptual variability and collect this variability by recording human perceptual judgments on the generated images.



**Fig. 3: Human facial expression recognition experiment procedure.** In each round of the experiment, the participant will first see a cross at the center of the screen for 300 ms. Following this, a facial image containing a specific expression will be presented for 200 ms. Next, a choice page with six buttons will appear, and the participant is required to judge the expression of the face in the image just shown and select the corresponding button. After the participant clicks a button, the current trial ends, and the next trial begins. Each participant will complete a total of 400 trials, which include 390 random trials and 10 sentinel trials.

#### A. Generating Images on ANN perceptual boundary

Many existing studies indicate that images which significantly affect the perception of artificial neural networks (ANNs) can also influence human perception ([11], [3], [10], [1]). This suggests that ANNs may share a similar perceptual space with humans. Based on this, we propose that stimuli generated by sampling the perceptual boundaries of ANNs could similarly induce perceptual variability in humans, leading to differences in how individuals perceive and judge the stimuli.

#### B. Facial expression recognition experiment

Inspired by the two-stage generation method proposed in DALL-E 2 [9], where image embeddings are first generated and then used to create the image, we also adopt a two-stage approach for image generation. In the first stage, we apply a diffusion model to add noise and remove noise from the input image embeddings, simultaneously implementing guidance in this process: uncertainty guidance. The goal of uncertainty guidance is to sample from the perceptual boundaries of the classifier. The loss function for uncertainty is as follows:

$$\text{loss}(x, y) = -p(y|x) * q(y) \quad (1)$$

where  $p(y)$  is the classifier's predicted distribution, and  $q(y)$  is the guiding target distribution. This loss function is inspired by GANs [35] and aims to maximize the probability of the target distribution (e.g., ‘fear’ and ‘happiness’) while minimizing the probability of non-target distributions, thereby generating controllable high-uncertainty images.

In previous studies, researchers attempting to use synthetic images to investigate human and model cognition often encountered the issue of unnatural or unrealistic generated images ([11], [10], [13], [12]). This made it difficult for participants to recognize the images, severely impacting the effectiveness of human experiments. Recent research has shown that using diffusion models as regularizers can introduce natural image priors during the generation process ([21], [36], [22], [37], [25]), making the images significantly more natural and realistic, which in turn helps evoke the intrinsic variability in human perception. Based on this, we employed a two-stage diffusion process to generate the final images. This approach ensures that the generated images better align with the real distribution of natural images, effectively enhancing their impact on human perception. The function for sampling process with diffusion model in first stage is as follows:

$$x_{t-1} = DDPM^-(x_t) - \gamma \nabla_{x_t} \text{loss}(x_t, y) \quad (2)$$

where  $DDPM^-(x_t)$  represents the reverse diffusion step,  $\text{loss}(x_t, y)$  is the uncertainty loss, and  $\gamma$  is the hyperparameter of the guidance strength. In our experiment, we chose U-ViT [38] as the diffusion model in the first stage generation,  $\gamma = 0.5$ , and stable diffusion XL in the second stage generation.

#### C. Filtering Generated Images

Since we applied guidance only during the first-stage diffusion process, the images generated in the second stage exhibit considerable randomness. To ensure that the generated images align with the expected distribution, we filtered the images based on specific criteria. The filtering criterion is:

$$p_{\text{emotion1}} > k_{\text{emotion1}} \& p_{\text{emotion2}} > k_{\text{emotion2}} \quad (3)$$

where  $p_{\text{emotion1}}$  and  $p_{\text{emotion2}}$  are the activation values of the two emotions (emotion1, emotion2) predicted by the ANN for the generated image, and  $k_{\text{emotion1}}$  and  $k_{\text{emotion2}}$  are the 75th percentiles of the activation values for the corresponding emotions (emotion1, emotion2) across all images in the RAF-DB dataset. Through this filtering process, we ensure that the generated images effectively induce perceptual variability in the ANN. Activation values distribution of RAF-DB dataset can be found in Figure 8.

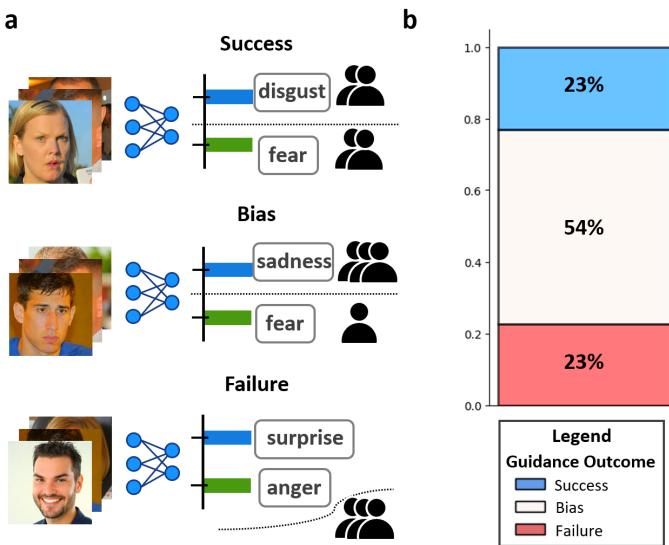
## IV. COLLECTING HUMAN PERCEPTUAL VARIABILITY

We used the filtered synthetic images as experimental materials, with the aim of collecting human participants' choices regarding these images. The experiment was approved by the local university's ethics committee before its commencement. The experiment was implemented using jsPsych and conducted online through the NAODAO platform, with a total of 100 participants. Prior to participating in the experiment, each participant read an informed consent form detailing the potential risks. Participants were free to withdraw from the experiment at any time, and no personal information was collected. During data processing, we retained feedback data

from 66 participants with sentinel trial accuracy greater than 70%, resulting in a total of 1,678 images and 22,450 trials. We used the retained participant data to construct a facial expression dataset with high perceptual variability, named varEmotion.

**Evaluation metrics.** To effectively evaluate the efficacy of guidance in generative methods, we propose three types of guidance outcomes, as illustrated in Figure 4(a): *success*, *bias*, and *failure*. For the guiding targets *emotion1* and *emotion2*, we define  $p_1$  and  $p_2$  to represent the probabilities of *emotion1* and *emotion2*, respectively. If the result is  $\min(p_1, p_2) > 0.25$  and  $p_1 + p_2 > 0.6$ , it indicates that the generated stimulus leads individuals to select the guiding targets evenly, and we classify the guidance as *success*. If the result is  $\min(p_1, p_2) < 0.25$  and  $p_1 + p_2 > 0.6$ , it suggests that all subjects tend to choose a specific target, and we classify the guidance as *bias*. If  $p_1 + p_2 < 0.6$ , it indicates that the stimulus does not effectively influence the subjects' choices, and we classify the guidance as *failure*.

#### A. Quantitative Analysis of varEmotion



**Fig. 4: Quantitative Analysis of varEmotion.** (a) Examples of three guidance outcome:*success*, *bias*, *failure*. (b) Guidance outcome across the varEmotion dataset. The sum of **success** and **bias** rates approaches 80% .

**ANN variability can arouse human variability.** To examine whether the images generated by the ANN perception boundary sampling effectively evoke human subjects' perception variability, we calculated the entropy of the probability distribution of subjects' choices for emotion images and plotted the corresponding entropy distribution in Figure 7. It is evident that the entropy for the vast majority of images is greater than 0, indicating that these images effectively triggered varying responses among different subjects. Furthermore, as shown in Figure 4(b), nearly 80% of all generated images fall under the categories of *success* or *bias*. This indicates that, in the

majority of cases, human choices aligned with either both or one of the guidance targets. This demonstrates that the generation method effectively guided human facial expression and emotion recognition behavior.

## V. PREDICTING HUMAN PERCEPTUAL VARIABILITY

### A. Model Finetuning For Human Alignment

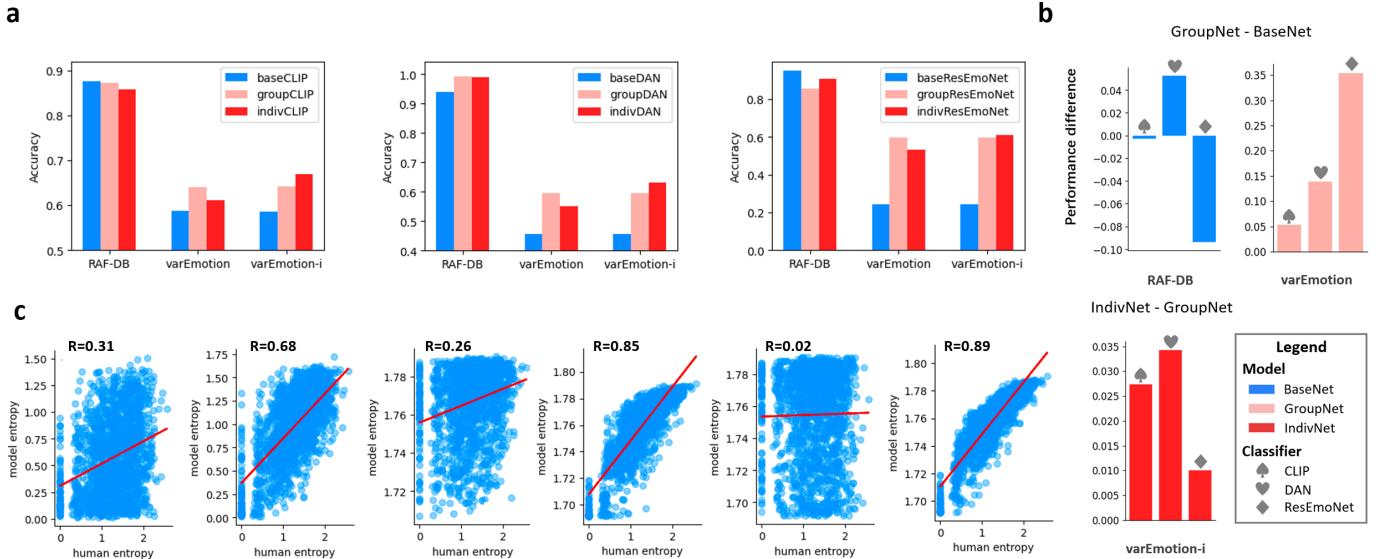
To fine-tune CLIP for the emotion classification task, we added an MLP head at the end of CLIP and aligned it with human perception through fine-tuning the MLP head. To align models with both group-level and individual-level performance, we adopted a mixed training approach with an 4:1 split for training and validation. For individual-level datasets (varEmotion-i), the validation set was designed to avoid overlap with the group validation set. For group-level training, we combined the varEmotion and RAF-DB datasets in a 2:1 ratio, ensuring performance on RAF-DB while fine-tuning for perceptual variability. For individual-level training, we mixed varEmotion-i, varEmotion datasets in a 2:1 ratio, ensuring the models performed effectively on individual-specific and group datasets.

For group-level fine-tuning, the original classifier models were trained on mixed RAF-DB, varEmotion datasets. The pictures were normalized with 'ToTensor' transformation. For Individual-level fine-tuning, the initial model is the group model. Training and testing sets were loaded with a batch size of 128, and the models were implemented with 3 different configurations to map input images to 6 output classes. Training was performed on NVIDIA GPU using Adam optimizer ( $lr = 1 \times 10^{-4}$ ) for 15 epochs, and CrossEntropyLoss function was used to compute the classification loss.

### B. Alignment Analysis

**Fine-tuning improves both group-level and individual-level prediction performance.** As illustrated in Figure 5a, on the RAF-DB dataset, the performance of BaseNet, GroupNet, and IndivNet is comparable, suggesting that fine-tuning at both the group and individual levels did not result in a significant decrease in prediction accuracy for this dataset. In contrast, on the varEmotion dataset, both GroupNet and IndivNet show improved prediction accuracy compared to BaseNet, indicating that fine-tuning at both levels effectively enhanced the model's ability to align with human perceptual boundaries. Moreover, on the individual dataset varEmotion-i, IndivNet, which is fine-tuned using individual data, demonstrates an average prediction accuracy improvement of 3% over the group model GroupNet. This highlights the model's capacity to effectively capture individual perceptual differences and better align with individual perceptual boundaries. We anticipate that in future research, IndivNet, which models individual perception effectively, will play a pivotal role in advancing our understanding of individual perception and behavior regulation.

**Different classifiers exhibit inconsistent performance.** Figure 5(b) compares the changes in predictive performance of various models on the RAF-DB, varEmotion, and varEmotion-i datasets after fine-tuning at both the group and individual



**Fig. 5: Human alignment results.** (a) Accuracy of BaseNet, GroupNet and IndivNet on RAF-DB, varEmotion and varEmotion-i. On varEmotion, GroupCLIP and IndivCLIP improve 5% over baseNet, GroupDAN and IndivDAN improve 14%, GroupResEmoNet and IndivResEmoNet improve 35%. On varEmotion-i, IndivCLIP outperform 2.5% over GroupCLIP, IndivDAN outperform 3.5% over GroupDAN, and IndivResEmoNet outperform 1% over GroupResEmoNet. After fine-tuning different models at the individual and group levels, performance differences were observed, indicating that the model architecture is related to the model’s ability to fit the boundaries of human perception. (b) Finetuning result for 3 classifiers. On varEmotion, all classifiers improved, with ResEmoNet showing the largest gains and CLIP the smallest. Individual fine-tuning further improved all classifiers with the same trend. All classifiers, after fine-tuning, were able to better predict human behavior, suggesting that fine-tuning with human data effectively enables the model to align with human perception. (c) For DAN, Spearman rank correlation between model and human entropy increased from  $\rho = 0.26$  to  $\rho = 0.85$  after group fine-tuning. After fine-tuning DAN at the group level, the entropy distribution of the predicted images became closer to that of human predictions, indicating that the model has captured the uncertainty in human perception.

levels. On the varEmotion dataset, the prediction accuracy of all models shows significant improvement after fine-tuning, though the degree of improvement varies across models. ResEmoNet achieves the highest accuracy gain, while CLIP shows the smallest. On the varEmotion-i dataset, accuracy continues to improve post-fine-tuning, with DAN exhibiting the greatest improvement and ResEmoNet the least. These discrepancies in accuracy changes suggest that differences may exist in the perceptual boundaries at the group, individual, and model levels, which contribute to variations in model fitting performance across different architectures.

**Human variability can be predicted by models.** To assess the alignment between model and human perceptual variability, we analyzed the correlation between model and human entropy, as shown in Figure 5c. Taking DAN as an example, group fine-tuning increases the Spearman rank correlation between model and human entropy from  $\rho = 0.26$  to  $\rho = 0.85$ . This significant improvement indicates that fine-tuning allows the model to better capture human uncertainty, aligning model predictions more closely with human perceptual behavior.

## VI. CONCLUSION

This study demonstrates that ANN decision boundaries serve as meaningful indicators of inter-individual perceptual

variability in facial expression recognition. By leveraging a perceptual boundary sampling approach, we systematically generate stimuli that challenge both ANN classifiers and human observers, revealing a strong correspondence between machine and human perceptual uncertainty. The varEmotion dataset, constructed from large-scale behavioral experiments, provides empirical evidence that ambiguous ANN samples also evoke divergent interpretations among individuals, reinforcing the hypothesis that ANN-confusing stimuli capture key dimensions of human perceptual variability. Beyond dataset generation, our findings underscore the feasibility of aligning ANN representations with individual-level human perceptual patterns. Through fine-tuning on behavioral data, we successfully adapt ANN models to account for subject-specific differences, highlighting the potential for more personalized affective computing systems. This research paves the way for future studies on human-machine alignment in emotion perception, suggesting that ANN decision boundaries can serve as a valuable tool for studying perceptual variability and enhancing adaptive AI-driven emotion recognition.

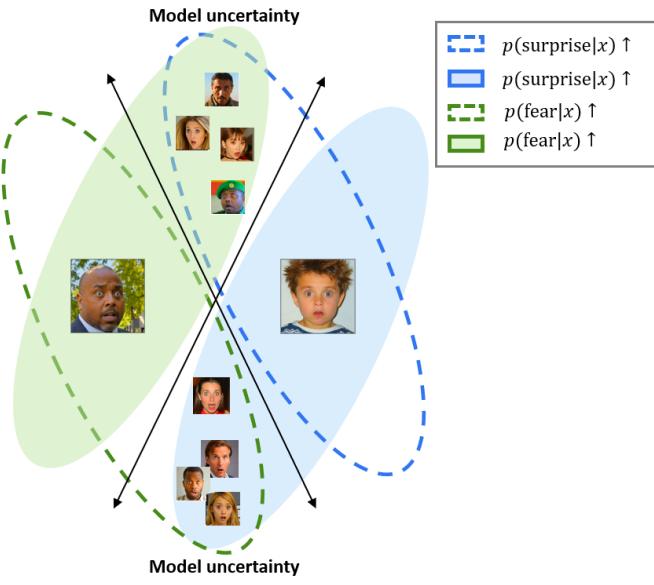
## ACKNOWLEDGMENT

This work was supported by the National Natural Science Foundation of China (62472206), Shenzhen Science

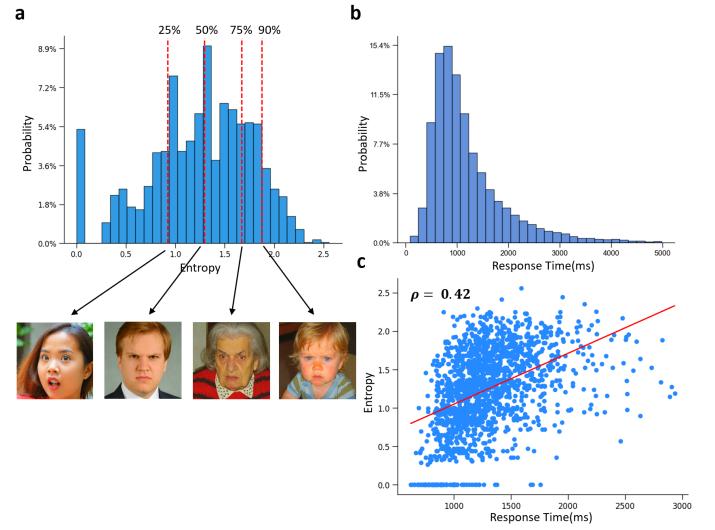
and Technology Innovation Committee (2022410129, KJZD20230923115221044, KCXFZ20201221173400001), GuangDong Basic and Applied Basic Research Foundation (2025A1515011645 to Z.C.L.), Shenzhen Doctoral Startup Project (RCBS20231211090748082 to X.K.S.), Guangdong Provincial Key Laboratory of Advanced Biomaterials (2022B1212010003), and the Center for Computational Science and Engineering at Southern University of Science and Technology.

## APPENDIX

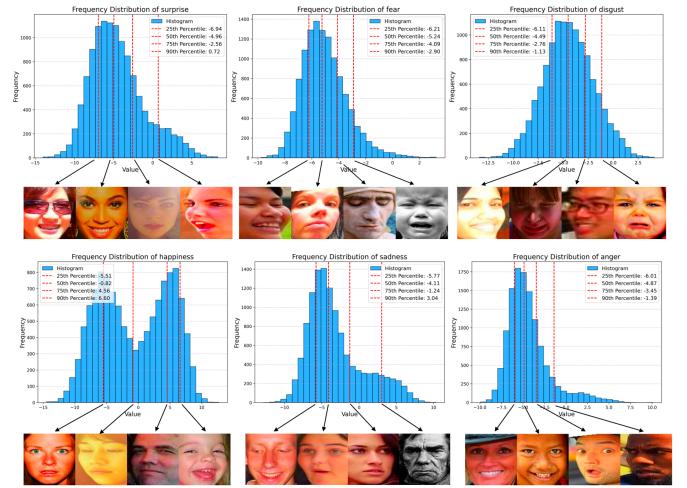
In the appendix, we present supplementary figures that complement the experimental results and data analysis in the main text. Figure 6 illustrates a schematic diagram of perceptual boundary sampling, demonstrating how uncertainty guidance is employed to sample at the ANN perceptual boundary. Figure 7 provides a detailed analysis of human behavioral data, including the entropy distribution of human judgments on images, the distribution of human reaction times, and the correlation between entropy and reaction time. Figure 8 presents an analysis of images from the RAF-DB dataset, showing the distribution of activation values across six emotional dimensions for all images in the dataset.



**Fig. 6: Sampling on perceptual boundaries.** The perceptual space of ANN can be divided into four regions based on two classification axes. Taking the emotion pairs (fear, surprise) as an example. our objective is to generate images that induce uncertainty of ANN, as illustrated in the figure. The images in the upper and lower regions can lead the ANN to predict high probabilities for both fear and surprise. In the left region, the ANN predicts a high probability for fear but a low probability for surprise. Conversely, in the right region, the ANN predicts a high probability for surprise but a low probability for fear.



**Fig. 7: Behavioral results of the Digit recognition task.** (a) The entropy distribution of human judgments on images primarily concentrates between 0.5 and 2.0, approximately following a normal distribution. (b) The time humans take to judge the images is concentrated between 500 and 1500 ms, and it exhibits the characteristics of a heavy-tailed distribution. (c) Entropy and response time exhibit a positive correlation, with a Spearman rank correlation coefficient of 0.42.



**Fig. 8: Distribution of activation values varies across different emotions.** The distribution of emotion activation values in images from the RAF-DB dataset across various emotional dimensions. It can be observed that there are certain differences in the distribution of activation values for different emotions. The activation value distributions for the majority of emotions resemble a normal distribution, while the distribution for ‘happiness’ is distinctly bimodal.

## REFERENCES

- [1] C. Wei, J. Zou, D. Heinke, and Q. Liu, "Cocog: Controllable visual stimuli generation based on human concept representations," *arXiv preprint arXiv:2404.16482*, 2024.
- [2] ———, "Cocog-2: Controllable generation of visual stimuli for understanding human concept representation," *arXiv preprint arXiv:2407.14949*, 2024.
- [3] L. Muttenhaler, J. Dippel, L. Linhardt, R. A. Vandermeulen, and S. Kornblith, "Human alignment of neural network representations," *arXiv preprint arXiv:2211.01201*, 2022.
- [4] F. P. Mahner, L. Muttenhaler, U. Güçlü, and M. N. Hebart, "Dimensions underlying the representational alignment of deep neural networks with humans," *arXiv preprint arXiv:2406.19087*, 2024.
- [5] C. Y. Zheng, F. Pereira, C. I. Baker, and M. N. Hebart, "Revealing interpretable object representations from human behavior," *arXiv preprint arXiv:1901.02915*, 2019.
- [6] M. N. Hebart, C. Y. Zheng, F. Pereira, and C. I. Baker, "Revealing the multidimensional mental representations of natural objects underlying human similarity judgements," *Nature human behaviour*, vol. 4, no. 11, pp. 1173–1185, 2020.
- [7] L. Muttenhaler, C. Y. Zheng, P. McClure, R. A. Vandermeulen, M. N. Hebart, and F. Pereira, "Vice: Variational interpretable concept embeddings," *Advances in Neural Information Processing Systems*, vol. 35, pp. 33 661–33 675, 2022.
- [8] M. H. Goodwin and H. Loyd, "The face of noncompliance in family interaction," *Text & Talk*, vol. 40, no. 5, pp. 573–598, 2020.
- [9] A. Ramesh, P. Dhariwal, A. Nichol, C. Chu, and M. Chen, "Hierarchical text-conditional image generation with clip latents," *arXiv preprint arXiv:2204.06125*, vol. 1, no. 2, p. 3, 2022.
- [10] V. Veerabadran, J. Goldman, S. Shankar, B. Cheung, N. Papernot, A. Kurakin, I. Goodfellow, J. Shlens, J. Sohl-Dickstein, M. C. Mozer *et al.*, "Subtle adversarial image manipulations influence both human and machine perception," *Nature Communications*, vol. 14, no. 1, p. 4933, 2023.
- [11] G. Gaziv, M. Lee, and J. J. DiCarlo, "Strong and precise modulation of human percepts via robustified anns," *Advances in Neural Information Processing Systems*, vol. 36, 2024.
- [12] J. Feather, G. Leclerc, A. Madry, and J. H. McDermott, "Model metamers reveal divergent invariances between biological and artificial neural networks," *Nature Neuroscience*, vol. 26, no. 11, pp. 2017–2034, 2023.
- [13] T. Golan, P. C. Raju, and N. Kriegeskorte, "Controversial stimuli: Pitting neural networks against each other as models of human cognition," *Proceedings of the National Academy of Sciences*, vol. 117, no. 47, pp. 29 330–29 337, 2020.
- [14] T. Golan, M. Siegelman, N. Kriegeskorte, and C. Baldassano, "Testing the limits of natural language models for predicting human language judgements," *Nature Machine Intelligence*, vol. 5, no. 9, pp. 952–964, 2023.
- [15] J. Feather, A. Durango, R. Gonzalez, and J. McDermott, "Metamers of neural networks reveal divergence from human perceptual systems," *Advances in Neural Information Processing Systems*, vol. 32, 2019.
- [16] V. Nanda, T. Speicher, C. Kolling, J. P. Dickerson, K. Gummadi, and A. Weller, "Measuring representational robustness of neural networks through shared invariances," in *International Conference on Machine Learning*. PMLR, 2022, pp. 16 368–16 382.
- [17] V. Nanda, A. Majumdar, C. Kolling, J. P. Dickerson, K. P. Gummadi, B. C. Love, and A. Weller, "Do invariances in deep neural networks align with human perception?" in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 37, no. 8, 2023, pp. 9277–9285.
- [18] S. Fu, N. Tamir, S. Sundaram, L. Chai, R. Zhang, T. Dekel, and P. Isola, "Dreamsim: Learning new dimensions of human visual similarity using synthetic data," *arXiv preprint arXiv:2306.09344*, 2023.
- [19] L. Muttenhaler, K. Greff, F. Born, B. Spitzer, S. Kornblith, M. C. Mozer, K.-R. Müller, T. Unterthiner, and A. K. Lampinen, "Aligning machine and human visual representations across abstraction levels," *arXiv preprint arXiv:2409.06509*, 2024.
- [20] S. Sundaram, S. Fu, L. Muttenhaler, N. Y. Tamir, L. Chai, S. Kornblith, T. Darrell, and P. Isola, "When does perceptual alignment benefit vision representations?" *arXiv preprint arXiv:2410.10817*, 2024.
- [21] G. Jeanneret, L. Simon, and F. Jurie, "Adversarial counterfactual visual explanations," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 16 425–16 435.
- [22] X. Chen, X. Gao, J. Zhao, K. Ye, and C.-Z. Xu, "Advdifuser: Natural adversarial example synthesis with diffusion models," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 4562–4572.
- [23] G. Jeanneret, L. Simon, and F. Jurie, "Diffusion models for counterfactual explanations," in *Proceedings of the Asian Conference on Computer Vision*, 2022, pp. 858–876.
- [24] P. Vaeth, A. M. Frühwald, B. Paassen, and M. Gregorova, "Diffusion-based visual counterfactual explanations—towards systematic quantitative evaluation," *arXiv preprint arXiv:2308.06100*, 2023.
- [25] H. Atakan Bedel and T. Çukur, "Dreamr: Diffusion-driven counterfactual explanation for functional mri," *arXiv e-prints*, pp. arXiv–2307, 2023.
- [26] J. Yu, Y. Wang, C. Zhao, B. Ghanem, and J. Zhang, "Freedom: Training-free energy-guided conditional diffusion model," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 23 174–23 184.
- [27] J. Ma, T. Hu, W. Wang, and J. Sun, "Elucidating the design space of classifier-guided diffusion generation," *arXiv preprint arXiv:2310.11311*, 2023.
- [28] L. Yang, S. Ding, Y. Cai, J. Yu, J. Wang, and Y. Shi, "Guidance with spherical gaussian constraint for conditional diffusion," *arXiv preprint arXiv:2402.03201*, 2024.
- [29] P. Ekman and W. V. Friesen, "Facial action coding system," *Environmental Psychology & Nonverbal Behavior*, 1978.
- [30] D. T. Cordaro, R. Sun, D. Keltner, S. Kamble, N. Huddar, and G. McNeil, "Universals and cultural variations in 22 emotional expressions across five cultures." *Emotion*, vol. 18, no. 1, p. 75, 2018.
- [31] D. Keltner, D. Sauter, J. Tracy, and A. Cowen, "Emotional expression: Advances in basic emotion theory," *Journal of nonverbal behavior*, vol. 43, pp. 133–160, 2019.
- [32] D. Matsumoto, D. Keltner, M. N. Shiota, M. O'Sullivan, and M. Frank, "Facial expressions of emotion," *Handbook of emotions*, vol. 3, pp. 211–234, 2008.
- [33] R. E. Jack, O. G. Garrod, and P. G. Schyns, "Dynamic facial expressions of emotion transmit an evolving hierarchy of signals over time," *Current biology*, vol. 24, no. 2, pp. 187–192, 2014.
- [34] L. Snoek, R. E. Jack, P. G. Schyns, O. G. Garrod, M. Mittenbühler, C. Chen, S. Oosterwijk, and H. S. Scholte, "Testing, explaining, and exploring models of facial expressions of emotions," *Science advances*, vol. 9, no. 6, p. eabq8421, 2023.
- [35] I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial networks," 2014. [Online]. Available: <https://arxiv.org/abs/1406.2661>
- [36] C. Wei, J. Zou, D. Heinke, and Q. Liu, "Cocog-2: Controllable generation of visual stimuli for understanding human concept representation," 2024. [Online]. Available: <https://arxiv.org/abs/2407.14949>
- [37] P. Väth, A. M. Frühwald, B. Paassen, and M. Gregorova, "Diffusion-based visual counterfactual explanations—towards systematic quantitative evaluation," in *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*. Springer, 2023, pp. 120–135.
- [38] F. Bao, S. Nie, K. Xue, Y. Cao, C. Li, H. Su, and J. Zhu, "All are worth words: A vit backbone for diffusion models," 2023. [Online]. Available: <https://arxiv.org/abs/2209.12152>