1. The computational process of the interactive interaction flow

Human ratings are represented as acceptable intervals whose acceptability level follows a normal distribution relative to the deviation from the mean score. We introduce a *critical probability* $p(f)$ and a threshold $\mathcal{E}$: if the probability that a score is acceptable exceeds $\mathcal{E}$, the externally assigned score is considered accepted by users. Let the human-expected mean for item i be $\mu_i = \bar{f}_i^h$, and define the human-acceptable interval $[\,\bar{f}_i^h - \delta f_i^h,\ \bar{f}_i^h + \delta f_i^h\,]$. Based on the central-limit property of the normal distribution, an external (AI) score f_i^{AI} is regarded as acceptable if it falls in this interval, i.e.,

$$p\big(\bar{f}_i^h - \delta f_i^h < f_i^{AI} < \bar{f}_i^h + \delta f_i^h\big) \;=\; \mathcal{E}. \tag{1}$$

To align AI-generated scores with human acceptability while preserving personalization, we aim to efficiently converge to an AI score inside the human-accepted range using gradient descent.

Initially, the AI draws a score $f_{i,0}^{AI}$ from a candidate set $\mathcal{S}_i^{AI}$, while the human-expected mean is $\bar{f}_i^h$. If $f_{i,0}^{AI} \in [\bar{f}_i^h - \delta f_i^h, \bar{f}_i^h + \delta f_i^h]$, we take $f_{i,0}^{AI}$ as the final score. Otherwise, we iteratively adjust the score to bring it into the acceptable interval while minimizing the distance $\Delta f_i = f_i^{AI} - \bar{f}_i^h$ (see Figure S3).

The loss at iteration k is defined as

$$J(k) \;=\; \sum_i \big(f_{i,k}^{AI} - \bar{f}_i^h\big)^2. \tag{2}$$

Its partial derivative with respect to the i-th score is

$$\frac{\partial J(k)}{\partial f_{i,k}^{AI}} \;=\; 2\big(f_{i,k}^{AI} - \bar{f}_i^h\big). \tag{3}$$

With learning rate $\theta > 0$, the gradient-descent update becomes

$$f_{i,k+1}^{AI} \;=\; f_{i,k}^{AI} \;-\; \theta\,\big(f_{i,k}^{AI} - \bar{f}_i^h\big). \tag{4}$$

The procedure terminates when the AI score enters the human-acceptable interval:

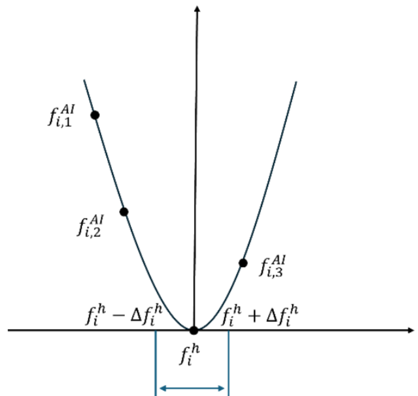
$$f_{i,k^\star}^{AI} \in \big[\,\bar{f}_i^h - \delta f_i^h,\ \bar{f}_i^h + \delta f_i^h\,\big]\,, \quad \text{and } k^\star \text{ is minimal.} \tag{5}$$

A pseudocode of the algorithm is provided in below:

Algorithm 1: Gradient Descent with Acceptable Range Check for AI Scoring

Input: Human ratings f_i^h, tolerance δf_i^h, learning rate θ, max iterations $K_{\max}$
Output: Final AI ratings $f_{i,f}^{AI}$

```
for i ← 1 to n do
    Randomly initialize f_{i,0}^{AI};
    if f_i^h − δf_i^h < f_{i,0}^{AI} < f_i^h + δf_i^h then
        f_{i,f}^{AI} ← f_{i,0}^{AI};
    else
        k ← 0;
        repeat
            e_i ← f_{i,k}^{AI} − f_i^h;
            f_{i,k+1}^{AI} ← f_{i,k}^{AI} − θ · e_i;
            k ← k + 1;
        until f_i^h − δf_i^h < f_{i,k}^{AI} < f_i^h + δf_i^h or k ≥ K_max;
        f_{i,f}^{AI} ← f_{i,k}^{AI};
return f_{i,f}^{AI} for all i
```



Pseudo-code and schematic for gradient descent computation

2. Confidence score explanation description

In our experimental design, uncertainty-based explanations serve as the baseline. Because this study involves *regression* prediction of aesthetic scores rather than a classification task, confidence is estimated using residual-based statistical methods. The residual is defined as the difference between the model's predicted value P_{pred} and the ground-truth label P_t; a smaller residual indicates higher confidence:

$$\text{Uncertainty} \;=\; \left|\, P_{\text{pred}} - P_t \,\right|. \tag{6}$$

The statistical properties of the residual distribution can be used to construct confidence intervals or to compute confidence scores for predictions on new samples. Here, we adopt a confidence-scoring approach, where confidence is defined as the probability that the prediction error falls within a specified range. Specifically, R denotes the maximum possible score range (in this study, $R = 5$). The confidence score is

$$\text{Confidence} \;=\; 1 - \frac{\left|\, P_{\text{pred}} - P_t \,\right|}{R}. \tag{7}$$

3. Expressed mathematic of advice-taking rate

$$\textbf{Advice taking rate} = \frac{\text{judge final estimate} - \text{judge initial estimate}}{\text{advisor recommendation} - \text{judge initial estimate}}. \tag{8}$$