Order Learning with Large Multimodal Models for Facial Beauty Prediction

基於大型多模態模型的順序式學習於顏值預測之應用

Group: B578 **Members：蔡信彰、林進成** **Supervisor：翁詠祿 教授**

Abstract

Facial beauty prediction (FBP) on datasets such as SCUT-FBP5500 is increasingly addressed using order-learning frameworks that model pairwise preferences and infer continuous scores via the Bradley–Terry (BT) model. While conventional neural networks directly regress a score from a single image (Fig. 1), large multimodal model (LMM)–based methods such as LMOL instead predict pairwise comparisons and aggregate them through BT-style inference (Fig. 2), achieving strong cross-dataset generalization. However, many implementation details and intermediate behaviors of these pipelines remain insufficiently examined.

In this project, Order Learning with Large Multimodal Models for Facial Beauty Prediction, we re-implemented an LMOL-style pipeline on SCUT-FBP5500 with fully specified training hyperparameters and BT estimation procedures. We analyze the LMM comparator not only through score metrics (PC, MAE, RMSE) but also as a three-class classifier *("First", "Second", "Similar"*). Beyond the balanced pair construction used in prior work, we introduce imbalanced three-class datasets and simplified two-class variants to probe the role of label distribution, evaluating loss functions such as weighted cross-entropy and a KL-based order-consistency loss enforcing symmetry for reversed pairs, and test the performance of checkpoint ensemble strategy.

Under a single-GPU budget, our best model achieves a Pearson Correlation of 0.9532, approaching LMOL's reported 0.9565. Our experiments reveal how three-class behavior—especially the treatment of *"Similar"* pairs—affects BT-derived scores, and provide a clearer understanding of how ambiguous comparisons, class imbalance, and loss design jointly shape FBP performance. Overall, our results show that LMOL-level accuracy is attainable with substantially lower computational cost while offering a more transparent view of the order-learning process.

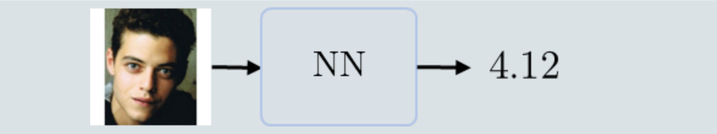

Fig. 1. Direct regression baseline: A conventional neural network predicts a beauty score directly from a single input image.

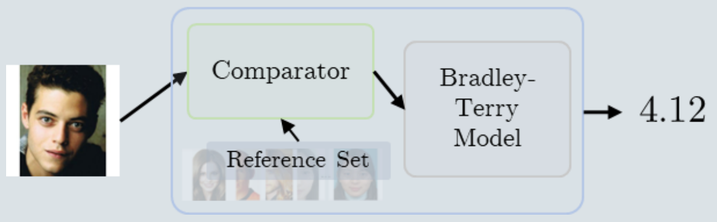

Fig. 2. Order-learning pipeline: The model compares the test image against a reference set, outputs pairwise labels, and infers the final score via the Bradley–Terry model.

Dataset Construction

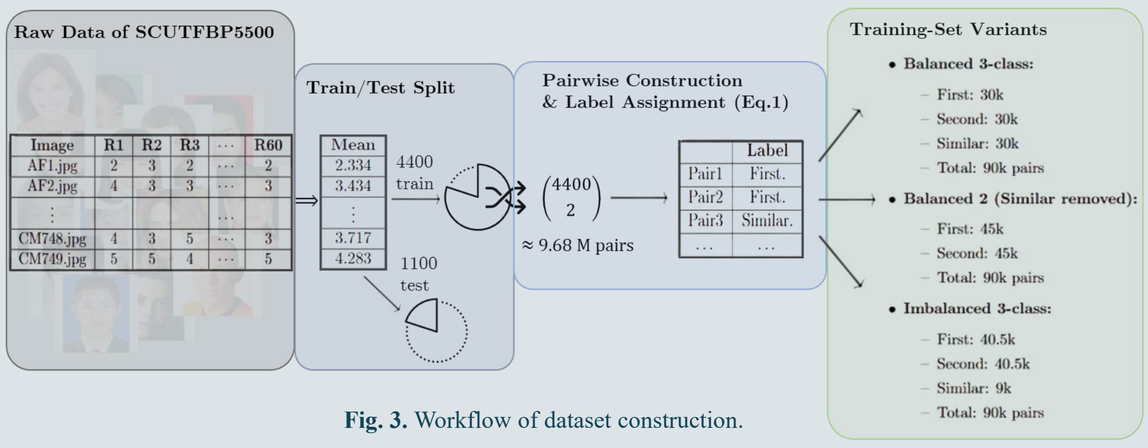

Fig. 3. Workflow of dataset construction.

We adopt an order-learning framework inspired by UOL/LMOL, predicting facial attractiveness through pairwise comparison rather than direct score regression. For any image pair (x_i, x_j) we assign a three-way label based on the score difference $\Delta y = y_i - y_j$ with a fixed margin $\theta = 0.2$: pairs within the margin are labeled Similar, while the remaining pairs identify the preferred face (First or Second). LMOL replaces the FC-based comparator with a large multimodal model (LMM) that outputs textual labels *"First", "Second",* or *"Similar".* Following the SCUT-FBP5500 protocol, we adopt a 4:1 split (4400 train / 1100 test). Although the training split yields 9.68M possible pairs, training an LMM comparator on all of them is infeasible. Therefore, we follow LMOL's sampling strategy and fix the training set to 90k labeled pairs under three controlled variants: (1) Balanced 3-class: 30k *"First"*, 30k *"Second"*, 30k *"Similar"*; (2) Balanced 2-class: remove *"Similar"*, 45k "*First"* and 45k *"Second"*; (3) Imbalanced 3-class (4:4:2): 36k *"First"*, 36k *"Second"*, 18k *"Similar"*. These variants allow a systematic study of how label distribution affects comparator accuracy and the stability of downstream BT-based beauty scores.

Results

Under a unified 4400/1100 split and 5-epoch training budget, our best configuration—balanced 3-class cross-entropy + lightweight checkpoint ensemble—reaches **PC = 0.9532, MAE = 0.1720, RMSE = 0.2141**, which approaching LMOL's reported performance (PC = 0.9565) while using significantly less compute. A comparison of different label constructions reveals the critical role of the *"Similar"* class: although imbalanced three-class and two-class settings push overall accuracy higher, they significantly degrade BT-based ranking quality due to reduced or unreliable *"Similar"* predictions, which otherwise act as soft constraints stabilizing local score gaps. The quantitative results are summarized below:

Training Setting	PC	MAE	RMSE
Balanced 3-class (CE)	95.26%	0.1730	0.2150
Imbalanced 3-class	94.76%	0.1760	0.2290
Balanced 2-class	93.35%	0.2345	0.2916
CE + Checkpoint Ensemble	**95.32%**	**0.1720**	**0.2141**

Loss-function analysis further confirms the importance of the *"Similar"* class: weighted CE increases overall accuracy but collapses *"Similar"* accuracy and worsens BT results, while KL consistency improves forward–reverse agreement yet provides limited gains once fewer consistent *"Similar"* pairs remain. Altogether, these findings indicate that **preserving clean and balanced *"Similar"* comparisons—combined with simple checkpoint smoothing—is more crucial for stable BT score estimation than maximizing raw classification accuracy**.

Conclusion

Our study shows that an LMOL-style order-learning pipeline can approach state-of-the-art facial beauty prediction under a strict single-GPU budget when used carefully rather than simply scaled up. Using an open-source 7B LMM with a balanced three-class setting and a lightweight checkpoint ensemble, we obtain stable BT scores and score-level metrics in the same regime as LMOL, without larger backbones or extensive hyperparameter search.

A central insight is the role of the *"Similar"* class: clean and sufficiently frequent "Similar" pairs act as soft constraints anchoring local score differences, so distorting their distribution—through imbalanced sampling or aggressive reweighting—can hurt BT stability even when pairwise accuracy increases. Together with our results on simple ensemble smoothing, this suggests a practical recipe for FBP and related order-learning tasks: prioritize well-designed pairwise data and training objectives, then use small, compute-friendly tricks to improve robustness instead of relying solely on more parameters and more GPUs.

Looking forward, our findings point to several directions for extending LMOL-style order learning. The strong role of *"Similar"* pairs suggests opportunities to design better strategies for collecting, filtering, or actively sampling informative comparisons to further stabilize BT estimation across domains. Since much of LMOL-level performance stems from data and loss design rather than scale, exploring more efficient comparators, alternative visual encoders, or lighter LMM families could push performance further while reducing resource demands. More broadly, these insights offer a foundation for applying LMM-based order learning to other preference-driven tasks—such as aesthetic ranking, product comparison, or educational content evaluation—where reliable relative judgments matter more than absolute regression.

Model Architecture & BT Inference

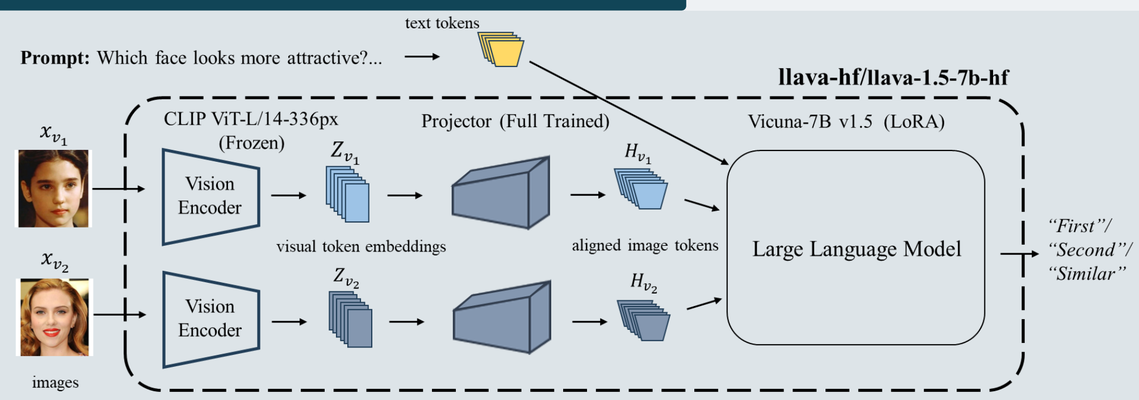

Fig. 4. Model Architecture of LMOL

Model Architecture

Our comparator follows the LMOL formulation and is instantiated using LLaVA-1.5-7B, which integrates a CLIP ViT-L/14 vision encoder and a Vicuna-7B language model. Each input face is encoded into visual tokens:

$$Z_{v_1} = g(x_{v_1}), \qquad Z_{v_2} = g(x_{v_2}) \tag{1}$$

These tokens are projected into the LLM embedding space:

$$H_{v_1} = W \cdot Z_{v_1}, \qquad H_{v_2} = W \cdot Z_{v_2} \tag{2}$$

They are concatenated to form the multimodal input:

$$H_v = [H_{v_1}, H_{v_2}] \tag{3}$$

The model receives the instruction prompt:

$$X_{\text{instruct}} = \langle image1 \rangle \langle image2 \rangle, Which\ face\ looks\ more\ attractive? \tag{4}$$

Only the first generated token is supervised, and vocabulary logits are masked to the three output labels: *"First", "Second", "Similar"*. The class probability is computed by masked SoftMax:

$$p_c = \frac{\exp(t_c)}{\sum_{k \in \mathcal{C}} \exp(t_k)}, \qquad c \in \mathcal{C} \tag{5}$$

Training Loss

(1) Cross-Entropy Baseline

$$\mathcal{L}_{\text{CE}} = -\frac{1}{B} \sum_{n=1}^{B} \sum_{c=1}^{3} (y_{n,c} \cdot \log p_{n,c}) \tag{6}$$

(2) Weigthed Cross Entropy, weights w_c emphasize the *"Similar"* class:

$$\mathcal{L}_{\text{WCE}} = -\frac{1}{B} \sum_{n=1}^{B} \sum_{c=1}^{3} (w_c \cdot y_{n,c} \cdot \log p_{n,c}) \tag{7}$$

(3) KL-Based Forward-Reverse Consistency

$$\mathcal{L}_{\text{KL}} = \frac{1}{2B} \sum_{b=1}^{B} \left[\text{KL}\left(M_{p_b}^{(f)} \parallel p_b^{(r)}\right) + \text{KL}\left(M_{p_b}^{(r)} \parallel p_b^{(f)}\right) \right] \tag{8}$$

where M is the label-reversal permutation matrix, and $p_b^{(f)}$, $p_b^{(r)}$ are forward / reversed predictions

The full objective:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{CE}} + \lambda_{\text{KL}} \cdot \mathcal{L}_{\text{KL}} \tag{9}$$

Bradley–Terry Score Estimation

After comparator training, continuous attractiveness scores are computed using a Bradley–Terry likelihood model. A uniformly sampled reference set (~350 images) is selected by binning the training-set scores (width 0.1, max 10 per bin). For a test image with unknown score s_t, and reference scores s_{r_i}, define:

$$\delta_i = s_t - s_{r_i} \tag{10}$$

Predicted labels determine which likelihood term P_i (refer to **Fig. 5.**) is used. The joint likelihood:

$$\mathcal{L}(s_t) = \prod_i P_i \tag{11}$$

Corresponding negative log-likelihood:

$$\text{NLL}(s_t) = -\sum_i \log P_i \tag{12}$$

We grid search over $s_t \in [1.0, 5.0]$, $\Delta s = 1e-3$ and final score estimate is given:

$$\hat{s}_t = \arg\min_{s_t} \text{NLL}(s_t) \tag{13}$$

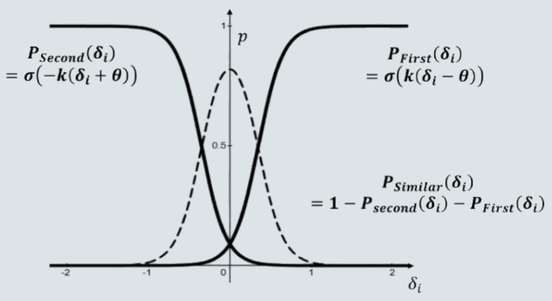

Fig. 5. BT likelihood functions over score difference δ_i.

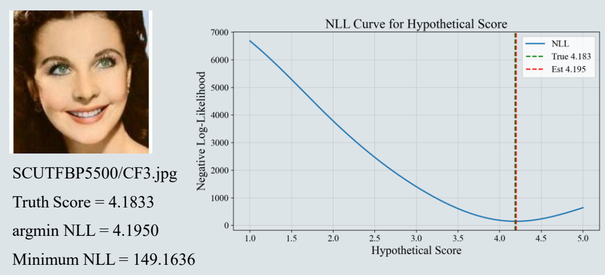

Fig. 6. Negative log-likelihood curve for a single test image