


# Large Multimodal Model is a Better Comparator on Facial Beauty Prediction

Zhenyou Liu<sup>1 2</sup>    Xuefeng Liang<sup>1 3</sup>     Jian Lin<sup>4</sup>

<sup>1</sup> Guangzhou Institute of Technology, Xidian University, China

<sup>2</sup> Information Technology Co., Ltd., China Mobile (Hangzhou), China

<sup>3</sup> School of Artificial Intelligence, Xidian University, China

<sup>4</sup> Research & Innovation Institute, China Mobile (Zhejiang), China

**Abstract**—Order learning has been proven to improve the generalization of facial beauty prediction (FBP) models. However, the scope for advancement remains due to the constraints of current FBP datasets and model scales. In this study, we propose a Large Multimodal Model based Order Learning (LMOL), pioneering the use of a Large Multimodal Model (LMM) as a comparator in order learning. Meanwhile, we develop an FB instruction dataset to fine-tune the LMM, thereby enabling LMOL to discern the FB order. Experiments on three datasets showcase substantial performance improvements in FBP tasks, especially with a notable 18% increase in the Pearson Correlation Coefficient on zero-shot tests (two unseen datasets). This study demonstrate that LMM is a better comparator for FBP tasks.

**Index Terms**—Large multimodal model, Order learning, Facial beauty prediction.

## I. INTRODUCTION

In contemporary society, marked by increasing diversity and inclusivity, it cannot be overlooked that facial beauty (FB) plays an undeniable role on career development, interpersonal relationships and social acceptance [1]. The task of facial beauty prediction (FBP) aims to simulate human cognition of FB to automatically evaluate the FB level. This technology has numerous applications, such as facial image beautification, social network recommendations, plastic surgery, to name a few [2]–[6]. Mainstream FBP methods model this task as a regression problem. However, they often exhibit weak generalization when applied to different datasets.

An emerging work, UOL [7], attributes the weak generalization of conventional methods to two inconsistencies: (1) the inconsistency of FB standards across datasets due to the non-alignment of FB reference bases; and (2) the inconsistency in beauty cognition among different individuals. UOL addresses these two inconsistencies by introducing order learning, which focuses on learning the FB order between faces, as FB order tend to be more consistent across datasets compared to the subjective FB scores. Although UOL improves generalization ability to some extent through the use of order learning, but there is still room for improvement. In addition, UOL is a small-scale FBP model with a limited upper bound on its generalization capability.

A common method to enhance generalization of a model is to expand datasets greatly. It is well-known that obtaining high-quality annotated training data is extremely costly

and time-consuming, especially for subjective data like facial beauty. This is the primary reason for the scarcity and small scale of current FB databases. Although it is possible to train models with different datasets together, the number of training samples across all existing datasets still falls far short of the requirements.

Another method is to employ backbone networks with stronger generalization. The latest Large Multimodal Models (LMMs) have demonstrated exceptional capabilities in data comprehension, including the understanding and analysis of subjective data, such as emotions, mental state. The study [8] shows that LMMs inherently possess a certain comparative capability, showing potential as order learning comparators. However, since LMMs are trained on general data, their comparative capability, especially in the domain of FB, are still insufficient to support their direct application to FBP tasks.

To address this issue, we introduce LMMs into the order learning paradigm for the first time and expand its architecture to replace the conventional comparator based on CNN and MLP in the original order learning, and name it as Large Multimodal Model based Order Learning (LMOL). Meanwhile, the FB order instruction dataset constructed using FB labels is used to fine-tune the LMM to enhance its vertical domain expertise in facial beauty comparison.

The contributions of this work are as follows:

- We introduce LMMs as comparators in the order learning paradigm for the first time, alleviating the issues of weak model generalization and inconsistent FB standards caused by the small scale of FB databases.
- We convert the data with FB scores into FB order instruction dataset to fine-tune the LMM for an enhanced comparative capability in FBP task.
- Extensive experiments on three benchmark datasets, especially the zero-shot evaluation on two unseen datasets, have shown that LMOL achieves significant improvements in performance and generalization compared to the regression models and the traditional order learning model, demonstrating that LMM is a superior comparator.

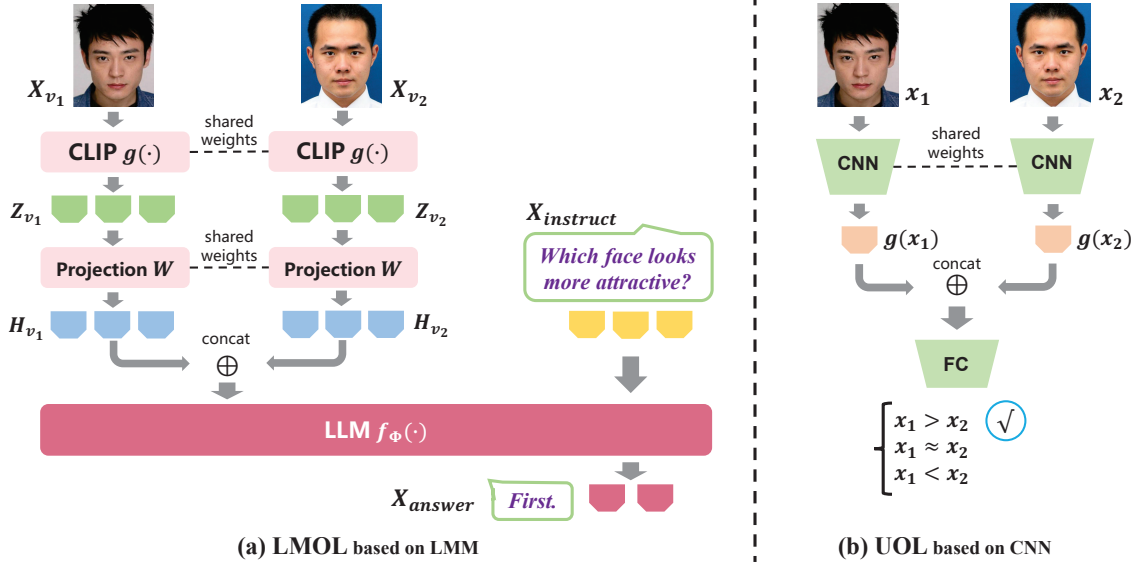


Fig. 1. Comparison between (a) our proposed LMOL and (b) UOL. Given two facial images, UOL encodes each image to a feature vector by a pretrained CNN, and concatenates these two vectors and input them to an MLP-based classifier to predict the FB order. Whereas, LMOL employs the shared-weights CLIP and projection to extract visual feature sequences of the images. These sequences are then processed by the large language model along with the instruction prompts. Finally, LLM outputs the FB order of two images in natural language.

## II. METHOD

### A. Order Learning

Order learning aims to train the model to learn the FB order relations between two facial images. As FB order is independent on the reference base, it can largely circumvent the inconsistencies in FB standards across facial beauty datasets. This approach facilitates more accurate prediction of FB scores when given a reference sample.

Given two facial images,  $x_i$  and  $x_j$ , with their FB scores  $y_i$  and  $y_j$  respectively, the order between  $x_i$  and  $x_j$  can be defined and encoded by a one-hot label,

$$Y = \begin{cases} x_i \approx x_j : [1, 0, 0], & \text{if } |y_i - y_j| \leq \theta, \\ x_i < x_j : [0, 1, 0], & \text{if } y_i - y_j < -\theta, \\ x_i > x_j : [0, 0, 1], & \text{if } y_i - y_j > \theta, \end{cases} \quad (1)$$

where  $\theta = 0.2$  is the threshold that represents the discrimination of FB orders.

The original order learning model extracts visual features of  $x_i$  and  $x_j$  through a siamese visual encoder  $\bar{g}(\cdot)$ , and uses a classifier  $\bar{f}(\cdot)$  composed of linear layers to predict the order relation  $\hat{Y}$ , as shown in Fig. 1 (b):

$$\hat{Y} = \bar{f}(\bar{g}(x_i), \bar{g}(x_j)). \quad (2)$$

### B. The Architecture of LMOL

To enable LMM to learn order relations, we propose the LMOL shown in Fig. 1 (a). It consists of three components: *the visual encoder, the projection matrix, and the backbone LLM*. Unlike UOL, the visual encoder employs a CNN to extract facial visual features, LMOL utilizes CLIP [9] as the visual encoder, combined with a projection matrix to extract

FB features. For two input images  $x_{v1}$  and  $x_{v2}$ , the visual encoder  $g(\cdot)$  is used to encode the features:

$$Z_{v1} = g(x_{v1}), \quad Z_{v2} = g(x_{v2}). \quad (3)$$

Then the projection matrix  $W$  is used to transform the visual features  $Z_{v1}$  and  $Z_{v2}$  into the text embedding space, obtaining the visual token sequences  $H_{v1}$  and  $H_{v2}$ :

$$H_{v1} = W \cdot Z_{v1}, \quad H_{v2} = W \cdot Z_{v2}. \quad (4)$$

Afterward, they need to be fed into an order comparator to obtain their FB ordinal relation. In our LMOL, the backbone LLM,  $f_\phi(\cdot)$ , replaces the MLP classifier used in UOL. Note that the MLP classifier compares two visual features by concatenating them in the feature space. However, in the LMM paradigm, the visual token  $H_{vi}$  of each facial image is a sequence of length  $N$ , where  $N$  is the number of image patches fed into the visual encoder. Therefore, we concatenate the features of two images in the sequence dimension, obtaining the token sequence

$$H_v = [H_{v1}, H_{v2}], \quad (5)$$

as the visual feature that is fed into LMM.

### C. Learn the Order from Instructions

To enable LMM to comprehend the FB comparison task, it is necessary to construct text instructions that LMM can understand, typically in the form of image-text instruction. Then, we design the following instruction for every FB-image pair,  $X_{instruct}$ :

$$X_{instruct} = \langle \text{image1} \rangle \langle \text{image2} \rangle, \quad \text{which face looks more attractive?} \rangle, \quad (6)$$

where  $\langle image1 \rangle$  and  $\langle image2 \rangle$  are placeholders for facial images  $x_{v_1}$  and  $x_{v_2}$ . Similarly, the FB order labels are also converted into the natural language,  $X_{answer}$ :

$$Y = \begin{cases} \text{"Similar."}, & \text{if } x_{v_1} \approx x_{v_2}, \\ \text{"Second."}, & \text{if } x_{v_1} < x_{v_2}, \\ \text{"First."}, & \text{if } x_{v_1} > x_{v_2}. \end{cases} \quad (7)$$

Therefore, the complete order learning instruction for FB-image pairs,  $d_{order}$ , is formulated as:

$$d_{order} = \{x_{v_1}, x_{v_2}, X_{instruct}, X_{answer}\}. \quad (8)$$

It is worth noting that we do not adopt the online sampling method used by UOL, as the huge computational overhead of LMMs often precludes the use of larger batch sizes. Sampling FB-image pairs in a batch may result in an overly random number of pairs for each ordinal category, leading to an imbalanced data distribution. In contrast, we employ an offline sampling method to select two different facial images from the image training dataset, and construct all possible FB-image pairs, which expands the data construction space from a single batch to the entire dataset. We then select  $M$  FB-image pairs from each ordinal relation to construct the instruction dataset. This strategy can ensure each category with sufficient FB-image pairs and maintain a balanced distribution of the selected training data.

#### D. Training Objective

The objective is to train the LLM,  $f_\phi(\cdot)$ , to give the correct answer  $X_{answer}$  using the given two image tokens  $H_{v_1}$  and  $H_{v_2}$ , as well as the order learning instruction  $X_{instruct}$ . Assuming the length of the answer is  $L$ , the conditional probability of the predicted answer  $\hat{X}_{answer}$  can be formulated as:

$$p(\hat{X}_{answer} | H_{v_1}, H_{v_2}, X_{instruct}) = \prod_{i=1}^L f_\phi(X_i | H_{v_1}, H_{v_2}, X_{instruct}, X_{answer < i}), \quad (9)$$

where  $X_i$  is the  $i$ -th token in  $X_{answer}$ , and  $X_{answer < i}$  are the answer tokens before the current prediction token  $X_i$ .

We optimize the backbone LLM's parameters by maximizing the standard condition language modeling loss  $\mathcal{L}$ .

$$\mathcal{L} = \sum_{j=1}^L \log \prod_{i=1}^j f_\phi(X_i | H_{v_1}, H_{v_2}, X_{instruct}, X_{answer < i}). \quad (10)$$

### III. EXPERIMENTS

#### A. Dataset and Evaluation Metrics

To verify the effectiveness and generalization capability of our LMOL, we follow UOL [7]'s setting and train our model only on the SCUT-FBP5500 [10] dataset, and conduct zero-shot evaluations on the Hot-Or-Not [11] and MEBeauty [12] datasets.

SCUT-FBP5500 [10] is currently one of the most widely used datasets for FBP tasks, containing 5500 frontal facial

images, which were scored in the range of  $[1, 5]$  by 60 human volunteers. Hot-Or-Not [11] and MEBeauty [12] are both unseen datasets in this study. Hot-Or-Not contains 2056 frontal female facial images scored in the range of  $[-3, 3]$  by 30 volunteers. MEBeauty consists of 2550 frontal facial images scored in the range  $[1, 10]$  by 300 volunteers. We follow UOL's evaluation metrics and select *Pearson Correlation Coefficient* (PC), *Mean Absolute Error* (MAE), and *Root Mean Squared Error* (RMSE) to evaluate the performance of all methods, especially their generalization capabilities.

#### B. Experimental Settings

Our LMOL is based on the LLaVA-1.5-7B [13], a pre-trained CLIP image encoder ViT-L/14-336px [9] is used as the visual encoder, with an input resolution of  $336 \times 336$ ; The projection matrix is also consistent with LLaVA-1.5, containing two fully connected layers; the backbone LLM is Vicuna-1.5 [14].

During the training phase, we fine-tune LLaVA-1.5-7B using the instruction data constructed in the Sec. II-C on one NVIDIA A100  $\times$  40G GPU. We optimize all parameters of the projection matrix  $W$  and fine-tune the backbone LLM,  $f_\phi(\cdot)$ , with 4-bit Qlora [15], where the LoRA parameter  $r$  is set to 8 and  $\lambda_{scale}$  is set to 4. For each cross-validation folder of SCUT-FBP5500, we construct 90K training pairs for one epoch. The optimizer is AdamW with varied learning rates for the projection matrix and LoRA. The initial learning rates of the projection matrix and LoRA are  $2e-5$  and  $2e-4$ , respectively. Both are adjusted using cosine annealing strategy, with the minimum learning rate 0.

In the evaluation stage, we adopt the Bradley-Terry-based score estimation method proposed by UOL [7], and convert the FB orders predicted by the model into FB scores for performance comparison with the competing methods.

#### C. Performance Evaluation

We compare our method with traditional regression methods  $R^3CNN$  [16], AaNet [17], Co-attention [18], CRNet [19], ComboLoss [20], CNN-ER [21], and the state-of-the-art UOL [7] based on order learning. Table I lists the results of all methods on the dataset SCUT-FBP5500. It can be seen that our LMOL achieves the best results in all evaluation metrics and also gains a significant improvement of 3% on PC, 0.042 on MAE and 0.06 on RMSE compared to the second best, UOL, indicating the state-of-the-art performance on the seen dataset. Figure 2 shows three examples for the intuitive comparisons.

#### D. Generalization Evaluation

To demonstrate the robust generalization capability of LMOL, we follow the protocol of UOL's generalization experiment, by constructing the training instruction dataset only from SCUT-FBP5500, and evaluate the zero-shot performance on two unseen datasets (Hot-Or-Not, MEBeauty). The results in Table II demonstrate our LMOL achieves the best performance again and outperforms UOL with great performance gains. On the Hot-Or-Not dataset, our method improve the PC by 13%.

TABLE I  
PERFORMANCE COMPARISON ON SCUT-FBP5500. THE RESULTS OF  
COMPETING METHODS ARE FROM THEIR PAPERS.

Methods	PC (%) $\uparrow$	MAE $\downarrow$	RMSE $\downarrow$
$R^3$ CNN [16]	91.42	0.2120	0.2800
CRNet [19]	88.69	0.2397	0.3186
AaNet [17]	90.55	0.2236	0.2954
ComboLoss [18]	91.99	0.2050	0.2704
Co-Attention [20]	92.60	0.2020	0.2660
CNN-ER [21]	92.50	0.2009	0.2650
UOL [7]	92.40	0.1975	0.2633
<b>LMOL</b>	<b>95.65</b>	<b>0.1552</b>	<b>0.2032</b>
$\Delta_{second\ best}$	$\uparrow 3.05$	$\downarrow 0.0423$	$\downarrow 0.0601$

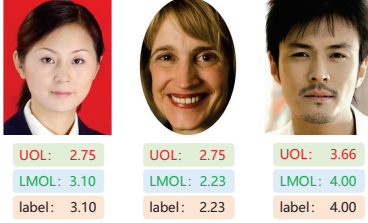


Fig. 2. Intuitive comparison of three examples with the FB scores predicted by LMOL and UOL.

On the MEBeauty dataset, the improvement on PC is up to 18%, indicating that LMOL has much better generalization capability than traditional models.

#### E. Ablation Studies

LMOL and UOL primarily have three distinctions. To validate their individual impacts on the model, we conduct three ablation studies. The models are trained on SCUT-FBP5500 and evaluated through zero-shot testing on MEBeauty. The results are listed in Table III.

**Different visual encoders:** LMOL uses CLIP-ViT, which is transformer-based and pre-trained with more than 400 million vision-language pairs, as the visual encoder, while UOL utilizes a VGG16 pre-trained on ImageNet. Therefore, we replace the features output by VGG16 in UOL with the class embedding features from CLIP-ViT (which can be considered as the pooling features of ViT), and employ UOL’s MLP classifier to predict the FB order, named as setting (1). Comparing the results of setting (1) with UOL, we can observe a 14% improvement of the PC by replacing the visual encoder. This indicates that CLIP ViT possesses a stronger ability to extract FB feature, and provides a major contribution to the model’s generalization performance. This is likely due to CLIP’s training process, which incorporates vast amounts of vision-language data, thereby endowing it with superior generalization capabilities compared to VGG.

**Different encoding and aggregation methods of visual features:** LMOL encodes visual features into sequences and aggregate them by a transformer. In contrast, UOL encodes an image to one feature vector, and processes the concatenated features with an MLP. In this experiment, we first use the class embedding features encoded by CLIP ViT as visual features, employing a 3-layer transformer to predict the FB

TABLE II  
ZERO-SHOT EVALUATION ON HOT-OR-NOT AND MEBeauty. A BETTER  
PERFORMANCE INDICATES A STRONGER GENERALIZATION ABILITY.

Methods	Hot-Or-Not			MeBeauty		
	PC(%) $\uparrow$	MAE $\downarrow$	RMSE $\downarrow$	PC(%) $\uparrow$	MAE $\downarrow$	RMSE $\downarrow$
$R^3$ CNN [16]	35.55	0.5741	0.7140	50.39	0.5329	0.6691
CRNet [19]	32.50	0.5811	0.7294	43.80	0.5645	0.9019
AaNet [17]	28.93	0.5923	0.7399	37.46	0.6102	0.7548
ComboLoss [18]	33.29	0.6154	0.7677	50.78	0.5481	0.6888
Co-Attention [20]	26.97	0.5613	0.7080	49.76	0.5476	0.6907
CNN-ER [21]	35.13	0.5269	0.6653	49.11	0.5753	0.6973
UOL [7]	40.73	0.5410	0.6779	55.32	0.5230	0.6489
<b>LMOL</b>	<b>53.85</b>	<b>0.4927</b>	<b>0.6283</b>	<b>73.66</b>	<b>0.4274</b>	<b>0.5371</b>
$\Delta_{second\ best}$	$\uparrow 13.12$	$\downarrow 0.0483$	$\downarrow 0.0496$	$\uparrow 18.34$	$\downarrow 0.0956$	$\downarrow 0.1118$

TABLE III  
ABLATION STUDIES ON MODEL ARCHITECTURE AND FEATURE ENCODING  
METHODS.

Methods	PC (%) $\uparrow$	MAE $\downarrow$	RMSE $\downarrow$
UOL [7]	55.32	0.5230	0.6489
(1) CLIP class + MLP	69.94	0.4574	0.5631
(2) CLIP class + transformer	69.33	0.4545	0.5649
(3) CLIP patch + transformer	72.05	0.4386	0.5456
<b>LMOL (ours)</b>	<b>73.66</b>	<b>0.4274</b>	<b>0.5371</b>

order, referred to as setting (2). Comparing the results of setting (1) and (2), it is observed that using the transformer is not as effective as directly utilizing an MLP for prediction. Secondly, we adopt the same visual encoding method as LMM, using CLIP-ViT’s patch embeddings as the sequence of encoded visual features, and use the same transformer as in setting (2) to predict the FB order, referred to as setting (3). Comparing settings (1) and (3), setting (3) improves the PC by approximately 2%. This is mainly due to two reasons: 1) The sequence features contain richer FB information; 2) The transformer is more suitable for processing long-range sequences. It can be concluded that the feature based on visual sequences and the transformer architecture in LMMs are more suitable for enhancing the generalization of FBP models.

**Different scale of model parameters:** The architecture of setting (3) is similar to that of LMOL, but LMOL has a larger scale of parameters, amounting to 7B. Compared to setting (3), LMOL increases PC by about 1.6%, indicating a larger scale of parameter is also conducive to the improvement of model generalization.

#### IV. CONCLUSION

We propose LMOL, a novel order learning paradigm to break the limitation of small-scale FB datasets by leveraging the knowledge of the LMM. Beyond replacing of the CNN-based comparator by LMM, a FB instruction data is constructed from existing FB datasets and used to fine-tune the LMM, for enabling LMOL to compare the FB order. Thus, it demonstrates significant gains and better generalization over SOTA approaches on seen and unseen FB datasets, respectively. Ablation analysis further reveals the reasons of such great improvements: the combination of vision-language pretrained encoder, transformer-based feature aggregation and large parameter scale makes LMM a better comparator. In future work, we will strive to extract more accurate FB features by adjusting the structure of LMMs.

## REFERENCES

- [1] Aldo Laurentini and Andrea Bottino, “Computer analysis of face beauty: A survey,” *Computer Vision and Image Understanding*, 125:184–199, 2014.
- [2] Taleb Alashkar, Songyao Jiang, Shuyang Wang, and Yun Fu, “Examples- rules guided deep neural network for makeup recommendation,” *AAAI*, volume 31, 2017.
- [3] Lingyu Liang, Lianwen Jin, and Xuelong Li, “Facial skin beautification using adaptive region-aware masks,” in *IEEE transactions on cybernetics*, 44(12):2600–2612, 2014.
- [4] Luoqi Liu, Junliang Xing, Si Liu, Hui Xu, Xi Zhou, and Shuicheng Yan, “Wow! you are so beautiful today!” in *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, 11(1s):1–22, 2014.
- [5] Xinyu Ou, Si Liu, Xiaochun Cao, and Hefei Ling, “Beauty emakeup: A deep makeup transfer system,” in *ACMMM*, pages 701–702, 2016.
- [6] Kristina Scherbaum, Tobias Ritschel, Matthias Hullin, Thorsten Thormählen, Volker Blanz, and Hans-Peter Seidel, “Computer-suggested facial makeup,” in *Computer Graphics Forum*, volume 30, pages 485–492, 2011.
- [7] Xuefeng Liang, Zhenyou Liu, Jian Lin, Xiaohui Yang, and Takatsune Kumada, “Uncertainty-oriented order learning for facial beauty prediction,” *arXiv preprint arXiv:2409.00603*, 2024.
- [8] Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, Kai Chen, and Dahua Lin, “Mmbench: Is your multi-modal model an all-around player?” *arXiv preprint arXiv:2307.06281*, 2023.
- [9] Hieu Pham, Zihang Dai, Golnaz Ghiasi, Kenji Kawaguchi, Hanxiao Liu, Adams Wei Yu, Jiahui Yu, Yi-Ting Chen, Minh-Thang Luong, Yonghui Wu, et al, “Combined scaling for open-vocabulary image classification,” *arXiv preprint arXiv:2111.10050*, 2021.
- [10] Lingyu Liang, Luoju Lin, Lianwen Jin, Duorui Xie, and Mengru Li, “Scut-fbp5500: A diverse benchmark dataset for multi-paradigm facial beauty prediction,” in *International conference on pattern recognition (ICPR)*, pages 1598–1603, 2018.
- [11] Douglas Gray, Kai Yu, Wei Xu, and Yihong Gong, “Predicting facial beauty without landmarks,” in *ECCV*, pages 434–447, 2010.
- [12] Irina Lebedeva, Yi Guo, and Fangli Ying, “Mebeauty: a multi-ethnic facial beauty dataset in-the-wild,” in *Neural Computing and Applications*, pages 1–15, 2021.
- [13] Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee, “Improved baselines with visual instruction tuning,” in *CVPR*, pages 26296–26306, 2024.
- [14] Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E Gonzalez, et al, “Vicuna: An open-source chatbot impressing gpt-4 with 90%\* chatgpt quality,” website, Apr. 2023. [Online]. Available: <https://vicuna.lmsys.org>
- [15] Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer, “Qlora: Efficient finetuning of quantized llms,” in *Advances in Neural Information Processing Systems*, 36, 2024.
- [16] Luoju Lin, Lingyu Liang, and Lianwen Jin, “Regression guided by relative ranking using convolutional neural network ( $r^3$ cnn) for facial beauty prediction,” in *IEEE Transactions on Affective Computing*, 13(1):122–134, 2019.
- [17] Luoju Lin, Lingyu Liang, Lianwen Jin, and Weijie Chen, “Attribute-aware convolutional neural networks for facial beauty prediction,” in *IJCAI*, pages 847–853, 2019.
- [18] Shengjie Shi, Fei Gao, Xuanton Meng, Xingxin Xu, and Jingjie Zhu, “Improving facial attractiveness prediction via co-attention learning,” in *ICASSP*, pages 4045–4049, 2019.
- [19] Lu Xu, Jinhai Xiang, and Xiaohui Yuan, “Crnet: classification and regression neural network for facial beauty prediction,” in *Pacific Rim Conference on Multimedia*, pages 661–671, 2018.
- [20] Lu Xu and Jinhai Xiang, “Comboloss for facial attractiveness analysis with squeeze-and-excitation networks,” *arXiv preprint arXiv:2010.10721*, 2020.
- [21] F. Bougourzi and F. Dornaika and A. Taleb-Ahmed, “Deep learning based face beauty prediction via dynamic robust losses and ensemble regression,” in *Knowledge-Based Systems*, 242:pages 108246, 2022.