Rethinking Personalized Aesthetics Assessment: Employing Physique Aesthetics Assessment as An Exemplification

Haobin Zhong† Shuai He† Anlong Ming†,* Huadong Ma
Beijing University of Posts and Telecommunications
{zhonghaobin2023, hs19951021, mal, mhd}@bupt.edu.cn

Abstract

The Personalized Aesthetics Assessment (PAA) aims to accurately predict an individual's unique perception of aesthetics. With the surging demand for customization, PAA enables applications to generate personalized outcomes by aligning with individual aesthetic preferences. The prevailing PAA paradigm involves two stages: pre-training and fine-tuning, but it faces three inherent challenges: 1) The model is pre-trained using datasets of the Generic Aesthetics Assessment (GAA), but the collective preferences of GAA lead to conflicts in individualized aesthetic predictions. 2) The scope and stage of personalized surveys are related to both the user and the assessed object; however, the prevailing personalized surveys fail to adequately address assessed objects' characteristics. 3) During application usage, the cumulative multimodal feedback from an individual holds great value that should be considered for improving the PAA model but unfortunately attracts insufficient attention. To address the aforementioned challenges, we introduce a new PAA paradigm called PAA+, which is structured into three distinct stages: pre-training, fine-tuning, and continual learning. Furthermore, to better reflect individual differences, we employ a familiar and intuitive application, physique aesthetics assessment (PhysiqueAA), to validate the PAA+ paradigm. We propose a dataset called PhysiqueAA50K, consisting of over 50,000 annotated physique images. Furthermore, we develop a PhysiqueAA framework (PhysiqueFrame) and conduct a large-scale benchmark, achieving state-of-the-art (SOTA) performance. Our research is expected to provide an innovative roadmap and application for the PAA community. The code and dataset are available in here.

1. Introduction

The saying "beauty is in the eye of the beholder" implies that each individual possesses a distinct perception of aesthetics. While the Generic Aesthetics Assessment (GAA) partially captures subjectivity, it primarily reflects collective aesthetic judgments. The Personalized Aesthetics Assessment (PAA) aims to accurately predict an individual's unique perception of aesthetics, thereby catering to the increasing demand for customization. PAA enables applications such as recommendations or enhancements to generate personalized outcomes that align with individual aesthetic preferences.

The prevailing PAA paradigm [1] typically consists of two stages, namely pre-training and fine-tuning (Fig. 2 (a)). In *the pre-training stage*, GAA datasets are used for supervised training to build a foundational backbone that imparts prior knowledge to the PAA model. In *the fine-tuning stage*, the user's PAA dataset is employed to fine-tune the PAA model by incorporating individual aesthetic experiences. However, the practical application of the prevailing paradigm gradually reveals three inherent questions:

Q1: Is a GAA Model a Superior Choice for PAA Pre-training? The GAA model possesses an inherent characteristic: the collective preferences of GAA cannot satisfy the transitivity conditions as rational individual preferences do. This limitation can be proved by the voting paradox [2], which highlights issues of non-transitivity or obstacles in translating individual choices to collective decisions based on the "majority rule."

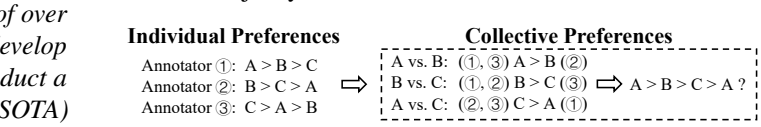
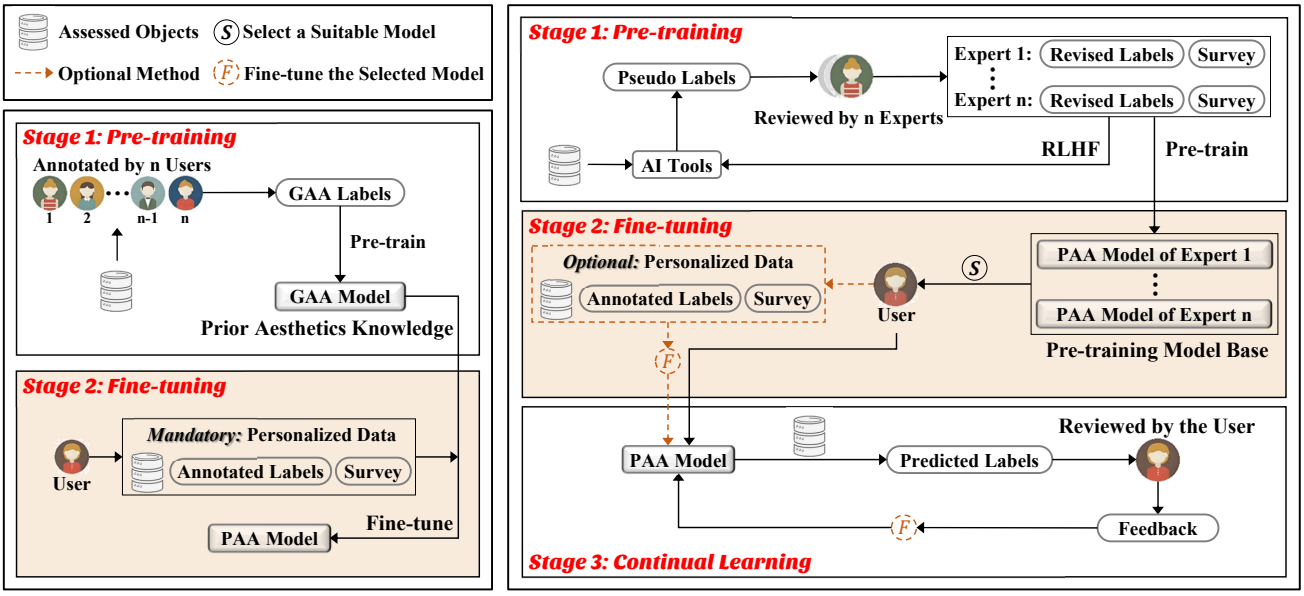


Figure 1. Simple example of the voting paradox in GAA, where three annotators do aesthetic assessments for images A, B, and C. If $A > B$, it implies A's aesthetic score is higher than B's.

In GAA, as shown in Fig. 1, collective preferences create a preference order of $A > B > C > A$, resulting in an inevitable cyclical majority when the number of assessed images reaches a certain large quantity (where the likelihood of a cyclical majority occurring is 100% [3, 4]), regardless of the number of annotators. This example demonstrates that, in GAA datasets, collective preferences fail to satisfy the transitivity conditions as rational individual preferences

† Equal contribution. * Corresponding author.



(a) Prevailing PAA paradigm (b) Our PAA+ paradigm

Figure 2. Paradigm comparison between the Prevailing PAA and our PAA+.

do. Thus, the incorporation of prior knowledge pre-trained on GAA datasets into the PAA model may result in conflicts, potentially leading to personalized issues.

Q2: Are the Scope and Stage of Personalized Surveys Sufficiently Comprehensive? Prevailing personalized surveys consider both the user and the assessed object, which plays a critical role in PAA. However, prevailing surveys come with notable limitations. ***1) Preferences should be Static or Changeable?*** To our knowledge, prevailing personalized surveys are conducted only at the fine-tuning stage and do not account for the timeliness of individual preferences. Given that individual preferences can evolve over time, employing static surveys may lead to a misalignment with users' current tastes. ***2) Capturing Comprehensive Aesthetic Experiences?*** Prevailing surveys typically summarize users' private data into users' attributes, such as age, gender, and big-five personality traits [5, 6]. However, these attributes are less correlated with the objects' characteristics. (e.g., preferences for specific color combinations)

Q3: Does Insufficient Accumulation of User Feedback During Usage Influence Model Performance? The prevailing PAA paradigm typically incorporates user feedback during the fine-tuning stage; however, this type of feedback lacks temporal sensitivity and fails to reflect changes or support modification in user preferences over time. Additionally, the form of feedback is usually restricted to aesthetic scores or rankings, thereby constraining the PAA model's ability to capture a broader understanding of user preferences. Actually, user feedback during usage offers crucial personalized insights that can significantly enhance the PAA model. Regrettably, the prevailing PAA paradigm fails to adequately leverage this data.

This study aims to establish a more comprehensive and advanced PAA paradigm, called PAA+, which addresses the aforementioned issues while employing physique aesthetics assessment (PhysiqueAA) as an exemplification. Our contributions can be summarized as follows:

- This study identifies and analyzes three challenges in PAA, discussing their implications for PAA and emphasizes the necessity for innovative solutions. Notably, we present ***a novel finding*** suggesting that GAA can adversely impact the transitivity of personalized aesthetic preferences.
- This study establishes our PAA+ paradigm (Fig. 2 (b)) that addresses the aforementioned crucial challenges. The proposed paradigm, structured into three distinct stages: pre-training, fine-tuning, and continual learning, offers ***an innovative roadmap for the PAA community***.
- To validate the proposed PAA+ paradigm, we employ a personalized PhysiqueAA task. Guided by our paradigm, we establish a dataset called PhysiqueAA50K (Fig. 4), consisting of over 50,000 fully annotated physique images. Additionally, a Human-AI collaborative annotation method is adopted to enhance the efficiency of the annotation process. Furthermore, we develop a PhysiqueAA framework (PhysiqueFrame), and conduct a large-scale benchmark, achieving state-of-the-art (SOTA) performance. Our exemplification of the new paradigm offers ***an innovative application for the PAA community***.

2. Related Work

2.1. Personalized Aesthetics Assessment

The concept of PAA was initially introduced by Ren *et al.* [7], with a primary focus on image aesthetics assessment.

They proposed training a GAA model to establish fundamental prior knowledge, which was further refined through the incorporation of aesthetic attributes and content features using residual learning techniques, thereby enabling personalized aesthetic "offset" learning. Additionally, they employed the FLICKR-AES [7], REAL-CUR [7], and AADB [8] datasets for PAA.

Subsequently, many outstanding studies have built upon this work, further advancing the field of PAA. Prevailing PAA methods utilize user-annotated labels (scores, ranks, or aesthetic reviews) and personalized surveys (age, gender, personality traits, etc.) to develop PAA models. Wang *et al.* [9] first designed a convolutional neural network (CNN) with user-image relation embedding to train a GAA model. They then employed a collaborative matrix to analyze aesthetic correlations among users and fine-tune the GAA model into a PAA model for a specific user. Other studies have utilized users' interactive behaviors to enhance PAA. For example, Lv *et al.* [10, 11] collected feedback by having users re-rank the aesthetic order of images to explore their preferences. With ongoing developments in PAA research, more studies recognize the importance of personalized surveys for PAA. Li *et al.* [6] proposed a method for modeling aesthetic differences between GAA and PAA scores by utilizing personality traits obtained through personalized surveys. Yang *et al.* [5] introduced a new PAA dataset, PARA, including users' attributes from personalized surveys, such as age, gender, education, photographic experience, and Big-Five personality traits. Their study aims to demonstrate, for the first time, how image aesthetics and individual user attributes interact to shape intricate personalized tastes.

Prevailing PAA methods have several common aspects: 1) they generally utilize GAA models for pre-training; 2) prevailing personalized surveys do not adequately address preferences directly related to the characteristics of the assessed objects; 3) there is insufficient attention given to feedback data during usage, which is primarily focused on the fine-tuning stage; and 4) the assessed objects are images. ***However***, as analyzed in the introduction, the prevailing PAA paradigm encounters three challenges. To address these challenges, we propose a novel PAA+ paradigm structured into three stages: pre-training, fine-tuning, and continual learning. ***Furthermore***, prevailing PAA methods primarily assess the overall aesthetics of images, but we argue that lots of images fail to adequately reflect personalized differences. In contrast, physique aesthetics is closely intertwined to human health and well-being, as individuals possess an inherent desire to evaluate their own physiques, which highlights aesthetic differences among users. Additionally, the new PAA+ paradigm necessitates substantial personal data for effective pre-training while also requiring continuous feedback during the continual learning stage. ***Regrettably***, the prevailing PAA datasets do not satisfy these requirements.

2.2. Physique Aesthetics Assessment

"Physique aesthetics" predominantly focuses on the study of aesthetics in relation to the physical appearance, form, and build of the body [12, 13]. Physique aesthetics is closely tied to individuals, reflecting strong subjective preferences and individual aesthetic differences, rendering PhysiqueAA particularly suitable for validating our new PAA paradigm. The classical literature of aesthetics [14], authored by Shusterman, argues that the aesthetics of the body should not solely focus on the body's appearance but should also encompass considerations of health and behavioral performance. Therefore, we summarize the evaluative dimensions of physique aesthetics as ***appearance, health, and posture***: appearance represents the expressive nature of the physique; health emphasizes its overall well-being; and posture reflects structural control and balance, serving as a crucial factor in harmonizing appearance with health (cf. Appendix A.2).

Building on Shusterman's perspective, PhysiqueAA primarily focuses on individuals' perceptions and preferences toward diverse body shapes and postures. It holds significant application value in various fields, such as healthcare and medical aesthetics [15] as well as recommendation systems [16]. However, PhysiqueAA encounters substantial challenges: 1) the assessed objects exist in a 3D space, complicating evaluations beyond traditional 2D image aesthetics assessment (IAA); 2) it entails a highly subjective task without explicitly designed datasets for PhysiqueAA due to difficulties in deconstructing physique aesthetics.

Potential Approaches for Investigating PhysiqueAA. To our knowledge, there is currently no publicly available learning-based PhysiqueAA model. However, potential approaches for investigating PhysiqueAA can be found in three tasks: ***PAA***, ***IAA***, and ***human-centric visual tasks***. Prevailing PAA [17–19] and IAA [19–26] approaches primarily focus on assessing the overall visual aesthetics of images. Although these approaches are not specifically designed for PhysiqueAA, the underlying principles of aesthetic assessment can also contribute to differentiating the aesthetic qualities of physique. On the other hand, human-centric visual task [27] approaches possess strong perception capabilities that effectively capture crucial physique-related information; hence they hold promise as valuable methods for exploring PhysiqueAA. Building on these foundational insights, we have established a benchmark for PhysiqueAA by selecting approaches based on the following criteria: 1) ***classical architectures with available code***; 2) ***excellence in aesthetic assessment or human-centric perception***.

In this paper, we propose the PhysiqueAA50K dataset, which is collected across different personality types based on the 16 MBTI personality types [28], and introduce the PhysiqueFrame framework, tailored specifically for PhysiqueAA to validate our novel paradigm.

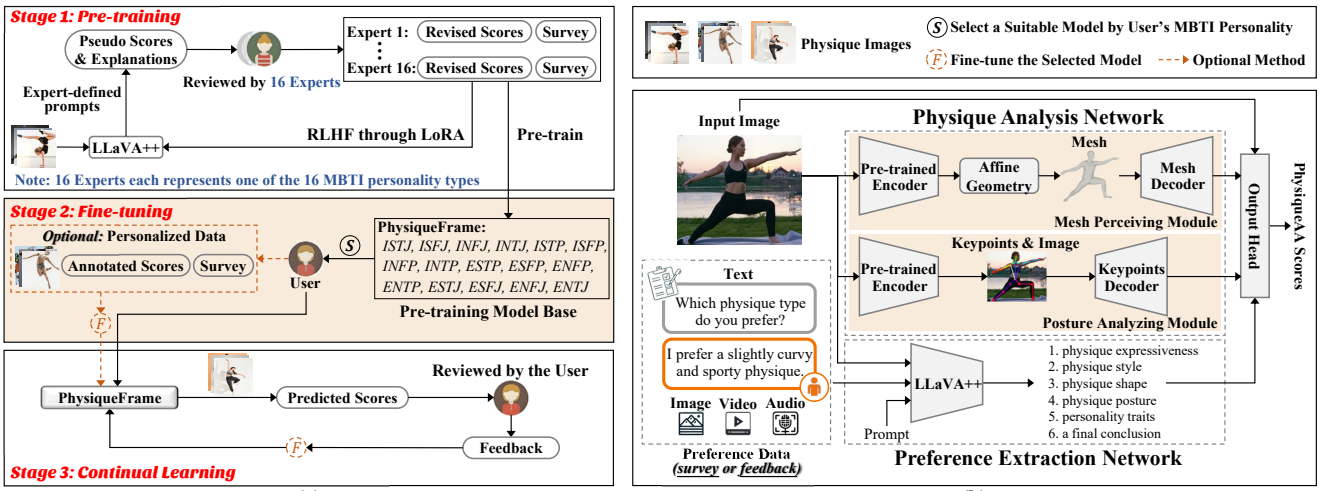


Figure 3. (a) Validation for our PAA+ paradigm. (b) PhysiqueFrame architecture (comprising two networks: PANet, which utilizes a dual-branch module to extract 3D physique-related features such as shape, posture, and facial expression; and PENet, designed to effectively capture user preferences from multimodal preference data).

3. Validation for our PAA+ Paradigm

As illustrated in Fig. 3, the proposed PAA+ paradigm is structured into three stages.

3.1. Stage 1: Pre-training

In this stage, we construct datasets corresponding to each of the 16 MBTI personality types [28], and develop personalized models for all 16 types to replace the traditional GAA model as the pre-training model.

Data Organization. Firstly, we compile a dataset comprising 40,000 images. Subsequently, to facilitate PhysiqueAA annotations, we employ a Large Language Model (LLM), LLaVA++ [29], augmented with Reinforcement Learning from Human Feedback (RLHF) to generate pseudo-labels, thereby alleviating the burden on annotation experts. The RLHF workflow is iterative and involves three primary steps for each cycle:

1) A subset of 1,000 images is selected. The LLaVA++ utilizes expert-defined prompts to generate PhysiqueAA scores (pseudo-labels) for each personality type, accompanied by corresponding explanations.

2) Each annotation expert, representing one of the 16 MBTI personality types, independently reviews and verifies the pseudo-labels corresponding to their respective type. Experts rectify any identified errors by adjusting the PhysiqueAA scores accordingly.

3) The revised scores are subsequently utilized to fine-tune the LLaVA++ using the Low-Rank Adaptation (LoRA) method [30], thereby augmenting the model's predictive accuracy for subsequent annotation iterations.

This iterative process is repeated for each new batch of 1,000 images until the complete 40,000-image dataset is annotated. Through this process, we have generated a comprehensive ***PAA-16-personality*** dataset comprising annotated data for 40,000 images across 16 personality types. This dataset covers 3 assessment dimensions: ***appearance score, health score, and posture score***. Additionally, to enhance the comprehensiveness of personalized surveys, we collected additional information on the annotation experts, including their characteristics (e.g., age, gender) as well as their preferences regarding physique (e.g., preference for a slender or curvy physique). For more details on data organization and the 16 MBTI personality types, *cf.* Appendix A.1, A.2, B.

Training Strategy. Utilizing PhysiqueFrame (*cf.* Sec. 3.4), we employ pre-training to develop a set of 16 personalized models ($F_1, F_2, \ldots, F_{16}$), with each model tailored specifically to one of the 16 distinct personality types based on the ***PAA-16-personality*** dataset.

To optimize the prediction of PhysiqueAA scores across the three assessment dimensions, an automated weight adjustment strategy is applied. The objective function for the i-th model ($i = 1, 2, \ldots, 16$) is formulated as follows:

$$\mathcal{L}_i = \sum_{j}^{3} \left(\frac{1}{\sigma_j^2} \cdot \left(S_{i,j} - F_{i,j}\left(X, P_i\right) \right)^2 + log\ \sigma_j^2 \right), \quad (1)$$

where X is the input image, $S_{i,j}$ indicates the target score for the j-th dimension, and P_i refers to the survey annotation data relevant to the i-th personality type. Here, σ_j is a learnable parameter that dynamically adjusts the weighting of the loss for each PhysiqueAA score, and $log\ \sigma_j^2$ acts as a regularization term to prevent overfitting.

3.2. Stage 2: Fine-tuning

In this stage, an individual user U_i has two options: U_i can either ***directly*** select the pre-trained model most compatible with U_i's personality as the personalized model, or ***alternatively***, fine-tune the selected model using U_i's personalized

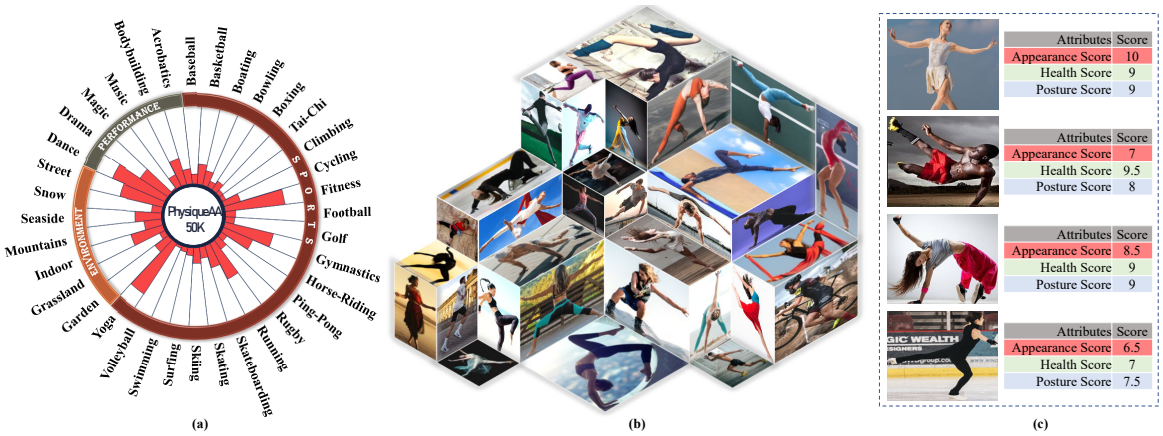


Figure 4. Overview of the PhysiqueAA50K dataset: (a) distribution; (b) visualization; (c) examples with the annotations.

data to achieve a closer alignment with U_i's specific preferences. Here, we focus on the latter approach. The main steps are as follows:

1) Collect a sufficient amount of personalized image data from U_i, including annotated scores and surveys.

2) Use the 16 pre-trained personalized models to predict scores for U_i's images, then select the model that aligns best with U_i annotated scores as the base model F_i. Note: the user also has the option to choose any model they prefer directly.

3) Fine-tune the selected base model F_i with the user-specific data, obtaining the user-specific model F_i^*.

Specifically, we collected 10,000 images for validation, each paired with personalized data from three users, ***User-ISFJ***, ***User-ESFJ*** and ***User-ISTJ***, representing the most common MBTI personality types: ISFJ, ESFJ, and ISTJ. This dataset, referred to as the ***PAA-3-User*** dataset, was used to fine-tune the models with the same loss function as specified in Equation (1). For more details, *cf.* Appendix A.1 and A.2.

3.3. Stage 3: Continual Learning

Due to potential limitations in the sampling of training data or changes in user preferences over time, prediction errors are inevitable during the model's usage. In this stage, we continuously refine the user-specific model F_i^* by collecting user feedback and personalized surveys throughout usage, allowing the model to adapt to the user's evolving aesthetic preferences.

For the user U_i and the user-specific model F_i^*, this stage involves the following three main steps in each update epoch:

1) Use the model F_i^* to assess images and provide scores to U_i;

2) Allow U_i to provide various forms of feedback (e.g., thinking the scores are higher, lower, or about the same, revised scores, texts, images, videos) on the provided scores;

3) Collect a batch of feedback and integrate it with U_i's personalized surveys, then update F_i^*.

For validation, we collected feedback data from ***User-ISFJ***, ***User-ESFJ***, and ***User-ISTJ*** on a set of 400 images, covering 4 update epochs. Each epoch contained 100 feedback entries. We applied the same loss function as defined in Equation (1) to optimize the model.

3.4. Architecture of PhysiqueFrame

Overview. The primary objective of the PhysiqueAA task is to perform a regression analysis that assesses physique aesthetics. Our model is designed to 1) extract physique-related features and 2) capture user preferences and adapt them for personalized assessments.

Fig. 3 (b) illustrates the pipeline of the proposed framework. Given the input image, the Physique Analysis Network (PANet) extracts 3D physique-related features such as shape, posture, and facial expression through a dual-branch module. Meanwhile, the Preference Extraction Network (PENet) captures user preferences such as physique style, body proportions, and muscle definition from personalized surveys and feedback.

3.4.1. Physique Analysis Network (PANet)

How can we effectively represent the human physique? Previous studies have shown that it can be deconstructed into two components: the shape, defined by the external features of flesh, and the posture, determined by the internal skeleton [31–33]. However, current IAA and PAA methods lack 3D analysis capabilities, limiting their ability to capture the spatial and volumetric characteristics of the human physique.

In this paper, the proposed PANet includes two modules. The Mesh Perceiving Module (MPM) leverages 3D human mesh data to analyze body shape, while simultaneously capturing facial features and expressions. The Posture Analyzing Module (PAM), on the other hand, utilizes skeleton keypoints and their connections to assess posture.

Mesh Perceiving Module (MPM). Human mesh data is extracted by an encoder [34], which is pre-trained on datasets such as MSCOCO [35], MPII [35], Human3.6M [36], and

Ubody [37]. Processing the mesh data is challenging due to its inherent irregularity and sparsity [38–40]. A key challenge is to standardize this data to create a coherent representation of a 3D human mesh. To tackle this issue, we propose an affine geometry-based method that normalizes local points while preserving their intrinsic geometric properties. Affine geometry facilitates geometric transformations, including scaling, translation, and rotation, which maintain essential spatial relationships within the data structure.

Specifically, for an initial point $v_i \in \mathbb{R}^m$, we define v_i 's local neighborhood as $\{v_{i,j}\}_{j=1,\cdots,k} \in \mathbb{R}^{m\times k}$, containing k neighboring points. Each neighbor point $v_{i,j}$ has dimension m. The transformation of these local points is represented by the following formula:

$$\{v^*_{i,j}\} = \frac{\{v_{i,j}\} - v_i}{\mu + \varepsilon} \odot \xi + \delta, \tag{2}$$

$$\mu = \sqrt{\frac{\sum_{i=1}^{n}\sum_{j=1}^{k}(v_{i,j} - v_i)^2}{n \times m \times k}}, \tag{3}$$

where $\{v^*_{i,j}\}$ represents the transformed points, $\odot$ indicates Hadamard product, and $\xi \in \mathbb{R}^m$ and $\delta \in \mathbb{R}^m$ are learnable parameters. The scalar μ quantifies the feature deviation across all local groups and channels, where ε is a small constant added for numerical stability. Finally, the normalized mesh data is processed by an MLP-based mesh decoder for further analysis to extract the feature f_{mpm} (*cf.* Appendix C.1 for details).

Posture Analyzing Module (PAM). Posture is a critical aspect of body language and nonverbal communication, reflecting the fluidity and grace of movement while conveying emotions and attitudes of the PhysiqueAA. Structurally, posture can be modeled as a skeleton, where joints are represented as nodes and bones as edges. This structural representation makes graphs a suitable tool for capturing posture features [41–43]. To exploit this representation, we developed the PAM integrated with a Graph Convolutional Network (GCN) to extract relevant posture features. Here, keypoints are the computational representation of human joints [41–43].

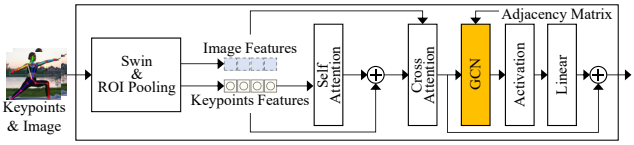


Figure 5. Structure of the keypoints decoder.

Human keypoints are extracted using a pre-trained encoder, which is pre-trained on datasets [35, 35–37], and then processed with the input image through a keypoints decoder to generate features. This decoder first utilizes the Swin Transformer V2 [44] and ROI Pooling [45] to obtain both image and keypoints features. Subsequently, these features are fed into a GCN built upon an adjacency matrix that reflects the human skeletal framework. A residual structure is incorporated to preserve critical low-level details, allowing these details to be retained through successive transformation layers (Fig. 5). Finally, the output undergoes further processing through activation and linear layers, producing the feature f_{pam}.

3.4.2. Preference Extraction Network (PENet)

Our framework integrates data from user surveys and feedback, encompassing a range of intricate modalities, including images, text, scores, and videos. This multimodal data captures comprehensive user preferences to support the final prediction of the PhysiqueAA scores. However, multimodal data is inherently diverse and complex, lacking a unified representation. Understanding the relationship between these multimodal data and users' preferences for physique images is challenging.

To solve the above issue, we implement the PENet based on an LLM, LLaVA++ [29], and utilize a Chain of Thought (CoT) methodology [46] to consolidate diverse modalities into a standardized, quantitative format. The key steps in this process are as follows:

1) **Task Decomposition and Factor Scoring.** The primary task is to understand user preferences as reflected by multimodal data. We decompose the task into a series of subtasks, with each structured as a question-answer prompt directed to ChatGPT. These prompts assess sentiment polarity regarding five specific factors: physique expressiveness, physique style, physique shape, physique posture, and personality traits (*cf.* Appendix C.2 for details). Each factor is assigned a probability score, where scores close to 1 indicate a positive user preference, while scores near 0 suggest a negative preference. Moreover, we prompt ChatGPT to provide a final conclusion, asking: "Does the user like this image?" This is represented by a binary output, scored as 1 for "yes" and 0 for "no."

2) **Applying CoT to Train PENet.** After reorganizing the data, we apply the CoT approach to train PENet. The training is supported by a Q&A mechanism based on the question: "Q: Does the following data suggest that the user likes this image?" An example of a ground-truth response could be: "The physique expressiveness score is 0.6, and the physique style score is 0.8... therefore, the answer is 1."

3) **Score Aggregation.** In the final stage, the scores are aggregated into the feature $f_{preference}$ with f_{mpm},f_{pam}, and the original image, to predict the PhysiqueAA scores.

4. Experiments

4.1. Experimental Settings

Benchmark Dataset. We conducted experiments on the proposed PhysiqueAA50K dataset (Fig. 4), which encompasses data across the three stages: PAA-16-Personality, PAA-3-User, and 400 images for stage 3 validation (see Sec. 3.1-3.3 for details). The PhysiqueAA50K dataset includes PhysiqueAA scores, personalized user surveys, and user

Metric		User-ISFJ GAA	User-ISFJ Ours	User-ESFJ GAA	User-ESFJ Ours	User-ISTJ GAA	User-ISTJ Ours
Appearance	$\mathcal{S}\uparrow$	.650	**.699**	.628	**.649**	.603	**.617**
	$\mathcal{L}\uparrow$	.676	**.724**	.650	**.678**	.631	**.650**
	$\mathcal{A}\uparrow$	.739	**.759**	.721	**.744**	.701	**.716**
Health	$\mathcal{S}\uparrow$	.601	**.625**	.516	**.540**	.509	**.546**
	$\mathcal{L}\uparrow$	.626	**.657**	.539	**.567**	.552	**.570**
	$\mathcal{A}\uparrow$	.747	**.761**	.725	**.740**	.704	**.726**
Posture	$\mathcal{S}\uparrow$	.652	**.664**	.632	**.657**	.568	**.597**
	$\mathcal{L}\uparrow$	.689	**.703**	.684	**.695**	.616	**.641**
	$\mathcal{A}\uparrow$	.744	**.764**	.740	**.761**	.691	**.701**
Satisfaction	$\overline{Sat}\uparrow$	75.3%	**84.5%**	70.6%	**81.0%**	68.9%	**80.9%**

Table 1. Performance comparison of the GAA model and personality-matched personalized models by comparing them as pre-training models, with each being fine-tuned on the three datasets (User-ISFJ, User-ESFJ, User-ISTJ) for stages 2 and 3 to evaluate performance.

Method	Appearance $\mathcal{S}\uparrow$	Appearance $\mathcal{L}\uparrow$	Health $\mathcal{S}\uparrow$	Health $\mathcal{L}\uparrow$	Posture $\mathcal{S}\uparrow$	Posture $\mathcal{L}\uparrow$	Satisfaction $\overline{Sat}\uparrow$
w/o preferences	.683	.704	.614	.639	.651	.686	79.4%
w/ preferences	**.699**	**.724**	**.625**	**.657**	**.664**	**.703**	**84.5%**

Table 2. Performance comparison in stages 2 and 3, when user preferences are included or excluded from the personalized survey.

feedback, with scores across three dimensions: appearance, health, and posture (*cf.* Appendix A.2 and A.3 for details).

Benchmark Models and Training Protocols. In the proposed paradigm, the PANet for physique feature extraction can be replaced with other models to establish a comprehensive PhysiqueAA benchmark. As described in Sec. 2.2, we selected 8 representative baseline models [17–23, 27]. Each model was trained with its recommended parameter settings (e.g., optimizer and batch size) and evaluated under the same training and testing conditions. Equation (1) was used as the loss function across all models, with the training strategy for each stage outlined in Sec. 3.1-3.3.

Evaluation Metrics. We employ three popular evaluation metrics to assess the model's capabilities in **stages 1 and 2**: Spearman's rank correlation coefficient (SRCC, $\mathcal{S}$), the linear correlation coefficient (LCC, $\mathcal{L}$), and binary classification accuracy (ACC, $\mathcal{A}$). SRCC and LCC provide fine-grained assessments of the model's performance in score regression, while ACC, based on a 5-point threshold, evaluates the model's coarse-grained ability to classify PhysiqueAA scores as aesthetically positive or negative.

In **stage 3**, however, user feedback is immediate and dynamic with each update epoch, making it impossible to establish a fixed testing dataset. As a result, the above metrics are not applicable. Instead, we assess performance based on user satisfaction, measuring the proportion of positive feedback. The average user satisfaction rate ($\overline{Sat}$) reflects the overall performance across this stage.

Stage 1	Stage 2	Stage 3	Appearance $\mathcal{S}\uparrow$	Appearance $\mathcal{L}\uparrow$	Health $\mathcal{S}\uparrow$	Health $\mathcal{L}\uparrow$	Posture $\mathcal{S}\uparrow$	Posture $\mathcal{L}\uparrow$	Satisfaction $\overline{Sat}\uparrow$
	✓	✓	.686	.710	.617	.642	.653	.690	78.5%
✓		✓	.682	.705	.613	.637	.651	.685	76.7%
✓	✓		.691	.719	.618	.651	.658	.696	82.3%
✓	✓	✓	**.699**	**.724**	**.625**	**.657**	**.664**	**.703**	**84.5%**

Table 3. Impact of incorporating personalized surveys in different stages on model performance.

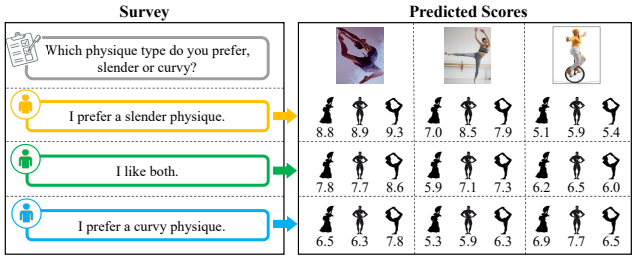


Figure 6. Impact of personalized surveys on PhysiqueAA predictions, with icons , and representing scores of appearance, health, and posture, respectively.

4.2. Ablation Study of our PAA+ Paradigm

Validation for Q1: Is a GAA Model a Superior Choice for PAA Pre-training? To answer this question, for three users with different personalities (ISFJ, ESFJ, ISTJ), we pre-trained both a GAA model and three personality-matched personalized models, as pre-training models in stage 1. For the GAA model, pre-training was conducted using the average scores across all personalities from our ***PAA-16-personality*** dataset, while each personalized model was pre-trained on personality-specific data (without personalized surveys). We then evaluated the final PAA performance on the ***PAA-3-User*** dataset. As shown in Table 1, the GAA model, when used as a pre-training model, performed worse than the personality-matched personalized models. This suggests that integrating prior knowledge from the GAA dataset into the PAA model may introduce conflicting influences.

Validation for Q2: Are the Scope and Stage of Personalized Surveys Sufficiently Comprehensive? The personalized surveys used in prevailing methods primarily assess user attributes without exploring their preferences. As shown in Table 2 (on the User-ISFJ dataset), excluding user preferences from the personalized survey leads to a performance decline, revealing a limitation in the scope of these methods' personalized surveys. Moreover, Table 3 demonstrates that incorporating personalized surveys across all three stages enhances the model's prediction accuracy, compared to applying personalized surveys in only a single stage. Fig. 6 provides qualitative examples illustrating the impact of personalized surveys on model predictions.

Validation for Q3: Does Insufficient Accumulation of User Feedback During Usage Influence Model Performance? In stage 3, the sufficiency of user feedback in-

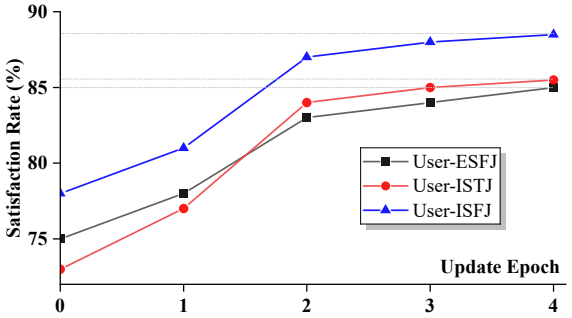


Figure 7. Relationship between the number of feedback update epochs (100 images each) from three users and their satisfaction rate during the continual learning stage.

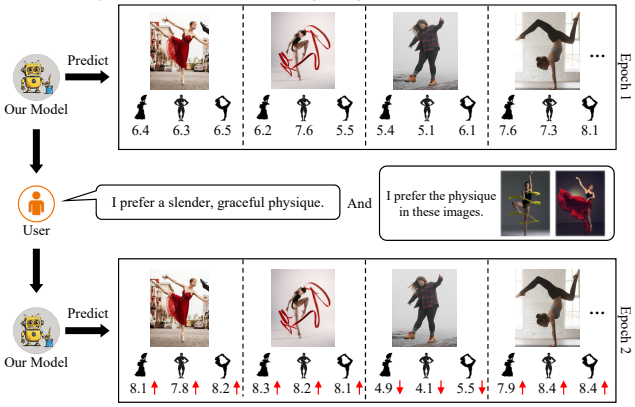


Figure 8. Feedback process and its impact on PhysiqueAA.

Metric		CNN-based models				Transformer-based models				Ours
		NIMA	BIAA	HGCN	TANet	MaxViT	TCFormer	EAT	Q-Align	
Appearance	$\mathcal{S}\uparrow$	.598	.567	.419	.601	.574	.452	.628	.619	**.699**
	$\mathcal{L}\uparrow$	.612	.583	.435	.625	.599	.477	.650	.634	**.724**
	$\mathcal{A}\uparrow$	.732	.694	.667	.712	.705	.673	.732	.731	**.759**
Health	$\mathcal{S}\uparrow$	.559	.485	.362	.528	.513	.383	.573	.555	**.625**
	$\mathcal{L}\uparrow$	.582	.513	.391	.569	.547	.417	.599	.575	**.657**
	$\mathcal{A}\uparrow$	.731	.665	.601	.705	.691	.612	.732	.726	**.761**
Posture	$\mathcal{S}\uparrow$	.581	.536	.486	.572	.563	.508	.592	.609	**.664**
	$\mathcal{L}\uparrow$	.616	.574	.502	.617	.596	.523	.629	.644	**.703**
	$\mathcal{A}\uparrow$	.736	.682	.618	.706	.695	.633	.729	.735	**.764**

Table 4. Comparison of the night models conducted on the User-ISFJ dataset for stage 2. We ***retrained the models to maximize performance,*** using the recommended parameters and settings.

creases with the number of update epochs in the continual learning process, which in turn enhances model performance. As shown in Fig. 7, user satisfaction improves as feedback epochs accumulate. Fig. 8 presents qualitative examples illustrating how specific user feedback impacts the outcomes of PhysiqueAA.

4.3. Ablation Study of proposed PhysiqueFrame

Performance Comparison with PANet Alternatives. As shown in Table 4, our model achieves the best performance across all metrics when compared with alternative models on the User-ISFJ dataset. Appendix D.1 provides experiments on the User-ESFJ and User-ISTJ datasets. This indicates that our model more effectively captures features related to physique aesthetics.

The GradCAM method [47] is applied to visualize the saliency maps in Fig. 9. Our model, with its comprehensive

PAM	MPM	PENet	Appearance $\mathcal{S}\uparrow$	Appearance $\mathcal{L}\uparrow$	Health $\mathcal{S}\uparrow$	Health $\mathcal{L}\uparrow$	Posture $\mathcal{S}\uparrow$	Posture $\mathcal{L}\uparrow$
✓			.663	.688	.594	.615	.648	.685
	✓		.676	.696	.608	.627	.641	.678
✓	✓		.680	.702	.611	.634	.647	.682
✓	✓	✓	**.699**	**.724**	**.625**	**.657**	**.664**	**.703**

Table 5. Ablation studies of PANet modules (PAM, MPM) and PENet, conducted on the User-ISFJ dataset for stage 2.

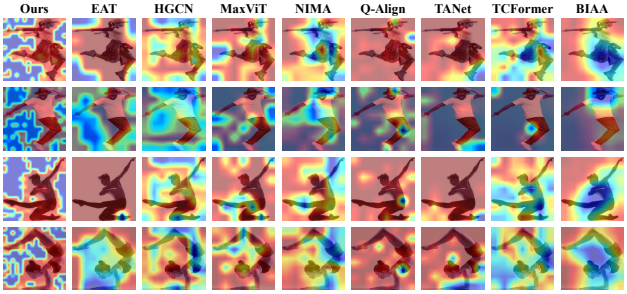


Figure 9. Saliency maps of the 9 models. The attention of our model is focused more on the human physique.

integration of physique aesthetics knowledge, enables a more human-centric assessment by directing the model's attention specifically toward the human body.

Ablation Study on Modules of PANet. As shown in lines 1–3 of Table 5, we examined the effectiveness of the PAM and MPM modules. The results suggest that incorporating PAM and MPM enables PhysiqueFrame to extract more physique-related features, thereby enhancing performance.

Ablation Study of PENet. A comparison between line 3 and line 4 in Table 5 shows that incorporating PENet improves the model's performance. These results show that PENet enhances the model's alignment with user preferences.

5. Conclusions

In this paper, we present the innovative PAA+ paradigm that addresses three challenges of prevailing approaches and incorporates multi-modal dynamic surveys across all stages to enhance model adaptability. We conduct a preliminary validation of the proposed PAA+ paradigm using a typical PAA task, PhysiqueAA. Following the PAA+ paradigm, we develop PhysiqueFrame, construct a physique-specific dataset, and establish the largest-scale benchmarks to date. Compared with potential models for investigating PhysiqueAA, PhysiqueFrame achieves SOTA performance and produces visually consistent results. We aim to make a valuable contribution to the PAA community through our work. However, we acknowledge that scoring alone is an insufficient means of assessment; therefore, we will optimize the PAA+ paradigm to enhance its effectiveness and applicability.

6. Acknowledgement

This work is supported by the National Natural Science Foundation of China under Grant U24B20176.

References

[1] Hancheng Zhu, Yong Zhou, Leida Li, Jiaqi Zhao, and Wenliang Du. The review of personalized image aesthetics assessment. *Journal of Image and Graphics*, 27(10):2937–2951, 2022. 1

[2] Kenneth J. Arrow. *Social Choice and Individual Values*. Yale University Press, 2012. 1

[3] William H Riker. Liberalism against populism. 1988. 1

[4] William V. Gehrlein. Condorcet's paradox and the condorcet efficiency of voting rules. *Mathematica Japonicae*, 45(1):173–199, 1997. 1

[5] Yuzhe Yang, Liwu Xu, Leida Li, Nan Qie, Yaqian Li, Peng Zhang, and Yandong Guo. Personalized image aesthetics assessment with rich attributes. In *CVPR*, pages 19861–19869, 2022. 2, 3, 12

[6] Leida Li, Hancheng Zhu, Sicheng Zhao, Guiguang Ding, and Weisi Lin. Personality-assisted multi-task learning for generic and personalized image aesthetics assessment. *TIP*, 29:3898–3910, 2020. 2, 3, 12

[7] Jian Ren, Xiaohui Shen, Zhe Lin, Radomir Mech, and David J Foran. Personalized image aesthetics. In *ICCV*, pages 638–647, 2017. 2, 3, 11

[8] Shu Kong, Xiaohui Shen, Zhe Lin, Radomir Mech, and Charless Fowlkes. Photo aesthetics ranking network with attributes and content adaptation. In *ECCV*, pages 662–679. Springer, 2016. 3

[9] Guolong Wang, Junchi Yan, and Zheng Qin. Collaborative and attentive learning for personalized image aesthetic assessment. In *IJCAI*, pages 957–963, 2018. 3

[10] Pei Lv, Meng Wang, Yongbo Xu, Ze Peng, Junyi Sun, Shimei Su, Bing Zhou, and Mingliang Xu. Usar: An interactive user-specific aesthetic ranking framework for images. In *ACMMM*, pages 1328–1336, 2018. 3

[11] Pei Lv, Jianqi Fan, Xixi Nie, Weiming Dong, Xiaoheng Jiang, Bing Zhou, Mingliang Xu, and Changsheng Xu. User-guided personalized image aesthetic assessment based on deep reinforcement learning. *TIM*, 25:736–749, 2021. 3

[12] Jean Baptiste Morelly. *Physique de la beauté ou pouvoir naturel de ses charmes*. aux dépens de la Compagnie, 1748. 3, 13

[13] Richard Coe. Sensualism and aesthetics: a study of morelly's physique de la beautè. *Australian Journal of French Studies*, 1(2):146–163, 1964. 3, 13

[14] Richard Shusterman. Somaesthetics: A Disciplinary Proposal. *The Journal of Aesthetics and Art Criticism*, 57(3):299–313, 06 1999. 3, 13

[15] Viren Swami and Adrian Furnham. *The psychology of physical attraction*. Routledge, 2007. 3

[16] Shuang Ma, Yangyu Fan, and Chang Wen Chen. Pose maker: A pose recommendation system for person in the landscape photographing. In *ACMMM*, pages 1053–1056, 2014. 3

[17] Hancheng Zhu, Leida Li, Jinjian Wu, Sicheng Zhao, Guiguang Ding, and Guangming Shi. Personalized image aesthetics assessment via meta-learning with bilevel gradient optimization. *TCYB*, 2020. 3, 7

[18] Zhengzhong Tu, Hossein Talebi, Han Zhang, Feng Yang, Peyman Milanfar, Alan Bovik, and Yinxiao Li. Maxvit: Multi-axis vision transformer. *ECCV*, 2022.

[19] Shuai He, Yongchang Zhang, Rui Xie, Dongxiang Jiang, and Anlong Ming. Rethinking image aesthetics assessment: Models, datasets and benchmarks. *IJCAI*, 2022. 3

[20] Hossein Talebi and Peyman Milanfar. NIMA: Neural image assessment. *TIP*, 27(8):3998–4011, 2018.

[21] Dongyu She, Yu-Kun Lai, Gaoxiong Yi, and Kun Xu. Hierarchical layout-aware graph convolutional network for unified aesthetics assessment. In *CVPR*, pages 8475–8484, 2021.

[22] Shuai He, Anlong Ming, Shuntian Zheng, Haobin Zhong, and Huadong Ma. Eat: An enhancer for aesthetics-oriented transformers. In *ACMMM*, pages 1023–1032, 2023. 11

[23] Haoning Wu, Zicheng Zhang, Weixia Zhang, Chaofeng Chen, Liang Liao, Chunyi Li, Yixuan Gao, Annan Wang, Erli Zhang, Wenxiu Sun, et al. Q-align: Teaching lmms for visual scoring via discrete text-defined levels. *arXiv preprint arXiv:2312.17090*, 2023. 7

[24] Shuai He, Anlong Ming, Yaqi Li, Jinyuan Sun, ShunTian Zheng, and Huadong Ma. Thinking image color aesthetics assessment: Models, datasets and benchmarks. In *ICCV*, pages 21838–21847, 2023.

[25] Shuai He, Yi Xiao, Anlong Ming, and Huadong Ma. Prompt-guided image color aesthetics assessment: Models, datasets and benchmarks. *Information Fusion*, 114:102706, 2025.

[26] Shuai He, Shuntian Zheng, Anlong Ming, Banyu Wu, and Huadong Ma. Rethinking no-reference image exposure assessment from holism to pixel: Models, datasets and benchmarks. *Advances in Neural Information Processing Systems*, 37:131596–131617, 2024. 3

[27] Wang Zeng, Sheng Jin, Wentao Liu, Chen Qian, Ping Luo, Wanli Ouyang, and Xiaogang Wang. Not all tokens are equal: Human-centric visual analysis via token clustering transformer. In *CVPR*, pages 11101–11111, 2022. 3, 7

[28] Carl G Jung. Psychological types the collected works of cg jung. *Tr. HG Baynes. Rev. RFC Hull. Princeton: Princeton UP*, 1971. 3, 4, 12

[29] Hanoona Rasheed, Muhammad Maaz, Salman Khan, and Fahad S Khan. Llava++: extending visual capabilities with llama-3 and phi-3 (2024). *URL https://github. com/mbzuai-oryx/LLaVA-pp*. 4, 6

[30] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *ICLR*, 2022. 4

[31] Ce Zheng, Wenhan Wu, Chen Chen, Taojiannan Yang, Sijie Zhu, Ju Shen, Nasser Kehtarnavaz, and Mubarak Shah. Deep learning-based human pose estimation: A survey. *ACM Comput. Surv.*, 56(1), August 2023. 5

[32] Georgios Pavlakos, Luyang Zhu, Xiaowei Zhou, and Kostas Daniilidis. Learning to estimate 3d human pose and shape from a single color image. In *CVPR*, pages 459–468, 2018.

[33] Federica Bogo, Angjoo Kanazawa, Christoph Lassner, Peter Gehler, Javier Romero, and Michael J Black. Keep it smpl: Automatic estimation of 3d human pose and shape from a single image. In *ECCV*, pages 561–578. Springer, 2016. 5

[34] Zhongang Cai, Wanqi Yin, Ailing Zeng, Chen Wei, Qingping Sun, Wang Yanjun, Hui En Pang, Haiyi Mei, Mingyuan Zhang, Lei Zhang, et al. Smpler-x: Scaling up expressive human pose and shape estimation. *Advances in Neural Information Processing Systems*, 36, 2024. 5

[35] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *ECCV*, pages 740–755. Springer, 2014. 5, 6

[36] Catalin Ionescu, Dragos Papava, Vlad Olaru, and Cristian Sminchisescu. Human3. 6m: Large scale datasets and predictive methods for 3d human sensing in natural environments. *IEEE transactions on pattern analysis and machine intelligence*, 36(7):1325–1339, 2013. 5

[37] Jing Lin, Ailing Zeng, Haoqian Wang, Lei Zhang, and Yu Li. One-stage 3d whole-body mesh recovery with component aware transformer. In *CVPR*, pages 21159–21168, 2023. 6

[38] Georgios Pavlakos, Vasileios Choutas, Nima Ghorbani, Timo Bolkart, Ahmed AA Osman, Dimitrios Tzionas, and Michael J Black. Expressive body capture: 3d hands, face, and body from a single image. In *CVPR*, pages 10975–10985, 2019. 6

[39] Mario Botsch. Polygon mesh processing. *AK Peters*, 2010.

[40] Xu Ma, Can Qin, Haoxuan You, Haoxi Ran, and Yun Fu. Rethinking network design and local geometry in point cloud: A simple residual mlp framework. *arXiv preprint arXiv:2202.07123*, 2022. 6

[41] Lei Shi, Yifan Zhang, Jian Cheng, and Hanqing Lu. Two-stream adaptive graph convolutional networks for skeleton-based action recognition. In *CVPR*, pages 12026–12035, 2019. 6

[42] Sijie Yan, Yuanjun Xiong, and Dahua Lin. Spatial temporal graph convolutional networks for skeleton-based action recognition. In *AAAI*, volume 32, 2018.

[43] Pengfei Zhang, Cuiling Lan, Wenjun Zeng, Junliang Xing, Jianru Xue, and Nanning Zheng. Semantics-guided neural networks for efficient skeleton-based human action recognition. In *CVPR*, pages 1112–1121, 2020. 6

[44] Ze Liu, Han Hu, Yutong Lin, Zhuliang Yao, Zhenda Xie, Yixuan Wei, Jia Ning, Yue Cao, Zheng Zhang, Li Dong, Furu Wei, and Baining Guo. Swin transformer v2: Scaling up capacity and resolution, 2022. 6

[45] RCNN Faster. Towards real-time object detection with region proposal networks. *NeurIPS*, 9199(10.5555):2969239–2969250, 2015. 6

[46] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837, 2022. 6

[47] Jacob Gildenblat and contributors. Pytorch library for cam methods. https://github.com/jacobgil/pytorch-grad-cam, 2021. 8

[48] Xuan Ju, Ailing Zeng, Jianan Wang, Qiang Xu, and Lei Zhang. Human-art: A versatile human-centric dataset bridging natural and artificial scenes. In *CVPR*, pages 618–629, 2023. 11

[49] Shikai Li, Jianglin Fu, Kaiyuan Liu, Wentao Wang, Kwan-Yee Lin, and Wayne Wu. Cosmicman: A text-to-image foundation model for humans. In *CVPR*, pages 6955–6965, 2024.

[50] Xiaodan Liang, Chunyan Xu, Xiaohui Shen, Jianchao Yang, Si Liu, Jinhui Tang, Liang Lin, and Shuicheng Yan. Human parsing with contextualized convolutional neural network. In *ICCV*, pages 1386–1394, 2015.

[51] Mykhaylo Andriluka, Leonid Pishchulin, Peter Gehler, and Bernt Schiele. 2d human pose estimation: New benchmark and state of the art analysis. In *CVPR*, June 2014. 11

[52] Somdev Sangwan and contributors. Roop: Real-time face swap tool. https://github.com/s0md3v/roop, 2024. 11

[53] Naila Murray, Luca Marchesotti, and Florent Perronnin. Ava: A large-scale database for aesthetic visual analysis. In *CVPR*. IEEE, 2012. 11

[54] Limin Liu, Shuai He, Anlong Ming, Rui Xie, and Huadong Ma. Elta: An enhancer against long-tail for aesthetics-oriented models. In *ICML*. 11

[55] WHO Consultation. Obesity: preventing and managing the global epidemic. *World Health Organization technical report series*, 894:1–253, 2000. 11

[56] Hancheng Zhu, Leida Li, Sicheng Zhao, and Hongyan Jiang. Evaluating attributed personality traits from scene perception probability. *Pattern Recognition Letters*, 116:121–126, 2018. 12

[57] Francesco Gelli, Tiberio Uricchio, Xiangnan He, Alberto Del Bimbo, and Tat-Seng Chua. Learning subjective attributes of images from auxiliary sources. In *ACMMM*, pages 2263–2271, 2019. 12

[58] Richard Shusterman. *Body Consciousness: A Philosophy of Mindfulness and Somaesthetics*. Cambridge University Press, 2008. 13

[59] Michael Freeman. The photographer's eye: Composition and design for better digital photos. *Focal Press*, 2007. 13

[60] Cynthia Freeland. Portraits in painting and photography. *Philosophical Studies*, 135:95–109, 2007. 13

[61] Fred Davis. *Fashion, culture, and identity*. University of Chicago Press, 1994. 13

[62] Alan M Klein. *Little big men: Bodybuilding subculture and gender construction*. State University of New York Press, 1993. 13

[63] Sondra Horton Fraleigh. *Dance and the lived body*. University of Pittsburgh Pre, 1996. 13

Rethinking Personalized Aesthetics Assessment: Employing Physique Aesthetics Assessment as An Exemplification

Supplementary Material

Appendix

A. More Details of Proposed Dataset

A.1. Data Collection

How to avoid selection bias by building a comprehensive PhysiqueAA dataset. Selection bias occurs when certain types of physiques or scenes are insufficiently collected in the dataset, which can negatively affect the model's generalizability. To address this issue, we collected various human physique images from five open-source datasets [7, 48–51], to ensure the richness of physiques in our dataset. Additionally, our dataset includes three major categories (performance, sports, environment) and 36 subcategories (dance, yoga, street scene, etc.), ensuring a wide variety of physical activities.

How to Avoid Privacy Violations in Human Images. When annotating data, annotators may be particularly concerned with images that involve human faces. To safeguard privacy, we employed face-swapping technology, Roop [52]. This technology enables us to swap human faces while preserving essential facial features, including expressions and contours. The process consisted of three steps:

1) We created a facial dataset that includes a diverse range of faces, such as open-source, AI-generated, and digital human images, annotated with relevant labels such as age and gender.

2) For each human image, we selected the top five faces from this dataset based on cosine similarity of features, while also matching faces based on similar age and gender.

3) The images generated through Roop were then manually reviewed to identify the most suitable and natural-looking face.

How to mitigate long-tailed distributions with a preassessment approach. Aesthetic datasets [53, 54] often exhibit long-tailed distributions, which leads to a model bias toward the majority of samples. Existing IAA models can assess the overall aesthetics of physique-related images, making them a useful reference for preliminary screening. To mitigate long-tailed distributions, we performed an initial assessment using EAT [22]. We categorized the aesthetic quality of physique images into three groups: good, fair, and poor, aiming to achieve an even distribution across these categories. Finally, we selected a balanced set of samples from each group.

A.2. Data Annotation

Annotators should provide assessments across the following dimensions:

1) The appearance score. The score evaluates the individual's ability to convey emotions and engage an audience through their physical presence, highlighting how their expressive qualities enhance the overall aesthetic impact of the images.

2) The health score. The score assesses the visual indicators of health and vitality as depicted in the images, focusing on attributes such as muscle definition, fitness, and balance that contribute to a striking aesthetic appeal.

3) The posture score. The score analyzes the individual's posture in the images, considering its aesthetic qualities, such as stability, creativity, and expressiveness. It measures how effectively the posture enhances visual elegance and artistic expression, showcasing the interplay between technical accuracy and emotional resonance.

The initial score range for the three dimensions is from 0 to 10. However, given the decisive role of the healthy score in the PhysiqueAA and the difficulty of assessing health solely from a visual perspective, we refined the healthy score S_h based on the objective BMI metric [55], which defines a BMI between 18.5 to 25 as healthy. Thus, we established the following formula:

$$S_h^{'} = S_h \cdot (1 - \sigma \cdot \mathbb{I}(\text{BMI} \notin [18.5, 25)) \,, \tag{4}$$

where $\mathbb{I}(\cdot)$ denotes the indicator function, taking a value of 1 when the condition is true (BMI $\notin [18.5, 25)$ and 0 otherwise, and σ is a scaling constant set to 0.7. Finally, The health score ranges from 0 to 10.

A.3. Data Partition

Our PhysiqueAA50K dataset consists of three subsets: PAA-16-personality (for stage 1 validation), PAA-3-User (for stage 2 validation), and 400 images (for stage 3 validation), as outlined below:

1) PAA-16-personality contains 40,000 images annotated with PhysiqueAA scores and surveys from three experts, representing ISFJ, ESFJ, and ISTJ personalities. Each expert's set of 40,000 images is split into 32,000 training images and 8,000 testing images.

2) PAA-3-User includes 10,000 images annotated with PhysiqueAA scores and surveys from three users (User-ISFJ, User-ESFJ, and User-ISTJ). Each user's set of 10,000 images is divided into 8,000 for training and 2,000 for testing;

3) In stage 3, user feedback is immediate and dynamic during each update epoch, making it impossible to establish a fixed testing dataset. Therefore, the 400 images for stage 3 validation cannot be divided into training and testing sets.

B. 16 MBTI Personality Types

The Myers-Briggs Type Indicator (MBTI) is a psychological tool developed based on Carl Jung's theory of personality types [28]. It identifies how individuals naturally perceive and make decisions, emphasizing inherent psychological preferences in these processes. The MBTI categorizes people into 16 distinct personality types, derived from four pairs of opposite preferences:

1) Introversion (***I***) vs. Extraversion (***E***): Focuses on whether you gain energy from the outside world or from within yourself;

2) Sensing (***S***) vs. INtuition (***N***): Describes whether you prefer concrete facts and details or look for patterns and possibilities;

3) Thinking (***T***) vs. Feeling (***F***): Refers to whether you make decisions based on logic and objectivity or personal values and emotions;

4) Judging (***J***) vs. Perceiving (***P***): Describes whether you prefer a structured, orderly lifestyle with careful planning to control your surroundings, or a flexible, spontaneous lifestyle focused on exploring and experiencing various ways of living without rigid control.

These preferences combine in different ways to form 16 unique types, each reflecting a stable pattern of cognitive and decision-making tendencies.

The user's personality can reflect their stable aesthetic subjectivity [5, 6, 56, 57]. We analyze three common MBTI personality types: ISFJ, ESFJ, and ISTJ. The following is a brief analysis of their characteristics:

- ISFJ (Introversion, Sensing, Feeling, Judging): ISFJs tend to value classic and timeless elements of physical aesthetics. They appreciate appearances that convey warmth, harmony, and subtlety, favoring styles that emphasize comfort and approachability. ISFJs are likely drawn to soft, natural designs that prioritize balance and evoke a sense of care and security.
- ESFJ (Extraversion, Sensing, Feeling, Judging): ESFJs evaluate physical aesthetics with an emphasis on social relevance and interpersonal harmony. Their preferences often align with popular trends and socially accepted norms. They tend to appreciate fashionable, detail-oriented styles that enhance social interactions and positively influence others' perceptions.
- ISTJ (Introversion, Sensing, Thinking, Judging): ISTJs prioritize practicality and precision when evaluating physical aesthetics. They tend to favor clean-cut, structured appearances that project order, reliability, and professionalism. Their preference leans towards classic and functional styles, prioritizing clarity and discipline over fleeting trends or extravagant designs.

There are two main reasons behind the rationale for linking MBTI types to physique preferences: 1) Previous studies, such as those in PAA [5, 6], have shown that personality traits reliably indicate stable aesthetic preferences. The widely recognized MBTI framework has also demonstrated a strong correlation with aesthetic preferences [57]. Therefore, physique preferences, as a type of aesthetic preference, are strongly correlated with MBTI types. 2) During data collection, we found that users with the same MBTI type tend to share similar physique preferences. For instance, around 80% of Extraverted (E) users preferred noticeable physiques, while only about 10% of Introverted (I) users did.

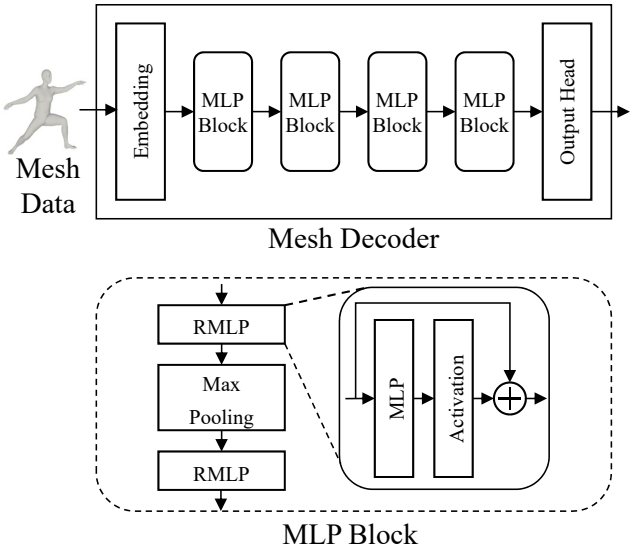


Figure 10. Comprehensive structure of the MPM.

C. More Details of PhysiqueFrame

C.1. Details of the Mesh Decoder

After processing the mesh data, we employ a mesh decoder (Fig. 10) for further analysis. This structure consists of an Embedding layer and four stages. Each stage is composed of two Residual Multilayer Perceptron (RMLP) blocks and a max-pooling layer alternately connected, which can be defined as:

$$f_{mpm} = \left\{ R_{2nd}\left(M\left(R_{1st}\left(v_{i,j}\right)\right)\right) \mid i \in \mathbb{R}^n, j \in \mathbb{R}^k, i \neq j \right\}, \tag{5}$$

The first residual block $R_{1st}(\cdot)$ learns shared weights from different local regions of the torso, enabling effective modeling of localized physique aesthetics. The max-pooling layer $M(\cdot)$ is used for feature aggregation. Then, the second residual block $R_{2nd}(\cdot)$ extracts deeply aggregated features f_{mpm}, which capture both local and global aspects of the physique. The alternating use of two RMLP blocks effectively balances the extraction of local features (such as limb details) and global features (such as overall posture). This

enables the model to capture the full characteristics of the physique while retaining crucial details.

C.2. Details of Five Specific Physique Factors

We analyzed literature on physique aesthetics [12, 13], somaesthetics [14, 58], and classical photography [59, 60]. Additionally, we referenced standards from real-world competitions, such as modeling [61], bodybuilding [62], and dancing [63]. These standards emphasize core elements such as posture, symmetry, and muscle definition, which are pivotal in evaluating physique aesthetics across different contexts. By synthesizing insights from these domains, we identified five factors highly relevant to PhysiqueAA: physique expressiveness, physique style, physique shape, physique posture, and personality traits.

The five specific physique factors differ from PhysiqueAA's dimensions (appearance, health, posture) by offering a completely objective framework to describe physique aesthetics. These factors facilitate the integration of user preference data, gathered from multimodal inputs such as surveys and feedback, into a structured and interpretable format. This approach establishes a coherent connection between subjective user preferences and the assessment of physique aesthetics.

C.3. Details of the feature aggregation from PENet and PANet

The features of two modules are concatenated and then passed through MLPs for non-linear mapping.

Metric		CNN-based models				Transformer-based models				Ours
		NIMA	BIAA	HGCN	TANet	MaxViT	TCFormer	EAT	Q-Align	
Appearance	$\mathcal{S}\uparrow$	.560	.503	.401	.548	.539	.405	.569	.549	**.649**
	$\mathcal{L}\uparrow$	.588	.531	.418	.564	.555	.428	.591	.571	**.678**
	$\mathcal{A}\uparrow$	.702	.684	.637	.714	.728	.643	.737	.716	**.744**
Health	$\mathcal{S}\uparrow$	.496	.425	.333	.454	.447	.345	.506	.482	**.540**
	$\mathcal{L}\uparrow$	.527	.457	.358	.492	.474	.382	.523	.521	**.567**
	$\mathcal{A}\uparrow$	.718	.674	.591	.700	.708	.598	.726	.716	**.740**
Posture	$\mathcal{S}\uparrow$	.584	.542	.467	.566	.577	.501	.604	.588	**.657**
	$\mathcal{L}\uparrow$	.631	.571	.503	.610	.625	.519	.624	.616	**.695**
	$\mathcal{A}\uparrow$	.707	.690	.603	.712	.703	.628	.744	.725	**.761**

Table 6. Comparison of the night models conducted on the User-ESFJ dataset for stage 2. We ***retrained the models to maximize performance,*** using the recommended parameters and settings.

Metric		CNN-based models				Transformer-based models				Ours
		NIMA	BIAA	HGCN	TANet	MaxViT	TCFormer	EAT	Q-Align	
Appearance	$\mathcal{S}\uparrow$	.527	.486	.409	.503	.502	.416	.532	.529	**.617**
	$\mathcal{L}\uparrow$	.548	.503	.425	.528	.524	.442	.560	.551	**.650**
	$\mathcal{A}\uparrow$	.708	.641	.630	.698	.659	.621	.692	.686	**.716**
Health	$\mathcal{S}\uparrow$	.499	.405	.318	.458	.451	.327	.508	.487	**.546**
	$\mathcal{L}\uparrow$	.525	.438	.334	.484	.477	.343	.529	.503	**.570**
	$\mathcal{A}\uparrow$	.704	.654	.555	.679	.671	.575	.705	.697	**.726**
Posture	$\mathcal{S}\uparrow$	.546	.487	.418	.488	.501	.426	.540	.518	**.597**
	$\mathcal{L}\uparrow$	.588	.523	.456	.532	.541	.462	.583	.566	**.641**
	$\mathcal{A}\uparrow$	.690	.610	.583	.669	.652	.589	.693	.678	**.701**

Table 7. Comparison of the night models conducted on the User-ISTJ dataset for stage 2. We ***retrained the models to maximize performance,*** using the recommended parameters and settings.

Metric		CNN-based models				Transformer-based models				Ours
		NIMA	BIAA	HGCN	TANet	MaxViT	TCFormer	EAT	Q-Align	
Appearance	$\mathcal{S}\uparrow$	.485	.489	.409	.509	.506	.422	.516	.509	**.580**
	$\mathcal{L}\uparrow$	.491	.501	.416	.512	.527	.447	.536	.514	**.603**
	$\mathcal{A}\uparrow$	.643	.654	.607	.661	.655	.610	.678	.674	**.706**
Health	$\mathcal{S}\uparrow$	.413	.415	.332	.428	.431	.352	.447	.455	**.539**
	$\mathcal{L}\uparrow$	.451	.457	.402	.463	.476	.405	.482	.494	**.599**
	$\mathcal{A}\uparrow$	.612	.622	.507	.608	.613	.539	.646	.635	**.687**
Posture	$\mathcal{S}\uparrow$	.493	.499	.377	.513	.518	.435	.524	.522	**.588**
	$\mathcal{L}\uparrow$	.518	.523	.396	.544	.557	.443	.561	.545	**.637**
	$\mathcal{A}\uparrow$	.729	.732	.680	.742	.743	.593	.752	.746	**.777**

Table 8. Comparisons of night models conducted on the PAA-16-personality (***ISFJ***) dataset for stage 1. We ***retrained the models for the best performance*** with the recommended parameters and settings.

Metric		CNN-based models				Transformer-based models				Ours
		NIMA	BIAA	HGCN	TANet	MaxViT	TCFormer	EAT	Q-Align	
Appearance	$\mathcal{S}\uparrow$	.447	.441	.401	.458	.469	.414	.473	.465	**.502**
	$\mathcal{L}\uparrow$	.456	.449	.412	.472	.481	.436	.486	.474	**.502**
	$\mathcal{A}\uparrow$	.633	.625	.605	.656	.652	.612	.671	.662	**.689**
Health	$\mathcal{S}\uparrow$	.385	.381	.322	.395	.400	.345	.420	.402	**.465**
	$\mathcal{L}\uparrow$	.423	.426	.375	.439	.451	.391	.478	.483	**.559**
	$\mathcal{A}\uparrow$	.577	.570	.516	.574	.592	.541	.603	.581	**.621**
Posture	$\mathcal{S}\uparrow$	.471	.479	.403	.483	.486	.425	.505	.497	**.545**
	$\mathcal{L}\uparrow$	.476	.485	.419	.497	.505	.436	.521	.505	**.551**
	$\mathcal{A}\uparrow$	.644	.655	.607	.660	.661	.615	.674	.663	**.698**

Table 9. Comparisons of night models on the PAA-16-personality (***ESFJ***) dataset for stage 1. We ***retrained the models for the best performance*** with the recommended parameters and settings.

Metric		CNN-based models				Transformer-based models				Ours
		NIMA	BIAA	HGCN	TANet	MaxViT	TCFormer	EAT	Q-Align	
Appearance	$\mathcal{S}\uparrow$	.517	.515	.449	.541	.514	.463	.555	.549	**.616**
	$\mathcal{L}\uparrow$	.531	.528	.466	.556	.531	.485	.578	.564	**.628**
	$\mathcal{A}\uparrow$	.684	.675	.635	.686	.683	.654	.701	.698	**.730**
Health	$\mathcal{S}\uparrow$	.501	.494	.389	.511	.503	.411	.536	.524	**.593**
	$\mathcal{L}\uparrow$	.536	.523	.408	.538	.530	.437	.561	.547	**.685**
	$\mathcal{A}\uparrow$	.615	.607	.551	.644	.635	.563	.672	.665	**.700**
Posture	$\mathcal{S}\uparrow$	.552	.563	.513	.590	.575	.535	.603	.605	**.660**
	$\mathcal{L}\uparrow$	.576	.585	.536	.623	.596	.561	.637	.640	**.683**
	$\mathcal{A}\uparrow$	.677	.682	.634	.695	.689	.656	.701	.704	**.732**

Table 10. Comparisons of night models on the PAA-16-personality (***ISTJ***) dataset for stage 1. We ***retrained the models for the best performance*** with the recommended parameters and settings.

D. More Details of Experiment

D.1. More Performance Comparison with PANet Alternatives

As shown in Table 6 and Table 7, our model achieves the best performance across all metrics when compared with alternative models on the User-ESFJ and User-ISTJ dataset. These results demonstrate our model's effectiveness in capturing physique aesthetics features.

Additionally, as shown in Table 8, Table 9, and Table 10, we compared our model with alternative models on PAA-16-personality datasets (ISFJ, ESFJ, ISTJ) for stage 1. Our model achieves state-of-the-art performance across all metrics.

D.2. More Quantitative Results

- Fig. 11 provides more qualitative examples illustrating the impact of personalized surveys on model predictions.
- Fig. 12 provides additional results from PANet Alterna-

tives. Our model predicts the results that are closest to the ground truth.

- Fig. 13 presents examples of images with the prediction results from PhysiqueFrame, illustrating the evaluation process from both a 3D (mesh) perspective and a posture perspective.

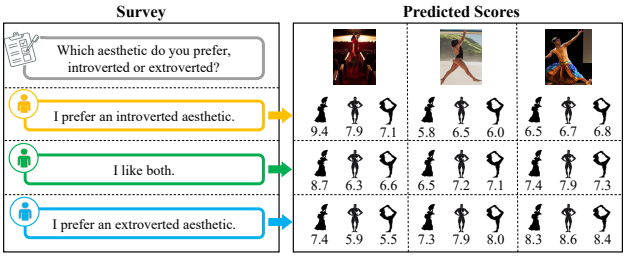


Figure 11. Impact of personalized surveys on PhysiqueAA predictions, with icons, and representing scores of appearance, health, and posture, respectively.

D.3. More ablation studies of different modules

As shown in Table 11, the scores from PENet alone perform worse than those from PANet + PENet. Besides, without GCN or Affine Geometry, the model's performance decreases.

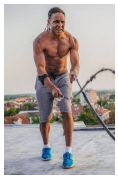
Method	Appearance		Health		Posture	
	$\mathcal{S}\uparrow$	$\mathcal{L}\uparrow$	$\mathcal{S}\uparrow$	$\mathcal{L}\uparrow$	$\mathcal{S}\uparrow$	$\mathcal{L}\uparrow$
w/o GCN in PAM	.675	.704	.604	.633	.643	.684
w/o Affine Geometry in MPM	.679	.711	.606	.630	.646	.688
only PENet	.687	.716	.611	.639	.657	.693
w/ preferences	**.699**	**.724**	**.625**	**.657**	**.664**	**.703**

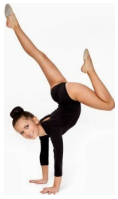
Table 11. Ablation studies on the performance of different modules, conducted on the User-ISFJ dataset for stage2.

E. Safeguards and Licenses for Existing Assets

The original owners of assets (e.g., code, data, models, personalized surveys) used in the paper are properly credited, and the licenses and terms of use are explicitly mentioned and properly respected, ensuring that there are no copyright issues and no risk of misuse.

Input	NIMA	BIAA	HGCN	TANet	MaxViT	TCFormer	EAT	Q-Align	Ours	Ground Truth
	6.9	6.8	7.3	7.6	7.8	6.8	7.5	8.2	9.1	9.2
	7.0	6.5	6.8	7.0	7.1	6.3	7.0	7.3	8.6	8.4
	7.3	7.5	6.8	7.7	7.5	5.5	7.6	8.1	9.5	9.5

Input	NIMA	BIAA	HGCN	TANet	MaxViT	TCFormer	EAT	Q-Align	Ours	Ground Truth
	6.0	4.2	4.6	5.6	5.5	4.8	6.3	5.4	7.5	7.4
	6.2	4.5	5.8	8.1	7.0	5.6	6.9	7.8	9.5	9.7
	4.7	4.1	3.9	4.8	5.0	5.5	5.2	4.3	6.1	6.3

Input	NIMA	BIAA	HGCN	TANet	MaxViT	TCFormer	EAT	Q-Align	Ours	Ground Truth
	6.5	6.0	6.6	7.3	6.8	6.3	7.3	7.1	8.0	8.2
	7.6	5.5	5.8	7.0	7.0	6.4	7.5	7.2	8.5	8.5
	5.2	7.1	6.9	8.3	7.3	6.8	8.2	7.7	9.4	9.5

Input	NIMA	BIAA	HGCN	TANet	MaxViT	TCFormer	EAT	Q-Align	Ours	Ground Truth
	6.0	5.8	3.9	5.7	6.0	4.8	5.4	6.1	7.6	7.8
	5.9	6.1	3.8	6.1	6.5	5.2	5.8	6.5	7.7	7.5
	4.8	4.9	4.8	3.7	5.4	4.6	5.1	4.8	6.3	6.1

Figure 12. The prediction results from the 9 models are displayed, with icons, and representing scores of appearance, health, and posture, respectively. We ***retrained the models on the User-ISFJ dataset for the best performance.***

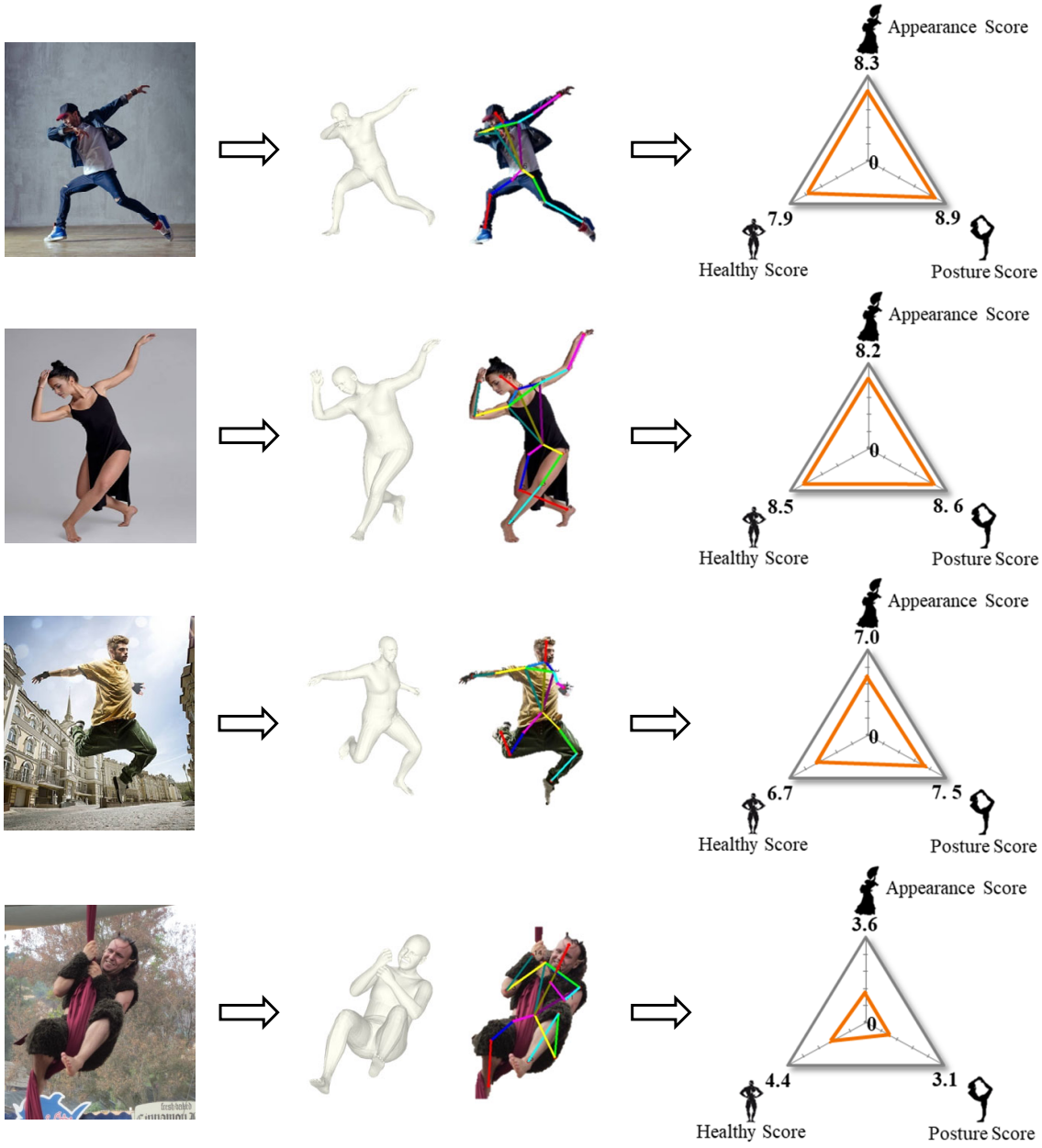


Figure 13. Examples of images are displayed with the prediction results from PhysiqueFrame. This section also illustrates the evaluation process from both a 3D (mesh) perspective and a posture perspective.