CHIAA-BENCH: ALIGNING CINEMATIC HUMAN IMAGE AESTHETIC ASSESSMENT WITH HUMAN AESTHETIC JUDGMENTS

Anonymous authors
Paper under double-blind review

ABSTRACT

Image aesthetic assessment has been widely studied for generic images, but remains underexplored for cinematic human images, whose aesthetic quality is mainly determined by human-centered cues such as facial clarity, expression, and atmosphere. To address this gap, we introduce a new task, **Cinematic Human Image Aesthetic Assessment (CHIAA)**, which aims to evaluate the aesthetic quality of cinematic human images from movies and TV series. We further present **CHIAA-Bench**, a new benchmark constructed from movie and TV frames as well as open-source cinematic datasets through human image filtering and a two-stage annotation pipeline with model scoring and expert refinement. Our annotation protocol covers four aesthetic dimensions and ten fine-grained sub-dimensions, producing interpretable five-level labels for cinematic human image aesthetics. Based on this benchmark, we develop **CineAesNet**, a lightweight model for efficient cinematic aesthetic assessment. Experiments show that CineAesNet consistently outperforms existing open-source aesthetic models and strong multimodal large language models on CHIAA-Bench while using substantially fewer parameters. In addition, our analysis shows that model-generated scores are useful for candidate filtering and low-resource warmup, whereas expert supervision remains essential for reliable learning of cinematic human aesthetics.

1 INTRODUCTION

Human images are among the most important visual assets in movies and TV series promotion. Their aesthetic quality can directly affect user attention and engagement. However, modern cinematic content contains massive numbers of video frames, making manual selection of high-quality human images inefficient and inconsistent. This motivates the need for automatic cinematic human image aesthetic assessment.

Prior work on Image Aesthetic Assessment (IAA) mainly focuses on generic natural images and evaluates aesthetics using cues such as composition, color harmony, and sharpness (Murray et al., 2012; Du et al., 2026; Cao et al., 2026; Yang et al., 2026). However, these approaches are not well aligned with cinematic human images. In movies and TV series frames, the human is usually the primary subject, and aesthetic judgment depends not only on global image quality but also on cues such as facial clarity and expression quality. Moreover, human images are shaped by narrative intent, shot design, and emotional context, which makes their aesthetic criteria different from those of generic photographs or portrait datasets. As a result, directly applying generic IAA methods to cinematic human image assessment often fails to capture the true promotional value and aesthetic quality of human images.

A key bottleneck in Human Image Aesthetic Assessment (HIAA) (Liao et al., 2025) is the lack of suitable data and evaluation criteria. As shown in Figure 1, existing human image datasets are mostly collected from web images or general photographs, which often provide limited visual quality and narrow viewpoints. In contrast, datasets in the cinematic domain are primarily designed for video understanding rather than aesthetic assessment (Song et al., 2024; Liu et al., 2026). Moreover, most HIAA works typically focus on relatively narrow aspects, such as facial attractiveness or portrait quality (Liang et al., 2018; Lebedeva et al., 2022), which are insufficient for cinematic aesthetic

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

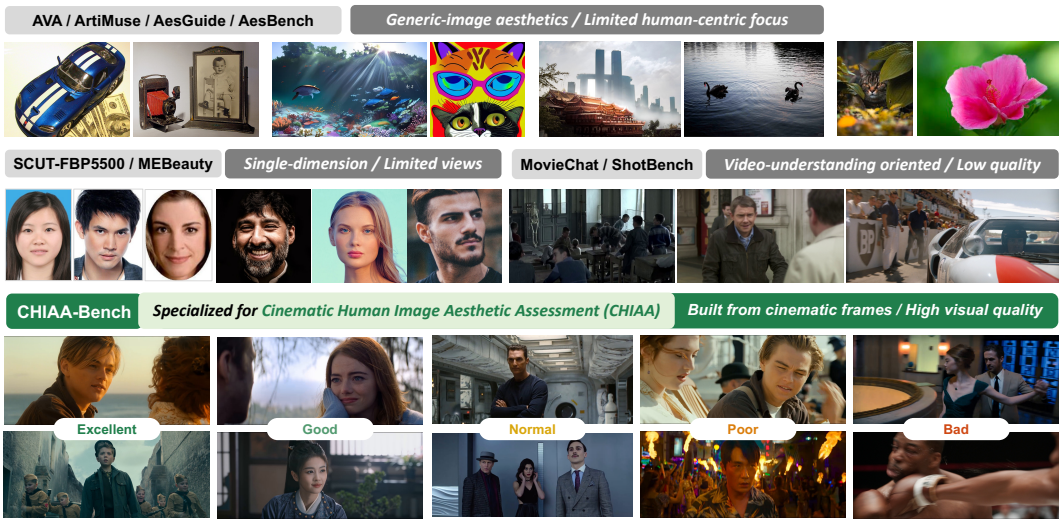

Figure 1: Comparison of CHIAA-Bench with related datasets. Existing datasets either focus on generic image aesthetics, provide only single-dimension evaluation, or are not designed for aesthetic assessment. In contrast, CHIAA-Bench is built from cinematic frames for Cinematic Human Image Aesthetic Assessment (CHIAA), featuring higher visual quality, and five-level annotations.

assessment. Although HumanAesExpert (Liao et al., 2025) provides a more comprehensive assessment of human images, it still does not explicitly consider cinematic factors such as emotion and atmosphere. As a result, existing benchmarks and criteria are still not well matched to the aesthetic assessment of cinematic human images.

To address these challenges, we collected high-quality frames from movies and TV series and introduce a new task, termed **Cinematic Human Image Aesthetic Assessment (CHIAA)**. To support this task, we construct a new benchmark, **CHIAA-Bench**. Specifically, we first build a clean candidate pool through a multi-stage filtering pipeline, including human detection, visibility filtering, and shot-level deduplication. We then develop an interpretable aesthetic assessment protocol with four major dimensions and ten fine-grained sub-dimensions, covering technical, facial, decorative, and atmospheric aesthetics. Based on this protocol, we adopt an annotation pipeline with two stages: candidate images are first scored using Multimodal Large Language Models (MLLM), existing aesthetic models, and quantitative visual cues, and are then refined by six cinematic experts under a five-level rating scheme (Wu et al., 2023). Each image is reviewed by at least two experts, and disagreements are resolved through discussion.

Based on CHIAA-Bench, we train a lightweight aesthetic assessment model, termed **CineAesNet**, for the CHIAA task. In addition to the in-domain subset built from movie and TV series, we further annotate samples from open-source cinematic datasets (Song et al., 2024; Liu et al., 2026) to construct an out-of-domain subset for evaluating cross-source generalization. Experimental results show that CineAesNet consistently outperforms existing image aesthetic models, and direct MLLM-based assessment, while remaining much more efficient at inference time.

In summary, our main contributions are as follows:

- We formulate **Cinematic Human Image Aesthetic Assessment (CHIAA)** as a new research task, filling a gap in existing aesthetic assessment by focusing on cinematic human images and establishing evaluation criteria that better align with human judgment.
- We construct **CHIAA-Bench**, the first human image aesthetic assessment dataset constructed from cinematic frames, and propose an annotation pipeline that combines multi-stage data filtering, model scoring, and expert refinement.
- We develop a lightweight CHIAA model, **CineAesNet**, and show that it outperforms existing image aesthetic methods and direct MLLM-based assessment on both in-domain and out-of-domain benchmarks.

2 Related Work

Generic Image Aesthetic Assessment. Image aesthetic assessment (IAA) has been extensively studied in generic natural images (Daryanavard Chounchenani et al., 2025). Representative benchmarks such as AVA and AADB (Murray et al., 2012; Kong et al., 2016) formulate the task as classification, regression, or score distribution prediction, mainly focusing on global visual factors such as composition, color, exposure, and sharpness. Existing methods have evolved from CNN-based predictors and distribution learning models, such as NIMA (Talebi & Milanfar, 2018), to more recent vision-language and multimodal approaches (Ke et al., 2023; Ma et al., 2026). However, these methods are primarily designed for general image aesthetics and are less suitable for human image aesthetics where human cues are more important.

Human Image Aesthetic Assessment. Existing work on human image aesthetics mainly focuses on measurable factors such as facial attractiveness, portrait quality, and facial detail preservation (Lebedeva et al., 2022; Sun et al., 2024). Representative datasets include SCUT-FBP5500 for facial attractiveness prediction (Liang et al., 2018) and PIQ23 for portrait quality assessment (Chahine et al., 2023). More recently, HumanAesExpert introduces Human Image Aesthetic Assessment (HIAA) with the HumanBeauty dataset and fine-grained human-related annotations (Liao et al., 2025). However, these datasets are mostly based on web portraits or ordinary photographs, and thus do not adequately reflect cinematic aesthetics. In contrast, CHIAA-Bench is built from movie and TV series, capturing richer lighting, composition, and emotional atmosphere.

MLLM-based aesthetic understanding and expert benchmarks. Recent work has explored Multimodal Large Language Models (MLLMs) and expert annotation for richer aesthetic understanding (Wu et al., 2023; Dai et al., 2025). AesBench (Huang et al., 2024b), AesExpert (Huang et al., 2024a), and ArtiMuse (Cao et al., 2026) support fine-grained aesthetic analysis with expert-aware or attribute-aware supervision, while other studies investigate aesthetic guidance, cropping, and ranking (Du et al., 2026; Yang et al., 2026). These works demonstrate the value of interpretable dimensions and model-assisted annotation, but mainly focus on generic image aesthetics. Moreover, the high inference cost of MLLMs limits their practicality for large-scale scoring of video frames. In contrast, we build a lightweight predictor for scalable assessment.

3 Method

3.1 Problem Formulation

We introduce **Cinematic Human Image Aesthetic Assessment (CHIAA)**, which aims to predict the overall aesthetic quality of cinematic human images. Given an input image x from movies or TV series, the goal is to learn a mapping function $f(\cdot)$ that outputs its aesthetic score:

$$s = f(x), \quad s \in \mathbb{R}, \tag{1}$$

where a higher score indicates better aesthetic quality. Unlike generic image aesthetic assessment, CHIAA focuses on cinematic human images, whose aesthetic quality depends on both global visual factors and human-centered cues. The aesthetic assessment protocol is introduced in Section 3.2.3.

3.2 CHIAA-Bench Construction

To support CHIAA, we construct **CHIAA-Bench**, the first benchmark for cinematic human image aesthetic assessment. As shown in Figure 2, the construction process involves three stages: Datasets Collection, Human Image Filtering, and Human-Aligned Aesthetic Annotation.

3.2.1 Datasets Collection

CHIAA-Bench is primarily constructed from movies and TV series collected from the Youku platform[1]. To ensure diversity of data, we carefully select 20 movies and 20 TV series as data sources covering diverse genres, countries, eras, and visual styles. For each TV series, we use the first three

[1] https://www.youku.com/

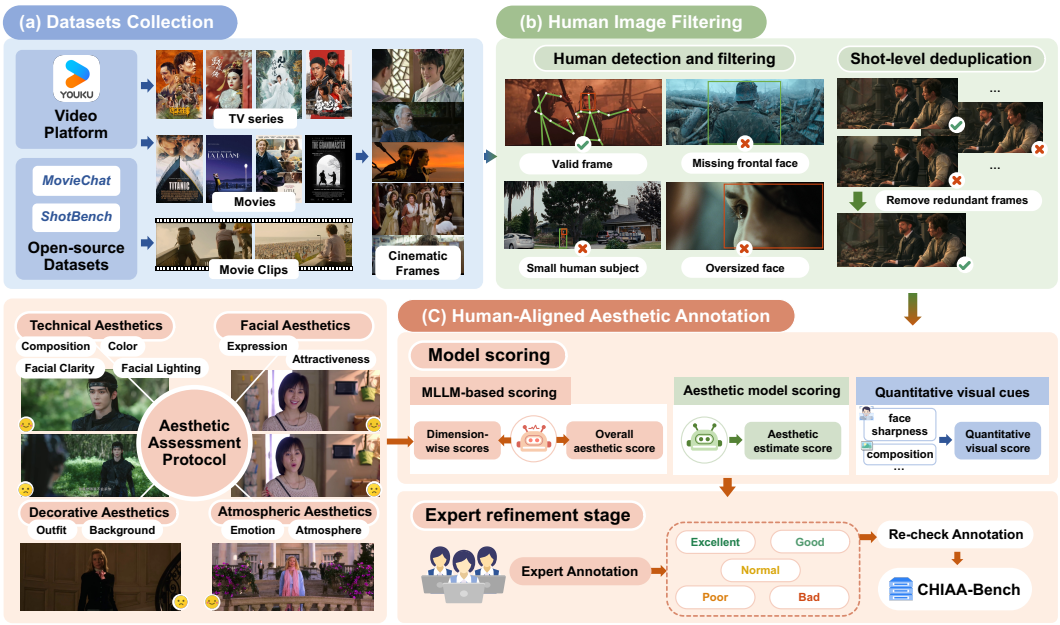

Figure 2: Pipeline for constructing CHIAA-Bench. (a) Frames are collected from the video platform and open-source cinematic datasets. (b) Human detection and shot-level deduplication are applied to obtain valid cinematic human images. (c) The remaining images are annotated through model scoring and expert refinement under the aesthetic assessment protocol to produce five-level labels.

episodes to make its duration comparable to that of a single movie, resulting in 80 long videos in total. In addition, we further collect cinematic videos from the open-source datasets MovieChat (Song et al., 2024) and ShotBench (Liu et al., 2026) to increase the diversity of data sources.

During video processing, we first manually remove irrelevant segments such as opening titles, ending credits, and advertisements, and retain only the main content of each movie or TV series. We then extract frames from all videos at 1 FPS, yielding a large pool of candidate cinematic frames.

3.2.2 Human Image Filtering

Raw cinematic frames contain many unsuitable samples for CHIAA, including frames without visible humans, repetitive neighboring frames, and noisy transition frames. To address this issue, we design a two-stage filtering pipeline to extract high-quality cinematic human images.

(1) Human detection and filtering. We first perform human filtering to retain frames with sufficiently visible and informative human subjects. Specifically, we apply YOLO11-pose (Fei, 2025) for human detection and pose estimation, and keep frames whose main detected person occupies at least 10% of the image area and contains at least 5 visible keypoints. We further apply RetinaFace (Deng et al., 2020) to verify facial visibility and discard frames without a reliable face. The remaining frames provide clear human visual cues for subsequent annotation.

(2) Shot-level deduplication. Frame sampling produces many highly similar neighboring frames. We therefore perform shot-level deduplication by comparing nearby frames with visual similarity cues to group samples. Detailed settings are provided in the appendix A.1.2. For each group, we retain only one frame. To further improve diversity and source balance, we cap samples from each video, retaining fewer than 900 frames per movie and fewer than 300 frames per TV episode.

3.2.3 Human-Aligned Aesthetic Annotation

We conduct annotation in two stages: model scoring and expert refinement. The first stage serves to evaluate pure model-based inference and assist manual annotation, while the second stage produces the final high-quality labels.

(1) Interpretable aesthetic assessment protocol. Before annotation, we establish an interpretable aesthetic assessment protocol tailored to CHIAA, so that model scoring can perform fine-grained evaluation from understandable aspects and expert annotation can follow the same unified reference. As shown in Figure 2, the protocol is organized into four major dimensions with ten fine-grained sub-dimensions. Specifically, **a) technical aesthetics** includes *composition*, *color*, *facial clarity*, and *facial lighting*, focusing on whether the image presents the human subject in a clear, balanced, and visually pleasing manner. **b) facial aesthetics** includes *expression* and *attractiveness*, capturing how facial appearance and expressiveness contribute to overall aesthetic preference. **c) decorative aesthetics** includes *outfit* and *background*, evaluating the visual appeal of styling and surrounding scene elements. **d) atmospheric aesthetics** includes *emotion* and *atmosphere*, reflecting the overall mood and affective tension. Detailed definitions of all sub-dimensions are provided in the appendix.

(2) Model scoring stage. We assign model scores to candidate images. This stage is used to examine the effectiveness of model-based inference, screen samples, and assist subsequent manual annotation and re-check. Specifically, we combine three types of automatic scoring signals.

a) MLLM-based scoring. We use Multimodal Large Language Models (MLLMs) to perform human image aesthetic assessment on candidate images. Specifically, we employ Qwen3.8-27B (Qiu et al., 2026) and Gemma4-31B (Team et al., 2026a) as the scoring models. Each model produces both a direct overall aesthetic score and a set of scores for the ten fine-grained sub-dimensions, where the latter offers a more detailed characterization of different human-centered aesthetic factors.

b) Existing aesthetic model scoring. We further use existing open-source aesthetic models, including OneAlign (Wu et al., 2023), HumanAesExpert (Liao et al., 2025), and FGAesQ (Yang et al., 2026), to provide assessments of overall aesthetics. These models serve as a supplement to the MLLM-based understanding of holistic aesthetic quality.

c) Quantitative visual cues. In addition to learning-based models, we construct six quantitative visual cues based on the technical aesthetics and facial aesthetics dimensions. These cues provide interpretable references for several human factors, such as using human area ratio to reflect composition and eye openness to reflect expression.

The final score is obtained by combining the above three types of signals with their respective weights. Detailed fusion settings and discussions are provided in the appendix. Based on the resulting model scores, we sample images from different score intervals to balance the sample distribution and construct a candidate set for subsequent expert annotation.

(3) Expert refinement stage. We invite six cinematic experts to perform fine-grained manual annotation on the candidate images. Following a unified annotation guideline, the experts assign each image to one of five categories:

$$y \in \{\texttt{bad}, \texttt{poor}, \texttt{normal}, \texttt{good}, \texttt{excellent}\}. \tag{2}$$

Compared with continuous-score annotation, this discrete grading scheme is more stable for manual labeling and more convenient for quality control (Wu et al., 2023). Each image is independently reviewed by at least two experts. If the reviewers agree, their label is provisionally adopted. Otherwise, the image is sent for further discussion until consensus is reached. After the initial annotation, we further conduct a re-check stage based on model scores, focusing on boundary samples near category transitions to improve annotation consistency and quality. Finally, we construct CHIAA-Bench with 20.8k annotated samples, including 16.1k samples from video platform, 4k samples from MovieChat, and 700 samples from ShotBench.

3.3 CINEAESNET

Based on CHIAA-Bench, we further train a lightweight cinematic human image aesthetic assessment model, termed **CineAesNet**, for efficient large-scale scoring of cinematic video frames.

Model architecture. CineAesNet consists of a visual backbone and an aesthetic prediction head. Given an input image, we apply standard preprocessing and feed it into an EVA02-448 backbone (Fang et al., 2024) with a patch embedding layer and 12 Transformer blocks. The backbone features are aggregated by combining the `[CLS]` token with pooled patch features, and the resulting

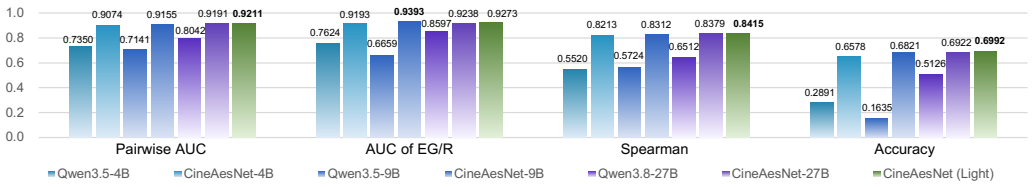

Figure 3: Effect of CHIAA-Bench on adapting large models for CHIAA.

representation is passed to a lightweight fully connected head. The head predicts a distribution over ordered score bins, from which the continuous aesthetic score is computed as the expectation.

Training data. CineAesNet is trained only on the annotated samples from the video platform, together with 5k low-quality samples selected during the previous filtering stage, resulting in about 18k training images in total. The five-level labels are mapped to five discrete score values in $[0, 1]$ for training. The remaining 3.1k platform samples are reserved for in-domain evaluation. The annotated samples from MovieChat and ShotBench are used only for testing, providing out-of-domain evaluation for cross-source generalization.

4 Experiments

4.1 Experimental Setup

Evaluation Data. As there is currently no suitable open-source benchmark for human image aesthetics, we conduct all experiments on CHIAA-Bench. The benchmark contains three test subsets: 3.1k samples from the video platform, 4k samples from MovieChat, and 700 samples from ShotBench. Unless otherwise stated, results are reported on the video platform subset by default.

Compared Methods. We compare CineAesNet with several representative open-source aesthetic assessment models, including ArtiMuse (Cao et al., 2026), HumanAesExpert (Liao et al., 2025), FGAesQ (Yang et al., 2026), and OneAlign (Wu et al., 2023). For these methods, we directly use their predicted aesthetic scores for evaluation. We further compare with several strong MLLMs, including HumanVLM (Dai et al., 2025), LLaVA-OneVision-2 (An et al., 2026), Qwen3.6-27B, Qwen3.8-27B (Yang et al., 2025; Qiu et al., 2026), Gemma4-31B (Team et al., 2026a), Qwen3.6-Plus, Qwen3.7-Plus, Qwen3.8-Max, and Kimi-K3 (Team et al., 2026b). We prompt these models to assess the aesthetic quality of each image and convert their outputs into scalar scores for evaluation.

Evaluation Metrics. We evaluate all methods using multiple metrics from complementary perspectives. *Pairwise AUC* (Hanley & McNeil, 1982) evaluates pairwise ranking ability, while boundary-based *AUC* metrics such as *EG/R* and *EG/N* assess discrimination between high-quality and lower-quality samples. *Spearman* (Spearman, 1904) and *Pearson* (Pearson, 1895) measure the agreement between predicted scores and expert judgments. We further report *Accuracy* and *Quadratic Weighted Kappa (QWK)* (Cohen, 1960) to measure category-level correctness and ordinal agreement under the five-level labeling protocol. For all metrics, higher values indicate better performance. Detailed definitions are provided in the appendix.

4.2 Main Results on Video Platform Series

Table 1 reports the main results on the video platform subset of CHIAA-Bench. Several observations can be drawn.

First, **existing open-source aesthetic models remain limited on CHIAA-Bench**. OneAlign is the strongest baseline, followed by FGAesQ and HumanAesExpert, while ArtiMuse performs the worst. This pattern suggests that CHIAA requires not only generic aesthetic perception, but also sensitivity to human-centered cues and cinematic context. Accordingly, models with stronger human-related or alignment-oriented design perform better than generic aesthetic predictors. However, even the best open-source aesthetic model still leaves a large gap to the best overall performance.

Table 1: Main results on the video platform subset of CHIAA-Bench. The best results are highlighted in bold. PwAUC denotes Pairwise AUC, EG/R denotes AUC of *excellent* + *good* vs. the rest, EG/N denotes AUC of *excellent* + *good* vs. *normal*, Sp denotes Spearman, Pe denotes Pearson, Acc denotes accuracy, and QWK denotes quadratic weighted kappa.

Method	#Params	PwAUC	EG/R	EG/N	Sp	Pe	Acc	QWK
Open-source aesthetic models								
ArtiMuse	8B	0.5131	0.5650	0.5882	0.0268	0.0767	0.3106	0.0134
HumanAesExpert-8B	8B	0.7017	0.7466	0.6352	0.4354	0.4365	0.4236	0.3956
FGAesQ	235.2M	0.7311	0.7821	0.6741	0.4952	0.4901	0.4634	0.4698
OneAlign	8B	0.7494	0.8192	0.7107	0.5312	0.5350	0.4776	0.5118
Open-source / proprietary MLLMs								
HumanVLM	8B	0.5480	0.5816	0.5595	0.2333	0.2244	0.0811	0.0625
LLaVA-OneVision-2	8B	0.6919	0.7731	0.6866	0.4558	0.4366	0.2800	0.3774
Qwen3.6-27B	27B	0.7920	0.8441	0.7436	0.6265	0.5883	0.5249	0.5617
Qwen3.8-27B	27B	0.8042	0.8597	0.7613	0.6512	0.6284	0.5126	0.5638
Gemma4-31B	31B	0.8063	0.8404	0.7300	0.6485	0.6567	0.5237	0.6083
Qwen3.6-Plus	-	0.7814	0.8349	0.7335	0.5935	0.5909	0.4907	0.5680
Qwen3.7-Plus	-	0.7924	0.8323	0.7235	0.6158	0.6168	0.4975	0.5639
Qwen3.8-Max	2.4T	0.7872	0.8585	0.7738	0.6034	0.5924	0.5189	0.5740
Kimi-K3	2.8T	0.8242	0.8542	0.7419	0.6767	0.6643	0.5374	0.6404
CineAesNet	86.4M	**0.9211**	**0.9273**	**0.8400**	**0.8415**	**0.8398**	**0.6992**	**0.8130**

Second, **MLLM-based assessment provides stronger baselines than most existing aesthetic models, but their performance is still limited**. Among the tested MLLMs, Kimi-K3 achieves the strongest overall results, while Gemma4-31B and Qwen3.8-27B are also competitive on several metrics. This indicates that MLLMs possess nontrivial aesthetic reasoning ability for cinematic human images. However, their predictions remain less stable and less aligned with expert judgments than those required for precise aesthetic assessment.

Finally, **CineAesNet delivers the strongest overall performance**. It surpasses all compared methods on every reported metric, including correlation, pairwise ranking, high-quality discrimination, and category agreement. Moreover, it achieves these results with only 86.4M parameters, far smaller than the 8B–31B open-source models and trillion-scale proprietary systems considered here. This confirms the advantage of task-specific modeling for cinematic human image aesthetic assessment, where alignment with expert judgment is more important than general multimodal capability.

4.3 Results on MovieChat and ShotBench

To further evaluate the cross-source generalization ability of our model, we conduct experiments on the MovieChat and ShotBench subsets of CHIAA-Bench, both of which come from open-source cinematic data sources. As reported in Table 2, CineAesNet achieves the best performance on both subsets across all reported metrics. On MovieChat, it reaches 0.9206 in pairwise AUC and 0.8365 in Spearman correlation, outperforming the strongest baseline Kimi-K3 by 0.0484 and 0.0767, respectively. On ShotBench, it further obtains 0.9016 in pairwise AUC and 0.8114 in Spearman correlation, surpassing Kimi-K3 by 0.0638 and 0.1087. We further observe that most methods perform worse on ShotBench than on MovieChat. This is likely because ShotBench is built from frames of Oscar-winning films, which generally present stronger cinematographic aesthetics and higher overall visual quality, making fine-grained quality assessment more difficult than on MovieChat. Nevertheless, CineAesNet remains consistently superior on both benchmarks, further demonstrating its robustness and transferability under cross-source distribution shifts.

4.4 Analysis of Model Scoring

We study the role of model scoring in CHIAA from three aspects: whether model scores can provide reliable aesthetic assessment, whether they can improve learning under limited human annotation, and how different components contribute to the quality of model scores.

Table 2: Results on the MovieChat and ShotBench subsets of CHIAA-Bench.

	MovieChat				ShotBench			
Method	PwAUC	EG/R	Sp	Acc	PwAUC	EG/R	Sp	Acc
Open-source aesthetic models								
ArtiMuse	0.5940	0.6573	0.2026	0.3635	0.4408	0.4678	-0.1281	0.2943
HumanAesExpert-8B	0.7644	0.7428	0.5645	0.4813	0.6184	0.6362	0.2609	0.3643
FGAesQ	0.7657	0.7297	0.5689	0.4823	0.6697	0.7185	0.3694	0.3986
OneAlign	0.8319	0.8345	0.6920	0.5435	0.7093	0.7459	0.4515	0.4117
Open-source / proprietary MLLMs								
HumanVLM	0.5498	0.5449	0.2345	0.0922	0.5318	0.5294	0.1730	0.1677
LLaVA-OneVision-2	0.7520	0.7436	0.5824	0.4828	0.6748	0.7053	0.4343	0.3398
Qwen3.6-27B	0.8590	0.8044	0.7457	0.5767	0.8259	0.8077	0.6984	0.4586
Qwen3.8-27B	0.8651	0.8080	0.7536	0.6008	0.8326	0.8049	0.7057	0.5229
Gemma4-31B	0.8637	0.8364	0.7499	0.6142	0.8228	0.8071	0.6734	0.5314
Qwen3.6-Plus	0.8605	0.8253	0.7417	0.6080	0.8158	0.7859	0.6675	0.4914
Qwen3.7-Plus	0.8565	0.8208	0.7411	0.5168	0.8080	0.7672	0.6580	0.4700
Qwen3.8-Max	0.8566	0.8472	0.7326	0.6022	0.7970	0.7867	0.6253	0.4914
Kimi-K3	0.8722	0.8691	0.7598	0.6300	0.8378	0.8180	0.7027	0.5150
CineAesNet	**0.9206**	**0.8805**	**0.8365**	**0.7017**	**0.9016**	**0.8840**	**0.8114**	**0.6357**

Table 3: Effect of model scoring and human annotations on CHIAA-Bench. *Model-score Inference* uses model scores directly for evaluation; *Model-score Pretraining* uses model scores for pretraining; *2k Human Labels* trains on 2,000 expert-labeled samples; *Pretraining + 2k Human Labels* adds model-score pretraining before 2k samples fine-tuning; and *Full Training* uses the full expert-refined training set.

Setting	PwAUC	EG/R	EG/N	Sp	Pe	Acc	QWK
Qwen3.8-27B Inference	0.8042	0.8597	0.7613	0.6512	0.6284	0.5126	0.5638
Model-score Inference	0.8449	0.8871	0.7893	0.7106	0.6872	0.5777	0.6751
Model-score Pretraining	0.8500	0.8791	0.7736	0.7209	0.6743	0.5931	0.6899
2k Human Labels	0.8369	0.8670	0.7489	0.7009	0.6896	0.5669	0.6685
Pretraining + 2k Human Labels	0.9000	0.9154	0.8208	0.8086	0.8033	0.6585	0.7810
Full Training (CineAesNet)	**0.9211**	**0.9273**	**0.8400**	**0.8415**	**0.8398**	**0.6992**	**0.8130**

As shown in Table 3, direct *Model-score Inference* already provides a strong baseline, showing that combining multiple models and quantitative cues produces more stable and reliable scores than any single MLLM alone, similar to aggregating judgments from multiple experts. At the same time, *Model-score Pretraining* substantially improves the low-resource setting: compared with training on *2k Human Labels* alone, *Pretraining + 2k Human Labels* yields clear gains across all metrics, showing that model scoring provides an effective initialization for cinematic aesthetic assessment. However, both inference and pretraining-based settings still remain below *Full Training*, indicating that existing models and quantitative cues are still insufficient to fully align with expert judgments, and that expert annotation remains essential for achieving the best human-aligned performance.

To better understand where the effectiveness of model scoring comes from, we further ablate the model score inference pipeline in Table 4. Among all components, removing the MLLM causes the largest performance drop, indicating that it is the core source of aesthetic reasoning in the pipeline. In addition, removing the 10-dimensional protocol or quantitative cues also degrades performance, suggesting that fine-grained scoring is valuable not only for interpretability, but also for helping the pipeline better capture the aesthetic criteria of cinematic human images.

4.5 Adapting Large Models with CHIAA-Bench

To further verify the value of our data, we adapt several large models on CHIAA-Bench by adding a regression head and a classification head for aesthetic prediction. As shown in Figure 3, all adapted

Table 4: Ablation study on the model scoring inference pipeline.

Setting	PwAUC	EG/R	EG/N	Sp	Pe	Acc	QWK
Model-score Inference	**0.8449**	**0.8871**	0.7893	**0.7106**	**0.6872**	0.5777	0.6751
w/o 10-dim Protocol	0.8403	**0.8871**	**0.7904**	0.7015	**0.6872**	**0.5792**	**0.6762**
w/o MLLM	0.7807	0.8356	0.7278	0.5940	0.5872	0.4994	0.5602
w/o Aesthetic Model	0.8434	0.8834	0.7850	0.7086	0.6809	0.5742	0.6720
w/o Quantitative Cues	0.8399	0.8780	0.7749	0.7018	0.6799	0.5679	0.6671



Figure 4: Qualitative comparison of the top-5 highest-scoring frames selected by different models from candidate frames. 🙂 denotes good frames, while 😟 denotes average or poor ones.

models outperform the base model by a large margin, showing that our annotated data can effectively align large models with human standards for CHIAA. Interestingly, the lightweight *CineAesNet* achieves the best overall performance, We conjecture that large models require additional cross-modal alignment, while our lightweight model directly learns the mapping from image features to aesthetic scores and is therefore better suited to capture aesthetic patterns in cinematic human images.

4.6 Qualitative Analysis

To better understand the behavior of CineAesNet, we provide a qualitative comparison in Figure 4. We sample candidate frames from a movie and show the top-5 frames selected by each model based on their predicted aesthetic scores. Compared with baseline methods, CineAesNet more consistently identifies frames with clearer human subjects, richer facial expression, and a more coherent cinematic atmosphere, indicating that CineAesNet better captures cinematic aesthetic cues than existing baseline methods.

5 Conclusion

In this paper, we introduce Cinematic Human Image Aesthetic Assessment (CHIAA), a new task for evaluating the aesthetic quality of cinematic human images, and construct CHIAA-Bench, a benchmark with rigorous filtering and expert annotation to better reflect human aesthetic standards. We also propose CineAesNet, a lightweight model tailored for this task. Experiments show that CineAesNet consistently outperforms existing open-source aesthetic models and strong multimodal large language models on CHIAA-Bench, and also reveals the gap between current models and expert judgments in cinematic human aesthetics. Our results further show that model scoring is effective in the inference setting, while our aesthetic image data helps enhance the aesthetic assessment ability of large models. We hope that CHIAA-Bench and CineAesNet will support future research on human-aligned aesthetic assessment.

AI USE STATEMENT

In this work, we use generative AI tools for limited assistance with manuscript editing, including language polishing, readability improvement, grammar checking, and minor coding support during experimentation. We do not use generative AI tools to generate the main research ideas, propose hypotheses, design the core methodology, conduct the experiments, produce the reported results, or determine the scientific conclusions of this work.

Most of the manuscript is written by the authors. All AI-assisted text and code suggestions are reviewed, corrected, and validated by the authors before use. All experiments are manually conducted by the authors, and all reported numbers are based on real experimental results. Where large-model scoring is involved in the data pipeline, this usage is explicitly described in the paper. We take full responsibility for the final content of this work, including any text or artifacts produced with the assistance of generative AI, and we affirm the authenticity and reproducibility of the reported research.

ETHICS STATEMENT

This work involves image data collected in part from publicly accessible videos on online video platforms and in part from existing datasets, including MovieChat and ShotBench. We use these data solely for academic research and do not intend to infringe upon the copyright of any cinematic video content. For all existing datasets used in this work, we follow their respective terms of use and data access requirements. In addition, all example images shown in the paper that are derived from video platforms have been marked and processed for presentation purposes, such as cropping and other necessary visual modifications. The aesthetic assessment in our pipeline is designed to better align model predictions with human judgments and to assist annotators in efficiently identifying high-quality cinematic human images; it is not intended for discriminatory profiling or harmful decision-making. In future data release, we will carefully consider copyright and licensing issues. We plan to prioritize releasing the evaluation subsets and annotations derived from ShotBench and MovieChat, while subset collected from video platforms will be released after copyright and compliance issues are properly addressed.

REPRODUCIBILITY STATEMENT

To support reproducibility, we plan to release the CineAesNet model, training code, and evaluation code. We also plan to release the evaluation subsets and annotations derived from ShotBench and MovieChat. The subset of CHIAA-Bench collected from video platforms will be released after copyright-related issues are resolved. In the paper and supplementary materials, we describe the dataset construction pipeline, model design, training settings, and evaluation protocol to facilitate reproduction of our experiments.

REFERENCES

Xiang An, Yin Xie, Feilong Tang, Yunyao Yan, Huajie Tan, Didi Zhu, Changrui Chen, Xiuwei Zhao, Bin Qin, Kaicheng Yang, et al. Llava-onevision-2: Towards next-generation perceptual intelligence. *arXiv preprint arXiv:2605.25979*, 2026.

Shuo Cao, Nan Ma, Jiayang Li, Xiaohui Li, Lihao Shao, Kaiwen Zhu, Yu Zhou, Yuandong Pu, Jiarui Wu, Jiaquan Wang, et al. Artimuse: Fine-grained image aesthetics assessment with joint scoring and expert-level understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 15313–15322, 2026.

Nicolas Chahine, Stefania Calarasanu, Davide Garcia-Civiero, Theo Cayla, Sira Ferradans, and Jean Ponce. An image quality assessment dataset for portraits. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9968–9978, 2023.

Jacob Cohen. A coefficient of agreement for nominal scales. *Educational and psychological measurement*, 20(1):37–46, 1960.

Dawei Dai, Long Xu, Yutang Li, Yuanhui Zhang, and Shuyin Xia. Humanvlm: Foundation for human-scene vision-language model. *Information Fusion*, 123:103271, 2025.

Maedeh Daryanavard Chounchenani, Asadollah Shahbahrami, Reza Hassanpour, and Georgi Gaydadjiev. Deep learning based image aesthetic quality assessment-a review. *ACM Computing Surveys*, 57(7):1–36, 2025.

Jiankang Deng, Jia Guo, Evangelos Ververas, Irene Kotsia, and Stefanos Zafeiriou. Retinaface: Single-shot multi-level face localisation in the wild. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5202–5211. IEEE, 2020.

Tianxiang Du, Hulingxiao He, and Yuxin Peng. Venus: Benchmarking and empowering multimodal large language models for aesthetic guidance and cropping. *arXiv preprint arXiv:2602.23980*, 2026.

Yuxin Fang, Quan Sun, Xinggang Wang, Tiejun Huang, Xinlong Wang, and Yue Cao. Eva-02: A visual representation for neon genesis. *Image and Vision Computing*, 149:105171, 2024.

He Fei. Human posture detection model based on improved yolo11-pose. In *2025 IEEE 5th International Conference on Electronic Technology, Communication and Information (ICETCI)*, pp. 1420–1424. IEEE, 2025.

James A Hanley and Barbara J McNeil. The meaning and use of the area under a receiver operating characteristic (roc) curve. *Radiology*, 143(1):29–36, 1982.

Yipo Huang, Xiangfei Sheng, Zhichao Yang, Quan Yuan, Zhichao Duan, Pengfei Chen, Leida Li, Weisi Lin, and Guangming Shi. Aesexpert: Towards multi-modality foundation model for image aesthetics perception. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pp. 5911–5920, 2024a.

Yipo Huang, Quan Yuan, Xiangfei Sheng, Zhichao Yang, Haoning Wu, Pengfei Chen, Yuzhe Yang, Leida Li, and Weisi Lin. Aesbench: An expert benchmark for multimodal large language models on image aesthetics perception. *arXiv preprint arXiv:2401.08276*, 2024b.

Junjie Ke, Keren Ye, Jiahui Yu, Yonghui Wu, Peyman Milanfar, and Feng Yang. Vila: Learning image aesthetics from user comments with vision-language pretraining. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10041–10051. IEEE, 2023.

Shu Kong, Xiaohui Shen, Zhe Lin, Radomir Mech, and Charless Fowlkes. Photo aesthetics ranking network with attributes and content adaptation. In *European conference on computer vision*, pp. 662–679. Springer, 2016.

Irina Lebedeva, Yi Guo, and Fangli Ying. Mebeauty: a multi-ethnic facial beauty dataset in-the-wild. *Neural Computing and Applications*, 34(17):14169–14183, 2022.

Lingyu Liang, Luojun Lin, Lianwen Jin, Duorui Xie, and Mengru Li. Scut-fbp5500: A diverse benchmark dataset for multi-paradigm facial beauty prediction. In *2018 24th International conference on pattern recognition (ICPR)*, pp. 1598–1603. IEEE, 2018.

Zhichao Liao, Xiaokun Liu, Wenyu Qin, Qingyu Li, Qiulin Wang, Pengfei Wan, Di Zhang, Long Zeng, and Pingfa Feng. Humanaesexpert: Advancing a multi-modality foundation model for human image aesthetic assessment. *arXiv preprint arXiv:2503.23907*, 2025.

Hongbo Liu, Jingwen He, Yi Jin, Dian Zheng, Yuhao Dong, Fan Zhang, Ziqi Huang, Yinan He, Weichao Chen, Yu Qiao, et al. Shotbench: Expert-level cinematic understanding in vision-language models. *Advances in Neural Information Processing Systems*, 38:129987–130019, 2026.

Xingyuan Ma, Shuai He, Anlong Ming, Haobin Zhong, and Huadong Ma. Regression over classification: Assessing image aesthetics via multimodal large language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 7827–7835, 2026.

Naila Murray, Luca Marchesotti, and Florent Perronnin. Ava: A large-scale database for aesthetic visual analysis. In *2012 IEEE conference on computer vision and pattern recognition*, pp. 2408–2415. IEEE, 2012.

Karl Pearson. Note on regression and inheritance in the case of two parents. *proceedings of the royal society of London*, 58:240–242, 1895.

Zihan Qiu, Zekun Wang, Xiao Li, Yanpeng Li, Yang Xu, Yixuan Wang, Huaqing Zhang, Rui Men, Bochao Mao, Chengruidong Zhang, et al. On the design of qwen3. 8-next architecture: Evaluation, efficiency, and training stability. *arXiv preprint arXiv:2608.30320*, 2026.

Enxin Song, Wenhao Chai, Guanhong Wang, Yucheng Zhang, Haoyang Zhou, Feiyang Wu, Haozhe Chi, Xun Guo, Tian Ye, Yanting Zhang, et al. Moviechat: From dense token to sparse memory for long video understanding. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 18221–18232. IEEE, 2024.

Charles Spearman. The proof and measurement of association between two things. *The American Journal of Psychology*, 15(1):72–101, 1904.

Wei Sun, Weixia Zhang, Yanwei Jiang, Haoning Wu, Zicheng Zhang, Jun Jia, Yingjie Zhou, Zhongpeng Ji, Xiongkuo Min, Weisi Lin, et al. Dual-branch network for portrait image quality assessment. *arXiv preprint arXiv:2405.08555*, 2024.

Hossein Talebi and Peyman Milanfar. Nima: Neural image assessment. *IEEE transactions on image processing*, 27(8):3998–4011, 2018.

Gemma Team, Sherif El Abd, Vaibhav Aggarwal, Robin Algayres, Alek Andreev, Olivier Bachem, Ian Ballantyne, Cormac Brick, Victor Cărbune, Michelle Casbon, et al. Gemma 4 technical report. *arXiv preprint arXiv:2607.02770*, 2026a.

Kimi Team, Tongtong Bai, Yifan Bai, Yiping Bao, Jianfeng Cai, Xinyuan Cai, Peizhou Cao, Yuxuan Cao, Ziwei Chai, Y Charles, et al. Kimi k3: Open frontier intelligence. *arXiv preprint arXiv:2607.24653*, 2026b.

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.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025.

Zhichao Yang, Jianjie Wang, Zhixianhe Zhang, Pangu Xie, Xiangfei Sheng, Pengfei Chen, and Leida Li. Fine-grained image aesthetic assessment: Learning discriminative scores from relative ranks. *arXiv preprint arXiv:2603.03907*, 2026.

A Details of CHIAA-Bench Construction

A.1 Human Image Filtering

A.1.1 Human detection and filtering

To remove frames that are unsuitable for human image aesthetic assessment, we perform human detection and visibility filtering on all candidate frames. Specifically, we apply YOLO11-pose (Fei, 2025) to detect the main person and estimate body keypoints. A frame is retained only when the main detected person occupies at least 10% of the image area and contains at least 5 visible keypoints, where a keypoint is regarded as visible if its confidence score is above 0.5. In addition, the person detection confidence threshold is set to 0.25. These criteria help remove frames without valid human subjects, with overly small persons, or with severe occlusion.

We further apply RetinaFace (Deng et al., 2020) to verify facial visibility and discard frames without a reliable face detection. In our implementation, the face detection threshold is set to 0.5. This step further excludes back-view, side-back, and other samples lacking clear facial cues. After this filtering stage, the remaining frames provide sufficiently clear human-centered visual information for subsequent aesthetic annotation.

A.1.2 SHOT-LEVEL DEDUPLICATION

To reduce redundancy caused by temporal continuity in cinematic videos, we perform shot-level deduplication on the filtered candidate frames. Specifically, we compare temporally nearby frames using a combination of complementary visual similarity cues, including HSV color histogram similarity, perceptual dHash distance, and grayscale correlation. Frames that are sufficiently similar within a local temporal window are grouped as duplicates from the same shot, while nearly identical frames across a larger temporal gap are also merged using a stricter similarity rule. In our implementation, the local comparison window is set to 100 frames, and the default similarity thresholds are 0.78 for histogram correlation, 30 for dHash Hamming distance, and 0.5 for grayscale correlation, with a stricter cross-window rule of 0.95 correlation and 10 dHash distance for nearly identical frames.

After grouping duplicate frames, we retain only one representative frame from each group. We further apply source-level sampling caps to improve diversity and prevent a small number of long videos or frequently appearing characters from dominating the candidate pool. Specifically, we keep fewer than 900 frames for each movie and fewer than 300 frames for each TV episode. This process effectively reduces redundancy while improving source balance and visual diversity in the final candidate set.

A.2 AESTHETIC ASSESSMENT PROTOCOL

CHIAA-Bench adopts an interpretable aesthetic assessment protocol organized into four major dimensions and ten fine-grained sub-dimensions, illustrated with representative examples in Figure 2. All sub-dimensions are scored independently on a 0–10 scale, and the assessment is human-centered: backgrounds, scenery, and person-less shots are not rewarded on their own, and when no person is present or the person is too small, back-facing, heavily occluded, or otherwise unrecognizable, the scores of the relevant dimensions are lowered accordingly. We detail the focus of each sub-dimension below.

(1) Technical aesthetics. This dimension captures how well the frame technically presents the human subject.

a) Composition attends to the spatial arrangement of the person in the frame, including the person's screen occupancy, position, cropping, occlusion, whether the subject stands out, and, for multi-person shots, whether the relationships and hierarchy among people are clear.

b) Color attends to the harmony and unity of the overall palette and how it flatters the person, with particular emphasis on the naturalness of skin tone, white-balance stability, the color relationship between the person and the environment, and the presence of a coherent, cinematic color style.

c) Facial clarity attends to how sharply the face is rendered, including whether the focus falls on the face, whether facial-part details are discernible, and whether blur, noise, motion smear, or compression artifacts are present; for multi-person shots, it is judged with respect to the primary person the lens targets.

d) Facial lighting attends to the illumination and exposure of the face, including whether facial brightness is appropriate, whether shadows are flattering and well shaped, whether highlight and shadow detail is retained, and whether the light sculpts the face (e.g., catchlights and rim light).

(2) Facial aesthetics. This dimension captures how the person's facial appearance and expressiveness contribute to overall aesthetic preference.

e) Expression attends to whether the captured expression and pose are attractive and worth freezing as a still, including eye state, mouth shape, head angle, and body posture, and whether the frame avoids discarded moments such as blinking or distorted expressions.

f) Attractiveness attends to the on-camera visual appeal of the person within this particular frame, including facial harmony, face shape and contour, and overall photogenic quality. It evaluates the appearance presented in the image rather than making a judgment of the real person's objective looks.

(3) Decorative aesthetics. This dimension captures the visual appeal of styling and surrounding scene elements.

g) Outfit attends to the completeness and quality of the person's costume, hair, makeup, and accessories, and whether the overall styling is coherent, photogenic, and consistent with the character.

h) Background attends to how the background, scene, props, and effects serve the person, including whether they enhance subject prominence, layering, and narrative context, rather than distracting from or overpowering the person.

(4) Atmospheric aesthetics. This dimension captures the overall mood and affective tension of the image.

i) Emotion attends to whether the emotion conveyed by the person's expression is clearly readable and whether it can resonate with or move the viewer, including its intensity and dramatic impact.

j) Atmosphere attends to the overall completeness, sense of story, and promotional appeal of the image as a character still, integrating person, scene, and cinematography under the criterion of whether the atmosphere strengthens the person's attractiveness.

A.3 Quantitative visual cues.

We construct several lightweight quantitative visual cues to provide interpretable references for human-centered aesthetic factors. These cues are mainly derived from face detection, face crops, and facial landmarks, and are normalized to a common 0–10 range before fusion.

For *composition*, we use the face boxes derived from YOLO-pose as the main human-centered object and compute three indicators: (1) a *rule-of-thirds* score based on the normalized distance between the mean face center and the nearest rule-of-thirds point; (2) a *visual-balance* score based on the horizontal offset of the mean face center from the vertical image centerline; and (3) a *face-ratio* score based on the total face area relative to the image area, where a moderate face proportion is preferred and overly small or overly large faces are penalized. These indicators are combined to form the final composition cue.

For *face sharpness*, we crop each detected face, convert it to grayscale, resize it to a fixed height to reduce scale effects, and compute the variance of the Laplacian. For images with multiple faces, we use the maximum value to represent the clearest visible face, and then apply robust normalization to obtain a 0–10 score.

For *face lighting*, we compute the mean grayscale intensity of each face crop and measure its distance to an empirically chosen ideal brightness level. Scores decrease for both under-exposed and over-exposed faces. This cue mainly reflects overall facial brightness and does not explicitly model more complex lighting properties such as contrast, cast shadows, or highlight clipping.

For *expression*, we use MediaPipe FaceLandmarker to estimate facial landmarks and compute the Eye Aspect Ratio (EAR) for both eyes. The average EAR is used to reflect eye openness, and the maximum value across faces is taken for multi-person images. The resulting score favors faces with open and readable eye states, although it may underestimate images with side faces, occlusion, sunglasses, or intentionally closed-eye expressions.

For *attractiveness*, we apply a pretrained SCUT-FBP5500 facial attractiveness predictor on the cropped face images. For multi-face images, we take the highest predicted attractiveness score and then normalize it to the 0–10 range. This cue is intended only as an auxiliary reference for on-image facial appeal rather than a definitive judgment of personal appearance.

A.4 Details of Model-Score Fusion

We fuse three types of automatic scoring signals to obtain the final model score used for candidate filtering and annotation assistance:

$$S_{\text{fused}} = 0.4\, S_{\text{10-dim}} + 0.5\, S_{\text{model}} + 0.1\, S_{\text{cue}}, \tag{3}$$

where all component scores are normalized before fusion.



Figure 5: More examples of the video platform subset of CHIAA-Bench.

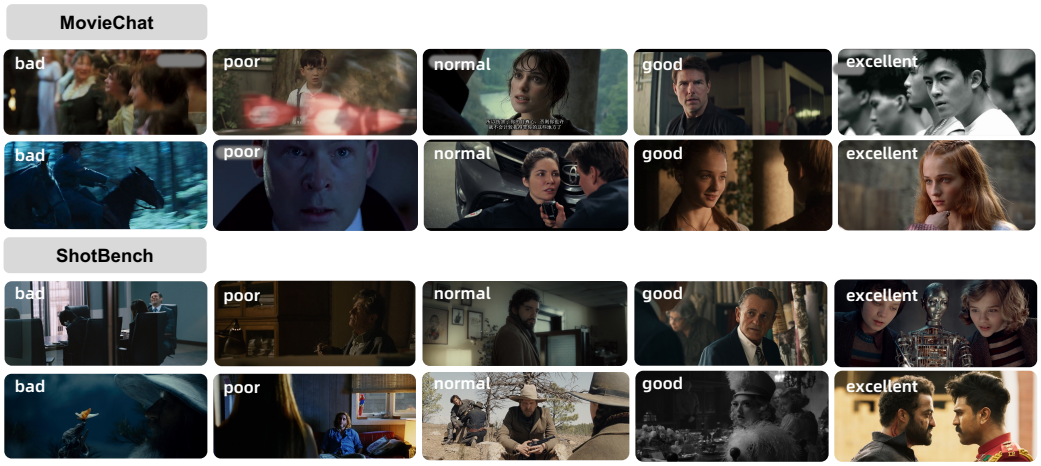

Figure 6: More examples of the Moviechat and ShotBench subset of CHIAA-Bench.

The 10-dimensional score is computed as a weighted combination of the ten fine-grained sub-dimensions:

$$S_{\text{10-dim}} = \sum_{i=1}^{10} w_i s_i, \tag{4}$$

with weights $\{0.100, 0.045, 0.250, 0.145, 0.140, 0.095, 0.045, 0.045, 0.065, 0.070\}$ for composition, color, face sharpness, face lighting, expression, attractiveness, outfit, background, emotion, and atmosphere, respectively.

The multi-model overall score is computed from five models:

$$S_{\text{model}} = 0.35\, S_{\text{Qwen3.8}} + 0.15\, S_{\text{HumanAes}} + 0.05\, S_{\text{FGAesQ}} + 0.15\, S_{\text{OneAlign}} + 0.30\, S_{\text{Gemma4}}, \tag{5}$$

where the larger weights are assigned to the stronger MLLM-based predictors.

The quantitative cue score is computed from five human-centered visual cues:

$$S_{\text{cue}} = 0.40\, S_{\text{sharpness}} + 0.20\, S_{\text{attractiveness}} + 0.10\, S_{\text{composition}} + 0.25\, S_{\text{expression}} + 0.05\, S_{\text{lighting}}. \tag{6}$$

These weights are selected through empirical exploration rather than manually fixed heuristics. For example, Appendix D.1 shows that weighted 10-dimensional scoring is more effective than equal-weight averaging, supporting the use of non-uniform aggregation.

Table 5: Category distribution of the training and test subsets of CHIAA-Bench. "Low-quality (as *bad*)" refers to the 5k discarded samples that are merged into the training set and labeled as *bad*.

Subset	*bad*	*poor*	*normal*	*good*	*excellent*	Total
Train / Video Platform	2,905	2,918	5,573	1,510	94	13,000
Train / Low-quality	5,000	–	–	–	–	5,000
Test / Video Platform	918	861	1,183	192	14	3,168
Test / MovieChat	1,023	1,483	1,275	210	9	4,000
Test / ShotBench	108	176	300	102	14	700

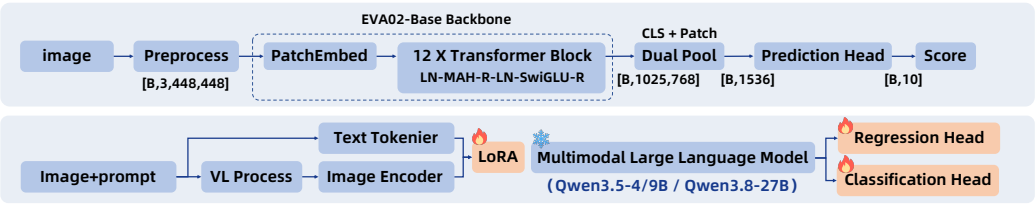

Figure 7: Architecture of CineAesNet and the training architecture for adapting MLLMs to CHIAA.

A.5 Source Videos and Data Scale

Source videos. The in-domain portion of CHIAA-Bench is collected from 40 popular titles on the Youku platform. The TV-series sources are all Chinese productions, dominated by costume dramas and supplemented by fantasy, modern urban, romance, and suspense themes. The movie sources consist of 10 Chinese films and 10 films from other countries and regions, covering multiple genres such as drama, romance, crime, comedy, and science fiction. Detailed information about the specific movie and TV titles will be disclosed after copyright and licensing issues are properly addressed. We have presented more cases in Figure 5 and Figure 6.

Data scale and category distribution. After multi-stage filtering and expert refinement, the in-domain video-platform subset contains 13.0k annotated training samples and 3,168 test samples. We additionally include 5k discarded low-quality samples as the *bad* category, yielding about 18k training images in total. Roughly 3k annotated training samples are selected through model-score-based stratified sampling to increase the proportion of high-quality images. The annotated MovieChat (4k) and ShotBench (700) subsets are used only for out-of-domain evaluation. The per-category distribution of all subsets is summarized in Table 5.

Label-to-score mapping. For training, the five aesthetic categories are mapped to five discrete target scores in $[0, 1]$,

$$\{\texttt{bad}, \texttt{poor}, \texttt{normal}, \texttt{good}, \texttt{excellent}\} \rightarrow \{0,\ 0.25,\ 0.5,\ 0.75,\ 1\}, \tag{7}$$

so that the targets follow the ordinal structure of the annotation scheme while allowing a non-uniform spacing that matches human perception across categories.

B Details of Experimental Setup

B.1 Implementation Details

As shown in Figure 7, CineAesNet is instantiated with EVA02-448 (Fang et al., 2024) and trained on 448×448 inputs. We map the five-level expert labels to discrete score targets in $[0, 1]$ and optimize the model with a weighted sum of EMD loss, PLCC loss, and pairwise rank loss, using weights 1.0, 0.5, and 0.3, respectively. We use AdamW with a base learning rate of 2×10^{-5}, a $10\times$ larger learning rate for the prediction head, weight decay 0.05, and layer-wise decay 0.75. The model is trained for 10 epochs with a batch size of 16, using 10% linear warmup, cosine decay, `bfloat16` mixed precision, gradient clipping of 1.0, and a drop path rate of 0.2. We use 10% of the training data for validation and select the best checkpoint by label-level Spearman correlation.

B.2 Evaluation Metrics

Pairwise AUC. Pairwise AUC measures whether a method can correctly rank image pairs with different aesthetic labels. Specifically, we construct sample pairs from images with different expert categories, and regard a pair as correctly ranked if the image with the higher ground-truth label also receives a higher predicted score. This metric reflects the pairwise ranking ability of a method over the full benchmark.

AUC for quality boundaries. To evaluate discrimination performance at practically important quality levels, we further compute several binary AUC metrics under different category partitions. These include distinguishing *excellent* + *good* from the rest, and distinguishing *excellent* + *good* from *normal*. These metrics specifically measure how well a method separates high-quality cinematic human images from medium- or low-quality ones.

Spearman. Spearman's rank correlation coefficient measures the consistency between the predicted ranking and the ground-truth ranking. Given n samples, let r_i and $\hat{r}_i$ denote the ranks of the ground-truth label and the predicted score of the i-th sample, respectively. Spearman is computed as

$$\rho = 1 - \frac{6 \sum_{i=1}^{n} (r_i - \hat{r}_i)^2}{n(n^2 - 1)}. \tag{8}$$

A higher Spearman value indicates that a method better preserves the relative ordering of image aesthetic quality.

Pearson. Pearson correlation coefficient measures the linear correlation between the predicted scores and the target scores. Given ground-truth targets y_i and predictions $\hat{y}_i$, Pearson is defined as

$$r = \frac{\sum_{i=1}^{n} (y_i - \bar{y})(\hat{y}_i - \bar{\hat{y}})}{\sqrt{\sum_{i=1}^{n} (y_i - \bar{y})^2} \sqrt{\sum_{i=1}^{n} (\hat{y}_i - \bar{\hat{y}})^2}}, \tag{9}$$

where $\bar{y}$ and $\bar{\hat{y}}$ are the means of the ground-truth and predicted scores. A higher Pearson value indicates stronger linear agreement with expert judgments.

Accuracy. We convert predicted continuous scores into five discrete categories using predefined thresholds, and compute the classification accuracy against the expert labels. This metric reflects the direct category-level prediction correctness under the five-level annotation protocol.

Quadratic Weighted Kappa. Quadratic weighted kappa measures the agreement between predicted labels and expert labels while accounting for their ordinal structure. Compared with plain accuracy, it assigns larger penalties to predictions that are farther away from the ground-truth category, and is therefore more suitable for ordered aesthetic labels ranging from *bad* to *excellent*.

For all metrics, higher values indicate better performance.

B.3 Comparison with Related Datasets

Table 6 compares CHIAA-Bench with several representative aesthetic assessment datasets. Unlike existing benchmarks that mainly focus on human portraits, general photography, artistic images, or AIGC content, CHIAA-Bench is specifically built for cinematic human images from movie and TV frames as well as cinematic open-source datasets, making it more suitable for our task. HumanBeauty (Liao et al., 2025) is closer to our setting in subject matter but is not publicly available for open evaluation, while datasets such as AesGuide (Du et al., 2026), FGAesthetics (Yang et al., 2026), ArtiMuse (Cao et al., 2026), and AesBench (Huang et al., 2024b) are more oriented toward general-purpose aesthetic assessment. Therefore, our experiments are mainly conducted on the three subsets of CHIAA-Bench, namely the video-platform, MovieChat, and ShotBench subsets, all annotated or refined by cinematic experts under a unified protocol.

Table 6: Comparison of CHIAA-Bench with representative aesthetic assessment datasets.

Dataset	Year	Size	Domain	Data Source	Annotators
AesBench (Huang et al., 2024b)	2024	2.8k	Natural, artistic, and AIGC images	Open-source dataset filtering	36 experts
ArtiMuse (Cao et al., 2026)	2025	10k	Artistic photography + AIGC	Expert collection and filtering	Experts
AesGuide (Du et al., 2026)	2026	10k	General photography	Web photos + photographers	20 experts
FGAesthetics (Yang et al., 2026)	2026	32k	General images + AIGC	Open-source datasets + AIGC	10 experts
HumanBeauty (Liao et al., 2025)	2025	108k	Human portraits	Open-source filtering + manual collection	300+ volunteers
CHIAA-Bench	2026	20.8k	Cinematic human images	Video platform + MovieChat + ShotBench	6 cinematic experts

C More Experiments

C.1 The complete results of the MovieChat and ShotBench

Due to space limitations, the main paper reports only a subset of the evaluation metrics on MovieChat and ShotBench. Here we provide the complete results for these two out-of-domain subsets in Table 7.

C.2 Effect of Prompting Strategies for Model Scoring

We further compare different prompting strategies for model-based scoring, including direct overall scoring and 10-dimensional scoring variants. As shown in Table 8, the effect of prompting strategy is model-dependent. For Qwen3.8-27B, 10-dimensional scoring improves most major metrics over direct overall scoring, including Spearman, Pearson, Pairwise AUC, accuracy, and QWK, indicating that decomposing aesthetic judgment into fine-grained dimensions can help the model produce more precise scores. For Gemma4-31B, however, the advantage of 10-dimensional scoring is less consistent, suggesting that different models benefit differently from structured prompting.

Another clear observation is that using equal weights over the 10 dimensions consistently degrades performance for both models. This shows that different aesthetic dimensions do not contribute equally to the final judgment, and simply averaging them cannot accurately reflect human aesthetic preference in cinematic human images. Overall, these results suggest that fine-grained prompting can improve model-based scoring, but the contribution of each dimension should be treated differently rather than assumed to be uniform.

C.3 Results on AVA and SCUT-FBP5500.

We further report results on two existing benchmarks, AVA (Murray et al., 2012) and SCUT-FBP5500 (Liang et al., 2018), in Tables 9 and 10. On AVA, OneAlign achieves the best performance, which is expected since it is explicitly designed and trained for general aesthetic assessment, and therefore transfers better to generic aesthetic benchmarks. In contrast, HumanAesExpert and the SCUT-FBP5500 attractiveness model perform poorly on AVA, indicating limited generalization beyond their more specific training objectives. CineAesNet remains competitive with Qwen3.8-27B, showing that it still preserves a certain level of general aesthetic capability. On SCUT-FBP5500, the dedicated attractiveness model and HumanAesExpert obtain higher scores, which is expected because both models have been trained on SCUT-FBP5500 and thus are better matched to this benchmark. However, in practice, high-scoring samples in SCUT-FBP5500 are not always high-quality images in an aesthetic sense, as some blurry or visually imperfect images may still receive high scores. This suggests that models such as HumanAesExpert may exhibit a certain degree of benchmark-specific overfitting by relying too heavily on narrow cues such as attractiveness. By

Table 7: Complete results on the MovieChat and ShotBench subsets of CHIAA-Bench.

Method	PwAUC	EG/R	E/R	Sp	Pe	Acc	QWK
MovieChat							
Open-source aesthetic models							
ArtiMuse	0.5940	0.6573	0.6400	0.2026	0.2484	0.3635	0.1814
HumanAesExpert-8B	0.7644	0.7428	0.7482	0.5645	0.5525	0.4813	0.5079
FGAesQ	0.7657	0.7297	0.8557	0.5689	0.5505	0.4823	0.5256
OneAlign	0.8319	0.8345	0.8614	0.6920	0.6769	0.5435	0.6342
Open-source / proprietary MLLMs							
HumanVLM	0.5498	0.5449	0.6283	0.2345	0.1917	0.0922	0.0743
LLaVA-OneVision-2	0.7520	0.7436	0.8473	0.5824	0.5465	0.4828	0.4598
Qwen3.6-27B	0.8590	0.8044	0.8306	0.7457	0.7267	0.5767	0.6667
Qwen3.8-27B	0.8651	0.8080	0.8984	0.7536	0.7462	0.6008	0.6281
Gemma4-31B	0.8637	0.8364	0.8804	0.7499	0.7640	0.6142	0.6704
Qwen3.6-Plus	0.8605	0.8253	0.9266	0.7417	0.7405	0.6080	0.6782
Qwen3.7-Plus	0.8565	0.8208	0.8574	0.7411	0.7395	0.5168	0.6309
Qwen3.8-Max	0.8566	0.8472	**0.9338**	0.7326	0.7262	0.6022	0.6622
Kimi-K3	0.8722	0.8691	0.9230	0.7598	0.7599	0.6300	0.7164
CineAesNet	**0.9206**	**0.8805**	0.8466	**0.8365**	**0.8294**	**0.7017**	**0.7791**
ShotBench							
Open-source aesthetic models							
ArtiMuse	0.4408	0.4678	0.6401	-0.1281	-0.0405	0.2943	-0.1094
HumanAesExpert-8B	0.6184	0.6362	0.4645	0.2609	0.2862	0.3643	0.2514
FGAesQ	0.6697	0.7185	0.9005	0.3694	0.3712	0.3986	0.3622
OneAlign	0.7093	0.7459	**0.9325**	0.4515	0.4837	0.4117	0.4358
Open-source / proprietary MLLMs							
HumanVLM	0.5318	0.5294	0.5076	0.1730	0.1489	0.1677	0.0802
LLaVA-OneVision-2	0.6748	0.7053	0.7321	0.4343	0.4290	0.3398	0.3080
Qwen3.6-27B	0.8259	0.8077	0.8569	0.6984	0.6606	0.4586	0.5553
Qwen3.8-27B	0.8326	0.8049	0.8721	0.7057	0.7152	0.5229	0.6377
Gemma4-31B	0.8228	0.8071	0.9251	0.6734	0.7245	0.5314	0.6794
Qwen3.6-Plus	0.8158	0.7859	0.7777	0.6675	0.6799	0.4914	0.6462
Qwen3.7-Plus	0.8080	0.7672	0.7231	0.6580	0.6945	0.4700	0.6382
Qwen3.8-Max	0.7970	0.7867	0.8556	0.6253	0.6609	0.4914	0.6178
Kimi-K3	0.8378	0.8180	0.8477	0.7027	0.7334	0.5150	0.6833
CineAesNet	**0.9016**	**0.8840**	0.8873	**0.8114**	**0.8044**	**0.6357**	**0.7799**

Table 8: Effect of different prompting strategies for model-based scoring. We compare direct overall scoring and 10-dimensional scoring variants on two multimodal large language models. The best results are highlighted in bold.

Model	Strategy	Sp	Pe	PwAUC	EG/R	EG/N	Acc	QWK
Qwen3.8-27B	Overall scoring	0.6512	0.6284	0.8042	**0.8597**	**0.7613**	0.5126	0.5638
	10-dim scoring	**0.6800**	**0.6597**	**0.8270**	0.8487	0.7383	**0.5578**	**0.6421**
	10-dim equal scoring	0.6194	0.6164	0.7952	0.8322	0.7290	0.5215	0.5821
Gemma4-31B	Overall scoring	**0.6485**	**0.6567**	**0.8063**	**0.8404**	**0.7300**	**0.5237**	0.6083
	10-dim scoring	0.6339	0.5910	0.7981	0.8387	0.7231	0.4894	**0.6117**
	10-dim equal scoring	0.5832	0.5737	0.7716	0.8178	0.7114	0.4665	0.5646

contrast, CHIAA-Bench evaluates human image aesthetics from multiple dimensions rather than a single factor such as attractiveness. Therefore, lower performance on these existing datasets does not necessarily imply weaker model capability; instead, it reflects that current benchmarks

Table 9: Results on the AVA dataset.

Model	Pearson	Spearman	Kendall Tau
OneAlign	0.8870	0.8820	0.7040
SCUT-FBP5500	0.0500	0.0571	0.0382
HumanAesExpert	-0.0620	-0.0980	-0.0660
Qwen3.8-27B	0.3320	0.3680	0.2700
Gemma4-31B	0.4710	0.4890	0.3580
CineAesNet	0.3360	0.3290	0.2220

Table 10: Results on the SCUT-FBP5500 dataset.

Model	Pearson	Spearman	Kendall Tau
OneAlign	0.5210	0.4600	0.3190
SCUT-FBP5500	0.9019	0.8909	0.7148
HumanAesExpert	0.8080	0.8540	0.6480
Qwen3.8-27B	0.7030	0.6640	0.5210
Gemma4-31B	0.7640	0.7290	0.5840
CineAesNet	0.4160	0.4230	0.2890

and their associated training objectives may not fully capture human judgments of cinematic human image aesthetics. In comparison, CHIAA-Bench is better aligned with expert human evaluation, and CineAesNet performs more favorably in this setting.

C.4 Generalization to an external animation domain.

We further perform a qualitative generalization test on external animation videos, whose visual appearance differs substantially from the movie and TV series data used in CHIAA-Bench. For each method, we select the top-3 highest-scoring frames after shot-level deduplication to reduce redundancy. As shown in Figure 8, CineAesNet still tends to retrieve reasonable high-quality character frames under this noticeable domain shift. By contrast, some baseline methods show limited robustness. For instance, HumanAesExpert may assign high scores to poor transition frames such as nearly black screens. This suggests that existing huma image aesthetic models do not always generalize well to domains with substantially different visual characteristics, while CineAesNet remains comparatively more stable.

C.5 Effect of low-quality samples.

We study the effect of adding the 5k low-quality samples as auxiliary *bad* training data. As shown in Figure 9, the resulting model assigns clearly lower scores to bad-quality inputs, while the scores on normal samples are largely unaffected. Table 11 further shows that adding these samples reduces the mean, variance, and standard deviation of the predicted scores on the bad-sample test set. These results indicate that low-quality samples help the model better suppress noisy and undesirable frames, improving robustness and generalization without harming normal aesthetic assessment.

D Prompt

D.1 Detailed Scoring Prompt

We report below the full prompt used for MLLM-based aesthetic scoring, which operationalizes the protocol in Appendix A.2. The prompt is model-agnostic and is issued to each scoring MLLM.

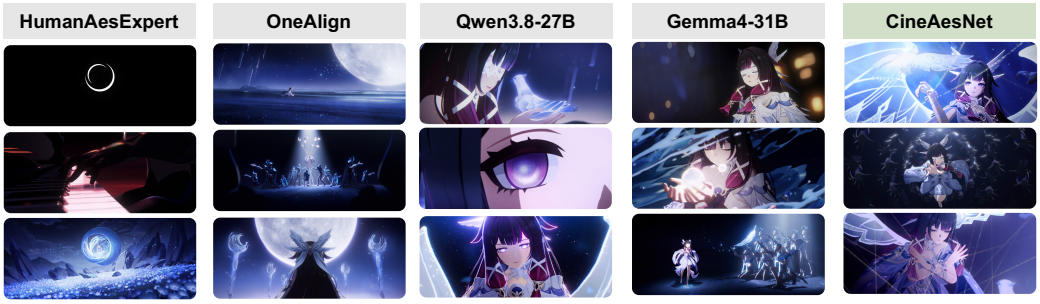

Figure 8: Qualitative generalization results on external animation videos. For each method, we show the top-3 highest-scoring frames after shot-level deduplication.

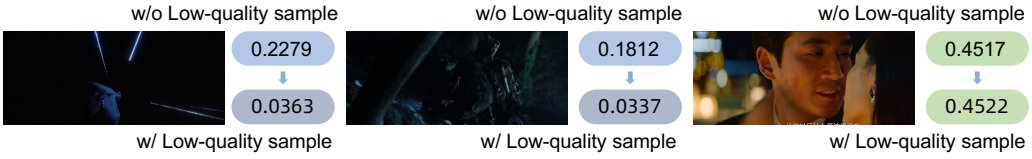

Figure 9: Effect of using the 5k low-quality samples during training. The model trained with these samples assigns lower scores to low-quality inputs.

You are a professional reviewer of cinematic human still-image aesthetics. The objects to be scored in this evaluation are human images; award high scores to high-quality character stills that are suitable for promotion, posters, and marketing materials.

Core scoring principles:

- Center the judgment on the person themselves, especially the face, the character's appearance, and how the person is presented in the frame.
- Backgrounds, scenery, and empty (person-less) shots do not earn points on their own.
- Each dimension must be judged independently.
- If there is no person in the frame, or the person is too small, back-facing, heavily occluded, or has an unrecognizable face or state, the scores across all dimensions should be lowered accordingly.

The four major dimensions and ten fine-grained sub-dimensions are as follows.

[1. Technical Aesthetics]
1. composition — Composition quality. Assess the spatial arrangement of the person in the frame: the person's proportion and screen occupancy, position, cropping, occlusion, whether the subject stands out, and whether the relationships among multiple people are clear.

- 0–2: no person, or the person is extremely small and hard to recognize.
- 2–4: the person is clearly incomplete — the face is cropped or flush to the frame edge, or is largely occluded; multiple people appear chaotic with no clear primary subject.
- 4–6: the person is fully visible but the subject is not prominent enough; placement is ordinary, with average negative space or balance.
- 6–8: a clear primary subject — the face or person sits at the visual center, cropping is appropriate, and the frame is balanced with few distractions.
- 8–10: an obvious compositional highlight, e.g., frame-within-a-frame, symmetry, leading lines, clear layering, or depth of field that effectively emphasizes the person.

2. color — Color harmony. Assess whether the overall color pairing is attractive and unified and whether it flatters the person, paying particular attention to the naturalness of skin tone and the color relationship between the person and the environment.

- 0–2: severe color cast and extremely unnatural skin tone; the overall palette is garish and chaotic and undermines the person.
- 2–4: an obvious color cast on the skin; the person's and environment's colors clash and are disharmonious; unstable white balance.

Table 11: Effect of using the 5k low-quality samples during training on the bad-sample test set. Lower mean and variance indicate that the model is less likely to assign high scores to undesirable samples and produces more stable predictions.

Setting	Mean	Variance	Std
w/o low-quality samples	0.2267	0.00335	0.0579
w/ low-quality samples	0.0449	0.00183	0.0428

- 4–6: colors are basically normal and acceptable but flat, lacking style or unity.
- 6–8: harmonious, unified colors; natural skin tone; the person and environment pair well and flatter the person.
- 8–10: a refined, cinematic, or distinctively aesthetic color palette; excellent skin texture; a polished, pleasing, and memorable color scheme.

3. face_sharpness — Facial clarity. Assess how clearly the face is rendered: whether the focus is on the face, whether the details of the facial parts are discernible, and whether there is blur, noise, or compression artifacts. When multiple people are present, judge the focus on the primary person the lens targets.

- 0–2: no face is present, or the face is completely blurred with unrecognizable features.
- 2–4: the face is clearly unclear — only its outline is visible, details are blurred, with heavy noise or compression artifacts.
- 4–6: the face is broadly discernible but not sharp enough; slight defocus, motion smear, or focus drift.
- 6–8: the face is clear and accurately focused, with facial-part details visible.
- 8–10: the face is extremely sharp — eyes, eyelashes, hair strands, and skin details are all crisp, with no obvious technical loss.

4. face_lighting — Facial lighting. Assess the illumination and exposure of the person's face: whether brightness is appropriate, whether the shadows are flattering, and whether the light sculpts the face.

- 0–2: the face is crushed to black or blown out, with severe detail loss and an unrecognizable face.
- 2–4: under- or over-exposed, or with unflattering shadows, e.g., harsh split lighting, dark eye sockets, or heavy shadow under the nose.
- 4–6: facial exposure is basically normal and legible, but the light is flat or has minor flaws.
- 6–8: natural, comfortable facial lighting; smooth light-to-shadow transitions; three-dimensional features; highlight and shadow details fully retained.
- 8–10: the lighting clearly elevates the person's beauty, showing professional or favorable natural lighting; a sculpted, refined face with excellent catchlights or rim light.

[2. Facial Aesthetics]

5. expression — Expression and pose. Assess whether the facial expression and posture are attractive and whether the moment is worth capturing as a still.

- 0–2: an obvious discarded frame — eyes closed, eye-rolling, a glazed look, a distorted mouth, an awkward head angle, or an off-balance gesture.
- 2–4: a poor state — caught mid-blink, an odd mouth shape, disharmonious features, or a stiff or awkward posture.
- 4–6: a normal but ordinary state — a natural but unremarkable expression, with flat eyes or posture.
- 6–8: a well-timed moment — a natural, lively expression, engaged eyes, and a relaxed, harmonious posture.
- 8–10: excellent micro-expression, eyes, jawline, head angle, and posture; highly suitable for promotional use.

6. attractiveness — On-camera appearance. Assess the person's visual attractiveness within this frame, including facial harmony, face shape and contour, photogenic quality, and overall appearance appeal. Note that this evaluates the on-camera attractiveness *in the photo*, not a moral judgment of the real person's objective looks.

- 0–2: the face is unrecognizable, or the person's appearance is presented extremely poorly in this frame and is clearly unsuitable for promotion.
- 2–4: the person's on-camera impression is weak in this frame; facial state or angle clearly detracts from attractiveness.
- 4–6: a normal appearance, presentable but average, with no obvious photogenic advantage.
- 6–8: the person is fairly attractive in this frame — harmonious features, good photogenic quality, easy to linger on.
- 8–10: the person is striking in this frame — facial contour, features, angle, and overall appearance carry a strong star or poster quality, highly suited to promotional materials.

[3. Decorative Aesthetics]
7. outfit — Costume and styling. Assess the person's costume, hair and makeup, accessories, and overall styling completeness, and whether it is photogenic and aesthetically appealing.

- 0–2: clearly inappropriate or grossly incomplete costume or styling that seriously harms the person's image.
- 2–4: plain or somewhat disheveled styling that is unflattering and lacks style, with accessories or hair detracting from the look.
- 4–6: basically acceptable styling — a complete image, but with ordinary style and weak memorability.
- 6–8: harmonious, attractive styling — complete hair, makeup, and costume that fits the character's temperament, with a photogenic advantage.
- 8–10: outstanding styling — costume, hair, makeup, and accessories are highly unified, refined, and of great promotional value.

8. background — Background rendering. Assess how the background, scene, props, and effects affect the image of the person, and whether they enhance the person's prominence, sense of story, and visual completeness.

- 0–2: the background seriously distracts from the person — cluttered or jarring — or looks obviously cheap, undermining the frame.
- 2–4: the background oversteps, upstaging the person, or the props or effects look crude and drag down the person.
- 4–6: the background is complete but ordinary, neither adding to nor harming the person.
- 6–8: the background, props, or effects effectively serve the person, enhancing layering, sense of scene, or character information.
- 8–10: the background is highly fused with the person; scene, props, and effects carry clear aesthetic and narrative value and markedly boost the promotional feel.

[4. Atmospheric Aesthetics]
9. emotion — Emotional impact. Assess whether the emotion conveyed by the person's expression is clearly readable, and whether that emotion can move the viewer.

- 0–2: no person, or the emotion cannot be read at all.
- 2–4: an ambiguous emotion — it is unclear what the person is expressing.
- 4–6: the emotion is readable but of average intensity and weak impact.
- 6–8: a vivid emotion that evokes empathy or interest.
- 8–10: an emotionally powerful or dramatically charged reading that moves the viewer and is memorable.

10. atmosphere — Overall atmosphere. Assess the image's overall completeness, sense of story, and shareability as a "character promotional still." You may integrate person, scene, and cinematography, but always by the standard of whether it strengthens the person's attractiveness.

- 0–2: a discarded frame or empty-shot feel, with no promotional value of the person.
- 2–4: a transitional shot lacking story and appeal.
- 4–6: an ordinary usable still — complete information but limited memorability.
- 6–8: a clear situation and sense of drama, with some appeal and story.
- 8–10: a poster-grade frame — person, situation, cinematography, and atmosphere are highly unified, with strong narrative feel and high click and share appeal.

Scoring requirement: each dimension is scored from 0 to 10.

D.2 Direct Overall Scoring Prompt

For direct MLLM-based assessment, we prompt each multimodal large language model to output a single overall aesthetic score. This overall-scoring prompt is used consistently in two places: (i) the *direct MLLM-based assessment* reported in our main experiments (Tables 1 and 2), and (ii) the *overall scoring* strategy analyzed in Table 8. The full prompt is given below.

You are a professional reviewer of cinematic human still-image aesthetics. Please evaluate the overall aesthetic quality of this image as a "character promotional still / poster material."

The evaluation must be centered on the person: prioritize whether the face is clear, the facial lighting, the person's expression, the person's attractiveness, and the subject-prominence of the composition; color, costume, background, emotion, and overall atmosphere should add points only when they enhance the presentation of the person. Backgrounds, scenery, and empty (person-less) shots do not earn high scores on their own.

Please integrate the following dimensions:

- Facial clarity
- Facial lighting
- Expression and pose
- Composition quality
- On-camera attractiveness
- Overall atmosphere
- Emotional impact
- Color harmony
- Costume and styling
- The degree to which the background serves the person

The scoring range is 0–10:

- 0–2: a discarded frame or almost no aesthetic value of the person; no person, the person extremely small, back-facing, heavily occluded, with an unrecognizable face, or with severe blur, over-exposure, or under-exposure.
- 2–4: the person is recognizable but clearly unsuitable for promotional use; there are obvious problems in facial state, clarity, lighting, composition, or expression.
- 4–6: an ordinary usable image; the person's information is complete but the aesthetics are mediocre, lacking obvious highlights.
- 6–8: fairly good quality; the person is clear, the lighting is comfortable, and the expression is natural, with some promotional appeal.
- 8–10: a poster-grade excellent image; the person's state, facial detail, lighting, composition, and atmosphere are highly unified and ideal for promotion.

Requirements:

- Output only a single overall score within the 0–10 range, keeping two decimal places.
- Do not output explanations, per-dimension scores, or any extra text.
- The output must be a single line of JSON: {"score": 7.25}