LATENT AESTHETIC DELIBERATION: LEARNING TO REFINE AESTHETIC JUDGMENTS

Anonymous authors
Paper under double-blind review

ABSTRACT

Image Aesthetic Assessment (IAA) requires weighing visual cues whose aesthetic contributions depend on image context. Although image–rating pairs are widely available, they specify judgment outcomes rather than how evidence should be integrated or assessments revised. Direct scoring with multimodal large language models (MLLMs) does not explicitly model judgment revision, whereas supervised verbalized reasoning requires additional reasoning-text targets and incurs token-by-token decoding overhead. Motivated by this, we introduce **Latent Aesthetic Deliberation** (LAD), a recurrent framework for learning judgment refinement from ratings without annotated reasoning trajectories. After a single multimodal encoding pass, an evolving judgment state repeatedly retrieves and integrates evidence from persistent image-token memory. Regression against human ratings trains the recurrent scorer, while internal self-distillation aligns intermediate score readouts with the final prediction without prescribing their latent representations. At inference, LAD fuses the recurrent score with a parallel language-aligned rating estimate without decoding an explanation. Extensive experiments demonstrate state-of-the-art performance on the evaluated benchmarks for holistic and aspect-specific aesthetic assessment.

1 INTRODUCTION

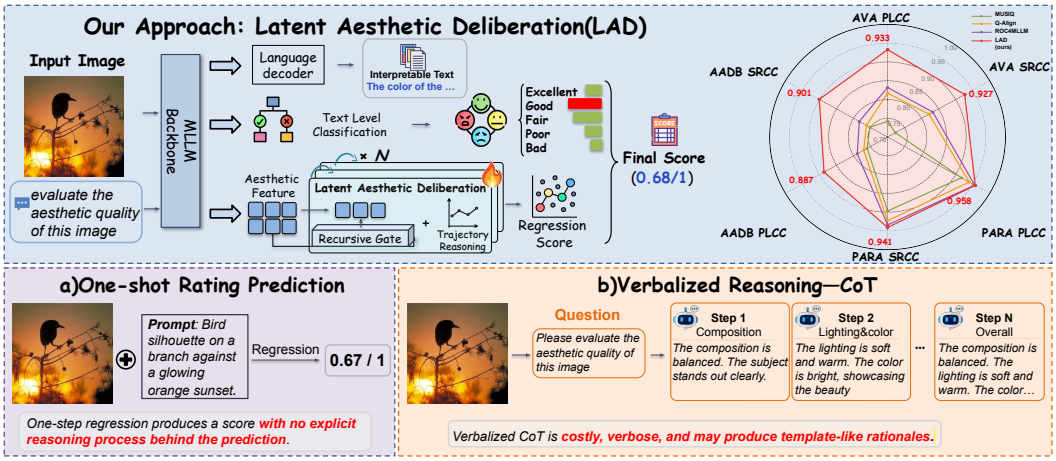


Figure 1: **Scoring paradigms for MLLM-based IAA. (a) Direct prediction decodes an aesthetic score from a multimodal representation. (b)** Verbalized Chain-of-Thought generates an aesthetic analysis before scoring. In the upper panel, LAD recurrently refines a latent regression score using visual evidence and fuses it with a language-aligned classification score. The text decoder is an optional output branch, so generating a textual assessment is not required for scoring.

Image Aesthetic Assessment (IAA) predicts human judgments of an image's aesthetic quality. Its difficulty lies in evaluating visual cues whose aesthetic roles depend on the image as a whole: shal-

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.

low focus may isolate a portrait subject yet obscure important landscape details. Such cues must be weighed in relation to composition and context. Widely used benchmarks provide human ratings or score distributions, sometimes accompanied by photographic attributes and rater information (Murray et al., 2012; Kong et al., 2016; Yang et al., 2022). These annotations describe outcomes and selected factors, but leave their integration unspecified. Learning how interdependent cues jointly support an overall judgment is therefore central to aesthetic scoring with the available supervision.

Multimodal large language models (MLLMs) support aesthetic scoring through two representative paradigms. One-shot methods directly predict scores through text-defined rating levels or continuous regression (Figure 1(a)) (Wu et al., 2024; Ma et al., 2026). This formulation leaves intermediate judgment formation and revision implicit, without explicitly modeling how visual cues are reconsidered as the overall assessment develops. Verbalized Chain-of-Thought (CoT) methods make intermediate assessments explicit through textual reasoning steps before scoring (Figure 1(b)) (Jiang et al., 2025; Liu et al., 2026). However, supervised CoT training relies on additional reasoning-text targets (Liu et al., 2026), while generating these assessments token by token adds computational overhead and latency at inference. Although established IAA datasets provide image–rating pairs at scale, detailed annotations of the reasoning underlying these judgments remain scarce (Murray et al., 2012; Liu et al., 2026). Acquiring reliable reasoning targets requires additional annotation and validation effort, motivating methods that learn judgment refinement directly from existing ratings (Liu et al., 2026). This raises a central question: can an MLLM learn to refine aesthetic judgments through repeated visual evidence integration, using ratings only to supervise the updates without requiring annotated reasoning trajectories or decoding an explanation for scoring?

Iterative latent attention and recurrent computation provide a foundation for repeated processing (Jaegle et al., 2021; Geiping et al., 2025), but naively adapting them to IAA requires connecting evidence selection to continuous aesthetic evaluation. A cue's contribution depends on its relationship to the whole image and may change as other evidence is integrated. This motivates retaining visual evidence separately from the evolving judgment, so the latter can guide what is revisited next. Learning these updates poses a second challenge: a mean opinion score (MOS) specifies an aggregate evaluation without prescribing an evidence order or intermediate representation. Different balances of visual strengths and weaknesses can yield the same score. Latent Visual Reasoning (LVR), for example, constructs visual reconstruction targets from preannotated query-relevant regions during supervised fine-tuning (Li et al., 2026), whereas MOS provides no such targets. Final-score regression can train the recurrence, but does not explicitly relate intermediate assessments. To tackle these challenges, our key insight is to constrain the scores read from intermediate states while allowing their underlying representations to retain different evidence. The judgment state can then connect evidence selection and learning: it guides subsequent retrieval, while its score readout provides supervision for this evolving process.

In this paper, we introduce *Latent Aesthetic Deliberation* (LAD), a recurrent scoring module after the MLLM that couples evidence selection with continuous aesthetic judgment (upper panel of Figure 1). A single multimodal encoding pass supplies persistent image-token memory and initializes a compact judgment state from image and task context. A shared recurrent block uses the current state to retrieve and integrate visual evidence, allowing subsequent queries to emphasize different cues as the assessment develops. Keeping the image tokens available throughout this process enables later updates to revisit individual details in relation to the accumulated visual context. A shared readout maps each updated state to a continuous score, connecting these evolving representations to a common aesthetic evaluation scale. Regression against human MOS trains the recurrent scorer, while internal self-distillation uses its final prediction to guide earlier score readouts. This provides supervision for intermediate assessments while allowing their latent states to retain different visual information for subsequent retrieval. At inference, LAD fuses the final recurrent prediction with a parallel language-aligned rating estimate without generating an explanation.

Extensive experiments demonstrate LAD's state-of-the-art performance on the evaluated benchmarks for holistic and aspect-specific aesthetic assessment. Controlled analyses support the contribution of internal score supervision and show that retaining token-level visual access improves prediction over recurrence on a mean-pooled global summary. Our contributions are:

- We propose LAD, a recurrent aesthetic framework in which an evolving judgment state repeatedly retrieves and integrates token-level visual evidence after multimodal encoding.

- We develop a learning scheme that combines final human-score supervision with internal self-distillation, training intermediate score readouts without annotated judgment trajectories while allowing latent representations to evolve.
- We demonstrate state-of-the-art performance on the evaluated benchmarks for holistic and aspect-specific aesthetic assessment, with controlled ablations supporting the benefits of token-level evidence retrieval and internal score supervision.

2 Related Work

2.1 MLLM-based Image Aesthetic Assessment

Aesthetic modeling draws on human ratings, attributes, and textual feedback. Vision-language pretraining on user comments learns aesthetic representations in VILA (Ke et al., 2023), while CALM combines multi-scale text-guided self-supervision with aesthetic instruction tuning (Liu et al., 2025). PhotoEye develops photographic understanding through critique data and language-guided visual fusion (Qi et al., 2025). Its successive fusion layers update queries from preceding outputs, making visual aggregation depend on the evolving representation. LAD instead places recurrent scoring after multimodal encoding and repeatedly accesses the resulting image tokens. For score prediction, Q-Align maps probabilities over verbal rating levels to continuous scores (Wu et al., 2024), ROC4MLLM introduces a regression output (Ma et al., 2026), and ArtiMuse combines fine-grained scoring with expert-level aesthetic understanding (Cao et al., 2026). These prediction interfaces can coexist with rich aesthetic training, so the form of the final readout alone does not characterize the supervision used to learn it.

Aesthetic interpretation and explicit reasoning serve related but distinct roles. AesExpert supports assessment and interpretation through aesthetic instruction tuning (Huang et al., 2024). Multi-stage IAA organizes textual analysis from low-level stimulus to holistic organization and high-level perception before scoring (Jiang et al., 2025). Aes-R1 combines CoT cold-start training with reinforcement learning for explanations followed by scores (Liu et al., 2026). These latter approaches use generated text to represent intermediate judgments. LAD performs its scoring updates in latent states, regularizing intermediate predictions with internally generated score targets. Aesthetic text can still supervise its other outputs during training, while the recurrent scoring process learns without annotated judgment trajectories and inference requires no generated explanation.

2.2 Recurrent Computation and Latent Visual Reasoning

Recurrent computation provides a foundation for iterative representation refinement. Universal Transformers share transformations across recurrent depth (Dehghani et al., 2019). Perceiver alternates input cross-attention with latent processing and supports weight sharing, allowing successive retrievals to depend on the evolving latent representation (Jaegle et al., 2021). Recurrent-depth language models also expand computation through repeated hidden-state updates without requiring specialized reasoning data (Geiping et al., 2025). These architectures establish recurrence and iterative attention as reusable mechanisms. In particular, Perceiver's repeated access to inputs is closely related to LAD's use of persistent visual memory. LAD adapts these mechanisms to aesthetic scoring, with a shared readout that maps each recurrent state to a continuous assessment. Self-distillation can transfer knowledge from deeper to shallower classifiers within one network (Zhang et al., 2019). LAD applies internal distillation to scalar readouts across shared recurrent steps, regularizing their agreement with the final prediction without matching the latent states themselves.

Latent reasoning also extends language generation beyond explicit textual steps. Coconut uses continuous hidden states as intermediate thoughts (Hao et al., 2025). In multimodal models, LVR generates latent states that reconstruct relevant visual tokens and interleaves them with text (Li et al., 2026). Monet learns continuous visual thoughts through distillation-based supervised fine-tuning and visual-latent policy optimization (Wang et al., 2026). These approaches develop latent reasoning within the language model's generation process. LAD uses a separate recurrent scoring module after multimodal encoding, with a fixed visual memory and a continuous rating target. Its intermediate judgments are trained through final-score gradients and internal distillation, without requiring target visual thoughts or judgment trajectories for those updates.

3 METHODOLOGY

LAD learns evidence integration in latent space: an MLLM pass creates persistent visual memory, and a shared recurrent block updates a compact aesthetic state. MOS regression and trajectory self-distillation supervise final and intermediate states without annotated judgment trajectories. Rating-position logits provide an ordinal estimate, the final recurrent state provides continuous regression, and their fusion yields the score; text generation remains optional. Figure 2 summarizes the framework, and Sections 3.1–3.5 detail its formulation and optimization.

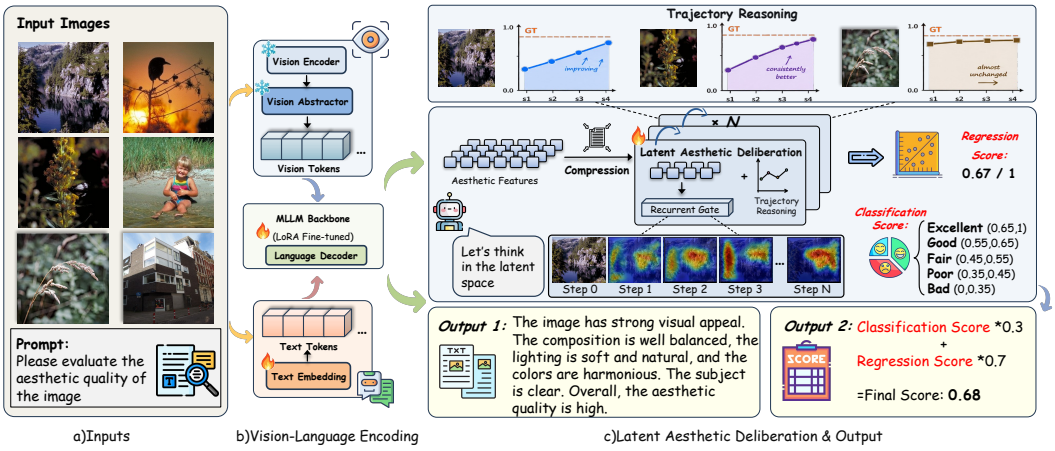


Figure 2: **Overview of Latent Aesthetic Deliberation (LAD).** One backbone pass produces contextualized image tokens, the score-context state for $\mathbf{Z}_0$, and rating-position logits. A shared block queries the persistent visual memory for T updates, progressively refining the latent state. Its final score token yields s_{reg}, while the original rating logits yield the language-aligned estimate s_{cls}. Their fusion gives the aesthetic score.

3.1 PROBLEM FORMULATION

Given an image I and a task-specific rating prefix P, LAD predicts continuous, rating-level, and textual outputs. Each aspect-specific target is trained separately using its corresponding attribute score; the complete rating prefixes and target construction are provided in Appendix D. For human ratings $\{r_n\}_{n=1}^{N_I}$ of the target under evaluation, where $r_n \in \mathcal{R}_D \subseteq [1, R_D]$, we normalize the mean opinion score (MOS) to $[0, 1]$:

$$m(I) = \frac{1}{N_I} \sum_{n=1}^{N_I} r_n, \qquad s(I) = \frac{m(I) - 1}{R_D - 1}, \tag{1}$$

Here, $R_D = 5$ for PARA and AADB and $R_D = 10$ for AVA and ICAA17K; histogram datasets use the vote-weighted mean. We learn $\hat{s} = f(I, P) \in [0, 1]$ together with the complementary outputs below.

3.2 MULTIMODAL INPUT ENCODING

LAD adopts the OneAlign-style mPLUG-Owl2 backbone (Wu et al., 2024; Ye et al., 2024). Image tokens from the vision encoder and visual abstractor are inserted at the <|image|> position and contextualized with a task-specific rating prefix. The same prefix is used during training and inference and ends with `is`. For sequence length N, width d, and vocabulary $\mathcal{V}$, one pass $F_\theta(I, P)$ yields hidden states $\mathbf{H} = [\mathbf{h}_1, \ldots, \mathbf{h}_N] \in \mathbb{R}^{N \times d}$ and logits $\mathbf{O} \in \mathbb{R}^{N \times |\mathcal{V}|}$. At the final-prefix position j_{sc}, we retain image states $\mathbf{V} = [\mathbf{v}_1, \ldots, \mathbf{v}_{N_v}] \in \mathbb{R}^{N_v \times d}$ as spatial evidence and set $\mathbf{h}_{\text{sc}} = \mathbf{h}_{j_{\text{sc}}}$ as the score context that initializes LAD. The logits $\mathbf{O}_{j_{\text{sc}}}$ supply the rating prediction, while $\mathbf{O}$ supports optional text generation. Causal masking prevents both quantities from conditioning on the ground-truth continuation or later responses.

3.3 Recurrent Latent Aesthetic Deliberation

Because IAA datasets lack step-wise rationales, LAD optimizes compact latent tokens from score supervision. Starting from an image- and task-conditioned state, it performs T recurrent updates over persistent image-token memory without decoding a rationale. The memory remains fixed while the latent state evolves, allowing later queries to revisit the same visual evidence in light of previously integrated information and revise the current judgment. The fixed memory also makes each refinement step directly comparable.

Visual memory and latent initialization. We project the contextualized image tokens to latent width d_r:

$$\mathbf{M} = \mathrm{LN}_v(\mathbf{V}W_v + \mathbf{b}_v) \in \mathbb{R}^{N_v \times d_r}, \tag{2}$$

where LN is layer normalization and $\mathbf{m}_i$ is the i-th row of $\mathbf{M}$. The memory is computed once and fixed during refinement, separating visual representation from the evolving latent judgment. Let $\mathbf{E} \in \mathbb{R}^{L \times d_r}$ contain L learnable embeddings, including a score token. We form score and image contexts $\mathbf{c}_{\mathrm{sc}} = W_{\mathrm{sc}}\mathbf{h}_{\mathrm{sc}}$ and $\mathbf{c}_{\mathrm{img}} = W_{\mathrm{img}}(N_v^{-1}\sum_i \mathbf{m}_i)$ in $\mathbb{R}^{d_r}$ and broadcast them over the learnable tokens:

$$\mathbf{Z}_0 = \mathrm{LN}_0\big(\mathbf{E} + \mathbf{1}_L\mathbf{c}_{\mathrm{sc}}^\top + \mathbf{1}_L\mathbf{c}_{\mathrm{img}}^\top\big). \tag{3}$$

The contexts $\mathbf{c}_{\mathrm{sc}}$ and $\mathbf{c}_{\mathrm{img}}$ provide task semantics and a global visual prior. Broadcasting them gives all latent tokens a shared reference, while distinct embeddings permit specialization without handcrafted attribute assignments. We use $L = 5$: four deliberation tokens integrate evidence and one score token supports the shared readout. Subsequent recurrent states refine the judgment from this common initialization.

Recurrent evidence aggregation and refinement. At step t, self-attention first exchanges accumulated information among latent tokens. The resulting state then queries $\mathbf{M}$ through cross-attention, using dynamic latent queries with fixed visual keys and values:

$$\begin{aligned}\widetilde{\mathbf{Z}}_t &= \mathbf{Z}_{t-1} + \mathrm{Drop}(\mathrm{MHA}_{\mathrm{self}}(\mathrm{LN}_s(\mathbf{Z}_{t-1}))),\\ \overline{\mathbf{Z}}_t &= \widetilde{\mathbf{Z}}_t + \mathrm{Drop}\Big(\mathrm{MHA}_{\mathrm{cross}}(\mathrm{LN}_c(\widetilde{\mathbf{Z}}_t), \mathbf{M}, \mathbf{M})\Big),\end{aligned} \tag{4}$$

where $\mathrm{MHA}_{\mathrm{cross}}(Q, K, V)$ is multi-head cross-attention. Because the query depends on $\mathbf{Z}_{t-1}$, retrieval is conditioned on integrated evidence rather than fixed pooling. Self-attention consolidates accumulated information, while cross-attention retrieves visual evidence for the current latent context. Thus, although the memory is fixed, the evolving state changes which evidence is incorporated.

A gated feed-forward update then regulates the post-attention state transformation:

$$\begin{aligned}\mathbf{U}_t &= \mathrm{FFN}(\mathrm{LN}_f(\overline{\mathbf{Z}}_t)),\\ \mathbf{G}_t &= \sigma\big([\overline{\mathbf{Z}}_t; \mathbf{U}_t]W_{\mathrm{gate}} + \mathbf{1}_L\mathbf{b}_{\mathrm{gate}}^\top\big),\\ \mathbf{Z}_t &= \overline{\mathbf{Z}}_t + \mathrm{Drop}(\mathbf{G}_t \odot \mathbf{U}_t), \qquad t = 1, \ldots, T.\end{aligned} \tag{5}$$

Here, $[\,\cdot\,;\,\cdot\,]$ is row-wise concatenation, $\mathbf{1}_L$ broadcasts the bias, and σ is sigmoid. The gate scales each feed-forward residual, and the updated state defines the next query. It therefore controls how retrieved evidence modifies the state, forming a recurrent cycle of evidence retrieval and refinement. Sharing the block makes its parameter count independent of T.

Step-wise score states. A shared MLP maps the score token $\mathbf{z}_t^{\mathrm{score}} \in \mathbb{R}^{d_r}$ to $s_t = \sigma(f_s(\mathbf{z}_t^{\mathrm{score}})) \in [0, 1]$. The score is read out rather than fed back; recurrence is carried by $\mathbf{Z}_t$. Applying the same readout to evolving states makes $\{s_t\}_{t=1}^T$ a judgment trajectory rather than independent predictions. We use $s_{\mathrm{reg}} = s_T$ for final regression and earlier readouts for trajectory supervision. This shared readout keeps every intermediate estimate on the same normalized scale, so trajectory supervision regularizes refinement without introducing step-specific heads. Because all steps share parameters, additional refinement depth changes computation without introducing a separate predictor at each stage.

3.4 Output Decoding and Prediction

LAD uses two complementary scoring paths and an optional text decoder. The rating path preserves the ordinal semantics of verbal quality levels, the recurrent path models continuous variation, and their fusion yields the scalar prediction.

Rating-level classification. Following Wu et al. (2024), classification uses $K = 5$ ordered tokens $\mathcal{R} = \{\rho_k\}_{k=1}^{K} = \{\textit{bad}, \textit{poor}, \textit{fair}, \textit{good}, \textit{excellent}\}$ with fixed anchors $(a_1, \ldots, a_5) = (0.28, 0.40, 0.50, 0.60, 0.72)$. At the score-context position defined in Section 3.2, $\ell_k = \mathbf{O}_{j_{\mathrm{sc}}, \rho_k}$ denotes the vocabulary logit for candidate rating token ρ_k. Restricted softmax probabilities $p_k = \exp(\ell_k) / \sum_j \exp(\ell_j)$ yield $s_{\mathrm{cls}} = \sum_k p_k a_k$ and rating $\rho_{\arg\max_k p_k}$.

Continuous score decoding. We combine the recurrent prediction s_{reg} with the ordinal estimate s_{cls}:

$$\hat{s} = \gamma s_{\mathrm{cls}} + (1 - \gamma) s_{\mathrm{reg}}, \tag{6}$$

where $\gamma \in [0, 1]$ is fixed, combining ordinal semantics with continuous refinement.

Aesthetic text decoding. Beyond the target-specific verbal rating, the optional decoder produces structured descriptions and aesthetic comments supervised by GLM-4.6V-generated targets. These outputs add semantic supervision but are not decoded for scalar prediction; their grouping and loss routing are detailed in Appendix C.

3.5 Training Objective and Inference

Training aligns each output with its available target. Let $s = s(I)$ be the normalized MOS from Eq. 1. A fixed quantizer with boundaries $0 = \tau_0 < \cdots < \tau_K = 1$ assigns $c(s) = \rho_k$ for $\tau_{k-1} \le s < \tau_k$, including $s = 1$ in the final interval; the numerical thresholds are reported in Eq. 13. The first assistant turn completes the task-specific prefix with rating word $c(s)$, which is the supervised next-token target. For $q \in \{\mathrm{core}, \mathrm{str}, \mathrm{com}\}$, $\mathcal{L}_q$ is the mean causal cross-entropy over its supervised positions and is zero when none are observed. With no intermediate judgment labels, the human score supervises the final state and its detached prediction supplies internal targets for earlier readouts:

$$\begin{aligned} \mathcal{L}_{\mathrm{reg}} &= (s_{\mathrm{reg}} - s)^2, \\ \mathcal{L}_{\mathrm{traj}} &= \sum_{t=1}^{T-1} \omega_t \, \mathrm{SmoothL1}(s_t, \mathrm{sg}(s_T)). \end{aligned} \tag{7}$$

Here, $\mathrm{sg}(\cdot)$ prevents earlier predictions from shifting the final-step teacher. The weights ω_t increase toward later states (Appendix Eq. 11), while $\mathcal{L}_{\mathrm{reg}}$ backpropagates the human MOS through all recurrent updates. Because trajectory supervision acts on the shared score readout rather than every latent dimension, the state can retain richer visual information for subsequent retrieval.

The complete training objective is

$$\mathcal{L} = \lambda_{\mathrm{core}} \mathcal{L}_{\mathrm{core}} + \lambda_{\mathrm{reg}} \mathcal{L}_{\mathrm{reg}} + \lambda_{\mathrm{traj}} \mathcal{L}_{\mathrm{traj}} + \lambda_{\mathrm{str}} \mathcal{L}_{\mathrm{str}} + \lambda_{\mathrm{com}} \mathcal{L}_{\mathrm{com}}, \tag{8}$$

where all λ terms are nonnegative; rating-only samples set $\mathcal{L}_{\mathrm{str}} = \mathcal{L}_{\mathrm{com}} = 0$.

At inference, one encoding pass provides $\mathbf{M}$, $\mathbf{h}_{\mathrm{sc}}$, and the rating logits. After T shared updates, s_T gives s_{reg} and the rating logits give s_{cls}; Eq. 6 returns their fused score. Text is decoded only when requested.

4 Experiments

We evaluate holistic and aspect-specific scoring, followed by recurrent score and visual-evidence refinement. More results such as component and depth ablations appear in the Appendix.

4.1 Experimental Setup

Datasets and metrics. Holistic assessment uses the standard splits of AVA (Murray et al., 2012) (235,579 training; 19,929 test images), PARA (Yang et al., 2022) (28,220 training; 3,000 test), and AADB (Kong et al., 2016) (8,500 training; 500 validation; 1,000 test). Aspect-specific assessment uses PARA composition and lighting and the official ICAA17K color split (He et al., 2023a) (15,734 training; 1,617 test). PARA and AADB use 1–5 ratings; AVA and ICAA17K use 1–10. We average ratings and normalize MOS by Eq. 1. We report Spearman's rank correlation (SRCC) and Pearson's linear correlation (PLCC); the iteration study also reports MAE on the original 1–5 PARA scale.

Reporting both correlation measures distinguishes rank preservation from linear agreement with normalized human scores across datasets.

Implementation details. LAD uses the OneAlign-style mPLUG-Owl2 backbone in Section 3.2 and the task-specific rating prefixes defined in Section 3.1; each prefix is held fixed within its corresponding target. Images are square-padded, resized to 448×448, and processed by CLIP. We freeze the vision encoder and visual abstractor, apply LoRA to the language model (rank 128, scale 256, dropout 0.05), and train all LAD modules. Training runs for three epochs with AdamW at 2×10^{-4}, zero weight decay, cosine decay, 0.03 warmup, bfloat16 precision, and gradient checkpointing. The per-device batch size is 12 with 16-step accumulation. Unless stated otherwise, $T = 4$ and $\gamma = 0.3$ in Eq. 6. Hyperparameters are shared across datasets, with remaining settings in Appendix B. Ablations hold the task-specific prefix, preprocessing, and evaluation fixed. All LAD targets are trained independently under the same architecture, optimization settings, and evaluation protocol.

Compared methods. We compare LAD with task-specific visual predictors NIMA (Talebi & Milanfar, 2018), AesMamba (Gao et al., 2024), TMCR (Yang et al., 2024), Charm (Behrad et al., 2025), and PPA-T (Hara & Otsuka, 2026), and with MLLM-based scorers Q-Align (Wu et al., 2024), ROC4MLLM (Ma et al., 2026), and ArtiMuse (Cao et al., 2026). ICAA (He et al., 2023a) is included for color. Complete holistic benchmark results are reported in Appendix Table 7.

4.2 Quantitative Evaluation

We evaluate holistic prediction on AVA, PARA, and AADB and aspect-specific prediction for composition, lighting, and color.

4.2.1 General Image Aesthetic Assessment

Table 1 shows that LAD obtains the best SRCC on all three holistic benchmarks and the best or tied-best PLCC. Against the strongest reported methods, it improves SRCC/PLCC by 0.094/0.101 on AVA and 0.118/0.098 on AADB; on PARA, SRCC improves by 0.005 and PLCC matches the best result of 0.958. The gains across datasets support recurrent latent deliberation for holistic aesthetic judgment.

The improvements on datasets with different sizes and rating scales indicate that the benefit is not confined to one benchmark setting. They are also substantial on AVA and AADB, where existing MLLM-based scorers already provide strong baselines.

Table 1: Main comparison with task-specific and MLLM-based aesthetic scorers. Higher SRCC and PLCC are better, and best results are bold.

Method	AVA		PARA		AADB	
	SRCC↑	PLCC↑	SRCC↑	PLCC↑	SRCC↑	PLCC↑
Task-specific visual predictors						
NIMA (Talebi & Milanfar, 2018)	0.611	0.635	0.878	0.863	0.701	0.710
AesMamba (Gao et al., 2024)	0.751	0.760	0.902	0.936	0.768	0.774
TMCR (Yang et al., 2024)	0.782	0.790	0.915	0.945	0.775	0.773
Charm (Behrad et al., 2025)	0.777	0.779	0.905	0.938	0.754	0.767
PPA-T (Hara & Otsuka, 2026)	0.780	0.782	0.919	0.947	0.768	0.773
MLLM-based scorers						
Q-Align (Wu et al., 2024)	0.822	0.817	0.923	0.940	0.762	0.770
ROC4MLLM (Ma et al., 2026)	0.833	0.832	0.934	0.956	0.783	0.789
ArtiMuse (Cao et al., 2026)	0.827	0.826	0.936	**0.958**	–	–
LAD (ours)	**0.927**	**0.933**	**0.941**	**0.958**	**0.901**	**0.887**

4.2.2 Aspect-specific Aesthetic Assessment

We further test whether the holistic advantage extends to individual aesthetic dimensions. These benchmarks separately evaluate composition, lighting, and color, providing a more fine-grained view of aesthetic prediction.

Table 2 shows the best performance on all three targets: SRCC/PLCC reaches 0.919/0.934, 0.923/0.935, and 0.908/0.913 on PARA-Composition, PARA-Light, and ICAA17K-Color. Relative to ROC4MLLM, the respective gains are 0.019/0.004, 0.023/0.005, and 0.014/0.010. Together with the holistic results, these consistent gains show that the same deliberation mechanism transfers across distinct aesthetic targets.

The consistent improvements show that LAD is not limited to a single overall score. The same architecture remains effective as the target changes across composition, lighting, and color, supporting both holistic and fine-grained assessment. This pattern remains consistent across datasets with different scales and targets.

Table 2: Quantitative comparison on aspect-specific aesthetic targets. Higher scores are better.

Method	PARA-Composition		PARA-Light		ICAA17K-Color	
	SRCC↑	PLCC↑	SRCC↑	PLCC↑	SRCC↑	PLCC↑
NIMA (Talebi & Milanfar, 2018)	0.740	0.720	0.740	0.770	0.809	0.815
MUSIQ (Ke et al., 2021)	0.760	0.770	0.770	0.760	0.835	0.841
ICAA (He et al., 2023a)	–	–	–	–	0.873	0.890
Q-Align (Wu et al., 2024)	0.890	0.910	0.890	0.910	0.761	0.781
ROC4MLLM (Ma et al., 2026)	0.900	0.930	0.900	0.930	0.894	0.903
LAD (ours)	**0.919**	**0.934**	**0.923**	**0.935**	**0.908**	**0.913**

4.3 Analysis of Iterative Visual Deliberation

The benchmark results establish final-prediction accuracy; to examine how the recurrent process evolves, we use two complementary views: score trajectories track scalar refinement, while spatial attention maps track changes in retrieved visual evidence across steps.

4.3.1 Step-wise Score Refinement

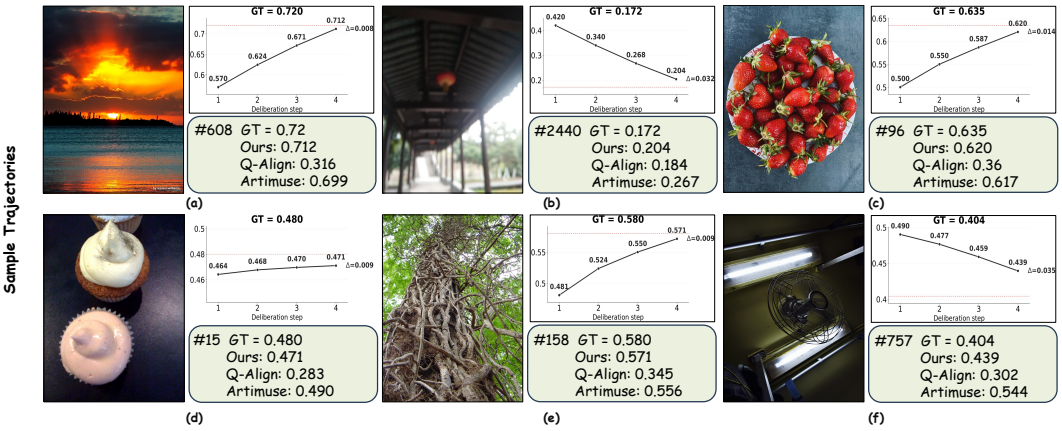


Figure 3: **Step-wise score refinement trajectories on the PARA dataset.** All aesthetic scores are normalized to $[0, 1]$ using the same MOS normalization as in the main evaluation. Each curve traces LAD's continuous regression branch, $s_{\text{reg}}^{(t)}$, over four recurrent deliberation steps. The step-4 endpoint is the final regression estimate s_{reg} reported as *Ours* for the corresponding sample, whereas the dashed horizontal line indicates the ground-truth MOS. For reference, Q-Align and ArtiMuse scores are listed in the corresponding sample panels.

We consider six representative cases spanning high-, mid-, and low-scoring images. Across these cases, LAD predictions progressively approach the ground-truth MOS over successive refinement steps, whereas the one-shot predictions of Q-Align and ArtiMuse exhibit larger deviations in several

cases. Differences use the unrounded predictions; displayed values are rounded. These examples qualitatively illustrate the progressive score calibration enabled by recurrent deliberation; the effect of recurrent depth is further evaluated on the full PARA benchmark in Appendix I. Together, the trajectories show that refinement changes the magnitude and direction of corrections according to each image.

4.3.2 Evolution of Visual Evidence

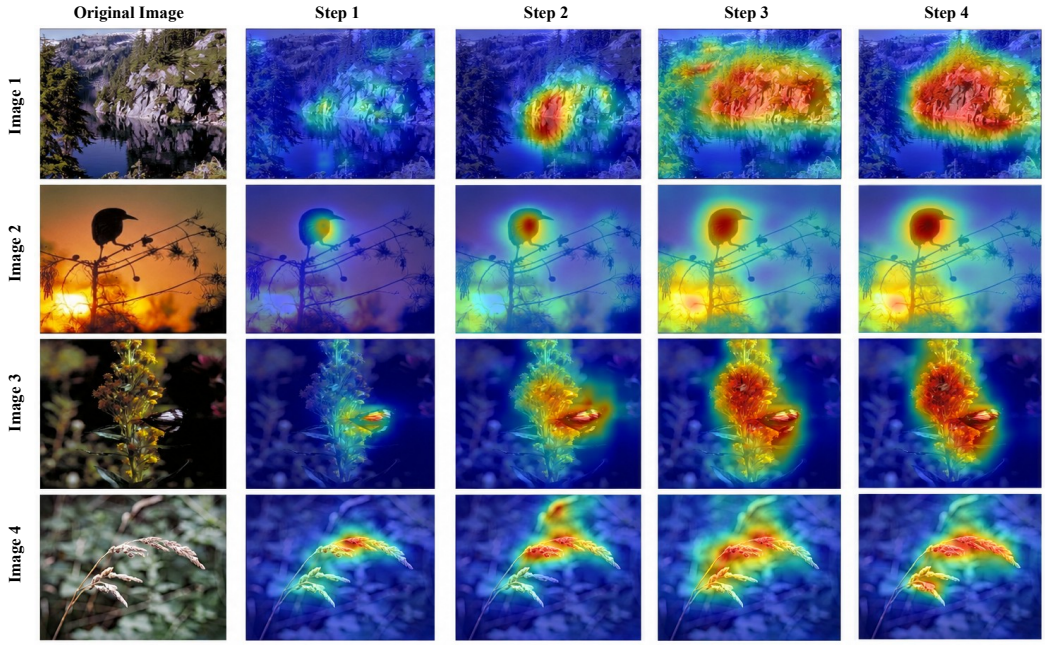


Figure 4: **Visualization of visual-memory attention across four recurrent refinement steps.** At each step, the visualization shows the head-averaged cross-attention from the score token to the persistent visual memory, illustrating how the retrieved visual evidence changes during recurrent refinement. The visualization is qualitative and does not imply a causal interpretation of attention or a verified pixel-level localization.

At each recurrent refinement step, we extract the cross-attention weights from the score token to the persistent visual memory. The attention is first averaged over the attention heads and then indexed at the score-token position, rather than aggregated across the five latent queries. This yields a step-dependent attention vector over the visual-memory tokens. For qualitative visualization, the resulting weights are normalized and mapped to the image using the same padding and cropping procedure as the input preprocessing. The visualization is intended to illustrate changes in visual-memory retrieval during recurrent refinement, rather than to establish a causal interpretation of attention or a verified pixel-level correspondence between abstractor tokens and image regions.

5 Conclusion

We proposed Latent Aesthetic Deliberation, a rationale-free MLLM framework that refines continuous scores through recurrent latent updates while preserving verbal-rating ordering. LAD repeatedly queries contextualized image tokens, fuses regression and language-aligned classification scores, and leaves text generation optional. Across AVA, PARA, AADB, and ICAA17K, one recurrent formulation aligns strongly with human judgments for holistic, composition, lighting, and color targets. Ablations attribute its gains to query-conditioned access to token-level visual memory, gated updates, and intermediate trajectory supervision. LAD thus enables iterative scoring without explicit reasoning sequences.

Ethics Statement

This work uses only publicly released aesthetic-assessment benchmarks (AVA, PARA, AADB, ICAA17K) under their existing licenses/terms, and does not involve new human-subject data collection. We are not aware of specific ethical concerns arising from this work.

Reproducibility Statement

Implementation details needed to reproduce our results are provided in Section 4.1, Appendix B, and the architecture description in Section 3. Dataset list sizes, score normalization, preprocessing, trainable modules, loss weights, and optimization settings are reported explicitly.

AI Use Statement

Generative AI tools were used to assist with training-strategy planning, experimental organization, implementation and debugging, manuscript structuring, language polishing, formatting, LaTeX checks, and spell-checking. Generative AI was also used to generate aesthetic comment text for a small portion of the training data, as described in Appendix C. All research decisions, experimental implementations, result verification, interpretations, and final manuscript content were reviewed and determined by the authors.

References

Fatemeh Behrad, Tinne Tuytelaars, and Johan Wagemans. Charm: The missing piece in ViT fine-tuning for image aesthetic assessment. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7815–7824, 2025.

Shuo Cao, Nan Ma, Jiayang Li, Xiaohui Li, Lihao Shao, Kaiwen Zhu, Yu Zhou, Yuandong Pu, Jiarui Wu, Jiaquan Wang, Bo Qu, Wenhai Wang, Yu Qiao, Dajuin Yao, and Yihao Liu. 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*, 2026.

Mostafa Dehghani, Stephan Gouws, Oriol Vinyals, Jakob Uszkoreit, and Lukasz Kaiser. Universal transformers. In *International Conference on Learning Representations*, 2019.

Fei Gao, Yuhao Lin, Jiaqi Shi, Maoying Qiao, and Nannan Wang. AesMamba: Universal image aesthetic assessment with state space models. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pp. 7444–7453, 2024. doi: 10.1145/3664647.3681011.

Jonas Geiping, Sean McLeish, Neel Jain, John Kirchenbauer, Siddharth Singh, Brian Bartoldson, Bhavya Kailkhura, Abhinav Bhatele, and Tom Goldstein. Scaling up test-time compute with latent reasoning: A recurrent depth approach. In *Advances in Neural Information Processing Systems*, volume 38, 2025. doi: 10.52202/085713-1380. URL https://proceedings.neurips.cc/paper_files/paper/2025/hash/3b01972cf31e6fa0fe29e4b8b5c2a0a1-Abstract-Conference.html.

Shibo Hao, Sainbayar Sukhbaatar, DiJia Su, Xian Li, Zhiting Hu, Jason Weston, and Yuandong Tian. Training large language models to reason in a continuous latent space. In *Proceedings of the Conference on Language Modeling*, 2025.

Takayuki Hara and Yuya Otsuka. Probabilistic prompt adaptation for unified image aesthetics and quality assessment. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 37236–37246, 2026.

Shuai He, Yongchang Zhang, Rui Xie, Dongxiang Jiang, and Anlong Ming. Rethinking image aesthetics assessment: Models, datasets and benchmarks. In *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence*, pp. 942–948, 2022. doi: 10.24963/ijcai.2022/132.

Shuai He, Anlong Ming, Yaqi Li, Jinyuan Sun, ShunTian Zheng, and Huadong Ma. Thinking image color aesthetics assessment: Models, datasets and benchmarks. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 21838–21847, 2023a.

Shuai He, Anlong Ming, Shuntian Zheng, Haobin Zhong, and Huadong Ma. EAT: An enhancer for aesthetics-oriented transformers. In *Proceedings of the 31st ACM International Conference on Multimedia*, pp. 1023–1032, 2023b. doi: 10.1145/3581783.3611881.

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*, 2024. doi: 10.1145/3664647.3680649.

Andrew Jaegle, Felix Gimeno, Andy Brock, Oriol Vinyals, Andrew Zisserman, and Joao Carreira. Perceiver: General perception with iterative attention. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 4651–4664, 2021. URL https://proceedings.mlr.press/v139/jaegle21a.html.

Zhenglang Jiang, Jianhao Liu, Hongming Li, Yue Liu, and Yu Song. Multi-stage image aesthetic assessment via chain-of-thought reasoning. *IEEE Access*, 13, 2025. doi: 10.1109/ACCESS.2025.3590925.

Junjie Ke, Qifei Wang, Yilin Wang, Peyman Milanfar, and Feng Yang. MUSIQ: Multi-scale image quality transformer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021.

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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10041–10051, 2023. URL https://openaccess.thecvf.com/content/CVPR2023/html/Ke_VILA_Learning_Image_Aesthetics_From_User_Comments_With_Vision-Language_Pretraining_CVPR_2023_paper.html.

Shu Kong, Xiaohui Shen, Zhe Lin, Radomir Mech, and Charless Fowlkes. Photo aesthetics ranking network with attributes and content adaptation. In *Proceedings of the European Conference on Computer Vision*, 2016.

Bangzheng Li, Ximeng Sun, Jiang Liu, Ze Wang, Jialian Wu, Xiaodong Yu, Emad Barsoum, Muhao Chen, and Zicheng Liu. Latent visual reasoning. In *International Conference on Learning Representations*, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/efa0939caa8a906fd0a61a9c60e46e8b-Abstract-Conference.html.

Leida Li, Yipo Huang, Jinjian Wu, Yuzhe Yang, Yaqian Li, Yandong Guo, and Guangming Shi. Theme-aware visual attribute reasoning for image aesthetics assessment. *IEEE Transactions on Circuits and Systems for Video Technology*, 33(9):4798–4811, 2023. doi: 10.1109/TCSVT.2023.3249185.

Mingxing Li, Rui Wang, Lei Sun, Yancheng Bai, and Xiangxiang Chu. Next token is enough: Realistic image quality and aesthetic scoring with multimodal large language model. *arXiv preprint arXiv:2503.06141*, 2025.

Boyang Liu, Yifan Hu, Senjie Jin, Shihan Dou, Gonglei Shi, Jie Shao, Tao Gui, and Xuanjing Huang. Unlocking the essence of beauty: Advanced aesthetic reasoning with relative-absolute policy optimization. In *International Conference on Learning Representations*, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/93397b489222f1e66a862f795fc32b96-Abstract-Conference.html.

Limin Liu, Shuai He, Anlong Ming, Rui Xie, and Huadong Ma. ELTA: An enhancer against long-tail for aesthetics-oriented models. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235, pp. 31106–31118, 2024.

Yuti Liu, Shice Liu, Junyuan Gao, Peng-tao Jiang, Hao Zhang, Jinwei Chen, and Bo Li. Advancing comprehensive aesthetic insight with multi-scale text-guided self-supervised learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 5748–5756, 2025. doi: 10.1609/aaai.v39i6.32613.

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*, 2026.

Naila Murray, Luca Marchesotti, and Florent Perronnin. AVA: A large-scale database for aesthetic visual analysis. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2408–2415, 2012.

Daiqing Qi, Handong Zhao, Jing Shi, Simon Jenni, Yifei Fan, Franck Dernoncourt, Scott Cohen, and Sheng Li. The photographer's eye: Teaching multimodal large language models to see, and critique like photographers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 24807–24816, 2025. URL https://openaccess.thecvf.com/content/CVPR2025/html/Qi_The_Photographers_Eye_Teaching_Multimodal_Large_Language_Models_to_See_CVPR_2025_paper.html.

Dongyu She, Yu-Kun Lai, Gaoxiong Yi, and Kun Xu. Hierarchical layout-aware graph convolutional network for unified aesthetics assessment. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8475–8484, 2021.

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

Zhengzhong Tu, Hossein Talebi, Han Zhang, Feng Yang, Peyman Milanfar, Alan Bovik, and Yinxiao Li. MaxViT: Multi-axis vision transformer. In *Proceedings of the European Conference on Computer Vision*, 2022.

Qixun Wang, Yang Shi, Yifei Wang, Yuanxing Zhang, Pengfei Wan, Kun Gai, Xianghua Ying, and Yisen Wang. Monet: Reasoning in latent visual space beyond image and language. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 12030–12040, 2026. URL https://openaccess.thecvf.com/content/CVPR2026/html/Wang_Monet_Reasoning_in_Latent_Visual_Space_Beyond_Image_and_Language_CVPR_2026_paper.html.

Haoning Wu, Zicheng Zhang, Weixia Zhang, Chaofeng Chen, Liang Liao, Chunyi Li, Yixuan Gao, Annan Wang, Erli Zhang, Wenxiu Sun, Qiong Yan, Xiongkuo Min, Guangtao Zhai, and Weisi Lin. Q-Align: Teaching LMMs for visual scoring via discrete text-defined levels. In *Proceedings of the International Conference on Machine Learning*, pp. 54015–54029, 2024. URL https://proceedings.mlr.press/v235/wu24ah.html.

Yuzhe Yang, Liwu Xu, Leida Li, Nan Qie, Yaqian Li, Peng Zhang, and Yandong Guo. Personalized image aesthetics assessment with rich attributes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022.

Zhichao Yang, Leida Li, Pengfei Chen, Jinjian Wu, and Weisheng Dong. Semantics-aware image aesthetics assessment using tag matching and contrastive ranking. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pp. 2632–2641, 2024. doi: 10.1145/3664647.3680972.

Qinghao Ye, Haiyang Xu, Jiabo Ye, Ming Yan, Anwen Hu, Haowei Liu, Qi Qian, Ji Zhang, Fei Huang, and Jingren Zhou. mPLUG-Owl2: Revolutionizing multi-modal large language model with modality collaboration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024.

Linfeng Zhang, Jiebo Song, Anni Gao, Jingwei Chen, Chenglong Bao, and Kaisheng Ma. Be Your Own Teacher: Improve the performance of convolutional neural networks via self distillation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 3713–3722, 2019. URL https://openaccess.thecvf.com/content_ICCV_2019/html/Zhang_Be_Your_Own_Teacher_Improve_the_Performance_of_Convolutional_Neural_ICCV_2019_paper.html.

Hancheng Zhu, Leida Li, Jinjian Wu, Sicheng Zhao, Guiguang Ding, and Guangming Shi. Personalized image aesthetics assessment via meta-learning with bilevel gradient optimization. *IEEE Transactions on Cybernetics*, 52(3):1798–1811, 2022. doi: 10.1109/TCYB.2020.2984670.