RLA-NET: RUBRIC-GROUNDED LEARNING FOR DECOMPOSABLE IMAGE AESTHETIC ASSESSMENT

Anonymous authors
Paper under double-blind review

ABSTRACT

A single aesthetic score summarizes preference but does not explain the visual strengths, weaknesses, and trade-offs behind it. Learning interpretable image aesthetic assessment (IAA) from aggregate ratings is challenging: the ratings leave individual aesthetic states underdetermined, prompt–feature similarity does not ensure attribute-specific visual responses, and rubric states alone do not specify their image-dependent importance. We introduce RLA-Net, a CLIP-based framework that jointly estimates readable aesthetic states and learns their adaptive scoring roles through rubric-grounded score composition. Local rubric grounding estimates grades by matching visual evidence to three language-defined reference states for each of eight aesthetic dimensions. Contrastive Rubric Alignment (CRA) complements overall-score supervision with source-relative, rubric-targeted synthetic pairs that constrain global image–anchor responses in prescribed directions. Gated Aesthetic Scoring combines rubric grades, visual descriptors, and image context into explicit prediction terms, allowing similar states to play different scoring roles across images. Training uses image-level ratings, fixed rubric anchors, and synthetic pairs, without per-image rubric annotations, comments, critiques, or teacher-generated targets. Experiments on AVA, PARA, TAD66K, and FLICKR-AES demonstrate competitive performance.

1 INTRODUCTION

Image aesthetic assessment (IAA) predicts human visual preferences and supports image retrieval, curation, and editing (Murray et al., 2012; Talebi & Milanfar, 2018). Similar overall ratings can conceal different aesthetic strengths and weaknesses: strong composition may compensate for poor lighting, or vice versa (Yang et al., 2022; Li et al., 2023). Understanding these trade-offs requires knowing both the state of each rubric and its role in the final judgment. Standard score-supervised IAA learns from image-level ratings, expressed as mean opinion scores (MOS) or rating distributions (Murray et al., 2012; Ren et al., 2017; He et al., 2022). These targets constrain overall preference without specifying the underlying aesthetic states or how they should combine.

Recent IAA research has explored several complementary directions (Fig. 1 left). Holistic discriminative scorers achieve strong aesthetic prediction but typically leave rubric-wise score composition implicit (Gao et al., 2024; Behrad et al., 2025; Yang et al., 2026). CLIP-based methods introduce aesthetic semantics and content-adaptive aggregation, yet often leave implicit how human-readable attribute states contribute to the final score (Li et al., 2023; Huang et al., 2024a; Sheng et al., 2023; Ke et al., 2023; Sheng et al., 2026; Hara & Otsuka, 2026; Wang et al., 2026). Large multimodal models (LMMs) and other critique-oriented approaches support verbal aesthetic reasoning. Still, they rely on computationally demanding backbones and use additional aesthetic-language supervision (Wu et al., 2024; Liu et al., 2026; Huang et al., 2024b; Cao et al., 2026). These limitations motivate a central question for interpretable CLIP-based IAA methods: how can we learn interpretable aesthetic scores and explicitly connect them to image-dependent score composition under aggregate image-level supervision without additional annotations?

Yet, answering this question via a naive CLIP-based IAA framework is challenging. Aggregate ratings leave aesthetic decomposition underconstrained: different states and combinations can predict the same preference (Locatello et al., 2019; Koh et al., 2020). Language priors enable visual concept learning without manual concept labels (Oikarinen et al., 2023), but prompt–feature similarity does

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.

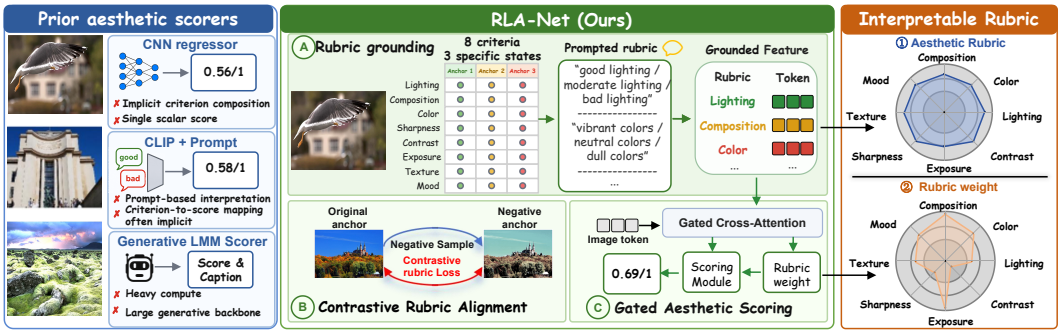

Figure 1: **Rubric-grounded aesthetic score composition with RLA-Net.** Left: representative approaches to image aesthetic assessment. Middle: local rubric grounding (A), CRA's source-relative directional guidance (B), and gated rubric composition (C) are coupled through joint training. Each of the eight rubrics uses three language-defined reference states to connect local visual evidence to a readable grading scale. Right: rubric grades and routing weights describe estimated attribute states and content-dependent aggregation, respectively. This separation allows similar rubric states to play different scoring roles depending on image content. Complete prediction terms also include contextual readouts and grade residuals. Training uses image-level ratings, fixed rubric phrases, and synthetic pairs without per-image rubric annotations.

not ensure attribute-specific responses, and predictive accuracy alone cannot establish their semantic meaning (Lai et al., 2024; Srivastava et al., 2024). Further, an aesthetic aspect (e.g., composition, lighting, or texture) of an image also differs from its scoring role, which depends on the image content (Li et al., 2023; Huang et al., 2024a). Therefore, learning the importance of each aesthetic aspect based on the image content is equally important as learning the decomposition.

To address these challenges, we reformulate CLIP-based IAA as *rubric-grounded score composition* and introduce RLA-Net (Fig. 1, middle and right). We define a *rubric* as an aesthetic dimension with ordered language anchors that specify a grading scale and a direction of semantic variation. This formulation links the above two challenges: estimating an image's state on each aesthetic scale and determining its role in the overall judgment. To learn visually grounded rubric scores without per-image attribute targets, we couple CLIP-based local grounding with Contrastive Rubric Alignment (CRA). Local grounding matches visual evidence to three anchors for each of eight rubrics. Drawing on distortion-based relative learning (Liu et al., 2017), CRA compares a source image with its rubric-targeted perturbation to constrain global image–anchor responses in prescribed directions, without requiring the source image's rubric grade. Shared encoders and anchors connect this directional guidance to local grounding. To account for the different scoring roles of similar aesthetic states across images, we introduce Gated Aesthetic Scoring, which combines rubric grades, visual descriptors, and image context into explicit, image-dependent prediction terms. Through this prediction path, overall-score supervision jointly refines rubric estimation and composition alongside CRA's directional guidance. Thus, semantic guidance and task-driven learning cooperate to produce interpretable rubric scores and their adaptive scoring roles, without additional annotations.

Experiments on AVA, PARA, TAD66K, and FLICKR-AES demonstrate strong performance against state-of-the-art methods. Ablation studies support the predictive benefits of combining rubric grounding, directional alignment, and adaptive score composition, while qualitative analyses illustrate how rubric states and aggregation weights vary with image content. Our contributions are:

- We reformulate IAA as rubric-grounded score composition, linking interpretable, language-referenced aesthetic grades to overall prediction under image-level supervision without additional per-image attribute annotations.
- We propose CRA, which uses rubric-targeted pairs to constrain global image–anchor responses in prescribed directions, supplying information absent from aggregate ratings.
- We develop Gated Aesthetic Scoring that combines rubric grades with contextual visual evidence into explicit prediction terms, allowing similar rubric states to play different scoring roles across images.

2 Related Work

2.1 Image Aesthetic Assessment

Deep IAA has advanced from convolutional regression and rating-distribution learning to content-aware representations, state-space models, and fine-grained ranking (Talebi & Milanfar, 2018; Hosu et al., 2019; He et al., 2022; Gao et al., 2024; Behrad et al., 2025; Yang et al., 2026). Attribute-aware models relate aesthetic factors through supervised attribute learning and dynamic selection (Li et al., 2023; Huang et al., 2024a). Vision–language approaches use multi-attribute alignment, user comments, and generated critiques to enrich scoring (Sheng et al., 2023; Ke et al., 2023; Huang et al., 2025; Li et al., 2025a). Zero-shot prompt methods transfer aesthetic knowledge through model-generated targets and prompts (Sheng et al., 2026). Recent work also learns adaptive semantic combinations from overall scores: prompt-mixture models estimate suitability from image and task context (Hara & Otsuka, 2026), while attribute-subspace methods use semantic anchors and perturbation-estimated sensitivity to guide sample-wise routing (Wang et al., 2026). In the latter, perturbations estimate prediction sensitivity for routing; our relative objective instead constrains the direction of named rubric responses. LMMs offer a complementary route through verbal score levels, critiques, or aesthetic reasoning (Wu et al., 2024; Huang et al., 2024b; Liu et al., 2026; Cao et al., 2026). Our focus is jointly learning language-referenced rubric scores and their scoring roles, coupling targeted directional guidance with aggregate supervision and explicit rubric-wise composition.

2.2 Interpretable Visual Assessment

Interpretable visual assessment exposes explanatory variables or semantic inputs to prediction. Within IAA, attribute-based regressors support post hoc attribution (Soydaner & Wagemans, 2026), while tag-based descriptors expose photography-related image content (Hou et al., 2025). Tag relevance describes an image's relation to photographic concepts; our grounding instead locates each rubric relative to language-defined reference states. Concept bottlenecks provide a broader precedent for predicting through readable intermediate variables, including label-free variants (Koh et al., 2020; Oikarinen et al., 2023). For image quality assessment (IQA), language-prompt methods measure perceptual attributes (Wang et al., 2023), while attribute–scene reasoning and vision–language multitask models incorporate semantic context into quality prediction (Huang et al., 2023; Zhang et al., 2023). Ranking and contrastive learning provide complementary signals: synthetic distortions supply overall-quality rankings (Liu et al., 2017), contrastive objectives learn degradation-sensitive features (Madhusudana et al., 2022; Saha et al., 2023), and tag-guided aesthetic ranking compares images with similar semantic backgrounds (Yang et al., 2024). Perceptual representations and overall rankings do not by themselves specify the response direction of a named aesthetic rubric. RLA-Net couples source-relative directional guidance with image-level supervision to learn interpretable rubric scores and combine grades, rubric-conditioned visual information, and image context into explicit prediction terms.

3 Methodology

Our goal is to learn interpretable rubric scores together with their image-dependent roles in overall aesthetic assessment. In our framework, language anchors provide readable reference states, CRA supplies targeted directional guidance, and the image-level objective trains rubric estimation and content-dependent composition through the prediction path (Fig. 2). Each rubric is represented by a grade and a visual descriptor, allowing the scoring head to use both its estimated state and supporting visual context without per-image rubric annotations.

3.1 Feature Encoding

As shown in the feature encoding block of Fig. 2, we adapt a pretrained CLIP ViT-L/14 through visual and textual low-rank adaptation (LoRA) modules, keeping the backbone parameters frozen (Radford et al., 2021; Hu et al., 2022). A 224×224 image yields a global feature $\mathcal{F}_z$ and $N = 256$ patch features $\mathcal{F}_p = \{\mathbf{p}_n\}_{n=1}^{N}$ in a shared $d = 768$ dimensional space. Patch fea-

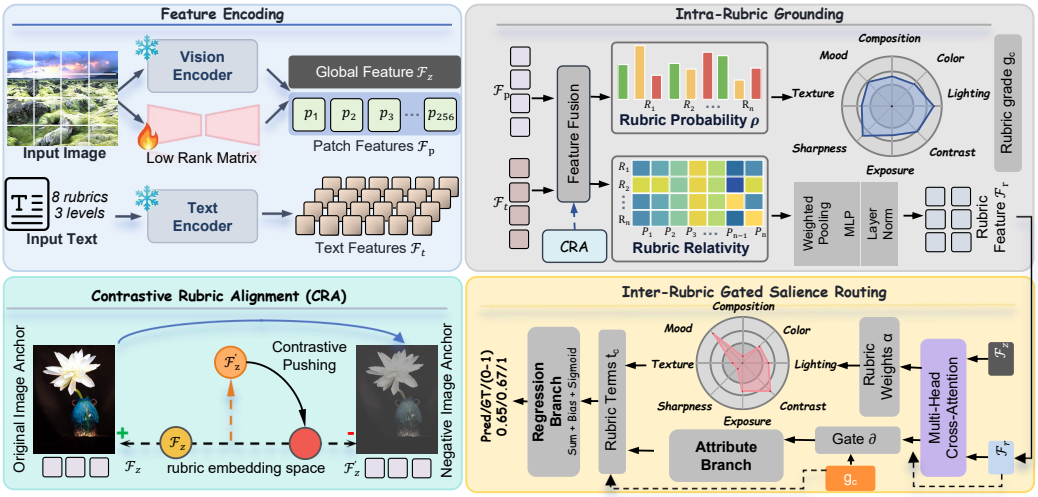

Figure 2: **Overview of RLA-Net.** *Feature encoding*: a CLIP backbone with LoRA adapters produces a global image feature $\mathcal{F}_z$, 256 patch features $\mathcal{F}_p$, and text embeddings $\mathcal{F}_t$ for the 8×3 rubric anchors. *Intra-rubric grounding*: patch–anchor similarities yield level probabilities, rubric grades g_c, local response maps, and pooled rubric descriptors $\mathcal{F}_r^{(c)}$. *Contrastive Rubric Alignment*: source-relative perturbation pairs provide directional guidance on global image–anchor responses through encoders and anchors shared with local grounding. *Gated salience routing*: the global feature queries the rubric descriptors, and a grade-conditioned gate combines each descriptor with shared image context. The shared readout and grade residual form the complete rubric terms t_c in Eq. (8); their sum and the output bias reconstruct the pre-sigmoid prediction. Image-level supervision jointly trains rubric estimation and score composition through this prediction path.

Rubric c	**Level** ℓ=1 (v_1=1)	**Level** ℓ=2 (v_2=0.5)	**Level** ℓ=3 (v_3=0)	**Primary perturbation**
1. Lighting	"good lighting"	"moderate lighting"	"bad lighting"	brightness $\times[0.15, 0.45]$
2. Composition	"good composition"	"average composition"	"bad composition"	resized crop, area $[0.3, 0.6]$, ratio $[0.75, 1.33]$
3. Color	"vibrant colors"	"neutral colors"	"dull colors"	saturation $\times[0.15, 0.45]$, hue ± 0.05
4. Sharpness	"sharp focus"	"slightly blurry"	"very blurry"	Gaussian blur, kernel 11, $\sigma \in [4, 8]$
5. Contrast	"high contrast"	"normal contrast"	"low contrast"	contrast $\times[0.2, 0.5]$
6. Exposure	"well exposed"	"normally exposed"	"poorly exposed"	brightness $\times[0.3, 0.8]$
7. Texture	"rich texture"	"moderate texture"	"smooth surface"	Gaussian blur, kernel 5, $\sigma \in [2, 4]$
8. Mood	"peaceful mood"	"neutral mood"	"dramatic mood"	saturation $\times[0.1, 0.3]$, hue ± 0.1

Table 1: **Rubric anchors and rubric-targeted perturbations.** Grades indicate position between the first and third anchors, not a universal aesthetic preference ordering; mood is a stylistic axis. Primary operations are listed here; additional jitter factors are detailed in Appendix A.3. These transformations prescribe training directions and need not isolate individual rubrics.

tures provide local evidence for grounding; the global feature supports CRA and content-dependent routing.

Our *rubric system* contains $C = 8$ *rubrics*, each defining one aesthetic dimension through $L = 3$ language anchors (Table 1) with values $(v_1, v_2, v_3) = (1, 0.5, 0)$. Anchor wording remains fixed, while textual LoRA adapts their embeddings $\mathcal{F}_t = \{\mathbf{t}_{c,\ell}\}$ to the shared space. These anchors specify reference states, with mood representing style rather than quality. Below, bars denote ℓ_2-normalized features used for cosine matching; unnormalized visual features are retained for pooling and routing.

3.2 Intra-Rubric Grounding

Overall ratings leave individual rubric states underconstrained (Locatello et al., 2019). Language anchors provide reference states, but matching alone does not prescribe responses to targeted visual changes. We couple local matching with CRA and image-level supervision: source-relative per-

turbations guide the shared representation in rubric-specific directions, while the scoring objective trains rubric estimation through its role in the overall prediction.

The grounding block in Fig. 2 produces two complementary outputs per rubric: a rubric grade g_c that scores the estimated rubric state on the anchor-defined scale, and a descriptor $\mathcal{F}_r^{(c)}$ retaining visual information for scoring. For patch–anchor similarities $S_{c,n,\ell} = \bar{\mathbf{p}}_n^\top \bar{\mathbf{t}}_{c,\ell}$, the level probabilities are

$$\pi_{c,n,\ell} = \mathrm{softmax}_\ell(S_{c,n,\ell}/\tau), \qquad \rho_{c,\ell} = \frac{1}{N}\sum_{n=1}^{N} \pi_{c,n,\ell}. \tag{1}$$

The grade averages the patch-level anchor expectations:

$$g_c = \sum_{\ell=1}^{L} \rho_{c,\ell} v_\ell = \frac{1}{N}\sum_{n=1}^{N} u_{c,n}, \qquad u_{c,n} = \sum_{\ell=1}^{L} \pi_{c,n,\ell} v_\ell. \tag{2}$$

Arranging $u_{c,n}$ on the token grid gives a spatial anchor-response map, rather than additive score attributions.

A scalar grade alone discards visual context that may matter to the final judgment. We therefore also pool rubric-relevant features using $R_{c,n} = \max_\ell S_{c,n,\ell}$ and $\omega_{c,n} = \mathrm{softmax}_n(R_{c,n}/\tau_p)$, and fuse them with the corresponding text representation:

$$\mathcal{F}_r^{(c)} = \mathrm{LN}\left(\mathbf{W}_f \left[\sum_n \omega_{c,n}\mathbf{p}_n \,\middle\|\, \sum_\ell \rho_{c,\ell}\bar{\mathbf{t}}_{c,\ell}\right] + \mathbf{b}_f\right), \tag{3}$$

where $\|$ denotes concatenation and τ, τ_p are learned temperatures. Pooling responds to a strong match with any anchor, including the third level. It thus retains evidence associated with either endpoint, allowing the scoring head to consider both the estimated state and the visual features underlying it.

CRA supplies a rubric-targeted directional signal during joint training (Fig. 2, CRA block). For each source image I, a targeted transformation gives $I'_c = T_c(I)$ (Table 1). Comparing I with I'_c supplies a response direction without specifying the source image's rubric grade. For example, blurring gives a training direction toward the low-sharpness anchor. The transformations prescribe response directions rather than per-image target grades; their effects can span several rubrics and depend on content, with mood treated as a stylistic dimension.

CRA operates on global responses $q_{c,\ell}(I) = \mathrm{softmax}_\ell(\bar{\mathcal{F}}_z(I)^\top \bar{\mathbf{t}}_{c,\ell}/\tau)$, sharing the assignment temperature with local grounding. Write q_ℓ, q'_ℓ for the source and transformed responses of the targeted rubric, and $\bar{q} = (q_1 + q_2)/2$. We use

$$\ell_{\mathrm{cra}}(I, I'_c; c) = \max(0, \bar{q}' - \bar{q} + m) + \max(0, q_3 - q'_3 + m), \tag{4}$$

with both margins $m = 0.03$; $\mathcal{L}_{\mathrm{cra}}$ averages over sampled pairs. Since $\bar{q} = (1 - q_3)/2$, both hinges constrain the same difference $\Delta q_3 = q'_3 - q_3$: the loss equals $\max(0, m - \Delta q_3/2) + \max(0, m - \Delta q_3)$ and vanishes for $\Delta q_3 \geq 2m$. The objective supplies a source-relative directional constraint on the global third-anchor response. It leaves redistribution between the first two levels unconstrained and does not impose local three-level ordering or rubric independence.

Shared adapted encoders, rubric embeddings, and assignment temperature connect this global guidance to local grounding. Image-level score gradients reach the same parameters through rubric descriptors and grades, jointly training the local rubric estimates and their scoring roles. Section 4.5 evaluates CRA's predictive effect.

3.3 Inter-Rubric Gated Salience Routing

Estimating rubric states leaves a second question unresolved: how should they enter the overall aesthetic judgment? A similar state can play different scoring roles across images, so its grade alone cannot determine its contribution. We address this with two complementary mechanisms: content-dependent routing weights control aggregation across rubrics, while a grade-conditioned gate controls how each rubric descriptor incorporates shared image context (Fig. 2, gated salience routing block).

To condition routing on image content, cross-attention (Vaswani et al., 2017) uses the global feature $\mathcal{F}_z$ as the query and rubric descriptors as keys and values. For each of $n_h = 4$ heads, affine projections produce $\mathbf{q}^{(h)}$, $\mathbf{k}_c^{(h)}$, and $\mathbf{v}_c^{(h)}$ of dimension d/n_h:

$$\alpha_c^{(h)} = \mathrm{softmax}_c\left(\frac{\langle \mathbf{q}^{(h)}, \mathbf{k}_c^{(h)} \rangle}{\sqrt{d/n_h}}\right), \qquad \mathbf{o}^{(h)} = \sum_{c=1}^{C} \alpha_c^{(h)} \mathbf{v}_c^{(h)}. \tag{5}$$

The routing weight $\alpha_c = n_h^{-1} \sum_h \alpha_c^{(h)}$ satisfies $\alpha_c \geq 0$ and $\sum_c \alpha_c = 1$. Its dependence on both global content and rubric descriptors allows the aggregation weights to vary across images. The head contexts are combined with a global residual to form shared scoring context:

$$\mathbf{o} = \mathrm{LN}\left(\mathbf{W}_O[\mathbf{o}^{(1)} \| \cdots \| \mathbf{o}^{(n_h)}] + \mathbf{b}_O + \mathcal{F}_z\right). \tag{6}$$

Routing determines the aggregation weights, but each rubric's evidence must also be interpreted in context. A shared MLP therefore takes the context, rubric descriptor, and grade to predict a mixing gate:

$$\begin{aligned} \beta_c &= \sigma\Big(\mathrm{MLP}([\mathbf{o} \| \mathcal{F}_r^{(c)} \| g_c])\Big), \\ \tilde{\mathcal{F}}_r^{(c)} &= \beta_c \mathbf{o} + (1 - \beta_c)\mathcal{F}_r^{(c)}. \end{aligned} \tag{7}$$

A smaller β_c retains more of the rubric descriptor; a larger value incorporates more shared context. Thus, β_c adjusts the representation used for scoring, whereas α_c weights its subsequent aggregation. This distinction allows similar grades to produce different scoring terms through their descriptors, context, gates, and routing weights. The gate is learned through the training objective without separate reliability labels.

A shared readout $\hat{s}_c = \mathbf{w}_s^\top \tilde{\mathcal{F}}_r^{(c)} + b_s$ converts the fused representation into a scalar. Together with the grade residual, it gives the complete rubric term and final prediction:

$$t_c = \eta \alpha_c (\hat{s}_c + \lambda_{\mathrm{res}} g_c), \qquad z = b_0 + \sum_{c=1}^{C} t_c, \qquad \hat{y} = \sigma(z), \tag{8}$$

where η, b_0, γ are learned and $\lambda_{\mathrm{res}} = |\gamma|$. This construction makes the score composition explicit: each t_c includes the contextual readout, grade residual, routing weight, and output scale, and the terms exactly reconstruct the pre-sigmoid prediction together with b_0. Score gradients train the shared grounding parameters through the rubric descriptors, the grade-conditioned gate, and the grade residual. Routing weights alone are not complete contributions or human importance scores. The grade-conditioned gate and unconstrained scale do not enforce grade monotonicity. With shared visual context, the decomposition does not establish independent causal effects.

3.4 Training Objective

The joint objective trains the adapters, grounding module, and routing head with smooth-ℓ_1 regression, listwise ranking (Cao et al., 2007), CRA's directional guidance, and regularization:

$$\begin{aligned} \mathcal{L}_t = \mathcal{L}_{\mathrm{reg}} &+ \lambda_{\mathrm{rank}} \mathcal{L}_{\mathrm{rank}} + \chi_t \lambda_{\mathrm{cra}} \mathcal{L}_{\mathrm{cra}} \\ &+ \lambda_{\mathrm{gate}} \mathcal{L}_{\mathrm{gate}} + \lambda_\mu \mathcal{L}_\mu + \lambda_{\mathrm{cond}} \mathcal{H}_{\mathrm{cond}} - \lambda_{\mathrm{marg}} \mathcal{H}_{\mathrm{marg}}. \end{aligned} \tag{9}$$

Regression fits the normalized MOS, and ranking supervises relative scores using the current batch and a queue of detached prediction–target pairs. The gate and mean penalties discourage gate saturation and batch prediction-mean mismatch, respectively. With the reported coefficient signs, both routing entropy terms favor distributed weights; they do not enforce variation across images. The indicator χ_t activates CRA in the same optimizer step on alternating mini-batches, starting with the first batch of each epoch. Detailed loss definitions are in Appendix A.2; coefficients and optimization settings are in Sec. 4.2.

4 Experiments

4.1 Datasets and Evaluation Metrics

We evaluate on AVA (Murray et al., 2012), PARA (Yang et al., 2022), TAD66K (He et al., 2022), and FLICKR-AES (Ren et al., 2017), training RLA-Net separately on each official training split

and evaluating on the corresponding test split. Training uses image-level ratings, without PARA attribute annotations or TAD66K theme labels. These experiments assess target-dataset prediction rather than unseen-dataset transfer. We report Spearman's rank correlation coefficient (SRCC) and Pearson's linear correlation coefficient (PLCC) for overall score prediction (higher is better). Supplementary output distances use mean absolute error (MAE) and rubric-averaged root mean squared error (macro RMSE); definitions are in Appendix B.

4.2 Implementation Details

We use CLIP ViT-L/14 with 224×224 inputs and visual/textual LoRA (rank 8, scale 16, dropout 0.2). AVA training uses AdamW with learning rate 5×10^{-5}, weight decay 10^{-4}, cosine decay, six epochs, batch size 32, seed 42, mixed precision, and gradient clipping at 1.0. The ranking queue holds 256 prediction–target pairs with $\tau_r = 0.2$. Four CRA pairs are sampled every other score mini-batch from a fixed training pool targeting 1,000 pairs per rubric; both margins are 0.03. In Eq. 9, we set $\lambda_{\text{rank}} = 10^{-3}$, $\lambda_{\text{cra}} = 10^{-2}$, $\lambda_{\text{gate}} = 10^{-3}$, $\lambda_{\mu} = 0.5$, $\lambda_{\text{cond}} = -10^{-4}$, and $\lambda_{\text{marg}} = 0.03$. AVA MOS is normalized as $(\text{MOS} - 1)/9$. We reserve 20% of the non-test pool for validation (seed 42) and report test results from the checkpoint with the highest validation SRCC. Preprocessing, initialization, and pair construction are detailed in Appendix A.3.

4.3 Comparison with State-of-the-Art Methods

We examine whether explicit rubric modeling can support competitive overall score prediction across four aesthetic benchmarks. The comparison covers CNN, transformer, and state-space scorers, together with CLIP-based vision–language methods (Table 2).

Starred entries are our direct evaluations of official AesCLIP checkpoints on TAD66K/FLICKR-AES and FGAesQ checkpoints on PARA/FLICKR-AES, without additional training or fine-tuning; other baseline results are cited reports. These checkpoint evaluations differ from RLA-Net's target-dataset training and should be interpreted accordingly. FGAesQ's reported AVA-to-TAD66K transfer is identified separately and excluded from highlighting.

Method	**AVA**		**PARA**		**TAD66K**		**FLICKR-AES**	
	SRCC	PLCC	SRCC	PLCC	SRCC	PLCC	SRCC	PLCC
CNN, transformer, and state-space scorers								
NIMA (Talebi & Milanfar, 2018)	0.612	0.636	0.882	0.922	0.390	0.405	0.478	0.513
HGCN (She et al., 2021)	0.665	0.687	0.865	0.881	0.486	0.493	–	–
MUSIQ (Ke et al., 2021)	0.726	0.738	0.882	0.918	0.489	0.517	0.533	0.569
TANet (He et al., 2022)	0.758	0.765	0.883	0.917	**0.513**	0.531	–	–
AesMamba-V (Gao et al., 2024)	0.751	0.760	0.902	0.936	0.475	0.503	–	–
CLIP-based vision–language scorers								
AesCLIP (Sheng et al., 2023)	0.771	0.779	0.926	**0.951**	0.094*	0.092*	0.219*	0.236*
VILA-R (Ke et al., 2023)	0.774	0.774	0.651	0.658	0.413	0.443	0.616	0.645
IAACLIP (Li et al., 2025b)	0.788	0.791	–	–	0.508	0.537	–	–
FGAesQ (Yang et al., 2026)	0.770	0.781	0.630*	0.593*	0.423†	0.455†	0.636*	0.675*
RLA-Net (Ours)	**0.813**	**0.812**	**0.927**	**0.951**	0.508	**0.539**	**0.812**	**0.832**

Table 2: **Image-level prediction.** Red bold/blue indicate the highest/second-highest distinct values, including ties. *: our official-checkpoint tests without additional training; other baseline values: cited reports. Highlighting ranks point estimates, not matched training conditions or statistical significance. †: FGAesQ transfer from AVA with additional fine-grained ranking data, excluded from highlighting. "–": unlisted result.

On AVA, RLA-Net reaches $0.813/0.812$ SRCC/PLCC, exceeding the strongest listed baseline, IAA-CLIP, by $0.025/0.021$. This result shows that exposing rubric states and their score composition is compatible with strong image-level prediction. The advantage is not uniform across benchmarks: on PARA, RLA-Net's $0.927/0.951$ is nearly tied with AesCLIP's $0.926/0.951$. Here, the model provides an explicit rubric-wise computation at comparable prediction performance, rather than a substantial correlation gain.

TAD66K further distinguishes linear association from ranking performance: RLA-Net obtains the highest listed PLCC (0.539), while its SRCC (0.508) ties IAACLIP and remains below TANet (0.513). Thus, the PLCC result does not imply uniformly better image ordering. On FLICKR-AES, RLA-Net reaches $0.812/0.832$ after target-dataset training; the comparison with starred checkpoint results also reflects different training protocols. Section 4.5 examines the effects of individual design changes. Additional overall-score comparisons appear in Appendix D, Fig. D.1.

4.4 Qualitative Analysis

To examine what RLA-Net reveals beyond an overall score, we visualize its rubric states and routing weights alongside predictions from AesCLIP, VILA, and FGAesQ. Figure 3 presents four examples with different visual content, exposing both the estimated states and their image-dependent aggregation.

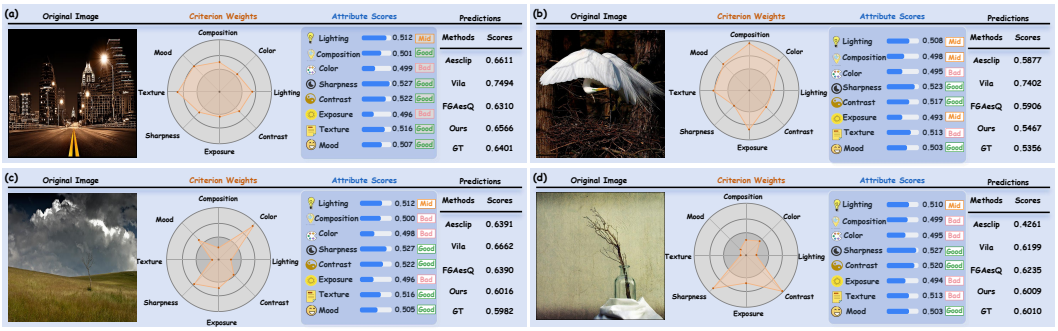

Figure 3: **Rubric states, routing profiles, and overall predictions.** Radar plots show routing weights α_c; bars show intermediate grades g_c used for scoring, and tags indicate modal anchors. Expectations and modes summarize different aspects of the anchor distributions, so similar grades can have different tags. Mood denotes a stylistic axis.

The landscape in (c) assigns its largest routing weight to color, whereas the still life in (d) emphasizes contrast and sharpness. Yet their color grades are similar (0.498 and 0.495). This illustrates the distinction between rubric state and scoring role: a rubric grade estimates the rubric state on the anchor-defined scale, while routing depends on image content and rubric descriptors. The two profiles therefore expose complementary aspects of the computation. Routing weights alone are not the complete score contributions, which also include the contextual readout and grade residual in Eq. 8.

RLA-Net is closest to GT among the displayed methods in (b)–(d), with absolute errors of 0.0111, 0.0034, and 0.0001, respectively; FGAesQ is closer in (a). These selected examples show how explicit rubric outputs accompany accurate overall predictions, without establishing that the routing profiles match human judgments.

4.5 Ablation Studies

We test the roles of local evidence, CRA, routing and gating, and anchor granularity in overall prediction. Table 3 reports AVA variants with matched splits, normalization, optimization budgets, seed, and validation-based selection, except for the specified changes. The frozen scalar probe (a) is a separate reference: the transition to (b) also changes adaptation and the prediction head, so it does not isolate the benefit of rubric modeling.

The effect of local grounding depends on the scoring configuration. Replacing it with global grounding lowers SRCC/PLCC from $0.813/0.812$ in (e) to $0.801/0.800$ in (g), whereas introducing local grounding in the simpler (b)–(c) comparison reduces performance. Local evidence therefore does not yield an unconditional gain; its predictive value depends on the surrounding design. Likewise, jointly adding routing and gating to (d) raises performance from $0.792/0.791$ to $0.813/0.812$. These comparisons support combining local evidence with content-dependent score composition, although they do not separate the effects of routing and gating.

Table 3: **AVA ablations.** (a)–(e): reference and progressive construction; (f)–(j): full-model variants. Routing and gate are varied jointly. (b)/(g) use global grounding; other dashes denote absent/inapplicable components. Bold marks the highest point estimates. Variant configurations are summarized in Appendix A.4.

Exp.	Local Grounding	Alignment	Rubric Routing + Gate	Rubric Anchors	SRCC ↑	PLCC ↑
Reference and progressive construction						
(a)	–	–	–	–	0.754	0.754
(b)	–	–	–	3-level	0.799	0.797
(c)	✓	–	–	3-level	0.791	0.790
(d)	✓	CRA	–	3-level	0.792	0.791
(e)	✓	CRA	✓	3-level	**0.813**	**0.812**
Variants of the full model						
(f)	✓	–	✓	3-level	0.794	0.793
(g)	–	CRA	✓	3-level	0.801	0.800
(h)	✓	CRA	✓	2-level	0.794	0.793
(i)	✓	InfoNCE	✓	3-level	0.791	0.790
(j)	✓	CRA (4 rubrics)	✓	3-level	0.794	0.793

CRA contributes beyond the grounding architecture alone. Removing it from the full model gives $0.794/0.793$ in (f), and replacing it with the evaluated InfoNCE objective gives $0.791/0.790$ in (i). The latter comparison supports the value of specifying a named rubric's response direction within a source–transformation pair, beyond the tested alternative contrastive objective. Restricting CRA to four rubrics also lowers performance to $0.794/0.793$ in (j), despite retaining all eight scoring rubrics. This result favors broader alignment coverage within the tested configuration.

The anchor-granularity comparison examines the reference-state structure. Using two levels in (h) yields $0.794/0.793$, below the full model's $0.813/0.812$ with three levels. Retaining rubric branches and learned aggregation is therefore insufficient to recover the full model's performance with the tested two-level configuration. This comparison supports the three-level design for overall prediction, but does not establish that three levels are universally optimal or that the estimated rubric states match human judgments. These ablations show predictive differences in the reported runs; they do not directly establish rubric-state accuracy. The variant configurations are summarized in Appendix A.4.

4.6 Exploratory Output Comparison

Appendix B reports an exploratory numerical comparison of RLA-Net and AesCLIP outputs against a fixed ArtiMuse reference. RLA-Net's post-processed grade profile is normalized across rubrics, whereas the reference and AesCLIP readouts are per-rubric anchor expectations. The reported distances therefore describe different output pipelines; they do not validate the absolute rubric states used for MOS prediction.

5 Conclusion

We presented RLA-Net, which jointly learns interpretable rubric scores and their image-dependent scoring roles. Language anchors define reference states, CRA guides shared representations in rubric-specific directions, and image-level supervision trains rubric estimation together with gated score composition. The complete prediction terms yield an explicit pre-sigmoid score decomposition. Training uses overall ratings, fixed anchors, and synthetic pairs without per-image rubric annotations. Four-benchmark results show competitive prediction, and AVA ablations support the complete design. This inspectable computation does not establish human-aligned or causal explanations. Future work should validate rubric states against human annotations and test robustness beyond the fixed rubric system and content-dependent perturbation priors.

Ethics Statement

This work studies image aesthetic assessment using existing benchmarks and their image-level ratings. Aesthetic judgments are subjective and context-dependent; the ratings, language anchors, and pretrained models may reflect cultural preferences and representation biases. Consequently, RLA-Net's predictions should not be interpreted as universal standards of aesthetic quality or judgments about the worth of individuals depicted in images. When used for image ranking or curation, such predictions may reinforce dominant visual preferences and reduce the visibility of underrepresented styles. Explicit rubric outputs make the computation inspectable but do not by themselves establish fairness or human-aligned explanations. The ArtiMuse-generated reference is used only for exploratory evaluation and is not treated as human-validated ground truth. Any downstream use should respect image copyrights, applicable data-use terms, and the privacy of depicted individuals, and should include context-specific bias assessment and appropriate human oversight.

AI Use Statement

Generative AI tools assisted with language polishing and organization. They also supported literature retrieval and discovery, including identifying relevant studies and locating sources for checking bibliographic information. ArtiMuse was used to generate rubric-conditioned reference labels for the exploratory evaluation on 3,000 AVA images described in Appendix B. These model-generated labels were used solely for evaluation, were not used to train RLA-Net, and are not treated as human-validated ground truth. The authors take responsibility for the final manuscript, including the accuracy of its text, citations, scientific claims, and AI-assisted content.

Reproducibility Statement

The Methodology section specifies the rubric system, CRA objective, and complete pre-sigmoid prediction terms. Section 4.2 reports the AVA optimization settings, loss coefficients, queue, validation split, and model selection. Appendix A provides detailed computations, preprocessing, pair construction, and ablation definitions. The experiments distinguish official-checkpoint tests from dataset-specific RLA-Net training. Appendix B specifies the external reference, rubric-output transformations, prompts, and error aggregation.

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 (CVPR)*, pp. 7815–7824, 2025. doi: 10.1109/CVPR52734.2025.00732.

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 (CVPR)*, pp. 15313–15322, 2026.

Zhe Cao, Tao Qin, Tie-Yan Liu, Ming-Feng Tsai, and Hang Li. Learning to rank: From pairwise approach to listwise approach. In *Proceedings of the International Conference on Machine Learning (ICML)*, pp. 129–136, 2007.

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

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 (CVPR)*, 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 International Joint Conference on Artificial Intelligence (IJCAI)*, pp. 942–948, 2022.

Vlad Hosu, Bastian Goldlücke, and Dietmar Saupe. Effective aesthetics prediction with multi-level spatially pooled features. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9375–9383, 2019.

Jingwen Hou, Weisi Lin, Yuming Fang, Haoning Wu, Chaofeng Chen, Liang Liao, and Weide Liu. Toward transparent deep image aesthetics assessment with tag-based content descriptors. *IEEE Transactions on Image Processing*, 34:3070–3085, 2025. doi: 10.1109/TIP.2023.3308852.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2022.

Yipo Huang, Leida Li, Yuzhe Yang, Yaqian Li, and Yandong Guo. Explainable and generalizable blind image quality assessment via semantic attribute reasoning. *IEEE Transactions on Multimedia*, 25:7672–7685, 2023. doi: 10.1109/TMM.2022.3225728.

Yipo Huang, Leida Li, Pengfei Chen, Jinjian Wu, Yuzhe Yang, Yaqian Li, and Guangming Shi. Coarse-to-fine image aesthetics assessment with dynamic attribute selection. *IEEE Transactions on Multimedia*, 26:9316–9329, 2024a. doi: 10.1109/TMM.2024.3389452.

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 ACM International Conference on Multimedia (ACM MM)*, pp. 5911–5920, 2024b.

Yipo Huang, Leida Li, Pengfei Chen, Haoning Wu, Weisi Lin, and Guangming Shi. Multi-modality multi-attribute contrastive pre-training for image aesthetics computing. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 47(2):1205–1218, 2025. doi: 10.1109/TPAMI.2024.3492259.

Junjie Ke, Qifei Wang, Yilin Wang, Peyman Milanfar, and Feng Yang. MUSIQ: Multi-scale image quality transformer. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pp. 5148–5157, 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 (CVPR)*, pp. 10041–10051, 2023.

Pang Wei Koh, Thao Nguyen, Yew Siang Tang, Stephen Mussmann, Emma Pierson, Been Kim, and Percy Liang. Concept bottleneck models. In *Proceedings of the International Conference on Machine Learning (ICML)*, pp. 5338–5348, 2020.

Songning Lai, Lijie Hu, Junxiao Wang, Laure Berti-Equille, and Di Wang. Faithful vision-language interpretation via concept bottleneck models. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2024. URL `https://openreview.net/forum?id=rp0EdI8X4e`.

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.

Leida Li, Xiangfei Sheng, Pengfei Chen, Jinjian Wu, and Weisheng Dong. Towards explainable image aesthetics assessment with attribute-oriented critiques generation. *IEEE Transactions on Circuits and Systems for Video Technology*, 35(2):1464–1477, 2025a. doi: 10.1109/TCSVT.2024.3470870.

Zhuo Li, Xingao Yan, Xuebin Wei, and Feng Shao. IAACLIP: Image aesthetics assessment via CLIP. *Electronics*, 14(7):1425, 2025b.

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 *Proceedings of the International Conference on Learning Representations (ICLR)*, 2026. URL https://openreview.net/forum?id=or3ZukbrKw.

Xialei Liu, Joost van de Weijer, and Andrew D. Bagdanov. RankIQA: Learning from rankings for no-reference image quality assessment. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pp. 1040–1049, 2017. doi: 10.1109/ICCV.2017.118.

Xiao-Chang Liu and Johan Wagemans. From concepts to judgments: Interpretable image aesthetic assessment. *arXiv preprint arXiv:2603.18108*, 2026.

Francesco Locatello, Stefan Bauer, Mario Lucic, Gunnar Raetsch, Sylvain Gelly, Bernhard Schoelkopf, and Olivier Bachem. Challenging common assumptions in the unsupervised learning of disentangled representations. In *Proceedings of the International Conference on Machine Learning (ICML)*, pp. 4114–4124, 2019.

Pavan C. Madhusudana, Neil Birkbeck, Yilin Wang, Balu Adsumilli, and Alan C. Bovik. Image quality assessment using contrastive learning. *IEEE Transactions on Image Processing*, 31:4149–4161, 2022.

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

Tuomas Oikarinen, Subhro Das, Lam M. Nguyen, and Tsui-Wei Weng. Label-free concept bottleneck models. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2023.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *Proceedings of the International Conference on Machine Learning (ICML)*, pp. 8748–8763, 2021.

Jian Ren, Xiaohui Shen, Zhe Lin, Radomir Mech, and David J. Foran. Personalized image aesthetics. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pp. 638–647, 2017.

Avinab Saha, Sandeep Mishra, and Alan C. Bovik. Re-IQA: Unsupervised learning for image quality assessment in the wild. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5846–5855, 2023.

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 (CVPR)*, pp. 8475–8484, 2021.

Xiangfei Sheng, Leida Li, Pengfei Chen, Jinjian Wu, Weisheng Dong, Yuzhe Yang, Liwu Xu, Yaqian Li, and Guangming Shi. AesCLIP: Multi-attribute contrastive learning for image aesthetics assessment. In *Proceedings of the ACM International Conference on Multimedia (ACM MM)*, pp. 1117–1126, 2023. doi: 10.1145/3581783.3611969.

Xiangfei Sheng, Leida Li, Pengfei Chen, Li Cai, and Giuseppe Valenzise. AesPrompt: Zero-shot image aesthetics assessment with multi-granularity aesthetic prompt learning. *IEEE Transactions on Multimedia*, 28:958–971, 2026. doi: 10.1109/TMM.2025.3632637.

Derya Soydaner and Johan Wagemans. Unveiling the factors of aesthetic preferences with explainable AI. *British Journal of Psychology*, 117(2):444–478, 2026. doi: 10.1111/bjop.12707.

Divyansh Srivastava, Ge Yan, and Tsui-Wei Weng. VLG-CBM: Training concept bottleneck models with vision-language guidance. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 37, 2024. doi: 10.52202/079017-2510.

Hossein Talebi and Peyman Milanfar. NIMA: Neural image assessment. *IEEE Transactions on Image Processing*, 27(8):3998–4011, 2018.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 30, 2017.

Jianyi Wang, Kelvin C. K. Chan, and Chen Change Loy. Exploring CLIP for assessing the look and feel of images. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, volume 37, pp. 2555–2563, 2023. doi: 10.1609/aaai.v37i2.25353.

Ye Wang, Maocai Dai, Jiang Xie, Xiuli Bi, Fei Tao, Xiao Li, and Hong Yu. When attributes disagree: Gradient conflict in image aesthetic assessment. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2026. URL https://icml.cc/virtual/2026/poster/65091.

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 (ICML)*, volume 235, pp. 54015–54029, 2024.

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 (CVPR)*, pp. 19861–19869, 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 ACM International Conference on Multimedia (ACM MM)*, pp. 2632–2641, 2024. doi: 10.1145/3664647.3680972.

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. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 145–155, 2026.

Weixia Zhang, Guangtao Zhai, Ying Wei, Xiaokang Yang, and Kede Ma. Blind image quality assessment via vision-language correspondence: A multitask learning perspective. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 14071–14081, 2023. doi: 10.1109/CVPR52729.2023.01352.