

UC Merced

Proceedings of the Annual Meeting of the Cognitive Science Society

Title

Learning to Look before Learning to Like: Incorporating Human Visual Cognition into Aesthetic Quality Assessment

Permalink

<https://escholarship.org/uc/item/3xw267sn>

Journal

Proceedings of the Annual Meeting of the Cognitive Science Society, 48(0)

Authors

Yu, Liwen

Liu, Chi

Han, Xiaotong

et al.

Publication Date

2026

Copyright Information

This work is made available under the terms of a Creative Commons Attribution License, available at <https://creativecommons.org/licenses/by/4.0/>

Peer reviewed

Learning to Look before Learning to Like: Incorporating Human Visual Cognition into Aesthetic Quality Assessment

Liwen Yu¹, Chi Liu (chiliu@cityu.edu.mo)^{1,†}, Xiaotong Han², Congcong Zhu¹, Minghao Wang¹ & Sheng Shen³

¹Faculty of Data Science, City University of Macau, Taipa, Macao, China

²Zhongshan Ophthalmic Center, Sun Yat-sen University, Guangzhou, China

³School of computer science, Torrens University Australia, Melbourne, Victoria, Australia

[†]Corresponding Author

Abstract

Automated Aesthetic Quality Assessment (AQA) treats images primarily as static pixel vectors, aligning predictions with human-rating scores largely through semantic perception. However, this paradigm diverges from human aesthetic cognition, which arises from dynamic visual exploration shaped by scanning paths, processing fluency, and the interplay between bottom-up saliency and top-down intention. We introduce AestheticNet, a novel cognitive-inspired AQA paradigm that integrates human-like visual cognition and semantic perception with a two-pathway architecture. The visual attention pathway, implemented as a gaze-aligned visual encoder (GAVE) pre-trained offline on eye-tracking data using resource-efficient contrast gaze alignment, models attention from human vision system. This pathway augments the semantic pathway, which uses a fixed semantic encoder such as CLIP, through cross-attention fusion. Visual attention provides a cognitive prior reflecting foreground/background structure, color cascade, brightness, and lighting, all of which are determinants of aesthetic perception beyond semantics. Experiments validated by hypothesis testing show a consistent improvement over the semantic-alone baselines, and demonstrate the gaze module as a model-agnostic corrector compatible with diverse AQA backbones, supporting the necessity and modularity of human-like visual cognition for AQA. Our code is available at github.com/keepgallop/AestheticNet

Keywords: Aesthetic Assessment; Gaze; Cognitive Modeling.

Introduction

Automated Aesthetic Quality Assessment (AQA) is a fundamental computer vision task that facilitates human-computer interaction, image understanding, AI cognition and game design. The standard paradigm trains a computational model to predict the aesthetic score of an image to match human ratings. Existing models, from CNN-based architectures like NIMA (Talebi & Milanfar, 2018) to recent Large Vision-Language Models like Q-Align (Wu et al., 2024), treat images as static pixel vectors and evaluate aesthetic quality by capturing pixel correlation and semantic content at only a single glance.

However, despite using human labels as ground truth, this paradigm misaligns with the human visual system (HVS), deviating from the true nature of human aesthetic cognition which is neither instantaneous nor static. Early eye-tracking studies (Yarbus, 1967) have proven that humans actively perceive scenes via scanpaths (Noton & Stark, 1971), driven by a dynamic interplay between bottom-up saliency (e.g., contrast, orientation) and top-down semantic intent. This active explo-

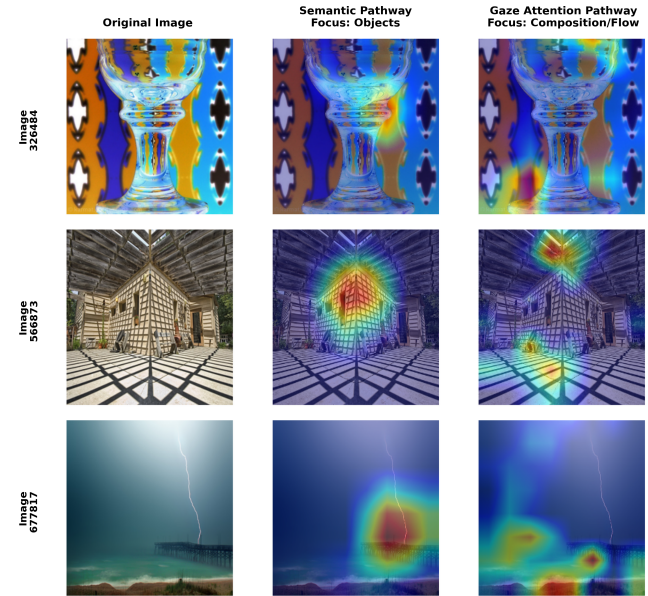


Figure 1: Attention divergence between two AQA pathways. The semantic pathway (middle) exhibits object-centric bias, while the gaze attention pathway (right) prioritizes scene compositional properties and visual flow, implying the synergy and complementary effect.

ration for “looking” strongly correlates with the outcome of “liking” (Shimamura & Palmer, 2012). Processing Fluency Theory further suggests that aesthetic pleasure arises from how easily visual information is structured and navigated by the eye-brain system (Reber et al., 2004): a compelling image that effectively guide eye gaze can reduce visual cognitive load substantially.

Building upon these theories, we raise the question that *whether it is both beneficial and feasible to incorporate human-like visual cognition into machine perception for AQA*. Models that overlook such a cognitive mechanism may risk semantic bias, overemphasizing semantic recognition (e.g., “sunset,” “dog”) while lacking an intrinsic understanding of how visual flow shapes perception. As illustrated in Figure 1, pure semantic encoders (e.g., CLIP (Radford et al., 2021)) exhibit a rigid object-centric preference. By contrast, a human-like cognition encoder modeled on eye behavior scans

composite scene properties such as foreground/background, chromatic stratification, luminance, illumination and inconsistencies, all of which significantly influence aesthetic judgment in addition to semantics.

Inspired by the brain’s Dual-Process Theory (Kahneman, 2011), we propose *AestheticNet*, a novel, cognitively inspired AQA paradigm that unites human-like visual cognition with semantic perception via a two-branch architecture, bridging the gap between “recognition” and “appreciation” of scenes. Aesthetic judgments emerge from the synergy between a dynamic, objective visual attention pathway and a static, subjective semantic interpretation pathway.

Concretely, *AestheticNet* augments semantic representations with HVS-derived visual attention captured by a dedicated gaze encoder, ensuring that aesthetic judgments reflect both human-viewing experience and machine-perceived semantics. This visual pathway utilizes a gaze-aligned visual encoder (GAVE), which is offline-pretrained on eye-tracking data to model how humans focus during aesthetic evaluation. We introduce Contrastive Gaze Alignment for pretraining, enabling resource-efficient learning that captures universal visual attentions with only about a hundred training images. The GAVE is incorporated as a complementary branch to the main semantic branch (e.g., a fixed CLIP encoder) (Radford et al., 2021) via cross-attention fusion, offering visual cognitive priors to enhance AQA. Notably, the human-like visual cognition module serves as an orthogonal cognitive complement, computationally decoupled from semantic perception. Hence, HVS-derived attention functions as a model-agnostic corrector, compatible with diverse AQA backbones as a plug-and-play cognitive prior that consistently improves performance.

To rigorously validate the effect of human-like visual cognition in AQA, we conduct a hypothesis test on its necessity and modularity. Across extensive experiments with *AestheticNet*, we provide empirical evidence supporting the hypotheses that incorporating HVS-derived attention generally enhances AQA to significantly outperform semantic-only baselines. Moreover, these learned gaze attentions function as a flexible, model-agnostic corrector, showing seamless compatibility with diverse AQA backbones, from CNNs to MLLMs, as a plug-and-play biological prior module that consistently yields performance gains.

Related Work

Computational Aesthetics

The development of AQA marks a shift from rules-based modeling to deep representation learning. Early methods used handcrafted features, such as the rule of thirds, color harmony, and depth of field to quantify photographic rules (Datta et al., 2006). But these features, while explainable, failed to capture high-level semantic context. More advanced methods predict the score distribution directly from the global image to match the human-rating ground truth (Su et al., 2020; Talebi & Milanfar, 2018). Recently, multimodal large language models such as Q-Align have utilized massive text-image pre-training

to align visual features with human aesthetic feedback (Wu et al., 2024). However, these end-to-end deep representation methods treat aesthetics as static, image-centric attributes and may introduce semantic bias by ignoring the cognitive dynamics of how humans perceive visual aesthetics.

Visual Attention in Aesthetics

Visual attention modeling has predominantly focused on generic Saliency Detection (Itti et al., 1998), which predicts fixation probability based on low-level feature contrast. However, aesthetic attention differs fundamentally from generic saliency; it is driven by artistic composition and visual balance rather than mere object distinctiveness. While prior works (Lu et al., 2015; Ma et al., 2017) have integrated saliency maps derived from salient object detection paradigms into AQA pipelines, they inherently introduce an object-centric bias and fail to reflect the exploratory nature of human aesthetic viewing behaviors. Differently, *AestheticNet* encodes human visual system-derived attention via a dedicated Gaze Encoder. This mechanism internalizes the oculomotor priors associated with aesthetic appreciation, enabling a more precise assessment that transcends simple object recognition.

Methods

Our computational framework (visualized in Figure 2) operationalizes the cognitive interaction between active vision mechanisms and semantic interpretation. The architecture centers on the *AestheticNet*, which integrates a pre-trained, resource-rational Gaze Encoder to synthesize perceptual structure with semantic intent.

The Gaze Encoder: Learning to Look

Standard saliency models typically rely on supervised pixel-wise regression trained on large-scale datasets. We argue that this “big data” paradigm lacks ecological validity. In human development, visual attention is an early-emerging capability; infants demonstrate sophisticated scanpath patterns driven by low-level features months before acquiring the semantic vocabulary to name objects. This suggests that the mechanism of visual attention is not a high-dimensional semantic function requiring millions of samples, but a low-dimensional biological prior. To model this, in the first stage, we propose a resource-rational learning strategy utilizing Contrastive Gaze Alignment (CGA).

We utilize DINOv2 (Oquab et al., 2024) (ViT-Small/14) as the computational proxy for the primary visual cortex (V1/V2). Unlike supervised CNNs that learn via explicit human annotation—analogue to formal instruction—DINOv2 (Oquab et al., 2024) derives representations solely from the statistical structure of visual data via self-supervision. Its self-attention mechanism naturally segments figure from ground without supervision, effectively providing a “phylogenetic” prior: a pre-wired sensitivity to visual structure that exists in the biological system prior to specific aesthetic training.

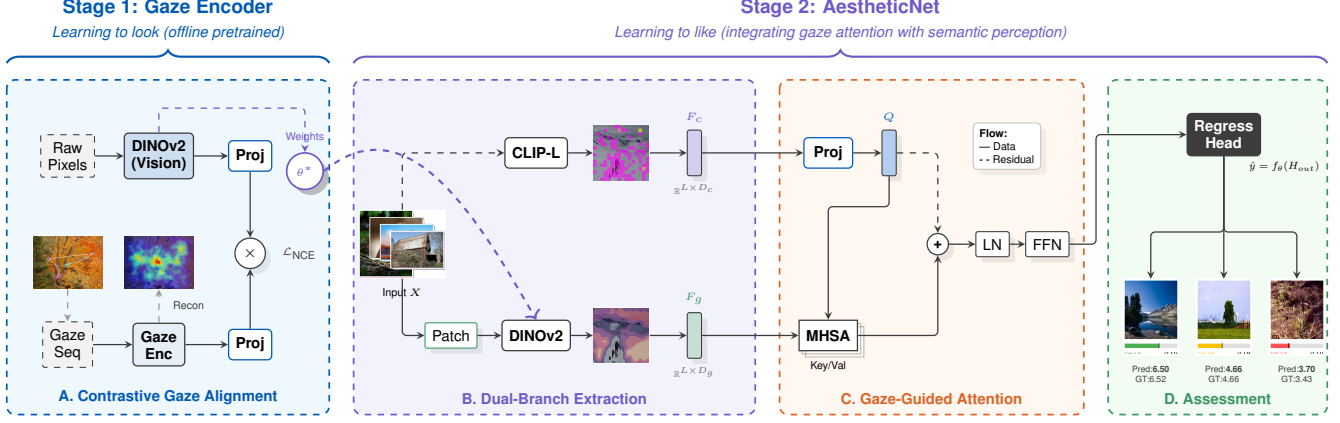


Figure 2: **The Cognitive Architecture of AestheticNet.** This framework achieves aesthetic perception through two stages and four steps. **(A) Contrastive Gaze Alignment (CGA):** The Gaze Encoder aligns raw pixels with eye-tracking sequences using contrastive loss ($\mathcal{L}_{CGA}$) to learn a general "gaze grammar". **(B) Dual-Branch Extraction:** A frozen Semantic Encoder and the Gaze-Aligned Visual Encoder (GAVE)—the visual backbone of the pre-trained Gaze Encoder—extract content and perceptual form. **(C) Gaze-Guided Attention:** Semantic representations (Q) actively query gaze-aligned features (K, V) via a directed attention mechanism. **(D) Assessment:** The synthesized representation is mapped to a scalar aesthetic score $\hat{y}$. (MHSA: Multi-Head Self-Attention; LN: Layer Normalization; FFN: Feed-Forward Network).

To fine-tune this biological backbone, we depart from data-hungry paradigms to reflect HVS resource rationality: evolving oculomotor priors minimizes the cognitive load of decoding every pixel semantically. We utilized a minimalist set of 109 images (Wilming et al., 2017). Although the image count is small, they provide extremely dense, high-fidelity behavioral scanpaths. By fitting universal human sequential behaviors rather than memorizing static pixels, the model effectively prevents overfitting. Consequently, the Image Encoder distills the universal grammar of looking—such as saccadic amplitude and center bias—robustly from these dense traces.

We employ a dual-tower architecture where the Image Encoder (E_I , initialized with DINOv2 (Oquab et al., 2024)) and a Gaze Transformer (E_G) are aligned via a symmetric InfoNCE loss. This maximizes the mutual information between the visual representation and the aggregate human eye movement data. Let $s_{i,j} = \text{sim}(v_i^I, v_j^G)$ denote the similarity between the i -th image and j -th gaze embedding. The loss is defined as:

$$\mathcal{L}_{CGA} = -\frac{1}{2N} \sum_{i=1}^N \left[\log \frac{e^{s_{i,i}/\tau}}{\sum_k e^{s_{i,k}/\tau}} + \log \frac{e^{s_{i,i}/\tau}}{\sum_k e^{s_{k,i}/\tau}} \right] \quad (1)$$

where $\tau = 0.05$ is the temperature parameter controlling the sharpness of the alignment. This objective forces the visual encoder to distinguish the specific oculomotor signature of an image from distractors, effectively "grounding" the DINOv2 (Oquab et al., 2024) features in biological behavior.

AestheticNet: Learning to Like

In the second stage, we construct the AestheticNet by synthesizing the learned gaze priors with high-level semantic

understanding using a dual-pathway model. The model architecture consists of two parallel pathways: the Visual Attention Branch, which utilizes a pre-trained GAVE to extract perceptual features $h_F \in \mathbb{R}^d$ representing human vision dynamics; and the Semantic Branch, which employs a frozen CLIP-ViT-L/14 (Radford et al., 2021) backbone to extract semantic embeddings $h_C \in \mathbb{R}^d$.

Gaze-Guided Attention Integration. Unlike passive concatenation strategies that treat form and content as static vectors, we model aesthetic judgment as an active interrogation process. We suggest that abstract semantic intents modulate the readout of perceptual features—a mechanism akin to top-down cognitive control. To operationalize this, we employ a Cross-Attention mechanism where the semantic latent vector governs the attentional focus over the perceptual stream. Formally, we define the semantic embedding as the Query (Q) and the perceptual features as both Key (K) and Value (V). The modulated representation and final prediction are computed as:

$$h_{attn} = \text{Softmax} \left(\frac{(h_C W_Q)(h_F W_K)^T}{\sqrt{d_k}} \right) (h_F W_V) \quad (2)$$

$$\hat{y} = \mathcal{M}_\phi (\text{Concat}[h_{attn}, h_C])$$

where W_Q, W_K, W_V are learnable projection matrices, d_k is the scaling factor, and $\mathcal{M}_\phi$ denotes the final regression Multi-Layer Perceptron. This formulation mathematically simulates active inference: the projection $(h_C W_Q)$ constructs a "cognitive search template" that re-weights the structural features (h_F) based on their relevance to the current semantic context, effectively filtering the "grammars of looking" through the lens of meaning.

Hybrid Optimization Objective. Human aesthetic judgment is characterized by a duality of absolute valuation (e.g., precise scoring) and ordinal ranking (e.g., preference sorting). To capture this, the network is optimized using a hybrid objective function. We combine Mean Squared Error (MSE) for numerical precision and a Pearson Linear Correlation Coefficient (PLCC) penalty for ranking linearity:

$$\mathcal{L}_{total} = \frac{1}{N} \sum_{i=1}^N (\hat{y}_i - y_i)^2 + \lambda \left(1 - \frac{\sum_i (\hat{y}_i - \mu_{\hat{y}})(y_i - \mu_y)}{\sqrt{\sum_i (\hat{y}_i - \mu_{\hat{y}})^2} \sqrt{\sum_i (y_i - \mu_y)^2}} \right) \quad (3)$$

where y_i and $\hat{y}_i$ represent the ground-truth and predicted scores, μ denotes the batch mean, and $\lambda = 0.5$ is a balancing hyperparameter. The inclusion of the differentiable PLCC term (second line) constrains the model to learn the monotonic relationship of human preference, mitigating the regression-to-the-mean effect often observed in pure MSE optimization.

Experiments

To rigorously validate the effect of human visual cognition in AQA and challenge the prevailing semantic-along perspective, we devise the following hypothesis test with two studies: Study 1 evaluates the necessity of the visual attention pathway (H_{1a}), and Study 2 assesses its independence and potential as a plug-and-play module across diverse AQA backbones (H_{1b}).

Hypothesis

H_0 (The Null Hypothesis: Semantic Sufficiency): AQA primarily depends on semantic content; incorporating HVS-derived attention does not significantly improve performance over semantic-only baselines ($p > 0.05$).

H_{1a} (The Necessity Hypothesis): Incorporating HVS-derived attention generally enhances AQA, as evidenced by AestheticNet, which fuses eye gaze attention with semantic representation, significantly outperforming semantic-only baselines ($p < 0.05$).

H_{1b} (The Independence Hypothesis): Human-like visual cognition serves as an orthogonal cognitive complement, computationally decoupled from semantic perception. Hence, HVS-derived attention functions as a model-agnostic corrector, evidenced by seamless compatibility with diverse AQA backbones as a plug-and-play module that consistently yields performance gains.

We formally reject H_0 only if both H_{1a} and H_{1b} are verified by empirical evidence.

Study 1: The Necessity of Human Visual Attention

This study investigates whether integrating GAVE with the Semantic Encoder yields a statistically significant improvement in AQA compared to semantic-only baselines.

Dataset Construction. We evaluated our framework on a category-aligned subset of the AVA dataset (Murray et al.,

2012). To ensure cognitive consistency with the Contrastive Gaze Alignment phase (which utilized compositional categories), we filtered the AVA dataset for 8 relevant tags (e.g., “Landscape”, “Architecture”, “Still Life”). This resulted in a total of 89,677 images, split into 71,741 for training and 17,936 for testing.

Experimental Setup. The AestheticNet was instantiated with a frozen CLIP-ViT-L/14 (Radford et al., 2021) (Semantic Encoder) and a DINOv2 (Oquab et al., 2024)-based GAVE (~21M trainable parameters). These streams were integrated via Gaze-Guided Attention. The model was optimized on the hybrid loss (MSE and PLCC constraints) using AdamW (batch size 64, cosine learning rate scheduling). To prevent overfitting, training was bounded by a 200-epoch maximum and a 50-epoch early stopping criterion based on validation PLCC.

Comparative Evaluation Having validated the necessity of the active vision mechanism, we benchmark AestheticNet against representative baselines: ResNet-50 (He et al., 2016), NIMA (Talebi & Milanfar, 2018), HyperIQA (Su et al., 2020), and CLIP-L (Radford et al., 2021). We also include Q-Align (Wu et al., 2024), but designate it as a *machine perception oracle* rather than a normal baseline. This is because Q-Align was trained on the full AVA dataset which covers the test subset in this experiment, rendering its performance a result of data memorization.

As detailed in Table 1, AestheticNet (0.747) consistently outperforms all baseline architectures. Most critically, it surpasses its own semantic backbone, CLIP-L (0.617), by a substantial margin of $\Delta + 0.130$. This empirical evidence directly supports H_{1a} , confirming that semantic representations alone are insufficient for accurate aesthetic assessment. Surpassing expert-level human benchmarks (SROCC ≈ 0.711 (Kong et al., 2016)), AestheticNet captures collective consensus effectively. As Figure 3 visualizes, compared to HyperIQA—selected as the strongest traditional baseline avoiding massive multimodal confounders—our dual-process model yields a tighter diagonal distribution, visually confirming this superior alignment.

Category-wise Analysis To verify that the gaze-guided improvement is robust across diverse semantic domains and not driven by a single class, we decomposed the performance by 8 distinct categories defined in the AVA dataset.

As shown in Table 2, AestheticNet yields statistically significant improvements ($p < .001$) across all 8 semantic categories, confirming that the benefit of the active vision pathway is universal rather than niche.

Notably, the performance gains are non-uniform. The most substantial improvements are observed in Still Life ($\Delta + 0.087$) and Architecture ($\Delta + 0.077$). These domains are inherently driven by geometric composition, spatial arrangement, and visual balance—features that are explicitly encoded by the

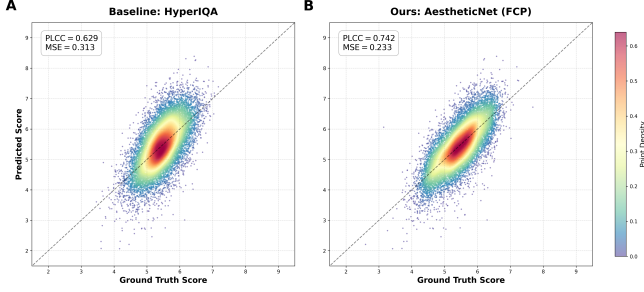


Figure 3: **Prediction alignment analysis.** (A) The baseline HyperlQA shows scattered predictions with higher variance around the central diagonal. (B) AestheticNet produces a tighter distribution along the diagonal line ($y = x$), particularly in the high-density regions (red/yellow). This confirms that our dual-process approach aligns more closely with human consensus than single-stream baselines.

Table 1: **Baseline comparison.** AestheticNet significantly outperforms all baselines. All metrics include 95% Confidence Intervals. * indicates $p < .001$ vs. AestheticNet.

Model	PLCC [95% CI]	SROCC [95% CI]	MSE [95% CI]
<i>Dataset: Full Test Set</i>			
ResNet-50	0.559* [0.549, 0.568]	0.548* [0.538, 0.559]	0.359* [0.351, 0.368]
NIMA	0.584* [0.574, 0.593]	0.575* [0.565, 0.586]	0.342* [0.334, 0.350]
CLIP-L	0.617* [0.607, 0.625]	0.618* [0.608, 0.628]	0.392* [0.383, 0.401]
HyperlQA	0.629* [0.619, 0.638]	0.615* [0.604, 0.624]	0.314* [0.306, 0.322]
AestheticNet	0.747 [0.740, 0.754]	0.740 [0.732, 0.747]	0.261 [0.254, 0.267]
<i>Q-Align (oracle)</i>	0.878 [0.874, 0.882]	0.875 [0.871, 0.879]	0.117 [0.114, 0.121]
<i>Human (expert-level)</i>		0.711	
<i>Dataset: Top-10k Lowest Error Samples</i>			
CLIP-L	0.655* [0.642, 0.666]	0.624* [0.611, 0.637]	0.067* [0.064, 0.069]
NIMA	0.859* [0.853, 0.865]	0.840* [0.832, 0.847]	0.057* [0.055, 0.059]
ResNet-50	0.869* [0.863, 0.874]	0.855* [0.848, 0.862]	0.059* [0.057, 0.061]
HyperlQA	0.913* [0.909, 0.916]	0.899* [0.894, 0.904]	0.050* [0.048, 0.052]
AestheticNet	0.953 [0.951, 0.955]	0.950 [0.947, 0.952]	0.042 [0.041, 0.043]
<i>Q-Align (oracle)</i>	0.978 [0.977, 0.979]	0.969 [0.968, 0.971]	0.018 [0.017, 0.018]

GAVE but often under-represented in semantic models (S-Only). For instance, in Still Life, the aesthetic value often derives from the specific framing of objects rather than the objects themselves. By capturing these compositional priors, AestheticNet effectively compensates for the structural blindness of the pure semantic backbone.

Ablation Study To isolate the contribution of the active vision pathway and test the Necessity Hypothesis (H_{1a}), we performed a component-wise ablation on the fully trained AestheticNet. Instead of training separate models, we utilized inference-time masking to measure the reliance of the joint architecture on each stream, following three configurations:

- *G-Only*: The Semantic stream is masked (zeroed out) during inference, forcing the model to rely solely on the GAVE.
- *S-Only*: The Gaze stream is masked during inference, relying solely on the frozen Semantic Encoder (H_0 baseline).
- *AestheticNet*: The full architecture with both streams active.

Table 2: **Category-wise performance.** Evaluated on the full test set, decomposed by semantic category. AestheticNet consistently outperforms the semantic baseline (S-Only) across all 8 categories. The improvement is particularly pronounced in structure-heavy domains like Still Life and Architecture. * indicates $p < .001$ compared to AestheticNet.

Category (N)	Model	PLCC [95% CI]	MSE [95% CI]
Nature (N=4464)	S-Only	0.664* [0.647, 0.681]	0.348* [0.332, 0.365]
	AestheticNet	0.737 [0.722, 0.752]	0.264 [0.250, 0.278]
Landscape (N=4062)	S-Only	0.711* [0.695, 0.725]	0.332* [0.317, 0.348]
	AestheticNet	0.763 [0.750, 0.776]	0.256 [0.244, 0.268]
Cityscape (N=1408)	S-Only	0.715* [0.690, 0.739]	0.332* [0.307, 0.358]
	AestheticNet	0.751 [0.729, 0.772]	0.271 [0.250, 0.295]
Architecture (N=1735)	S-Only	0.655* [0.625, 0.683]	0.304* [0.281, 0.328]
	AestheticNet	0.732 [0.708, 0.752]	0.233 [0.216, 0.251]
Still Life (N=3752)	S-Only	0.632* [0.611, 0.653]	0.359* [0.341, 0.378]
	AestheticNet	0.719 [0.701, 0.737]	0.265 [0.250, 0.282]
Water (N=732)	S-Only	0.673* [0.630, 0.711]	0.396* [0.352, 0.448]
	AestheticNet	0.755 [0.724, 0.787]	0.277 [0.247, 0.310]
Sky (N=1102)	S-Only	0.716* [0.685, 0.746]	0.335* [0.301, 0.365]
	AestheticNet	0.758 [0.731, 0.784]	0.297 [0.270, 0.326]
Rural (N=681)	S-Only	0.700* [0.661, 0.735]	0.296* [0.264, 0.331]
	AestheticNet	0.749 [0.715, 0.781]	0.228 [0.204, 0.255]

Table 3: **Ablation results.** Comparison of isolated branches (G-Only, S-Only) versus the complete model via inference-time masking. All metrics include 95% Confidence Intervals in brackets. * indicates $p < .001$ compared to AestheticNet.

Model	PLCC [95% CI]	SROCC [95% CI]	MSE [95% CI]
<i>Dataset: Full Test Set</i>			
G-Only	0.607* [0.598, 0.617]	0.596* [0.586, 0.606]	0.415* [0.405, 0.424]
S-Only	0.682* [0.674, 0.691]	0.678* [0.669, 0.687]	0.340* [0.333, 0.348]
AestheticNet	0.746 [0.740, 0.753]	0.740 [0.732, 0.747]	0.261 [0.255, 0.268]
<i>Dataset: Top-10k Lowest Error Samples</i>			
G-Only	0.912* [0.908, 0.916]	0.892* [0.887, 0.897]	0.069* [0.067, 0.071]
S-Only	0.941* [0.938, 0.943]	0.939* [0.936, 0.942]	0.056* [0.054, 0.058]
AestheticNet	0.953 [0.951, 0.955]	0.950 [0.947, 0.952]	0.042 [0.040, 0.043]

Table 3 presents the comparison. On the full test set, S-Only yields a PLCC of 0.682. Because our cross-attention employs residual connections, masking GAVE downgrades the network to a stable semantic regressor, averting out-of-distribution collapse. Re-activating both pathways (AestheticNet) significantly boosts performance to 0.746, confirming the GAVE provides indispensable structural information. The analysis of the *Top-10k Lowest Error Samples* further validates this dependency. Even in this high-precision regime, masking either branch (G-Only or S-Only) degrades performance compared to the full model (0.953). This implies that "peak aesthetic judgment" is functionally irreducible: it requires the simultaneous integration of visual flow and semantic content.

Study 2: The Independence of Human Visual Attention

In this study, we test the Independence Hypothesis (H_{1b}): if the GAVE captures an orthogonal biological signal, it should

Table 4: **Plug-and-play performance** evaluated on the full test set. Integrating the gaze attention yields statistically significant improvements across all architectures. * indicates $p < .001$, ** indicates $p < .01$ (Original vs. +Gaze).

Model / Config	PLCC [95% CI]	SROCC [95% CI]	MSE [95% CI]
ResNet-50			
Original	0.559* [0.549, 0.568]	0.548* [0.538, 0.559]	0.359* [0.351, 0.368]
+ Gaze	0.635 [0.625, 0.644]	0.624 [0.614, 0.635]	0.321 [0.313, 0.330]
CLIP-L			
Original	0.617* [0.607, 0.625]	0.618* [0.608, 0.628]	0.392* [0.383, 0.401]
+ Gaze	0.737 [0.728, 0.746]	0.730 [0.719, 0.739]	0.242 [0.233, 0.251]
NIMA			
Original	0.584* [0.574, 0.593]	0.575* [0.565, 0.586]	0.342* [0.334, 0.350]
+ Gaze	0.625 [0.615, 0.634]	0.616 [0.605, 0.626]	0.330 [0.322, 0.338]
HyperIQA			
Original	0.629* [0.619, 0.638]	0.615* [0.604, 0.624]	0.314* [0.306, 0.322]
+ Gaze	0.684 [0.674, 0.693]	0.675 [0.664, 0.684]	0.277 [0.269, 0.285]
Q-Align (13B)			
Original	0.878* [0.874, 0.882]	0.875** [0.871, 0.879]	0.117* [0.114, 0.121]
+ Gaze	0.879 [0.875, 0.883]	0.876 [0.871, 0.879]	0.116 [0.113, 0.120]

function as a universal cognitive patch, improving performance regardless of the host architecture.

Plug-and-Play Integration. We integrated the frozen GAVE into four diverse backbones using two strategies adapted to the host’s accessibility:

- *Feature-Level Fusion:* For standard backbones ResNet-50 (He et al., 2016), NIMA (Talebi & Milanfar, 2018), HyperIQA (Su et al., 2020), CLIP-L (Radford et al., 2021), we concatenated the gaze features directly with the host’s penultimate feature vector ($f_{fused} = f_{host} \oplus f_{gaze}$) before the final regression head.
- *Score-Level Correction:* For the large-scale Q-Align (Wu et al., 2024), we adopted a non-intrusive strategy by applying a lightweight residual correction to the final predicted scores ($S_{final} = S_{host} + \lambda \cdot S_{gaze}$).

Table 4 demonstrates that incorporating gaze prior information significantly improves the performance of all evaluation baselines. The magnitude of the performance improvement is related to the host model: the CLIP-L model (Radford et al., 2021), which heavily relies on semantic content, benefits the most ($\Delta PLCC + 0.120$), achieving a huge improvement simply by adding the GAVE.

More broadly, the universality of these gains across architecturally distinct baselines—from standard CNNs to specialized aesthetic regressors—supports a deeper conclusion. Despite their structural differences, these models all lack explicit oculomotor modeling. The GAVE functions as a model-agnostic cognitive corrector, compensating for the lack of active viewing dynamics in static deep learning models. This consistent enhancement verifies the Independence Hypothesis (H_{1b}), suggesting that the “grammar of looking” operates as an orthogonal cognitive prior computationally decoupled from visual feature encoding. Finally, regarding Q-Align (Wu et al., 2024), which serves as a *machine perception oracle* in

this context, we still observe a statistically significant refinement ($p < .001$). This confirms that the instinctual scanpath signatures captured by our module remain an orthogonal and irreducible component, indispensable even at the ceiling of machine perception.

Discussion

Our results formally reject the Null Hypothesis of Semantic Sufficiency (H_0). We validate the Necessity Hypothesis (H_{1a}) through three converging lines of evidence: comparative evaluations show AestheticNet significantly outperforms semantic-only baselines; category-wise analysis reveals robust gains across diverse compositional domains; and ablation studies verify the functional necessity of the Gaze Branch. Additionally, we confirm the Independence Hypothesis (H_{1b}), as the GAVE functions as a seamless, model-agnostic corrector across diverse architectures. Together, these findings prove that aesthetic judgment is not just a passive semantic readout, but an active integration process where oculomotor dynamics modulate semantic interpretation.

Although biological gaze intertwines diverse signals, our plug-and-play experiments demonstrate a clear computational decoupling. The GAVE’s ability to enhance architectures ranging from CNNs to Q-Align (Wu et al., 2024) confirms that the “grammar of looking” acts as an orthogonal cognitive complement. This suggests that robust aesthetic AI can be built by assembling specialized cognitive modules rather than relying solely on monolithic models, offering a path toward more interpretable systems.

However, our conclusions are bounded by current benchmarks. The AVA dataset follows a Gaussian distribution, creating data sparsity at aesthetic extremes. While AestheticNet handles these regions better than baselines, it remains constrained by this central tendency bias. Furthermore, the reliance on photography limits generalization to non-photographic domains, such as abstract art, where biological scanning strategies may differ.

Conclusion

This study establishes a resource-rational framework for computational aesthetics, effectively bridging active vision mechanisms with semantic interpretation. By operationalizing the synergy between the GAVE and the Semantic Encoder, we demonstrate that aesthetic perception relies not merely on high-dimensional feature mapping, but on the dynamic integration of innate oculomotor priors with semantic content. Our findings confirm that the GAVE functions as an independent, plug-and-play cognitive module, enabling significant performance gains with high data efficiency. Ultimately, AestheticNet advocates for a shift from monolithic modeling to biologically grounded, modular architectures.

Acknowledgement

This work was supported by the National Natural Science Foundation of China under Grant No. 62402009, and the

References

- Datta, R., Joshi, D., Li, J., & Wang, J. Z. (2006). Studying aesthetics in photographic images using a computational approach. *European Conference on Computer Vision (ECCV)*, 288–301. https://doi.org/10.1007/11744078_23
- He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 770–778. <https://doi.org/10.1109/CVPR.2016.90>
- Itti, L., Koch, C., & Niebur, E. (1998). A model of saliency-based visual attention for rapid scene analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20(11), 1254–1259. <https://doi.org/10.1109/34.730558>
- Kahneman, D. (2011). *Thinking, fast and slow*. Farrar, Straus; Giroux.
- Kong, S., Shen, X., Lin, Z., Mech, R., & Fowlkes, C. (2016). Photo aesthetics ranking network with attributes and content adaptation. *European Conference on Computer Vision (ECCV)*, 662–679. https://doi.org/10.1007/978-3-319-46448-0_40
- Lu, X., Lin, Z., Shen, X., Mech, R., & Wang, J. Z. (2015). Deep multi-patch aggregation network for image style, aesthetics, and quality estimation. *2015 IEEE International Conference on Computer Vision (ICCV)*, 990–998. <https://doi.org/10.1109/ICCV.2015.119>
- Ma, S., Liu, J., & Chen, C. W. (2017). A-Lamp: Adaptive layout-aware multi-patch deep convolutional neural network for photo aesthetic assessment. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 4535–4544.
- Murray, N., Marchesotti, L., & Perronnin, F. (2012). AVA: A large-scale database for aesthetic visual analysis. *2012 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2408–2415. <https://doi.org/10.1109/CVPR.2012.6247954>
- Noton, D., & Stark, L. (1971). Scanpaths in eye movements during pattern perception. *Science*, 171(3968), 308–311. <https://doi.org/10.1126/science.171.3968.308>
- Oquab, M., Darcet, T., Moutakanni, T., Vo, H., Szafraniec, M., Khalidov, V., Fernandez, P., Haziza, D., Massa, F., El-Nouby, A., et al. (2024). DINOv2: Learning robust visual features without supervision. *Transactions on Machine Learning Research*.
- Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al. (2021). Learning transferable visual models from natural language supervision. *International Conference on Machine Learning (ICML)*, 139, 8748–8763.
- Reber, R., Schwarz, N., & Winkielman, P. (2004). Processing fluency and aesthetic pleasure: Is beauty in the perceiver’s processing experience? *Personality and Social Psychology Review*, 8(4), 364–382. https://doi.org/10.1207/s15327957pspr0804_3
- Shimamura, A. P., & Palmer, S. E. (Eds.). (2012). *Aesthetic science: Connecting minds, brains, and experience*. Oxford University Press.
- Su, S., Yan, Q., Zhu, Y., Zhang, C., Ge, X., Sun, J., & Zhang, Y. (2020). Blindly assess image quality in the wild guided by a self-adaptive hyper network. *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 3667–3676.
- Talebi, H., & Milanfar, P. (2018). NIMA: Neural image assessment. *IEEE Transactions on Image Processing*, 27(8), 3998–4011. <https://doi.org/10.1109/TIP.2018.2831899>
- Wilming, N., Onat, S., Ossandón, J. P., Açık, A., Kietzmann, T. C., Kaspar, K., Gameiro, R. R., Vormberg, A., & König, P. (2017). An extensive dataset of eye movements during viewing of complex images. *Scientific Data*, 4(1), 160126. <https://doi.org/10.1038/sdata.2016.126>
- Wu, H., Zhang, Z., Zhang, W., Chen, C., Liao, L., Li, C., Gao, Y., Wang, A., Zhang, E., et al. (2024). Q-Align: Teaching large visual-language models to align with human judgments. *International Conference on Machine Learning (ICML)*.
- Yarbus, A. L. (1967). *Eye movements and vision*. Plenum Press.