A^3: Towards Advertising Aesthetic Assessment

Kaiyuan Ji[1,3] Yixuan Gao[2*] Lu Sun[1,4] Yushuo Zheng[1,2] Zijian Chen[1,2]
Jianbo Zhang[1,2] Xiangyang Zhu[1] Yuan Tian[1] Zicheng Zhang[1,2]
Guangtao Zhai[1,2,3*]
[1]Shanghai Artificial Intelligence Laboratory
[2]Institute of Image Communication and Network Engineering, Shanghai Jiao Tong University
[3]School of Information and Electronic Engineering, East China Normal University
[4]School of Computer Science and Technology, Xi'an Jiaotong University

Abstract

Advertising images significantly impact commercial conversion rates and brand equity, yet current evaluation methods rely on subjective judgments, lacking scalability, standardized criteria, and interpretability. To address these challenges, we present ***A^3 (Advertising Aesthetic Assessment)****, a comprehensive framework encompassing four components: a paradigm (****A^3-Law****), a dataset (****A^3-Dataset****), a multimodal large language model (****A^3-Align****), and a benchmark (****A^3-Bench****). Central to A^3 is a theory-driven paradigm, A^3-Law, comprising three hierarchical stages: (1) Perceptual Attention, evaluating perceptual image signals for their ability to attract attention; (2) Formal Interest, assessing formal composition of image color and spatial layout in evoking interest; and (3) Desire Impact, measuring desire evocation from images and their persuasive impact. Building on A^3-Law, we construct A^3-Dataset with 120K instruction-response pairs from 30K advertising images, each richly annotated with multi-dimensional labels and Chain-of-Thought (CoT) rationales. We further develop A^3-Align, trained under A^3-Law with CoT-guided learning on A^3-Dataset. Extensive experiments on A^3-Bench demonstrate that A^3-Align achieves superior alignment with A^3-Law compared to existing models, and this alignment generalizes well to quality advertisement selection and prescriptive advertisement critique, indicating its potential for broader deployment. Dataset, code, and models can be found at: https://github.com/euleryuan/A3-Align*

1. Introduction

Advertising imagery has become a pervasive part of daily life. However, with the explosive growth of digital media, consumers are now facing severe advertising clutter [30, 31] and overload [73]. This excessive bombardment of advertising messages has led to the dilution of consumer attention, an increase in ad avoidance behaviors, and has demonstrated a significant negative impact on brand recall [73]. In such a hyper-competitive environment, advertising aesthetics is no longer merely an enhancement; it has become a critical factor for cutting through the noise.

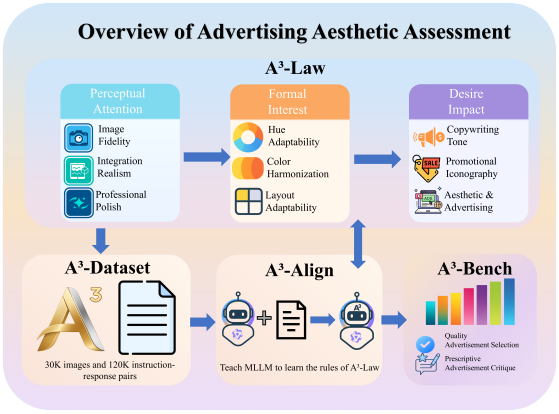

Figure 1. **Overview of A^3: Advertising Aesthetic Assessment.** A^3 centers on the A^3-Law, a three-stage paradigm with *Perceptual Attention*, *Formal Interest*, and *Desire Impact*. Built on this paradigm, **A^3-Dataset** contains 30K images and 120K instruction-response pairs with Chain-of-Thought (CoT); **A^3-Align** learns the rules of A^3-Law; and **A^3-Bench** evaluates MLLMs and two tasks such as quality selection and prescriptive critique.

Therefore, while understanding and measuring the aesthetic quality of advertising images is exceptionally important, current evaluation methods still suffer from significant limitations. First, mainstream quality assessment [23–25, 107] largely depends on manual subjective scoring [15, 48, 57, 58, 65], which struggles to achieve consistent consensus and lacks the scalability required to pro-

*Corresponding author.

cess massive data volumes. Second, existing automated systems [22, 40, 41] often act as mere threshold-based filters [49, 97, 98], failing to provide the systematic, diagnostic feedback necessary to identify specific shortcomings and guide minor refinements for near-standard advertisements.

In recent years, Multimodal Large Language Models (MLLMs) have demonstrated remarkable general understanding capabilities in image interpretation [18, 62, 106] and evaluation tasks [29, 113, 115]. However, their current applications in A^3 (Advertising Aesthetic Assessment) are largely confined to one-step holistic scoring that neglects progressive human cognition. Furthermore, these models often produce unstable, prompt-sensitive rationales with frequent misalignments between their reasoning [16, 42, 46, 95] and final outputs [9, 27, 44]. Therefore, there is an urgent need to develop a stepwise evaluation framework that provides suggestions, supported by datasets and training paradigms to ensure reliable and traceable assessments.

To address the theoretical and procedural gaps [60, 80, 99] in A^3, we propose a progressive evaluation paradigm named A^3-Law, inspired by the AIDA [82]. This paradigm deconstructs the A^3 into three stages. Perceptual Attention: evaluating perceptual image signals for their ability to attract attention. Formal Interest: assessing the formal composition of image color and spatial layout in evoking interest. Finally, Desire Impact: measuring desire evocation from images and their persuasive impact. The core contribution of A^3-Law lies in introducing the first A^3 framework that novelly operationalizes abstract theories into an executable hierarchy for annotation, training, and evaluation.

In order to validate the effectiveness of A^3-Law and support model training, we construct A^3-Dataset, a structured multimodal dataset comprising 120K instruction-response pairs from 30K advertising images with fine-grained labels, Chain-of-Thought (CoT) rationales, and visual annotations. Leveraging this dataset, we train A^3-Align via Supervised Fine-Tuning (SFT) and reinforcement learning (RL) to integrate domain knowledge and align evaluation signals. We then develop A^3-Bench to evaluate performance, with experiments showing that A^3-Align consistently surpasses existing MLLMs. Finally, A^3-Align demonstrates strong practical utility in real-world advertisement selection and prescriptive critique by reliably identifying high-quality ads and diagnosing issues with clear explanations.

As shown in Figure 1, our contributions are as follows:

- We propose paradigm **A^3-Law** for automated advertising aesthetics assessment, explicitly decomposing visual evaluation into three theory-driven stages.
- We construct **A^3-Dataset**, a large-scale dataset containing **120K** advertising-image annotations aligned with the **A^3-Law**, enabling structured and progressive model training.
- We establish **A^3-Bench**, a comprehensive benchmark that evaluates numerous mainstream MLLMs.
- We demonstrate that **A^3-Law** and **A^3-Align** can enhance two real-world applications: quality advertisement selection and prescriptive advertisement critique.

2. Related Works

Advertising Aesthetic Assessment. Current research on advertising images [58] primarily focuses on attractiveness prediction, click-through rate (CTR) modeling [5–7, 116], and aesthetic visual analysis [34, 65, 83, 98, 105], addressing whether ads are effective rather than why. Due to the absence of explicit modeling of aesthetic and persuasive mechanisms [1, 39], crucial factors like copywriting tone [21, 52, 75], layout [55, 59, 112], and color harmony [19, 32] remain overlooked, limiting interpretability and generalization [74, 77, 90, 92, 101]. To overcome this, we propose A^3-Law, a hierarchical framework for structured, progressive evaluation of advertising aesthetics.

Multimodal Large Language Models. MLLMs offer strong visual-language reasoning [45, 50, 56] for rule-based evaluation [108–111], but without domain alignment they often lack rule awareness [43, 61, 96, 114] and produce descriptive rather than judgmental outputs [17, 35, 88, 89, 91, 100]. To address this, we build A^3-Dataset directly around the rules of A^3-Law and leverage it to train A^3-Align.

Benchmarks. Existing benchmarks like AVA [65] and AADB [54] rely on single-dimensional metrics, overlooking semantic interpretation and inferential quality, and lack frameworks for assessing MLLMs in advertising aesthetics [37, 47, 71, 87]. To bridge the gap, we propose A^3-Bench.

3. Approach

3.1. A^3-Law: Hierarchical Paradigm

Perceptual Attention. It evaluates perceptual image signals for their ability to attract attention, constituting the first physiological threshold for advertising information processing and serving as the unconscious precondition for the "Attention" stage in the classic AIDA [10, 82] model. Rather than performing aesthetic judgment, it functions as an ultra-rapid preprocessing mechanism, completed within a hundred milliseconds [70, 86], in which the brain must first determine whether the incoming signal is "processable and valuable information" or should be immediately discarded as "visual noise." This initial screening is grounded in signal detection theory and information theory [26, 33, 78]: reliable perceptual admission requires a sufficient signal-to-noise ratio and low distortion, so that the visual input carries enough recoverable information to cross early physiological thresholds. Consequently, we posit the Image Fidelity principle, which states that images should be clear and minimally distorted to ensure effective reception before any higher-level aesthetic judgment can occur.

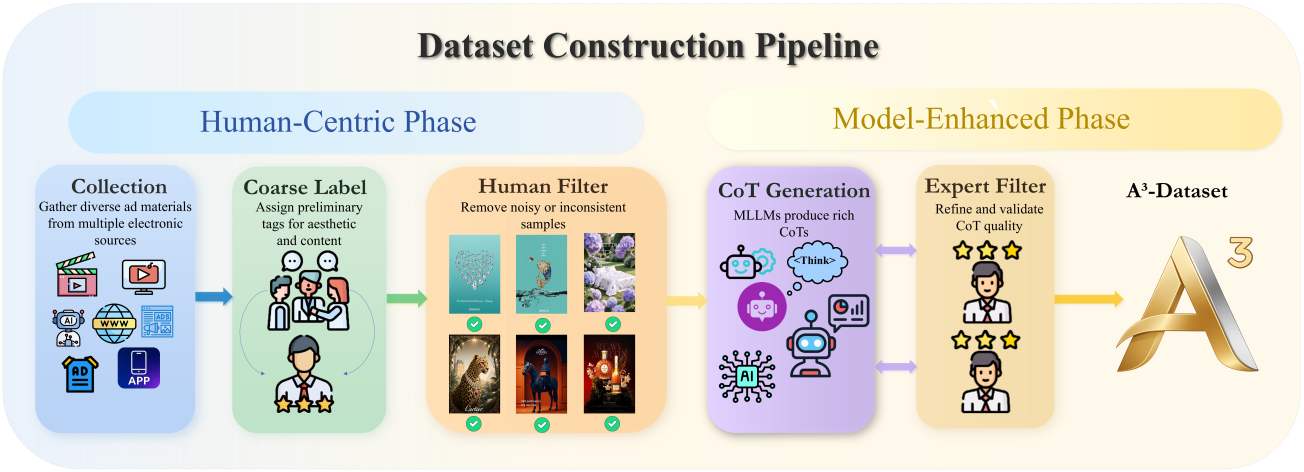

Figure 2. **A^3-Dataset construction pipeline.** The pipeline has two stages. In the *Human–Centric Phase* we collect diverse advertising images, assign preliminary aesthetic and content tags under A^3-Law, and remove noisy or inconsistent samples. In the *Model–Enhanced Phase* multimodal LLMs generate CoT rationales that are refined and validated by experts. The result is the A^3-Dataset with 30K images and 120K high–quality instruction-response pairs.

Second, Perceptual Fluency Theory [51, 72] holds that stimuli that are easier to parse elicit more credible judgments. We therefore operationalize fluency with two design constraints. Integration Realism requires physically coherent rendering, including consistent lighting, color temperature, shadows, and perspective, so that the scene accords with the visual system's prior expectations and reduces prediction error during early parsing. Professional Polish requires clean, artifact-free textures and legible micro-details that match commercial category prototypes, thereby minimizing processing friction and supporting rapid credibility judgments. Together with Image Fidelity, these constraints form a sequential dependency from input to parsing to trust, enabling signals to pass early physiological thresholds before any higher level aesthetic judgment.

Only by passing this physiological threshold is the advertising image confirmed as a qualified visual signal, thus permitting its entry into higher-order stages.

Formal Interest. It assesses the formal composition of image color and spatial layout in evoking interest. Following the Perceptual Attention physiological screening, this stage constitutes a higher–order cognitive task in which the visual system organizes disparate elements into a meaningful structure. Interest arises when a scene is sufficiently comprehensible yet moderately novel [12, 81]. In line with appraisal accounts [81], under these conditions, viewers engage a coherence-seeking drive that organizes the scene into a simple, orderly structure. This drive activates fundamental perceptual grouping mechanisms, as described by Gestalt psychology [53]. To execute this grouping task efficiently, the system must prioritize effective organizational cues.

Foundational work on visual perception establishes color as a dominant preattentive feature [93, 103]. Therefore, chromatic similarity functions as a primary, highly efficient grouping cue, allowing the system to rapidly carve the scene into coarse, color-based units before refining spatial relations. Anchored in these mechanisms, we formalize Color Construction, whereby similarity in hue, lightness, and saturation promotes stable grouping [19]. We operationalize this with two rules: Hue Adaptability, which constrains each hue's lightness and saturation so that color intensity remains pleasant rather than aversive; and Color Harmonization, which evaluates whether the palette is coherent or instead feels scattered and chaotic. The Hasler colorfulness [32, 69] metric serves as a reference for this assessment.

Once the color scheme is coordinated, attention shifts to Spatial Construction [94], which governs how elements are positioned so the design can be parsed quickly and comfortably. Key components (product, text, supporting visuals) are arranged in a clear hierarchy with a single focal point; related items are grouped by proximity, aligned to a simple grid, and spaced consistently. To accommodate aspect-ratio changes and routine platform crops, critical content is placed within safe regions, thereby minimizing information loss when the design is resized or trimmed. The result is a clean, readable layout that avoids clutter. We implement this as Layout Adaptability: a balanced, media-adaptive arrangement that preserves hierarchy, product visibility, and reading order under common format changes.

Together, these rules reduce parsing load and enforce structural clarity, enabling the image to pass coherence threshold and proceed to semantic and affective processing.

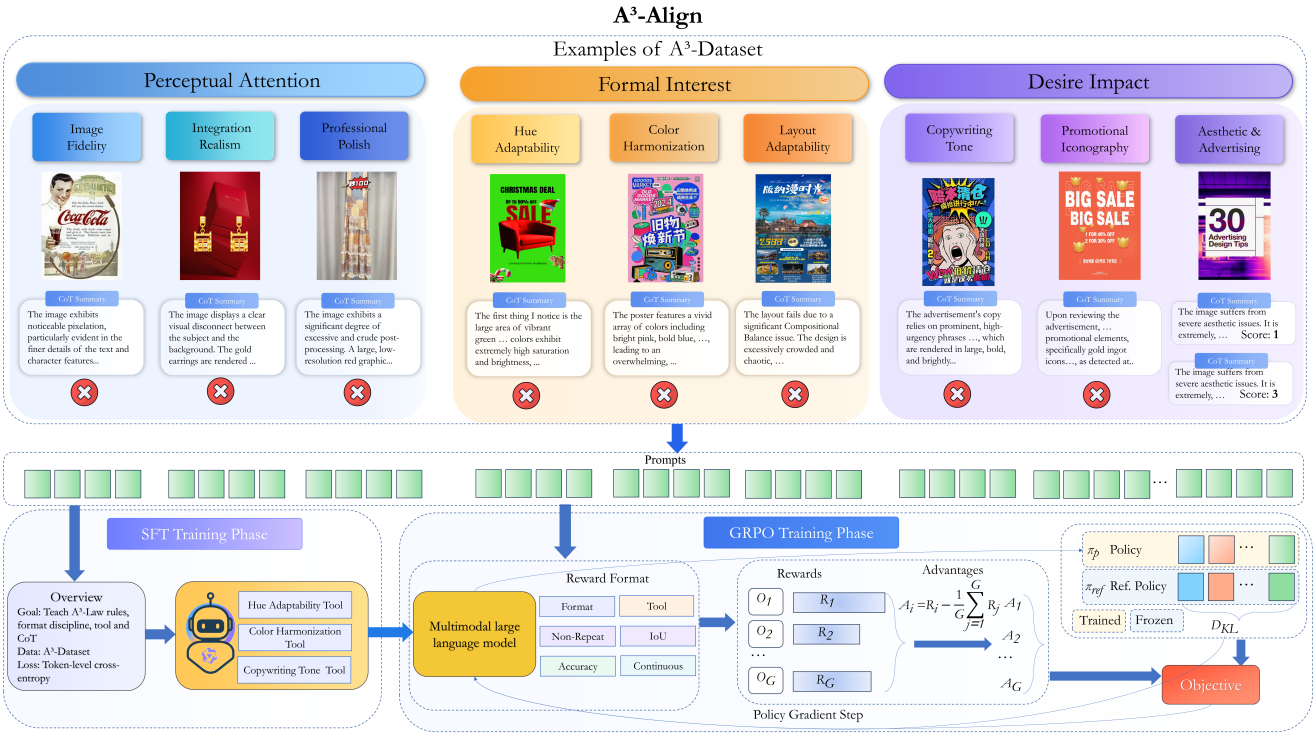

Figure 3. **A^3-Align under the A^3-Law.** The top panel shows examples from the A^3-Dataset organized by the three stages *Perceptual Attention*, *Formal Interest*, and *Desire Impact*, with subcriteria and CoT summaries. The bottom panel presents a two-phase training pipeline. In the SFT phase the multimodal LLM learns A^3-Law rules, structured output format, tool use, and CoT from the A^3-Dataset with token-level cross-entropy. In the GRPO phase, the model is optimized with multi-signal rewards, ultimately leading to A^3-Align, which produces rule-based judgments.

Desire Impact. This measures desire evocation from images and their persuasive impact. This stage, corresponding to the "Desire" stage of the AIDA model, follows the signal screening of the Perceptual Attention and the structural organization of the Formal Interest. At this stage, cognitive processing shifts from passive parsing to an active evaluation of the image's "semantic value" and "affective value," which constitutes the "value threshold."

Its construction is based on two theoretical pillars. First, Semiotics [11, 63] treats the advertisement as a "visual text" composed of signs. To assess its "semantic value," we derive the Copywriting Tone rule, which evaluates key "textual signs," and the Promotional Iconography rule, which evaluates "non-textual signs" such as promotional icons.

Second, in line with affective design [66, 67] and Appraisal Theory [76] of Emotion, decision propensity at this stage reflects integrated affect rather than structure alone. These accounts distinguish visceral impressions from reflective, meaning-based appraisals. Building on this distinction, it assesses affective value through an overall subjective evaluation rule with two components: Aesthetic Attribute, which quantifies visceral visual pleasure induced by formal features; and Advertising Attribute, which quantifies reflective, brand-anchored emotional connection and persuasion expectancy. Together, these components indicate whether the stimulus crosses the value threshold, complementing the semantic evaluation and advancing the AIDA Desire stage.

These three rules jointly evaluate whether an image can effectively convert "interest" into "desire". Desire Impact focuses on the clarity of persuasive signals, which is a universal commercial prerequisite, rather than on localized cultural symbols. Thus, we position A^3-Law as a framework, leaving specific cultural calibration for future work.

3.2. A^3-Dataset: Dataset Collection

Our dataset consists of three main components: images, CoT reasoning processes, and final decisions for each image. To systematically construct A^3-Dataset, we designed a rigorous workflow for the entire data collection process. First, we divided the rule-based QA pairs into three categories: (1) binary questions determining whether an image is suitable or unsuitable under a given rule; (2) object detection annotations specific to Promotional Iconography; and (3) subjective ratings for aesthetic and advertising at-

tributes. Based on expert-defined standards, we trained human annotators for image labeling. After each annotation batch, we randomly sampled a portion of the data for quality inspection, comparing the annotations with our gold-standard reference. Annotators were trained accordingly to evaluate images on a 1 to 5 rating scale. For objective metrics, only results with accuracy above 0.93 were accepted; for detection annotations, only batches with average IoU above 0.92 were retained; for subjective ratings, the acceptance threshold was an SRCC above 0.85. According to the specific rule descriptions, annotators provided binary answers, detection results, and rating scores.

After human-aligned annotation, we utilized multiple MLLMs to generate logically consistent CoTs, combining the rule-specific content with the answers. These generated CoTs were then evaluated in batches, with a subset reviewed by a panel of 5 human experts to determine final acceptance. We implemented a two-tiered validation process where, first, an individual CoT process was considered 'accepted' only if it received a majority vote (at least 3 out of 5) from the expert panel. Second, the MLLM generation and review process was iterated until the overall acceptance rate of the evaluated subset—meaning the proportion of CoTs meeting the majority vote—consistently exceeded 85%.

To enhance reasoning reliability, we introduce a tool-calling subset where MLLMs can access three lightweight analytical tools corresponding to rules with well-defined computational proxies. Specifically, a Hue Analysis Tool computes the central hue, lightness, and saturation of each hue cluster to assist judgments of Hue Adaptability; a Color Harmonization Tool provides the Hasler [32] colorfulness index as a quantitative reference for palette coherence; and DeepSeek-OCR [102] extracts textual content for assessing Copywriting Tone. These three rules are selected because their perceptual judgments can be meaningfully informed by quantitative cues, while other rules lack reliable automatic surrogates. Notably, we exclude external object detection for Promotional Iconography, encouraging models to recognize and interpret non-textual promotional symbols autonomously rather than outsourcing perception. In this subset, tool outputs serve as auxiliary evidence integrated into the model's CoT, ensuring decisions remain grounded yet not mechanically determined by the tools. The whole process of our pipeline is displayed in Figure 2.

To ensure robust generalization and prevent domain biases, the curated A^3-Dataset maximizes diversity in both commercial categories (e.g., electronics, cosmetics) for universal aesthetic learning, and data sources (e-commerce, social media, web banners), as detailed in the Supplementary.

3.3. A^3-Align: MLLM under A^3-Law

As shown in Figure 3, to align the MLLM's reasoning with the aesthetic framework of A^3-Law and endow it with autonomous tool use, we adopt SFT and Group Relative Policy Optimization (GRPO) [79]. SFT teaches the model to produce parseable outputs with a CoT and an answer. In the subsequent phase, we introduce a multi-source reward framework that jointly calibrates behavioral form, task accuracy, evidential grounding, and subjective value alignment. The rewards are divided into two categories:

General reward: (1) Format Reward. It enforces valid tag generation and serves as the foundation for all subsequent rewards: $R_{\text{format}} = \mathbf{1}_{\{\text{tags valid}\}}$

(2) Non-Repeat Reward. To inhibit repetition in the MLLM, we combine a sentence-level term and an n-gram term with equal weights. First, $R_{\text{sent}} = 1 - \frac{d}{N}$, where N is the total number of sentences, and d is the number of duplicate sentences. Then, $R_{\text{n-gram}} = \frac{|\text{uniq(n-Grams)}|}{|\text{n-Grams}|}$, where |n-Grams| is the total count of n-grams in the text, and |uniq(n-Grams)| is the count of unique n-grams. Finally, we define this Reward as: $R_{\text{nonrep}} = \frac{1}{2}(R_{\text{sent}} + R_{\text{n-gram}})$

These rewards target structural stability: R_{format} ensures parseability and R_{nonrep} prevents degenerative loops, guaranteeing the logical consistency of CoTs.

Rule-Specific Alignment Reward: (1) Accuracy Reward. This reward is applied to the eight binary rules under A^3-Law (*Image Fidelity*, *Integration Realism*, *Professional Polish*, *Hue Adaptability*, *Color Harmonization*, *Layout Adaptability*, *Copywriting Tone*, *Promotional Iconography*). It evaluates whether the model's predicted label ($\hat{y}$) matches the annotated ground truth (y): $R_{\text{acc}} = 1\{\hat{y} = y\}$,

(2) Tool Utilization Reward. This reward encourages tool-augmented reasoning within CoT: This reward is applied to the three tool-assisted rules Hue Adaptability, Color Harmonization, and Copywriting Tone. $R_{\text{tool}} =$ 1tool is invoked and referenced in reasoning

(3) IoU Reward. This reward is designed for Promotional Iconography. Given a set of N ground-truth boxes $G = \{g_1, \ldots, g_N\}$ and K predicted boxes $P = \{p_1, \ldots, p_K\}$, we first compute an optimal one-to-one matching (e.g., via Hungarian matching). We then reward predictions that match successfully and have an Intersection-over-Union (IoU) greater than 0.5.

For each matched pair (g, p) where $g \in G, p \in P$, the reward is defined as: $R_{\text{IoU}}(g, p) = \mathbf{1}_{\{\text{IoU}(g,p)>0.5\}}$

(4) Continuous Score Reward. This reward is applied to rules requiring continuous subjective scoring, specifically Aesthetic Attribute and Advertising Attribute.

Instead of using a binary 0/1 reward, we adopt a Gaussian-shaped function to softly reward predictions that closely match human ratings. This formulation encourages the model to produce fine-grained predictions that are as close as possible to the human-annotated ground-truth score $\hat{s}$, rather than merely hitting a coarse interval.

Formally, let s be the model's predicted score and $\hat{s}$ the human-provided reference score. The reward is defined as:

$$R_{\text{score}} = \exp\left(-\frac{(s-\hat{s})^2}{2\sigma^2}\right), \tag{1}$$

Here, σ is a tunable standard deviation controlling the sharpness of the reward curve.

Total Reward and Implementation Notes. Let $\mathcal{A}$ be the set of active rewards for the current sample , with weights $\alpha_i \geq 0$. To avoid scale drift due to different active subsets, we compute the total reward as a normalized weighted sum:

$$R_{\text{total}} = \frac{\sum_{i\in\mathcal{A}} \alpha_i R_i}{\sum_{i\in\mathcal{A}} \alpha_i}, \tag{2}$$

We always compute the general terms R_{format} and R_{nonrep}, and then include rule-specific terms $R_{\text{acc}}, R_{\text{IoU}}, R_{\text{tool}}, R_{\text{score}}$ only when applicable.

This reward framework allows the model to balance structural correctness, aesthetic accuracy, tool-grounded reasoning, and fine-grained subjective alignment. By integrating external tools into the CoT process and shaping outputs across multiple fidelity levels, the training process aligns the MLLM with the structure of aesthetic judgment.

4. Experiments

4.1. Experimental Setup

Training & Evaluation We select Qwen3-VL-8B-Instruct [85] as the base multimodal model, with all experiments conducted on an 8×H200 GPU cluster. Our training process encompasses two core stages: first, we perform SFT, followed by RL using Group Relative Policy Optimization (GRPO) integrated with a multi-signal feedback mechanism. During the GRPO stage, key parameters are set to $n = 4$ and $\sigma = 1.237$. For data partitioning, we split the annotated dataset into training, internal validation, and held-out test sets using an 8:1:1 ratio, ensuring no overlap between the subsets. All model performance evaluations are conducted on a unified test set. We employ specific metrics for different evaluation dimensions: for objective rules, we report Accuracy; for Promotional Iconography, we report both Accuracy and mAP@0.5; and for the two subjective rule categories, we report the Spearman's Rank Correlation Coefficient (SRCC) and the Pearson's Linear Correlation Coefficient (PLCC).

4.2. Results

A^3-Bench. The proposed A^3-Bench covers over 20 mainstream model variants, including both state-of-the-art (SOTA) open-source and closed-source models, and innovatively compares their performance under "non-thinking" (without reasoning chain) and "thinking" inference modes.

In Table 1, closed-source models exhibit consistently stronger performance compared to open-source counterparts. For instance, in the three perceptual attention subtasks, the best-performing closed-source model, Gemini-2.5-pro [20], achieves scores of 0.750, 0.732, and 0.806, significantly surpassing the best open-source model, Gemma-3-27B [84] (0.648, 0.574, and 0.722 respectively). Similarly, in subjective aesthetic evaluation tasks, the top closed-source model (Claude Opus 4.1 [3], 0.772) clearly outperforms the top open-source model (Gemma-3-27B, 0.677). Notably, both open-source and closed-source models struggle with promotional icon localization. The best open-source mAP is only 0.172, while the highest closed-source mAP is 0.283. This indicates a limitation among current models: they can determine whether a promotional icon exists but struggle with accurate spatial localization.

Analysis of models adopting a "thinking" approach reveals differing effects across model types. For open-source models, enabling "thinking" tends to be unstable or even harmful; for example, Qwen3-VL-32B-Instruct shows declines in subjective rating correlations and a slight reduction in icon localization mAP. Conversely, closed-source models consistently benefit from thinking-mode inference. For instance, Gemini-2.5-pro's subjective rating correlations improve from 0.743 to 0.759 after enabling it. Nonetheless, even these moderate improvements do not address the weak localization performance, indicating that although the "thinking" mechanism facilitates better evidence integration, it remains insufficient for solving localization tasks.

Our proposed A^3-Align significantly outperforms all existing models across every evaluation metric. Specifically, on perceptual attention subtasks, A^3-Align achieves scores of 0.870, 0.824, and 0.917, improving upon the strongest baselines by margins of 0.092, 0.074, and 0.111, respectively. In formal interest subtasks, A^3-Align's advantage is even more pronounced, with improvements of 0.130 and 0.120 on color harmonization and layout adaptability compared to the strongest existing models. Crucially, on the challenging promotional icon localization task, A^3-Align not only accurately classifies the presence of icons (accuracy improved to 0.926), but also significantly boosts spatial precision, achieving an mAP of 0.701, more than double the best closed-source baseline (0.283). Additionally, in continuous subjective scoring task for Aesthetic Attribute, A^3-Align attains an SRCC of 0.880, marking an improvement greater than 0.108 over prior best-performing models.

Overall, these substantial performance gains can be attributed to our specialized training designs: the model explicitly leverages tool utilization rewards for evidence-based reasoning, employs IoU-based rewards to fundamentally enhance spatial localization, and uses continuous score rewards to align precisely with human aesthetic judgments. Consequently, A^3-Align not only delivers SOTA perfor-

Table 1. **Performance across models under A^3-Law.** The table reports results for the three stages *Perceptual Attention*, *Formal Interest*, and *Desire Impact* with their subcriteria. For classification rules we show *Acc*. For icon grounding we show *mAP@0.5*. For continuous subjective prediction on *Aesthetic Attribute* and *Advertising Attribute* we show *SRCC* and *PLCC*. Models are grouped by type, including open source, open source with thinking, closed source, and closed source with thinking, with A^3-Align listed in the final row.

Model	**Perceptual Attention**			**Formal Interest**			**Desire Impact**						
	Image Fidelity	Integration Realism	Professional Polish	Hue Adaptability	Color Harmonization	Layout Adaptability	Copywriting Tone	Promotional Iconography		Aesthetic Attribute		Advertising Attribute	
	Acc↑	Acc↑	Acc↑	Acc↑	Acc↑	Acc↑	Acc↑	Acc↑	mAP@0.5↑	SRCC↑	PLCC↑	SRCC↑	PLCC↑
Open-source Models													
Qwen3-VL-8B-Instruct [85]	0.454	0.491	0.463	0.491	0.444	0.472	0.611	0.157	0.011	0.564	0.562	0.533	0.528
Gemma-3-27B-it [84]	0.648	0.574	0.722	0.639	0.583	0.694	0.667	0.333	0.001	0.677	0.657	0.660	0.660
Qwen3-VL-32B-Instruct [85]	0.602	0.583	0.518	0.694	0.574	0.482	0.880	0.259	0.172	0.640	0.617	0.728	0.701
Qwen2.5-VL-72B-Instruct [8]	0.509	0.380	0.481	0.676	0.463	0.518	0.824	0.343	0.032	0.544	0.512	0.628	0.597
Glm-4.5v-106B [36]	0.500	0.509	0.486	0.615	0.523	0.509	0.650	0.107	0.069	0.652	0.618	0.661	0.650
Llama-4-Scout-109B	0.546	0.546	0.491	0.546	0.472	0.519	0.694	0.417	0.019	0.636	0.613	0.565	0.562
Llama-4-Maverick-400B	0.518	0.491	0.509	0.630	0.472	0.491	0.676	0.259	0.066	0.490	0.478	0.539	0.519
Open-source Models - Thinking													
Qwen3-VL-32B-thinking	0.556	0.481	0.481	0.676	0.500	0.481	0.870	0.222	0.110	0.607	0.599	0.642	0.618
Qwen3-VL-235B-A22B-thinking	0.546	0.602	0.630	0.741	0.630	0.639	0.769	0.120	0.141	0.667	0.613	0.669	0.651
Closed-source Models													
ChatGPT 4.1	0.343	0.296	0.333	0.620	0.417	0.481	0.694	0.278	0.008	0.676	0.640	0.703	0.645
ChatGPT 4o [38]	0.518	0.529	0.472	0.620	0.481	0.537	0.778	0.222	0.001	0.680	0.648	0.781	0.756
Mistral medium 3.1 [64]	0.537	0.574	0.500	0.593	0.444	0.537	0.870	0.361	0.000	0.686	0.660	0.604	0.587
Doubao 1.5 vision pro [28]	0.509	0.639	0.495	0.732	0.602	0.611	0.806	0.046	0.227	0.699	0.663	0.587	0.539
Doubao seed 1.6 vision [14]	0.481	0.528	0.467	0.546	0.472	0.472	0.704	0.417	0.032	0.680	0.633	0.527	0.458
Grok 4 [104]	0.491	0.611	0.602	0.722	0.426	0.528	0.768	0.454	0.002	0.698	0.674	0.690	0.672
ChatGPT 5 [68]	0.509	0.648	0.500	0.694	0.500	0.537	0.917	0.250	0.013	0.732	0.698	0.752	0.703
Claude Haiku 4.5 [2]	0.556	0.639	0.589	0.778	0.509	0.537	0.843	0.407	0.001	0.571	0.542	0.680	0.658
Claude Sonnet 4.5 [4]	0.546	0.528	0.608	0.732	0.611	0.561	0.824	0.472	0.030	0.725	0.691	0.689	0.677
Claude Opus 4.1 [3]	0.528	0.565	0.757	0.768	0.704	0.630	0.833	0.352	0.036	0.772	0.749	0.749	0.751
Gemini 2.5 flash [20]	0.667	0.531	0.708	0.633	0.473	0.568	0.680	0.140	0.013	0.704	0.676	0.626	0.622
Gemini 2.5 pro	0.750	0.732	0.806	0.806	0.509	0.750	0.870	0.463	0.027	0.743	0.744	0.704	0.686
Closed-source Models - Thinking													
Doubao 1.5 vision pro-thinking	0.472	0.482	0.467	0.676	0.444	0.472	0.833	0.102	0.151	0.724	0.673	0.716	0.660
Doubao seed 1.6-thinking	0.537	0.454	0.579	0.694	0.509	0.602	0.880	0.361	0.097	0.688	0.651	0.729	0.704
ChatGPT o3	0.500	0.750	0.574	0.750	0.556	0.617	0.888	0.449	0.034	0.736	0.699	0.763	0.715
ChatGPT o4 mini high	0.537	0.750	0.648	0.822	0.574	0.608	0.879	0.780	0.036	0.711	0.667	0.720	0.702
ChatGPT 5 high	0.509	0.620	0.509	0.694	0.500	0.576	0.915	0.288	0.283	0.740	0.703	0.764	0.714
Claude Haiku 4.5-thinking	0.565	0.620	0.626	0.732	0.482	0.546	0.889	0.398	0.001	0.568	0.555	0.729	0.705
Claude Sonnet 4.5-thinking	0.556	0.546	0.689	0.757	0.574	0.626	0.815	0.472	0.024	0.732	0.697	0.713	0.696
Claude Opus 4.1-thinking	0.518	0.593	0.636	0.815	0.750	0.636	0.796	0.435	0.037	0.722	0.701	0.754	0.736
Gemini 2.5 flash-thinking	0.692	0.570	0.694	0.736	0.443	0.561	0.802	0.160	0.000	0.725	0.702	0.624	0.629
Gemini 2.5 pro-thinking	0.778	0.657	0.759	0.806	0.500	0.685	0.898	0.472	0.019	0.759	0.762	0.764	0.743
A^3-Align (Ours)	**0.870**	**0.824**	**0.917**	**0.824**	**0.880**	**0.870**	**0.991**	**0.926**	**0.701**	**0.880**	**0.880**	**0.838**	**0.836**

mance across individual tasks but also internalizes the core aesthetic principles defined by the A^3-Law framework.

Application I: Quality Advertisement Selection. To validate the effectiveness of A^3-Law in selecting high-quality advertisements, five raters scored 100 images from the A^3-Dataset on 1 to 5 scales for **Satisfaction** and **Action Intent**. Figure 4 reports cumulative gains from the unfiltered baseline *Start* (S = 2.956, A = 2.994) through *Perceptual Attention*, *Formal Interest*, and *Desire Impact*. Gains are monotonic for all models, confirming the value of A^3-Law. A^3-Align yields the largest improvement, reaching $+1.05$ in Satisfaction (35.5%) and $+1.20$ in Action Intent (40.1%) at *Desire Impact*, which is $3\times$ the mean gain of other models. The lift in Action Intent exceeds the lift in Satisfaction, indicating that A^3-Law improves perceived quality and more strongly increases the propensity to act.

Application II: Prescriptive Advertisement Critique. Figure 5 compares three open-ended critique dimensions–problem identification accuracy, depth of CoT, and overall clarity–averaged over 30 images. A^3-Align leads on all axes (mean 4.65), surpassing the next best by +0.38 (+8.9%). The largest margin appears in problem identifica-

Table 2. Ablation analysis under A^3-Law. Accuracy is reported for classification rules, mAP@0.5 for icon grounding, and SRCC/PLCC for continuous subjective prediction.

Method	Perceptual Attention			Formal Interest			Desire Impact						
	Image Fidelity	Integration Realism	Professional Polish	Hue Adaptability	Color Harmonization	Layout Adaptability	Copywriting Tone	Promotional Iconography		Aesthetic Attribute		Advertising Attribute	
	Acc↑	Acc↑	Acc↑	Acc↑	Acc↑	Acc↑	Acc↑	Acc↑	mAP@0.5↑	SRCC↑	PLCC↑	SRCC↑	PLCC↑
w/o CoT	0.848	0.796	0.880	0.769	0.833	0.806	0.962	0.889	0.678	0.805	0.811	0.792	0.788
w/o Accuracy Reward	0.833	0.788	0.870	0.778	0.824	0.787	0.935	0.833	0.692	0.862	0.859	0.824	0.817
w/o Tool Reward	0.861	0.824	0.898	0.751	0.824	0.833	0.962	0.917	0.681	0.850	0.843	0.827	0.824
w/o IoU Reward	0.861	0.815	0.907	0.806	0.870	0.870	0.981	0.861	0.624	0.870	0.873	0.822	0.825
w/o Continuous Reward	0.843	0.806	0.898	0.806	0.852	0.870	0.981	0.917	0.690	0.820	0.815	0.804	0.803
w/o Format Reward	0.852	0.806	0.889	0.796	0.852	0.824	0.972	0.898	0.683	0.870	0.855	0.817	0.827
w/o Non-Repeat Reward	0.843	0.815	0.889	0.806	0.861	0.861	0.972	0.907	0.688	0.866	0.871	0.824	0.822
Ours (Full)	**0.870**	**0.824**	**0.917**	**0.824**	**0.880**	**0.870**	**0.991**	**0.926**	**0.701**	**0.880**	**0.880**	**0.838**	**0.836**

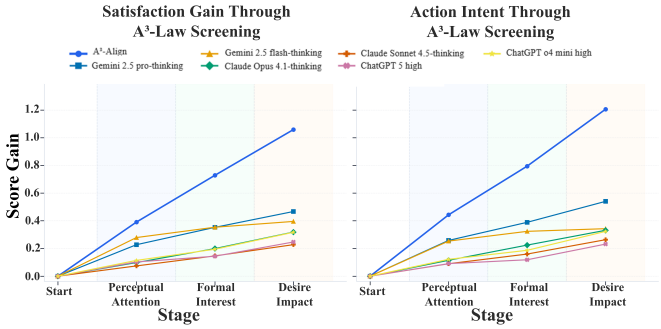

Figure 4. **Satisfaction and Action Intent Gains Through A^3-Law Screening.** The figure shows the cumulative score gains in *Satisfaction* (left) and *Action Intent* (right) across the three stages of A^3-Law screening: *Perceptual Attention*, *Formal Interest*, and *Desire Impact*.

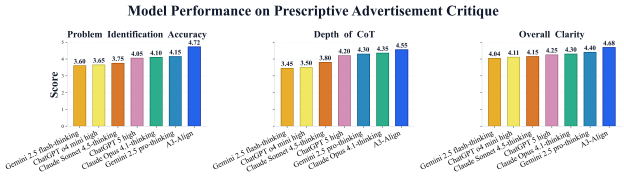

Figure 5. **Evaluation of Problem Identification Accuracy, Depth of CoT, and Overall Clarity.** The figure presents the scores of various models on three evaluation dimensions: *Problem Identification Accuracy*, *Depth of CoT*, and *Overall Clarity*.

tion (+0.57), while depth (+0.20) and clarity (+0.28) remain consistently higher, indicating that A^3-Align not only detects the right issues but also articulates deeper and clearer prescriptive guidance. Further evaluations demonstrating A^3-Align's robust generalization on the external AdImageNet [13] dataset and its potential for generative AI steering are detailed in the supplementary material.

Limitations and Failure Analysis. Despite its strong performance, A^3-Align exhibits two primary failure modes: **(1)** Spatial crowding, where dense layouts trigger icon hallucinations or merging; and **(2)** Attribute miscalibration, where the model occasionally underestimates minimalist premium designs by over-associating visual simplicity with weaker advertising effectiveness.

4.3. Ablation Study

In Table 2, ablation studies confirm the roles of explicit CoT and targeted RL rewards. Without CoT, the model shows limited drops on basic perceptual tasks but much larger declines on subjective alignment tasks, with Aesthetic Attribute SRCC decreasing from 0.880 to 0.805. The RL rewards are also complementary: removing the Accuracy, IoU, and Tool rewards lowers Promotional Iconography accuracy by 0.093, grounding mAP@0.5 by 0.077, and Hue Adaptability accuracy by 0.073, respectively. The Continuous Reward is crucial for subjective prediction, while the Format and Non-Repeat rewards provide smaller gains. Together, these components produce the best overall performance. Furthermore, we observe a hierarchical learning process in which format and non-repeat rewards converge first, followed by tool usage, accuracy, IoU, and finally the continuous score reward.

5. Conclusion

In this paper, we introduce A^3, a framework designed to address the subjectivity and limited scalability of advertising image evaluation by unifying a paradigm (A^3-Law), a dataset (A^3-Dataset), an MLLM (A^3-Align), and a benchmark (A^3-Bench). Our A^3-Law defines three stages—*Perceptual Attention*, *Formal Interest*, and *Desire Impact*—which guided the creation of A^3-Dataset (30K images, 120K instruction pairs with CoT). The A^3-Align model, trained with CoT-guided learning, demonstrates superior adherence to the law and transfers effectively to quality selection and prescriptive critique. This work shows broad potential for creative advertising aesthetic assessment.

6. Acknowledgment

This work was supported by the New Generation Artificial Intelligence National Science and Technology Major Project (Nos. 2025ZD0124104 and P25KK00221), in collaboration with the Shanghai Artificial Intelligence Laboratory, and by the National Natural Science Foundation of China (Grant Nos. 62571324, 62501337, and 62225112).

References

[1] Karuna Ahuja, Karan Sikka, Anirban Roy, and Ajay Divakaran. Understanding visual ads by aligning symbols and objects using co-attention. *arXiv preprint arXiv:1807.01448*, 2018. 2

[2] Anthropic. Introducing claude haiku 4.5. https://www.anthropic.com/news/claude-haiku-4-5, 2025. 7

[3] Anthropic. Claude opus 4.1. https://www.anthropic.com/news/claude-opus-4-1, 2025. 6, 7

[4] Anthropic. Claude sonnet 4.5 system card. https://www.anthropic.com/claude-sonnet-4-5-system-card, 2025. 7

[5] Javad Azimi, Ruofei Zhang, Yang Zhou, Vidhya Navalpakkam, Jianchang Mao, and Xiaoli Fern. The impact of visual appearance on user response in online display advertising. In *proceedings of the 21st international conference on World Wide Web*, pages 457–458, 2012. 2

[6] Javad Azimi, Ruofei Zhang, Yang Zhou, Vidhya Navalpakkam, Jianchang Mao, and Xiaoli Fern. Visual appearance of display ads and its effect on click through rate. In *Proceedings of the 21st ACM international conference on Information and knowledge management*, pages 495–504, 2012.

[7] Jing Bai, Xinyu Geng, Jiaqi Deng, Zhen Xia, Hongxia Jiang, Guoqiang Yan, and Jing Liang. A comprehensive survey on advertising click-through rate prediction algorithm. *The Knowledge Engineering Review*, 40:e3, 2025. 2

[8] Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, Jiabo Ye, Xi Zhang, Tianbao Xie, Zesen Cheng, Hang Zhang, Zhibo Yang, Haiyang Xu, and Junyang Lin. Qwen2.5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2025. 7

[9] Bowen Baker, Joost Huizinga, Leo Gao, Zehao Dou, Melody Y Guan, Aleksander Madry, Wojciech Zaremba, Jakub Pachocki, and David Farhi. Monitoring reasoning models for misbehavior and the risks of promoting obfuscation. *arXiv preprint arXiv:2503.11926*, 2025. 2

[10] Thomas E Barry. The development of the hierarchy of effects: An historical perspective. *Current issues and Research in Advertising*, 10(1-2):251–295, 1987. 2

[11] Roland Barthes. Rhetoric of the image. *Semiotics: An introductory anthology*, pages 192–205, 1985. 4

[12] Daniel E Berlyne. Novelty, complexity, and hedonic value. *Perception & psychophysics*, 8(5):279–286, 1970. 3

[13] Peter Brendan. AdImageNet. https://huggingface.co/datasets/PeterBrendan/AdImageNet, 2024. 8

[14] ByteDance Seed Team. Introduction to techniques used in seed1.6. https://seed.bytedance.com/en/blog/introduction-to-techniques-used-in-seed1-6, 2025. Official overview; closest public source for Seed 1.6 Vision. 7

[15] Linhan Cao, Wei Sun, Xiangyang Zhu, Kaiwei Zhang, Jun Jia, Yicong Peng, Dandan Zhu, Guangtao Zhai, and Xiongkuo Min. Towards generalized video quality assessment: A weak-to-strong learning paradigm. *arXiv preprint arXiv:2505.03631*, 2025. 1

[16] Yanda Chen, Joe Benton, Ansh Radhakrishnan, Jonathan Uesato, Carson Denison, John Schulman, Arushi Somani, Peter Hase, Misha Wagner, Fabien Roger, et al. Reasoning models don't always say what they think. *arXiv preprint arXiv:2505.05410*, 2025. 2

[17] Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, et al. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 24185–24198, 2024. 2

[18] Zijian Chen, Yuan Tian, Yuze Sun, Wei Sun, Zicheng Zhang, Weisi Lin, Guangtao Zhai, and Wenjun Zhang. Just noticeable difference for large multimodal models. *arXiv preprint arXiv:2507.00490*, 2025. 2

[19] Daniel Cohen-Or, Olga Sorkine, Ran Gal, Tommer Leyvand, and Ying-Qing Xu. Color harmonization. In *ACM SIGGRAPH 2006 Papers*, pages 624–630. 2006. 2, 3

[20] Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, et al. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*, 2025. 6, 7

[21] Arka Ujjal Dey, Suman K Ghosh, and Ernest Valveny. Don't only feel read: Using scene text to understand advertisements. *arXiv preprint arXiv:1806.08279*, 2018. 2

[22] Varun Dutt, Demetris Hadjigeorgiou, Lucas Galan, Faiyaz Doctor, Lina Barakat, and Kate Isaacs. Explainable digital creatives performance monitoring using deep feature attribution. In *2024 19th Annual System of Systems Engineering Conference (SoSE)*, pages 134–139. IEEE, 2024. 2

[23] Yixuan Gao, Xiongkuo Min, Yuqin Cao, Xiaohong Liu, and Guangtao Zhai. No-reference image quality assessment: Obtain mos from image quality score distribution. *IEEE Transactions on Circuits and Systems for Video Technology*, 35(2):1840–1854, 2024. 1

[24] Yixuan Gao, Xiongkuo Min, Yuqin Cao, Weisi Lin, Bu Sung Lee, and Guangtao Zhai. Blind image quality assessment by gaussian mixture distribution. *IEEE Transactions on Image Processing*, 2025.

[25] Yixuan Gao, Xiongkuo Min, Jinliang Han, Yuqin Cao, Sijing Wu, Yunze Dou, and Guangtao Zhai. Multi-dimensional text-to-face image quality assessment using llm: Database and method. In *Proceedings of the 33rd ACM International Conference on Multimedia*, pages 6948–6957, 2025. 1

[26] David Marvin Green, John A Swets, et al. *Signal detection theory and psychophysics*. Wiley, New York, 1966. 2

[27] Tianrui Guan, Fuxiao Liu, Xiyang Wu, Ruiqi Xian, Zongxia Li, Xiaoyu Liu, Xijun Wang, Lichang Chen, Furong Huang, Yaser Yacoob, et al. Hallusionbench: an advanced diagnostic suite for entangled language hallucination and visual illusion in large vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14375–14385, 2024. 2

[28] Dong Guo, Faming Wu, Feida Zhu, Fuxing Leng, Guang Shi, Haobin Chen, Haoqi Fan, Jian Wang, Jianyu Jiang, Jiawei Wang, et al. Seed1. 5-vl technical report. *arXiv preprint arXiv:2505.07062*, 2025. 7

[29] Yijin Guo, Kaiyuan Ji, Xiaorong Zhu, Junying Wang, Farong Wen, Chunyi Li, Zicheng Zhang, and Guangtao Zhai. Human-centric evaluation for foundation models. *arXiv preprint arXiv:2506.01793*, 2025. 2

[30] Louisa Ha. Digital advertising clutter in an age of mobile media. In *Digital advertising*, pages 69–85. Routledge, 2017. 1

[31] Louisa Ha and Kim McCann. An integrated model of advertising clutter in offline and online media. *International Journal of Advertising*, 27(4):569–592, 2008. 1

[32] David Hasler and Sabine E Suesstrunk. Measuring colorfulness in natural images. In *Human vision and electronic imaging VIII*, pages 87–95. SPIE, 2003. 2, 3, 5

[33] Michael J Hautus, Neil A Macmillan, and C Douglas Creelman. *Detection theory: A user's guide*. Routledge, 2021. 2

[34] Shuai He, Yongchang Zhang, Rui Xie, Dongxiang Jiang, and Anlong Ming. Rethinking image aesthetics assessment: Models, datasets and benchmarks. In *IJCAI*, pages 942–948, 2022. 2

[35] John Hewitt, Nelson F Liu, Percy Liang, and Christopher D Manning. Instruction following without instruction tuning. *arXiv preprint arXiv:2409.14254*, 2024. 2

[36] Wenyi Hong, Wenmeng Yu, Xiaotao Gu, Guo Wang, Guobing Gan, Haomiao Tang, Jiale Cheng, Ji Qi, Junhui Ji, Lihang Pan, et al. Glm-4.5 v and glm-4.1 v-thinking: Towards versatile multimodal reasoning with scalable reinforcement learning. *arXiv preprint arXiv:2507.01006*, 2025. 7

[37] Yipo Huang, Quan Yuan, Xiangfei Sheng, Zhichao Yang, Haoning Wu, Pengfei Chen, Yuzhe Yang, Leida Li, and Weisi Lin. Aesbench: An expert benchmark for multimodal large language models on image aesthetics perception. *arXiv preprint arXiv:2401.08276*, 2024. 2

[38] Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*, 2024. 7

[39] Zaeem Hussain, Mingda Zhang, Xiaozhong Zhang, Keren Ye, Christopher Thomas, Zuha Agha, Nathan Ong, and Adriana Kovashka. Automatic understanding of image and video advertisements. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1705–1715, 2017. 2

[40] Hyeongnam Jang, Yeejin Lee, and Jong-Seok Lee. Modeling, quantifying, and predicting subjectivity of image aesthetics. *arXiv preprint arXiv:2208.09666*, 2022. 2

[41] Kaiyuan Ji, Zhihan Wu, Jing Han, Jun Jia, Guangtao Zhai, and Jiannan Liu. Application of 3d nnu-net with residual encoder in the 2024 miccai head and neck tumor segmentation challenge. In *Challenge on Head and Neck Tumor Segmentation for MRI-Guided Applications*, pages 250–258. Springer, 2024. 2

[42] Kaiyuan Ji, Yijin Guo, Zicheng Zhang, Xiangyang Zhu, Yuan Tian, Ning Liu, and Guangtao Zhai. Medomni-45 {\deg}: A safety-performance benchmark for reasoning-oriented llms in medicine. *arXiv preprint arXiv:2508.16213*, 2025. 2

[43] Kaiyuan Ji, Jing Han, Guangtao Zhai, and Jiannan Liu. Assessing the capabilities of generative pretrained transformer-4 in addressing open-ended inquiries of oral cancer. *International Dental Journal*, 75(1):158–165, 2025. 2

[44] Kaiyuan Ji, Zhihan Wu, Jing Han, Guangtao Zhai, and Jiannan Liu. Evaluating chatgpt-4's performance on oral and maxillofacial queries: Chain of thought and standard method. *Frontiers in Oral Health*, 6:1541976, 2025. 2

[45] Kaiyuan Ji, Yixuan Gao, Lu Sun, Yushuo Zheng, Zijian Chen, Jianbo Zhang, Xiangyang Zhu, Yuan Tian, Zicheng Zhang, and Guangtao Zhai. A^3: Towards Advertising Aesthetic Assessment. *arXiv preprint arXiv:2603.24037*, 2026. 2

[46] Kaiyuan Ji, Yijin Guo, Zicheng Zhang, Xiangyang Zhu, Yuan Tian, and Ning Liu. MedOmni-45: A Safety–Performance Benchmark for Reasoning-Oriented LLMs in Medicine. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 35536–35544, 2026. 2

[47] Zhiwei Jia, Pradyumna Narayana, Arjun Akula, Garima Pruthi, Hao Su, Sugato Basu, and Varun Jampani. Kafa: Rethinking image ad understanding with knowledge-augmented feature adaptation of vision-language models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 5: Industry Track)*, pages 772–785, 2023. 2

[48] Yanwei Jiang, Wei Sun, Yingjie Zhou, Xiangyang Zhu, Yuqin Cao, Jun Jia, Yunhao Li, Sijing Wu, Dandan Zhu, Xingkuo Min, et al. Surveillance facial image quality assessment: A multi-dimensional dataset and lightweight model. *IEEE Transactions on Circuits and Systems for Video Technology*, 2026. 1

[49] Jian Jin, Jiangyong Ying, Huiyu Duan, Liu Yang, Sijing Wu, Yunhao Li, Yushuo Zheng, Xiongkuo Min, and Guangtao Zhai. Rgc-vqa: An exploration database for robotic-generated video quality assessment. In *Proceedings of the ACM International Conference on Multimedia*, 2025. 2

[50] Wenzhong Jin, Yilan Sun, Kaiyuan Ji, Xiaoyan Jiang, Yufeng Hu, Jinwu Wang, and Jiannan Liu. Medscreendental: Automated structured dental record generation via multimodal language model integration. *Displays*, 90:103119, 2025. 2

[51] William A Johnston, Veronica J Dark, and Larry L Jacoby. Perceptual fluency and recognition judgments. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 11(1):3, 1985. 3

[52] Kanika Kalra, Bhargav Kurma, Silpa Vadakkeeveetil Sreelatha, Manasi Patwardhan, and Shirish Karande. Understanding advertisements with bert. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7542–7547, 2020. 2

[53] Kurt Koffka. *Principles of Gestalt psychology*. routledge, 2013. 3

[54] Shu Kong, Xiaohui Shen, Zhe Lin, Radomir Mech, and Charless Fowlkes. Photo aesthetics ranking network with attributes and content adaptation. In *European conference on computer vision*, pages 662–679. Springer, 2016. 2

[55] Hsin-Ying Lee, Lu Jiang, Irfan Essa, Phuong B Le, Haifeng Gong, Ming-Hsuan Yang, and Weilong Yang. Neural design network: Graphic layout generation with constraints. In *European conference on computer vision*, pages 491–506. Springer, 2020. 2

[56] Chaoyu Lei, Kaiyuan Ji, Chen Zhao, Sisi Zhong, Chenyu Cao, Hao Chen, Chee Chew Yip, Sunisa Sintuwong, Jianbin Ding, PS Pandiyan, et al. Sequential sensitivity analysis of multimodal large language models for rare orbital disease detection. *Communications Medicine*, 2026. 2

[57] Chunyi Li, Jiaohao Xiao, Jianbo Zhang, Farong Wen, Zicheng Zhang, Yuan Tian, Xiangyang Zhu, Xiaohong Liu, Zhengxue Cheng, Weisi Lin, et al. Image quality assessment for embodied ai. *arXiv preprint arXiv:2505.16815*, 2025. 1

[58] Hairong Li and Nan Zhang. Computer vision models for image analysis in advertising research. *Journal of Advertising*, 53(5):771–790, 2024. 1, 2

[59] Jianan Li, Jimei Yang, Aaron Hertzmann, Jianming Zhang, and Tingfa Xu. Layoutgan: Generating graphic layouts with wireframe discriminators. *arXiv preprint arXiv:1901.06767*, 2019. 2

[60] Tongyang Li, Yuexin Su, Ziyi Yang, and Shengyu Zhang. Quantum approximate optimization algorithms for maximum cut on low-girth graphs, 2025. 2

[61] Bohan Liang, Zijian Chen, Qi Jia, Kaiwei Zhang, Kaiyuan Ji, and Guangtao Zhai. Priceseer: Evaluating large language models in real-time stock prediction. *arXiv preprint arXiv:2601.06088*, 2025. 2

[62] Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, et al. Mmbench: Is your multimodal model an all-around player? In *European conference on computer vision*, pages 216–233. Springer, 2024. 2

[63] David Glen Mick. Consumer research and semiotics: Exploring the morphology of signs, symbols, and significance. *Journal of consumer research*, 13(2):196–213, 1986. 4

[64] Mistral AI. Medium is the new large: Introducing mistral medium 3. https://mistral.ai/news/mistral-medium-3, 2025. Official announcement post. 7

[65] Naila Murray, Luca Marchesotti, and Florent Perronnin. Ava: A large-scale database for aesthetic visual analysis. In *2012 IEEE conference on computer vision and pattern recognition*, pages 2408–2415. IEEE, 2012. 1, 2

[66] Don Norman. Emotion & design: attractive things work better. *interactions*, 9(4):36–42, 2002. 4

[67] Donald A. Norman. *Emotional Design: Why We Love (or Hate) Everyday Things*. Basic Books, 2007. 4

[68] OpenAI. Gpt-5 system card. https://cdn.openai.com/gpt-5-system-card.pdf, 2025. 7

[69] Li-Chen Ou and M Ronnier Luo. A colour harmony model for two-colour combinations. *Color Research & Application: Endorsed by Inter-Society Color Council, The Colour Group (Great Britain), Canadian Society for Color, Color Science Association of Japan, Dutch Society for the Study of Color, The Swedish Colour Centre Foundation, Colour Society of Australia, Centre Français de la Couleur*, 31(3): 191–204, 2006. 3

[70] Mary C Potter, Brad Wyble, Carl Erick Hagmann, and Emily S McCourt. Detecting meaning in rsvp at 13 ms per picture. *Attention, Perception, & Psychophysics*, 76(2): 270–279, 2014. 2

[71] 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 Computer Vision and Pattern Recognition Conference*, pages 24807–24816, 2025. 2

[72] Rolf Reber and Norbert Schwarz. Effects of perceptual fluency on judgments of truth. *Consciousness and cognition*, 8(3):338–342, 1999. 3

[73] Abdur Rehman, Naveed Farooq, and Hazrat Bilal. Advertising overload: The impact of information overload on brand awareness: Case of university of swat students. *The Discourse*, 5(1):37–43, 2019. 1

[74] Iria Santos, Miguel A Casal, João Correia, Álvaro Torrente-Patiño, Penousal Machado, and Juan Romero. Towards robust evaluation of aesthetic and photographic quality metrics: Insights from a comprehensive dataset. *Complexity*, 2024(1):8223586, 2024. 2

[75] Andrey Savchenko, Anton Alekseev, Sejeong Kwon, Elena Tutubalina, Evgeny Myasnikov, and Sergey Nikolenko. Ad lingua: Text classification improves symbolism prediction in image advertisements. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 1886–1892, 2020. 2

[76] Klaus R Scherer, Angela Schorr, and Tom Johnstone. *Appraisal processes in emotion: Theory, methods, research*. Oxford University Press, 2001. 4

[77] Sven Schultze, Ani Withöft, Larbi Abdenebaoui, and Susanne Boll. Explaining image aesthetics assessment: An interactive approach. In *Proceedings of the 2023 ACM International Conference on Multimedia Retrieval*, pages 20–28, 2023. 2

[78] Claude E Shannon. A mathematical theory of communication. *The Bell system technical journal*, 27(3):379–423, 1948. 2

[79] Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, YK Li, Yang Wu, et al. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*, 2024. 5

[80] King-Siong Si, Lu Sun, Weizhan Zhang, Tieliang Gong, Jiahao Wang, Jiang Liu, and Hao Sun. Accelerating non-maximum suppression: a graph theory perspective. *Advances in Neural Information Processing Systems*, 37: 121992–122028, 2024. 2

[81] Paul J Silvia. What is interesting? exploring the appraisal structure of interest. *Emotion*, 5(1):89, 2005. 3

[82] Edward K Strong Jr. Theories of selling. *Journal of applied psychology*, 9(1):75, 1925. 2

[83] Hossein Talebi and Peyman Milanfar. Nima: Neural image assessment. *IEEE transactions on image processing*, 27(8): 3998–4011, 2018. 2

[84] Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, et al. Gemma 3 technical report. *arXiv preprint arXiv:2503.19786*, 2025. 6, 7

[85] Qwen Team. Qwen3 technical report, 2025. 6, 7

[86] Simon Thorpe, Denis Fize, and Catherine Marlot. Speed of processing in the human visual system. *nature*, 381(6582): 520–522, 1996. 2

[87] Yuan Tian, Guo Lu, Yichao Yan, Guangtao Zhai, Li Chen, and Zhiyong Gao. A coding framework and benchmark towards low-bitrate video understanding. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(8):5852–5872, 2024. 2

[88] Yuan Tian, Kaiyuan Ji, Rongzhao Zhang, Yankai Jiang, Chunyi Li, Xiaosong Wang, and Guangtao Zhai. Towards all-in-one medical image re-identification. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 30774–30786, 2025. 2

[89] Yuan Tian, Xiaoyue Ling, Cong Geng, Qiang Hu, Guo Lu, and Guangtao Zhai. Smc++: Masked learning of unsupervised video semantic compression. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2025. 2

[90] Yuan Tian, Shuo Wang, Rongzhao Zhang, et al. Semantic versus identity: A divide-and-conquer approach towards adjustable medical image de-identification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 20613–20625, 2025. 2

[91] Yuan Tian, Min Zhou, Yitong Chen, et al. Rofi: A deep learning-based ophthalmic sign-preserving and reversible patient face anonymizer. *npj Digital Medicine*, 2025. 2

[92] Jingru Tong, Guo Zhang, Peijie Kong, Yu Rao, Zhengkai Wei, Hao Cui, and Qing Guan. An interpretable approach for automatic aesthetic assessment of remote sensing images. *Frontiers in Computational Neuroscience*, 16: 1077439, 2022. 2

[93] Anne M Treisman and Garry Gelade. A feature-integration theory of attention. *Cognitive psychology*, 12(1):97–136, 1980. 3

[94] Edward R Tufte. Envisioning information. *Optometry and Vision Science*, 68(4):322–324, 1991. 3

[95] Miles Turpin, Julian Michael, Ethan Perez, and Samuel Bowman. Language models don't always say what they think: Unfaithful explanations in chain-of-thought prompting. *Advances in Neural Information Processing Systems*, 36:74952–74965, 2023. 2

[96] Chengbang Wang, Zijia Liu, Liangshi Hao, Shaohua Chen, Rongchang Guo, Lai Wei, Kaiyuan Ji, Fubo Wang, and Bin Xu. Quality evaluation of large language models in answering open-ended questions in the field of benign prostatic hyperplasia. *Displays*, 90:103144, 2025. 2

[97] Lijie Wang, Xueting Wang, Toshihiko Yamasaki, and Kiyoharu Aizawa. Aspect-ratio-preserving multi-patch image aesthetics score prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 0–0, 2019. 2

[98] Ruiyi Wang, Yushuo Zheng, Zicheng Zhang, Chunyi Li, Shuaicheng Liu, Guangtao Zhai, and Xiaohong Liu. Learning hazing to dehazing: Towards realistic haze generation for real-world image dehazing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2025. 2

[99] Shibo Wang. Bamnet: A brain area mapping-based multimodal saliency prediction method. In *Proceedings of the IEEE International Conference on Multimedia and Expo (ICME)*, pages 1–6, 2025. 2

[100] Shuangqing Wang, Kaiyuan Ji, Yushuo Zheng, Zhihan Wu, Xiaorong Zhu, Zijian Chen, Lu Sun, Shuo Wang, Jianbo Zhang, Zicheng Zhang, et al. Dental-qad: Reasoning-driven quality assessment and diagnosis in panoramic radiographs. *Displays*, page 103380, 2026. 2

[101] Shibo Wang, Yan Zhao, Shigang Wang, Jian Wei, and Shuo Li. A brain-inspired saliency prediction framework for human-ai cognitive consistency in aigc content via multi-region liquid neurons. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2026. 2

[102] Haoran Wei, Yaofeng Sun, and Yukun Li. Deepseek-ocr: Contexts optical compression. *arXiv preprint arXiv:2510.18234*, 2025. 5

[103] Jeremy M Wolfe and Todd S Horowitz. Five factors that guide attention in visual search. *Nature human behaviour*, 1(3):0058, 2017. 3

[104] xAI. Grok 4 model card. https://data.x.ai/2025-08-20-grok-4-model-card.pdf, 2025. 7

[105] Ran Yi, Haoyuan Tian, Zhihao Gu, Yu-Kun Lai, and Paul L Rosin. Towards artistic image aesthetics assessment: a large-scale dataset and a new method. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22388–22397, 2023. 2

[106] Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, et al. Mmmu: A massive multi-discipline multimodal understanding and reasoning bench-

mark for expert agi. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9556–9567, 2024. 2

[107] Zicheng Zhang, Wei Sun, Xiongkuo Min, Tao Wang, Wei Lu, and Guangtao Zhai. No-reference quality assessment for 3d colored point cloud and mesh models. *IEEE Transactions on Circuits and Systems for Video Technology*, 32 (11):7618–7631, 2022. 1

[108] Zicheng Zhang, Junying Wang, Yijin Guo, et al. Aibench: Towards trustworthy evaluation under the 45 law. *Displays*, page 103255, 2025. 2

[109] Zicheng Zhang, Junying Wang, Farong Wen, Yijin Guo, et al. Large multimodal models evaluation: A survey. *SCIENCE CHINA Information Sciences*, 68(12):221301–221369, 2025.

[110] Zicheng Zhang, Haoning Wu, Ziheng Jia, Weisi Lin, and Guangtao Zhai. Teaching lmms for image quality scoring and interpreting. *arXiv preprint arXiv:2503.09197*, 2025.

[111] Zicheng Zhang, Yingjie Zhou, Chunyi Li, Baixuan Zhao, Xiaohong Liu, and Guangtao Zhai. Quality assessment in the era of large models: A survey. *ACM Transactions on Multimedia Computing, Communications and Applications*, 21(7):1–31, 2025. 2

[112] Xinru Zheng, Xiaotian Qiao, Ying Cao, and Rynson WH Lau. Content-aware generative modeling of graphic design layouts. *ACM Transactions on Graphics (TOG)*, 38(4):1–15, 2019. 2

[113] Yushuo Zheng, Jiangyong Ying, Huiyu Duan, Chunyi Li, Zicheng Zhang, Jing Liu, Xiaohong Liu, and Guangtao Zhai. Geox-bench: Benchmarking cross-view geo-localization and pose estimation capabilities of large multimodal models. *arXiv preprint arXiv:2511.13259*, 2025. 2

[114] Yushuo Zheng, Zicheng Zhang, Xiongkuo Min, Huiyu Duan, and Guangtao Zhai. Lm fight arena: Benchmarking large multimodal models via game competition, 2025. 2

[115] Yushuo Zheng, Huiyu Duan, Zicheng Zhang, Xiaohong Liu, and Xiongkuo Min. Learning to wander: Improving the global image geolocation ability of lmms via actionable reasoning, 2026. 2

[116] Ke Zhou, Miriam Redi, Andrew Haines, and Mounia Lalmas. Predicting pre-click quality for native advertisements. In *Proceedings of the 25th International Conference on World Wide Web*, pages 299–310, 2016. 2