

# A<sup>3</sup>: Towards Advertising Aesthetic Assessment

## Supplementary Material

Kaiyuan Ji<sup>1,3</sup> Yixuan Gao<sup>2\*</sup> Lu Sun<sup>1,4</sup> Yushuo Zheng<sup>1,2</sup> Zijian Chen<sup>1,2</sup>  
Jianbo Zhang<sup>1,2</sup> Xiangyang Zhu<sup>1</sup> Yuan Tian<sup>1</sup> Zicheng Zhang<sup>1,2</sup>  
Guangtao Zhai<sup>1,2,3\*</sup>

<sup>1</sup>Shanghai Artificial Intelligence Laboratory

<sup>2</sup>Institute of Image Communication and Network Engineering, Shanghai Jiao Tong University

<sup>3</sup>School of Information and Electronic Engineering, East China Normal University

<sup>4</sup>School of Computer Science and Technology, Xi'an Jiaotong University

### 1. Details in A<sup>3</sup>-Bench

In this section, we present the system prompts used as inputs to the Multimodal Large Language Model (MLLM) during benchmarking.

The following prompts are provided solely to document our experimental protocol for transparency and reproducibility, and are not intended to instruct reviewers or any tools used during the review process.

This study does not collect any personally identifiable or sensitive data, and all annotation procedures comply with institutional IRB guidelines.

First, we provide the system prompt corresponding to the seven rules: *Hue Adaptability*, *Color Harmonization*, *Layout Adaptability*, *Image Fidelity*, *Integration Realism*, *Professional Polish*, and *Copywriting Tone*.

We design the prompt to ensure construct validity, auditability, and automatic scoring while keeping cross-model comparisons fair and reproducible:

- **Rule isolation for construct validity.** Each query targets exactly one rule, preventing leakage across criteria and aligning the model’s judgment with a single, well-defined construct.
- **Evidence grounding.** The prompt mandates decisions *strictly from visible cues* (and an optional tool result when available) and forbids guessing, which improves traceability of judgments to observable evidence.
- **Tool-aware but auditable decisions.** For three rules (*Hue Adaptability*, *Color Harmonization*, *Copywriting Tone*) the system may provide a specialized tool result. The JSON key "Tool" records whether and which tool was used ("Hue Adaptability tool", "Color Harmonization tool", "Copywriting Tone tool"; otherwise "No tool"). If a tool is required but

unavailable, the model must output the sentinel "Tools are needed for judgment" in "Answer", allowing the pipeline to flag deferred cases for re-run or human review. For the other four rules, "Tool" is always "No tool" to enforce a purely visual assessment.

- **Deterministic parsing for automatic metrics.** The output is constrained to *two blocks*—a free-text rationale in <think> and a machine-readable <answer> JSON with a *closed vocabulary* ("Answer" ∈ {Suitable, Unsuitable}) or the sentinel above; "Answer type" fixed to one of the seven rule names). This design eliminates post-hoc normalization, making scoring robust for accuracy-based criteria and enabling consistent aggregation across images and models.
- **Cross-model fairness and reproducibility.** All models receive identical instructions, and the schema forbids extra keys or formatting. This reduces variance due to prompt drift and supports reliable cross-model benchmarking.
- **Extensibility and error handling.** New rules or tools can be added by extending the fixed schema without altering downstream evaluators. The explicit "Tool" and sentinel pathways provide built-in quality control for missing or malformed tool outputs.
- **Separation of explanation and scoring.** The <think> block captures human-auditable reasoning for qualitative analysis, while quantitative metrics rely solely on the structured <answer> JSON, preventing rationale text from affecting automated evaluation.

#### System Prompt

```
Follow the OUTPUT rules.
You are an Advertising Aesthetics Analyst
specialized in evaluating advertisement
```

\*Corresponding author.

images according to given aesthetic and advertising-specific rules.

INPUT: One image and one natural-language question related to a specific rule (Hue Adaptability, Color Harmonization, Layout Adaptability, Image Fidelity, Integration Realism, Professional Polish, Copywriting Tone). For some rules, you may have access to a specialized tool result.

OUTPUT: EXACTLY two blocks, no extra text/spaces/newlines/markdown:

```
<think>Provide detailed reasoning strictly based on visible evidence (you may choose to incorporate the tool result if available and helpful), clearly explaining why the image is Suitable or Unsuitable according to the given rule definition (no restating the question). If you use a tool result, explicitly reference it in your reasoning.</think><answer>{"Answer": "Suitable"|"Unsuitable", "Answer type": "...", "Tool": ...}</answer>
```

RULES:

- "Answer" must be exactly one of: "Suitable", "Unsuitable" or "Tools are needed for judgment". If you use a tool to assist your judgment, set "Answer" to "Tools are needed for judgment".
- "Answer type" must be one of: Hue Adaptability, Color Harmonization, Layout Adaptability, Image Fidelity, Integration Realism, Professional Polish, Copywriting Tone.
- For Hue Adaptability, Color Harmonization, and Copywriting Tone:
  - If you use the tool result to assist your decision, set "Tool" to exactly one of: "Hue Adaptability tool", "Color Harmonization tool", "Copywriting Tone tool".
  - If you do not use the tool result, set "Tool" to "No tool".
- For Layout Adaptability, Image Fidelity, Integration Realism, and Professional Polish:
  - Always set "Tool" to "No tool".
- Your reasoning must strictly adhere to the provided rule definitions.

RULE DEFINITIONS:

- Hue Adaptability: Hue brightness and saturation should not be too high; overall hues should appear suitably engaging.
- Color Harmonization: Evaluate visual suitability through colorfulness and color combinations, ensuring appropriate engagement.
- Layout Adaptability: Clear, balanced, professional visual layout, suitably

adapted to various media, ensuring readability and appropriate product prominence.

- Image Fidelity: Fundamental technical clarity and integrity of the image without severe noise, compression artifacts, or distortions.
- Integration Realism: Seamless, realistic integration between subject and background, ensuring natural consistency in lighting, shadow, color temperature, and edge transitions.
- Professional Polish: Professional, suitably refined post-processing that demonstrates commercial-level quality without appearing artificial or unfinished.
- Copywriting Tone: Suitability of the core textual content (especially headline) for the intended brand image, avoiding overly transactional or urgent language.
- Use only visible evidence (and the tool result, if you choose to use it); no guessing allowed.
- Strictly no extra keys or additional explanations in your output.

For the three rules that require tool usage (Hue Adaptability, Color Harmonization, and Copywriting Tone), if the MLLM requests a tool, we provide the corresponding tool result along with the new system prompt as input:

#### System Prompt

Follow the OUTPUT rules.

You are an Advertising Aesthetics Analyst specialized in evaluating advertisement images according to given aesthetic and advertising-specific rules, with support from specialized tool results.

INPUT: One image, one natural-language question related to a specific rule (Hue Adaptability, Color Harmonization, Copywriting Tone), and an associated tool result.

OUTPUT: EXACTLY two blocks, no extra text/spaces/newlines/markdown:

```
<think>Provide detailed reasoning strictly based on visible evidence (including tool results), clearly explaining why the image is Suitable or Unsuitable according to the given rule definition (no restating the question).</think><answer>{"Answer": "Suitable"|"Unsuitable", "Answer type": "Hue Adaptability"|"Color Harmonization"|"Copywriting Tone", "Tool": "Hue Adaptability tool"|"Color Harmonization tool"|"Copywriting Tone
```

```
tool"}</answer>
```

**RULES:**

- "Answer" must be exactly one of: Suitable, Unsuitable.
- "Answer type" must be exactly one of: Hue Adaptability, Color Harmonization, Copywriting Tone.
- "Tool" must match the evaluated rule: Hue Adaptability tool, Color Harmonization tool, Copywriting Tone tool.
- Your reasoning must strictly adhere to the provided rule definitions, and explicitly incorporate tool result information.

**RULE DEFINITIONS:**

- Hue Adaptability: Hue brightness and saturation should not be too high; overall hues should appear suitably engaging.
- Color Harmonization: Evaluate visual suitability through colorfulness and color combinations, ensuring appropriate engagement.
- Copywriting Tone: Suitability of the core textual content (especially headline) for the intended brand image, avoiding overly transactional or urgent language.
- Use only visible evidence (including the provided tool result); no guessing allowed.
- Strictly no extra keys or additional explanations in your output.

```
y2]], ...] | []}</answer>
```

**RULES:**

- "Answer" must be exactly one of: Suitable, Unsuitable.
- "Answer type" must be Promotional Iconography.
- "Detector result" must list ALL detected promotional visual elements (e.g., Red Envelope, Coupon, Gift Box, Promotional Sticker), each with:
  - "label": exact icon category.
  - "box\_2d": coordinates [x1, y1, x2, y2] precisely outlining the detected element in the image.
- If no promotional visual elements are detected, set "Detector result" to an empty list.
- Your reasoning must strictly adhere to the provided Promotional Iconography definition:

**RULE DEFINITION:**

- Evaluate suitability based on the usage and appropriateness of non-textual, stylized visual symbols that serve as cognitive shortcuts to concepts such as "value", "discount", "urgency", or "luck" in advertising contexts.
- Use only visible evidence; no guessing allowed.
- Strictly no extra keys or additional explanations in your output.

For the rule Promotional Iconography, the requirements differ from those of the other rules; therefore, the system prompt is as follows:

**System Prompt**

Follow the OUTPUT rules.

You are an Advertising Aesthetics Analyst specialized in evaluating Promotional Iconography in advertisement images.

INPUT: One image and one natural-language question specifically about Promotional Iconography.

OUTPUT: EXACTLY two blocks, no extra text/spaces/newlines/markdown:

```
<think>Provide detailed reasoning strictly based on visible evidence, clearly explaining why the promotional icons or symbols usage is Suitable or Unsuitable according to the provided rule definition (no restating the question).</think><answer>{"Answer": "Suitable"|"Unsuitable", "Answer type": "Promotional Iconography", "Detector result": [{"label": "Icon_Label", "box_2d": [x1, y1, x2,
```

For evaluating the two subjective rules, the following system prompts are used.

We design this prompt to ensure validity, reliability, and machine-robust scoring while keeping cross-model comparisons fair and reproducible.

- **Construct validity via rule isolation.** Each query elicits a judgment for exactly one target attribute (*Aesthetic* or *Advertising*), preventing construct leakage between visual appeal and persuasive communication.
- **Anchored 5-point rubric.** The compressed, anchor-based descriptors for levels 1–5 stabilize usage of the ordinal scale, mitigating leniency/severity drift and improving inter-rater consistency.
- **Evidence grounding.** The prompt explicitly mandates decisions from *visible evidence only* and forbids guessing, which ties judgments to observable cues (composition, clarity, message salience) rather than external knowledge.
- **Deterministic, machine-parseable output.** The two-block format separates a human-auditable `<think>` rationale from a strict `<answer>` JSON *single-element array* with a closed vocabulary ("Answer" ∈ {1, 2, 3, 4, 5}; "Answer\_type" ∈ {*Aesthetic attribute*, *Advertising attribute*}). This eliminates ad-hoc postprocessing and reduces parsing failures.

- **Ordinal integrity and metric compatibility.** Returning a single scalar (1–5) preserves the ordinal nature of the judgment and enables downstream statistics such as inter-rater reliability (Cronbach’s  $\alpha$ ) across raters and human–model agreement via rank or linear associations (SRCC or PLCC).
- **Cross-model fairness.** All models receive identical, schema-locked instructions; the rationale text is *not* used for scoring, preventing verbosity or style from contaminating quantitative evaluation.
- **Reusability and extensibility.** The fixed schema allows drop-in addition of new subjective dimensions or additional analyses without changing the evaluator; failures to meet the schema can be programmatically flagged for re-run or human review.

```

System Prompt

Follow the OUTPUT rules.
You are Advertising Quality Assessment
Assistant for evaluating the aesthetic
and advertising quality of advertisement-
style images.

INPUT: one image and one natural-language
question.

OUTPUT: return EXACTLY two blocks, no extra
text/spaces/newlines/markdown:
<think>Some reasoning process (no restating
the question).</think><answer>[{"Answer":
1|2|3|4|5, "Answer_type": "Aesthetic
attribute"|"Advertising attribute"}]</
answer>

RULES:
- "Answer" must be exactly one of: 1, 2, 3,
4, 5.
- "Answer_type":
- Aesthetic attribute visual appeal,
artistry, clarity, composition.
- Advertising attribute message clarity,
persuasiveness, brand emphasis.
- Use only visible evidence; do not guess.
- The JSON inside <answer> must be a single-
element array with exactly these keys: "
Answer", "Answer_type". No extra keys.

EVALUATION CRITERIA (compressed)

AESTHETIC ATTRIBUTE
- 1: Chaotic/unclear; harsh or dull colors;
blur/noise; crowded; cheap look.
- 2: Weak focus; minor defects; loose layout;
low appeal.
- 3: Clear but plain; safe colors; basic
retouch; orderly basics.
- 4: Creative flow; harmonious colors;
refined detail; balanced design.
- 5: Artistic/high-impact; pro-grade detail;
bold yet coherent style.

```

Table 1. Satisfaction scoring rubric (5-point Likert scale).

Level	Satisfaction Description
1	Misleading or irrelevant to intent; copy hard to read; clutter or technical defects distract; evokes negative emotion; would skip immediately.
2	Weak relevance or unclear value proposition; visual/copy feels noisy or unbalanced; low credibility; unlikely to consider.
3	Acceptable quality; message is understandable but unremarkable; basic brand fit; limited emotional pull.
4	Clear benefit and hierarchy; harmonious visuals; credible brand fit; positive emotion; likely to consider or learn more.
5	Instantly appealing and easy to understand; visuals are impressive and everything looks polished; fits the brand perfectly and feels emotionally engaging.

Table 2. Action Intent scoring rubric (5-point Likert scale).

Level	Action Intent Description
1	Actively ignores or avoids; no perceived need or trust; will not click, save, or inquire.
2	Mild curiosity at best; message/value unclear; unlikely to click or engage; no purchase impulse.
3	Understands the offering; may compare or read briefly; weak purchase impulse; deferred decision.
4	Clear value and credible proof; noticeable purchase impulse; inclined to click, add to cart, or request details soon.
5	Strong desire to buy right away; the ad feels urgent and highly relevant; you’re likely to click, sign up, or make a purchase on the spot.

```

ADVERTISING ATTRIBUTE
- 1: Purpose unclear; text unreadable; scene
off-brand.
- 2: Vague message; overloaded or unfocused
text; weak scene-product link.
- 3: Intent recognizable; selling point not
prominent; limited emotion.
- 4: Selling point obvious; concise
persuasive text; scene builds relevance/
emotion.
- 5: Visuals alone convey brand; aligns with
deep needs/values; strong emotion.

```

## 2. Details in Application I

**Why these two subjective endpoints (Satisfaction & Action Intent).** We include *Satisfaction* and *Action Intent* as complementary, human-centric [7] endpoints to judge whether A<sup>3</sup>-Law genuinely improves ads that people would want to act on. **Action Intent** operationalizes the final AIDA [5] stage (*Action*). It directly probes a viewer’s propensity to click, save, sign up, or purchase when presented with a screened image. Because our pipeline is ultimately designed to curate images *for humans*, this decision-proximal signal is the most policy-relevant outcome. We additionally measure **Satisfaction** as an attitudinal, holistic judgment of perceived quality (clarity, polish, brand fit, emotional resonance). It serves as a stable, low variance anchor that reflects overall liking and credibility, providing face validity for improvements that may not immediately translate into overt action.

### Design principles.

- **Construct alignment.** *Action Intent* captures behavioral propensity (decision readiness), whereas *Satisfaction* captures a global attitude toward the ad’s quality; together they separate *liking* from *willingness to act*.
- **Complementarity and diagnostic power.** Some changes (e.g., cleaner hierarchy, refined retouching) may raise *Satisfaction* without strongly shifting *Action Intent*; conversely, clearer value propositions or urgency cues may lift *Action Intent* more than *Satisfaction*. Joint reporting disambiguates these pathways.
- **Reliability and robustness.** Both endpoints use anchored 5-point rubrics to reduce leniency/severity drift across raters. We aggregate decisions across models and samples and report inter-rater consistency via Cronbach’s  $\alpha$ , ensuring the subjective protocol is auditable and reproducible.

**Inter-rater Consistency.** Table 1 presents the detailed scoring criteria used for evaluating the Satisfaction dimension, whereas Table 2 outlines the corresponding criteria for assessing Action Intent. In our study, multiple annotators independently rated the images based on these criteria. To assess the reliability and consistency of their evaluations, we calculated Cronbach’s  $\alpha$  [2] for both dimensions. The resulting inter-rater consistency scores were 0.771 for Satisfaction and 0.802 for Action Intent, respectively, which demonstrate a high degree of agreement among the annotators. These values confirm the robustness and reliability of our scoring procedures for both subjective evaluation tasks.

## 3. Details in Application II

Table 3 presents the scoring rubric for *Problem Identification Accuracy*, Table 4 for *Depth of Chain of Thought*

Table 3. Problem Identification Accuracy scoring rubric (5-point Likert).

Level	Description
1	Misses the main problems; focuses on irrelevant details; often misreads the image; feedback conflicts with what is visibly wrong.
2	Notices a few surface issues; misses part of the core problems; mixes correct and incorrect observations; only loosely aligned with the visible defects.
3	Identifies the main problem; overlooks some secondary issues; adds a bit of noise or minor misreads; overall direction is usable but incomplete.
4	Correctly captures most key problems; only small omissions or mild overstatements; observations are clearly grounded in what is visible.
5	Comprehensively pinpoints all major and most minor issues; no clear misinterpretations; priorities are sensible and tightly aligned with visible evidence.

Table 4. Depth of CoT scoring rubric (5-point Likert).

Level	Description
1	Gives almost no reasoning; mostly labels or repeats the question; does not explain why the issues matter.
2	Provides short and shallow reasoning; mentions one or two factors without clear links; limited insight into causes or impact.
3	Explains the main issues with some causal links; touches on a few aspects but stops early; reasoning is linear and somewhat brief.
4	Builds multi-step reasoning that connects design elements and intent; clearly explains why problems matter and how they affect performance.
5	Offers rich, well-structured reasoning from multiple angles; ties concrete visual evidence to rules and user impact; suggests clear directions for improvement.

(*CoT*), and Table 5 for *Overall Clarity*. Following these rubrics, the researchers evaluated the CoT process of each MLLM, and the resulting inter-rater consistency coefficients Cronbach’s  $\alpha$  are 0.744, 0.772, and 0.712, respectively, indicating a high level of agreement among annotators.

**Why these three CoT-focused indicators.** We assess the *process* rather than only the final labels because downstream users (designers/marketers) need *actionable diagnoses*. The three indicators target complementary aspects of usefulness:

Table 5. Overall Clarity scoring rubric (5-point Likert).

Level	Description
1	Hard to follow; disorganized or repetitive; vague wording; key points are buried or ambiguous.
2	Partly understandable but wordy or scattered; mixes problems and suggestions without structure; some confusing or imprecise phrases.
3	Main points are understandable; some redundancy or mild ambiguity; feedback is usable but not very polished.
4	Clear and well organized; key points stand out; sentences are concise and concrete; designers can follow the feedback without effort.
5	Extremely clear and easy to scan; structure highlights priorities; wording is precise and succinct; feedback feels directly actionable with minimal interpretation.

- **Problem Identification Accuracy (what).** Measures the correctness and coverage of issues the model flags in the image, anchoring judgments to visibly verifiable evidence and  $A^3$ -Law rule definitions [4]. This penalizes hallucinations and off-topic remarks while rewarding precise, grounded detection of defects.
- **Depth of CoT (why/how).** Evaluates whether the model builds multi-step causal links from concrete visual cues to violated rules and expected user impact [6]. It discriminates shallow tagging from analytic diagnosis, and supports prioritization and remediation planning.
- **Overall Clarity (how to use it).** Assesses structure, concision, and readability so that a designer can follow the feedback with minimal effort [3]. This favors actionable wording over verbosity and reduces cognitive load and review latency.

Together, these form a minimal, orthogonal basis spanning *content correctness*, *explanatory rigor*, and *communicative quality*—mirroring the practical design-review loop. The anchored 1–5 descriptors are model- and length-agnostic: scores must be grounded in visible evidence rather than response length or style. This design improves construct validity, stabilizes inter-rater judgments, and enables fair cross-model comparisons and downstream correlations while remaining easy to audit.

#### 4. Performance on the external dataset

Table 6. Performance on the external dataset.

Dataset Name	Aesthetic Attribute		Advertising Attribute	
	SRCC	PLCC	SRCC	PLCC
AdImageNet [1]	0.822	0.815	0.807	0.793

The proposed  $A^3$ -Law is designed to serve as an interpretable operationalization of established advertising aes-

thetic theories. While the efficacy of this formulation is demonstrated in the main text through its success on two independent downstream tasks (Selection and Critique), we provide additional quantitative validation here to assess its generalization capabilities. To this end, we evaluated our model on an external dataset by manually annotating a subset of 100 images sampled from the AdImageNet dataset. As detailed in Table 6, the model maintains a strong correlation with human judgments across both aesthetic and advertising attributes, demonstrating the robust applicability of the  $A^3$ -Law beyond the primary dataset.

### 5. Application as a Reward Model for Generative Tasks

Table 7. Generalization to AI-Generated Content Models.

Generative Model	Aesthetic Attribute			Advertising Attribute		
	SRCC	PLCC	Gains	SRCC	PLCC	Gains
Gemini-2.5-Flash-image	0.753	0.742	0.88	0.715	0.711	0.72
GPT-Image-1.5	0.771	0.762	0.94	0.726	0.723	0.83

To further demonstrate the versatility of our approach, we explored the potential of  $A^3$ -Align to serve as a reward model for guiding AI-generated content (AIGC). Specifically, we applied  $A^3$ -Align to perform Best-of-5 rejection sampling on images synthesized by two state-of-the-art generative models: Gemini-2.5-Flash-image and GPT-Image-1.5. For this evaluation, we collected 100 generated samples and compared our model’s scoring against human ratings. As detailed in Table 7,  $A^3$ -Align exhibits a strong correlation with human judgments and achieves notable selection gains over random sampling across both aesthetic and advertising attributes. These results suggest that our model can effectively steer generative processes toward higher-quality advertising imagery without the need for task-specific retraining.

### 6. Fine-Grained Visualizations of the $A^3$ -Law Framework

To provide a more intuitive understanding of the abstract theoretical dimensions defined in our  $A^3$ -Law framework, this section presents fine-grained visual examples from our dataset. These visualizations demonstrate exactly how our model evaluates specific visual and semantic characteristics. By examining the model’s CoT reasoning alongside these images, the interpretability and precision of our evaluation criteria become highly evident.

Each user query should be accompanied by its corresponding **SYSTEM PROMPT**, as specified in the preceding sections.

#### 6.1. Image Fidelity

Positive Example (Suitable): The positive example demonstrates optimal commercial clarity. The edges of the subject



**User:**

Evaluate whether the technical quality and integrity of this image are suitable.

**Assistant:**

**Think**

The image exhibits noticeable pixelation, particularly evident in the finer details of the text and character features. This pixelation manifests as a lack of sharpness where edges and lines appear jagged instead of smooth. The text, especially smaller fonts, becomes less legible due to these irregularities. Such degradation substantially compromises the overall clarity, making it challenging for viewers to engage fully with the content as intended. The visual noise introduced by pixelation detracts from the realism, reducing the effectiveness of the image as a clear and reliable visual communication tool.

**Answer**

{ "Answer": "Unsuitable", "Answer type": "Image Fidelity", "Tool": "No tool" }

Figure 1. An example from the A<sup>3</sup>-Dataset for Image Fidelity.

are sharp, textures are rendered with high definition, and there is no visible visual noise or compression artifacts, ensuring the product is presented flawlessly.

Negative Example (Unsuitable): As shown in Figure 1, this image is penalized for noticeable pixelation, particularly in the text and character features. The edges appear jagged rather than smooth, introducing visual noise that substantially compromises clarity and reduces the image's reliability as a communication tool.

**6.2. Integration Realism**

Positive Example (Suitable): The subject and background share a cohesive lighting environment. Cast shadows are physically plausible, and color reflections match the surroundings, grounding the subject naturally within the scene.

Negative Example (Unsuitable): As shown in Figure 2, there is a stark visual disconnect. The gold earrings are lit with sharp specular highlights, while the background box features soft, diffuse lighting. The complete absence of cast shadows from the earrings onto the box makes the subject look like a floating, hyper-real digital cutout, undermining physical credibility.

**6.3. Professional Polish**

Positive Example (Suitable): The post-processing enhances the product without distracting the viewer. Any text or graphical overlays are cleanly integrated, using high-resolution assets that align with professional e-commerce standards.

Negative Example (Unsuitable): As shown in Figure 3, the image suffers from crude post-processing. A low-resolution, pixelated red graphic has been haphazardly superimposed, obscuring the product. This aggressive, unrefined overlay gives the image an amateurish appearance typical of informal marketplaces, severely detracting from its commercial credibility.

**6.4. Hue Adaptability**

Positive Example (Suitable): The color palette uses balanced saturation and brightness levels that are comfortable for the viewer while effectively drawing attention to the product.

Negative Example (Unsuitable): As shown in Figure 4, the image utilizes extremely high saturation and brightness, pairing fluorescent greens with intense reds. This jarring combination causes visual discomfort and eye strain, conveying a cheap aesthetic that repels rather than attracts the viewer.

**6.5. Color Harmonization**

Positive Example (Suitable): The image employs a controlled color scheme (e.g., complementary or analogous colors) that guides the viewer's eye toward a clear focal point without visual confusion.

Negative Example (Unsuitable): As shown in Figure 5, the sheer variety of competing, highly saturated hues (bright pink, bold blue, neon green) creates overwhelming visual clutter. This lack of color harmony fragments the overall



**User:**

Evaluate whether the integration between the subject and background in this image appears suitably realistic.

**Assistant:**

**Think**

The image displays a clear visual disconnect between the subject and the background. The gold earrings are rendered with extremely sharp, crisp edges and are illuminated by a bright, hard light source that creates strong specular highlights, making them appear almost digitally perfect. In stark contrast, the red background and the jewelry box behind them are lit with a much softer, more diffuse light. This mismatch is most evident in the complete absence of any cast shadows from the earrings onto the box, which defies physical plausibility given their proximity. The earrings seem to float in a separate visual plane rather than occupying the same space as the box, making them look like a high-quality digital cutout pasted onto a separately rendered background. This lack of realistic interaction between the lighting and shadows of the subject and its environment creates a synthetic, hyper-real look that undermines the scene's physical credibility and gives it an artificial quality.

**Answer**

{ "Answer": "Unsuitable", "Answer type": "Integration Realism", "Tool": "No tool" }

Figure 2. An example from the A<sup>3</sup>-Dataset for Integration Realism.



**User:**

Evaluate whether the post-processing of this image suitably meets professional commercial standards.

**Assistant:**

**Think**

The image exhibits a significant degree of excessive and crude post-processing. A large, low-resolution red graphic with text has been digitally superimposed over the top portion of the photograph, obscuring part of the product and its hanger. This added element is jarring, with pixelated edges and a design that looks like a basic sticker from a simple editing application. This type of aggressive, unrefined graphic overlay gives the image an amateurish and cheap appearance, typical of informal online marketplaces rather than professional e-commerce. The crude nature of this digital addition severely detracts from the product's appeal, creating a visual experience that lacks sophistication and undermines the advertisement's commercial credibility, making it appear untrustworthy and of low quality.

**Answer**

{ "Answer": "Unsuitable", "Answer type": "Professional Polish", "Tool": "No tool" }

Figure 3. An example from the A<sup>3</sup>-Dataset for Professional Polish.

message and results in an unprofessional, confusing aesthetic.

## 6.6. Layout Adaptability

Positive Example (Suitable): The design features a clear visual hierarchy. Ample negative space allows the primary selling points to breathe, making the information easy to



**User:**

Evaluate whether the hue saturation and brightness levels in this image are suitable.

**Assistant:**

**Think**

The first thing I notice is the large area of vibrant green that dominates the background, paired with the equally striking red hues of the chair and the lettering. These colors exhibit extremely high saturation and brightness, creating a jarring effect. The green appears almost fluorescent, while the red is so intense it verges on unnatural. This combination leads to a visually aggressive advertisement that can cause significant eye strain and discomfort for viewers. The high contrast between the colors also fails to convey sophistication; instead, it gives off a cheap, low-quality feel. Such overwhelming color choices detract from the overall aesthetic appeal and could potentially repel potential customers rather than attracting them.

**Answer**

{ "Answer": "Unsuitable", "Answer type": "Hue Adaptability", "Tool": "No tool" }

Figure 4. An example from the A<sup>3</sup>-Dataset for Hue Adaptability.



**User:**

Evaluate whether the colorfulness of this image, as measured by visual complexity (e.g., Hasler's formula), is suitable.

**Assistant:**

**Think**

The poster features a vivid array of colors including bright pink, bold blue, neon green, and white. The multitude of these competing hues is too high, leading to an overwhelming level of color richness and visual clutter. The bright pink and bold blue, in particular, clash due to their intense saturation levels, creating a jarring effect. This combination of colors results in a confusing and unprofessional aesthetic for the viewer. The excessive color variety makes it difficult to identify a clear focal point, fragmenting the overall message and visual appeal.

**Answer**

{ "Answer": "Unsuitable", "Answer type": "Color Harmonization", "Tool": "No tool" }

Figure 5. An example from the A<sup>3</sup>-Dataset for Color Harmonization.



Figure 6. An example from the A<sup>3</sup>-Dataset for Layout Adaptability.

process at a glance across different media formats.

Negative Example (Unsuitable): As shown in Figure 6, the layout is chaotic and suffers from severe text overload. Competing blocks of information, varied font sizes, and decorative graphics destroy any compositional balance, making it exceedingly difficult for consumers to identify key selling points.

### 6.7. Copywriting Tone

Positive Example (Suitable): The copy uses persuasive, brand-appropriate language that highlights product value without resorting to aggressive or desperate sales tactics.

Negative Example (Unsuitable): As shown in Figure 7, the use of hyperbolic phrases like “Losing money clearance” and loud, comic-book typography creates an overwhelmingly transactional atmosphere. This cliché sales language positions the brand as low-value and generic.

### 6.8. Promotional Iconography

Positive Example (Suitable): Promotional badges are used sparingly and strategically to support the core offer without dominating the visual space.

Negative Example (Unsuitable): As shown in Figure 8,

#### User:

Evaluate whether the visual structure and layout of this image are suitable for adaptation to different media formats.

#### Assistant:

##### Think

The layout fails due to a significant Compositional Balance issue. The design is excessively crowded and chaotic, suffering from text overload. Numerous distinct blocks of information, multiple images, varied font sizes and colors, and decorative graphics all compete for the viewer’s attention simultaneously. This dense arrangement lacks a clear visual hierarchy, making it difficult for a potential customer to quickly process the key selling points. The overall effect is overwhelming and visually cluttered, which detracts from the professional quality of the advertisement and hinders effective communication.

##### Answer

{“Answer”: “Unsuitable”, “Answer type”: “Layout Adaptability”, “Tool”: “No tool”}

the visual field is cluttered with an excessive number of gold ingot icons. This random, overwhelming placement distracts from the core typography and cheapens the overall aesthetic quality.

### 6.9. Subjective Assessment

Beyond the binary evaluation of fine-grained flaws, the A<sup>3</sup>-Dataset also encompasses holistic subjective evaluations scored on a 1–5 scale. These assessments measure the overall visual appeal (Aesthetic attribute) and the commercial persuasiveness (Advertising attribute) of the generated images, guided by strict system evaluation criteria.

**Aesthetic Attribute:** As demonstrated in Figure 9, the model evaluates the overall artistry, visual clarity, and composition. Aligning precisely with the defined criteria for a score of 2 (“Weak focus; minor defects; loose layout; low appeal”), the model’s Chain-of-Thought (CoT) correctly identifies the advertisement’s plain composition and weak focal point. It critically notes the slightly mismatched bright red and yellow color scheme, local defects such as heavy shadows, and a loose layout lacking a clear visual hierarchy, ultimately justifying the low aesthetic score.

**Advertising Attribute:** As shown in Figure 10, the



**User:**

Evaluate whether the tone conveyed by the core copywriting in this image is suitable for the brand.

**Assistant:**

**Think**

The advertisement's copy relies on prominent, high-urgency phrases such as "Losing money clearance" and "Frantic rush in progress!!", which are rendered in large, bold, and brightly colored fonts. These textual elements, combined with aggressive calls to action like "Lowest 20% off" and the English word "Sale," create an overwhelmingly transactional and urgent atmosphere. The use of hyperbolic and cliché sales language, such as claiming to lose money on a clearance, alongside the loud, comic-book style typography, positions the offering as low-value and generic. This aggressive, sales-focused tone undermines any potential for brand sophistication or trustworthiness, ultimately making the advertisement feel cheap and overly desperate rather than appealing or premium.

**Answer**

{ "Answer": "Unsuitable", "Answer type": "Copywriting Tone", "Tool": "No tool" }

Figure 7. An example from the A<sup>3</sup>-Dataset for Copywriting Tone.



**User:**

Evaluate whether the promotional icons or visual symbols used in this image are suitably appropriate.

**Assistant:**

**Think**

Upon reviewing the advertisement, it is clear that promotional elements, specifically gold ingot icons, are abundantly scattered across the visual field, as detected at positions including [145, 169, 285, 239], [451, 173, 545, 222], [698, 190, 844, 267], and [174, 558, 333, 631]. The sheer quantity and random placement of these icons create a visually overwhelming and cluttered composition. Instead of enhancing the promotional message, their excessive use results in a distracting and cheap visual effect, drawing attention away from the core offer and typography. This overuse of iconography ultimately diminishes the advertisement's aesthetic quality, making it appear overly transactional and low-quality, which can negatively impact consumer perception of the brand's value and trustworthiness.

**Answer**

{ "Answer": "Unsuitable", "Answer type": "Promotional Iconography", "Detector result": [{"label": "Gold Ingot", "box\_2d": [145, 169, 285, 239]}, {"label": "Gold Ingot", "box\_2d": [451, 173, 545, 222]}, {"label": "Gold Ingot", "box\_2d": [698, 190, 844, 267]}, {"label": "Gold Ingot", "box\_2d": [364, 454, 448, 485]}, {"label": "Gold Ingot", "box\_2d": [611, 457, 689, 491]}, {"label": "Gold Ingot", "box\_2d": [698, 529, 873, 607]}, {"label": "Gold Ingot", "box\_2d": [246, 466, 345, 510]}, {"label": "Gold Ingot", "box\_2d": [174, 558, 333, 631]}] }

Figure 8. An example from the A<sup>3</sup>-Dataset for Promotional Iconography.



**User:**

Given the definitions for each level, which score (1–5) best matches the aesthetic quality of this ad?

**Assistant:**

**Think**

The advertisement exhibits a plain and uninspired composition with a weak focal point. The central chef figure is somewhat prominent but lacks creativity in its presentation. The color scheme uses bright reds and yellows, which can come across as slightly mismatched and overwhelming. While the image is generally clear, there are local defects such as heavy shadows around the chef and a distracting background with multiple elements that do not contribute to a cohesive narrative. The layout appears loose and lacks a strong design sense, with elements scattered without a clear visual hierarchy. Overall, the advertisement feels unattractive and fails to engage the viewer effectively.

**Answer**

[{"Answer": "2", "Answer\_type": "Aesthetic attribute"}]

Figure 9. An example from the  $A^3$ -Dataset for the Aesthetic attribute.



**User:**

Given the definitions for each level, which score (1–5) best matches the advertising quality of this ad?

**Assistant:**

**Think**

The advertisement effectively communicates its promotional purpose by clearly showcasing the product, a bottle of peach-flavored calcium milk, with prominent text highlighting its flavor and nutritional benefits. The text is large, legible, and uses contrasting colors that stand out against the pink background, ensuring readability. The core selling points, such as the product's flavor and health benefits (calcium and vitamin D), are emphasized through bold fonts and strategic placement. The visual scene, featuring cute peach illustrations and a playful cartoon character, aligns well with the brand's youthful and nostalgic tone, creating an engaging and emotionally resonant experience for the target audience. The overall composition is clean and cohesive, reinforcing the brand's identity and enhancing the advertisement's appeal.

**Answer**

[{"Answer": "4", "Answer\_type": "Advertising attribute"}]

Figure 10. An example from the  $A^3$ -Dataset for the Advertising attribute.

model assesses message clarity, brand emphasis, and marketing persuasiveness. Matching the criteria for a score of 4 (“Selling point obvious; concise persuasive text; scene builds relevance/emotion”), the model recognizes how effectively the product (peach-flavored calcium milk) is showcased. The CoT highlights the prominent, legible text that emphasizes core selling points (flavor and health benefits), and praises the playfully cute visual scene for

building emotional resonance that perfectly matches the brand’s youthful tone.

## 7. Dataset Domain Variety

Crucially, the  $A^3$ -Dataset encompasses a highly diverse range of commercial domains, ensuring the  $A^3$ -Law framework is robust, unbiased, and universally applicable. By analyzing the model’s CoT across the provided examples, it is

evident that the evaluation adapts seamlessly to the specific visual and semantic contexts of different industries:

- **Food & Beverage:** As seen in the peach-flavored milk ad (Fig. 10) and the restaurant chef poster (Fig. 9), the model evaluates domain-specific elements such as nutritional messaging, appetite appeal, and brand resonance.
- **Jewelry & Luxury Goods:** In the evaluation of the gold earrings (Fig. 2), the model correctly shifts its focus to high-end physical plausibility, meticulously analyzing specular highlights, diffuse lighting, and the physical interaction between the subject and its environment.
- **E-commerce & Apparel:** When assessing the product on a hanger (Fig. 3), the model applies professional e-commerce standards, criticizing amateurish graphical overlays that damage digital storefront credibility.
- **Promotional Retail:** For the clearance sale banner (Fig. 7), the model demonstrates a deep understanding of retail marketing psychology, identifying how aggressive, hyper-urgent copy and loud comic-book typography create a low-value, transactional tone.

**Extensive Category Coverage:** Beyond the specific representative examples visualized above, it is important to note that the  $A^3$ -Dataset is constructed to reflect the true breadth of the real-world advertising ecosystem. The dataset systematically covers a vast array of additional commercial categories, including but not limited to:

- **Automotive Industry:** Evaluating everything from dynamic driving shots to detailed, high-end interior show-cases.
- **3C Electronics:** Encompassing computers, smartphones, and smart home devices, which require the model to assess sleek, minimalist aesthetics and precise textural rendering (e.g., metallic and glass surfaces).
- **Daily Necessities & FMCG:** Covering essential household goods and basic commodities (such as rice, salt, and cleaning supplies), testing the model’s ability to evaluate pragmatic, utilitarian, and mass-market advertising styles.
- **Beauty & Skincare:** Prioritizing the evaluation of skin texture preservation, soft lighting, and elegant typography.

This extensive cross-domain coverage guarantees that our framework extends far beyond generic image quality assessment, successfully capturing the nuanced, industry-specific aesthetic and advertising requirements necessary for practical commercial applications.

## 8. Limitations and Failure Case Analysis

While our model demonstrates robust evaluation capabilities across most domains, it occasionally struggles with highly minimalist, premium designs, often referred to as the “minimalism penalty.” Figure 11 and Figure 12 illustrate these two typical failure cases.

**Analysis of Failure 1: Spatial crowding.** As discussed in our limitations, the model occasionally struggles with severe spatial crowding, particularly in high-density e-commerce promotional materials. Figure 11 illustrates a typical failure case during the Promotional Iconography evaluation. In this “6.18” mega-sale advertisement, numerous promotional badges (e.g., the “Limited time discount 150” tags and gold coins) are densely packed and physically overlapping. Due to this extreme visual clutter and lack of negative space, the model suffers from degraded spatial resolution. Instead of accurately segmenting each distinct promotional tag, it erroneously merges multiple overlapping badges into a single, enlarged bounding box. This demonstrates that while the model successfully recognizes the presence of promotional semantics, highly dense and chaotic layouts remain a challenge for precise spatial grounding.

**Analysis of Failure 2: Attribute miscalibration.** In this example, the advertisement features diamond rings elegantly arranged in a heart shape against a classic, unadorned “Tiffany blue” background. The model’s CoT accurately praises the image’s aesthetic execution, noting the harmonious color scheme, perfect balance, and refined, luxurious feel. However, it assigns an Advertising attribute score of 4, whereas the ground truth is a perfect 5. This discrepancy occurs because the model often conflates visual simplicity with a lack of advertising depth. According to our criteria, a score of 5 requires the visuals alone to convey the brand, align with deep values, and evoke strong emotion. In luxury advertising, extreme minimalism, abundant negative space, and the use of iconic brand colors (like the signature blue shown here) are highly deliberate strategies used to project prestige and brand confidence without relying on explicit text.

The model, however, tends to expect more explicit semantic cues—such as overt promotional text, complex lifestyle scenes, or overtly stated selling points—to justify a maximum score of 5 for advertising persuasiveness. Consequently, it slightly penalizes premium minimalist designs, failing to fully grasp that in the luxury sector, the “simple” absence of clutter is precisely what maximizes the advertisement’s high-end emotional impact.

## References

- [1] Peter Brendan. AdImageNet. <https://huggingface.co/datasets/PeterBrendan/AdImageNet>, 2024. 6
- [2] Lee J Cronbach. Coefficient alpha and the internal structure of tests. *psychometrika*, 16(3):297–334, 1951. 5
- [3] Q Vera Liao, Daniel Gruen, and Sarah Miller. Questioning the ai: informing design practices for explainable ai user experiences. In *Proceedings of the 2020 CHI conference on human factors in computing systems*, pages 1–15, 2020. 6
- [4] Jakob Nielsen and Rolf Molich. Heuristic evaluation of user



**User:**

Evaluate whether the promotional icons or visual symbols used in this image are suitably appropriate.

**Assistant:**

**Assistant's Think**

Upon reviewing the image, a multitude of promotional elements are explicitly detected, including stickers, coins, and coupons scattered throughout the composition. The sheer abundance of these icons, such as the "LIMITED OFFER" badge, "SAVE MORE" sticker, and numerous dollar sign coins, creates a visually overwhelming and cluttered aesthetic. This excessive use of promotional iconography, combined with their prominent placement, results in an overly aggressive and cheap visual effect that distracts from the core message. Consequently, this overuse significantly undermines the advertisement's visual quality and premium feel, making the promotion appear low-quality and overly transactional rather than appealing.

**Assistant's Answer**

```
{ "Answer": "Unsuitable", "Answer type": "Promotional Iconography", "Detector result": [{"label": "Promotional Sticker", "box_2d": [829, 1, 999, 158]}, {"label": "Burst Tag", "box_2d": [879, 517, 999, 646]}, {"label": "Coupon", "box_2d": [836, 716, 999, 887]}, {"label": "Coupon", "box_2d": [704, 886, 958, 999]}, {"label": "Coupon", "box_2d": [35, 851, 202, 996]}, {"label": "Coupon", "box_2d": [29, 165, 214, 348]}, {"label": "Gold Coin", "box_2d": [249, 275, 334, 353]}, {"label": "Gold Coin", "box_2d": [685, 308, 767, 387]}, {"label": "Gold Coin", "box_2d": [270, 694, 364, 783]}, {"label": "Gold Coin Cluster", "box_2d": [590, 670, 809, 999]}]}
```

**Ground Truth Answer**

```
{ "Answer": "Unsuitable", "Answer type": "Promotional Iconography", "Detector result": [{"label": "Promotional Sticker", "box_2d": [829, 1, 999, 158]}, {"label": "Burst Tag", "box_2d": [879, 517, 999, 646]}, {"label": "Coupon", "box_2d": [836, 716, 999, 887]}, {"label": "Coupon", "box_2d": [704, 886, 879, 999]}, {"label": "Coupon", "box_2d": [35, 851, 202, 996]}, {"label": "Coupon", "box_2d": [29, 165, 214, 348]}, {"label": "Gold Coin", "box_2d": [249, 275, 334, 353]}, {"label": "Gold Coin", "box_2d": [685, 308, 767, 387]}, {"label": "Gold Coin", "box_2d": [270, 694, 364, 783]}, {"label": "Gold Coin Cluster", "box_2d": [590, 670, 730, 809]}, {"label": "Gold Coin Cluster", "box_2d": [864, 890, 958, 999]}]}
```

Figure 11. A failure example from the A<sup>3</sup>-Dataset for the Promotional Iconography.

interfaces. In *Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 249–256, 1990. 6

- [5] Edward K Strong Jr. Theories of selling. *Journal of applied psychology*, 9(1):75, 1925. 5
- [6] Jason Wei, Xuezi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35: 24824–24837, 2022. 6
- [7] Alice Xiang, Jerone TA Andrews, Rebecca L Bourke, William Thong, Julianne M LaChance, Tiffany Georgievski, Apostolos Modas, Aida Rahmattalabbi, Yunhao Ba, Shruti Nagpal, et al. Fair human-centric image dataset for ethical ai benchmarking. *Nature*, pages 1–12, 2025. 5



**User:**

Given the definitions for each level, which score (1–5) best matches the advertising quality of this ad?

**Assistant:**

**Assistant's Think**

*The advertisement exhibits a high degree of advertising quality. The composition is creative, with diamond rings arranged in a heart shape that draws the viewer's attention immediately. This arrangement follows a clear visual flow, leading the eye naturally through the image. The color scheme is harmonious, featuring a classic Tiffany blue background that complements the silver and white of the jewelry, enhancing the brand's identity. The image is polished, with excellent clarity and sharpness, allowing for detailed observation of the rings' textures and facets. Lighting is well-managed, creating subtle reflections and highlights that give the diamonds a lifelike sparkle. The balance is perfect, with each ring placed thoughtfully to avoid overcrowding while maintaining a sense of abundance. The visual hierarchy is strong, with the larger central diamond acting as a focal point, surrounded by smaller rings that add depth and interest. Overall, the advertisement achieves a refined and luxurious feel.*

**Assistant's Answer**

[{"Answer": "4", "Answer\_type": "Advertising attribute"}]

**Ground Truth Answer**

[{"Answer": "5", "Answer\_type": "Advertising attribute"}]

Figure 12. A failure example from the A<sup>3</sup>-Dataset for the Advertising attribute.