Open World Image Aesthetic Assessment

Supplementary Material

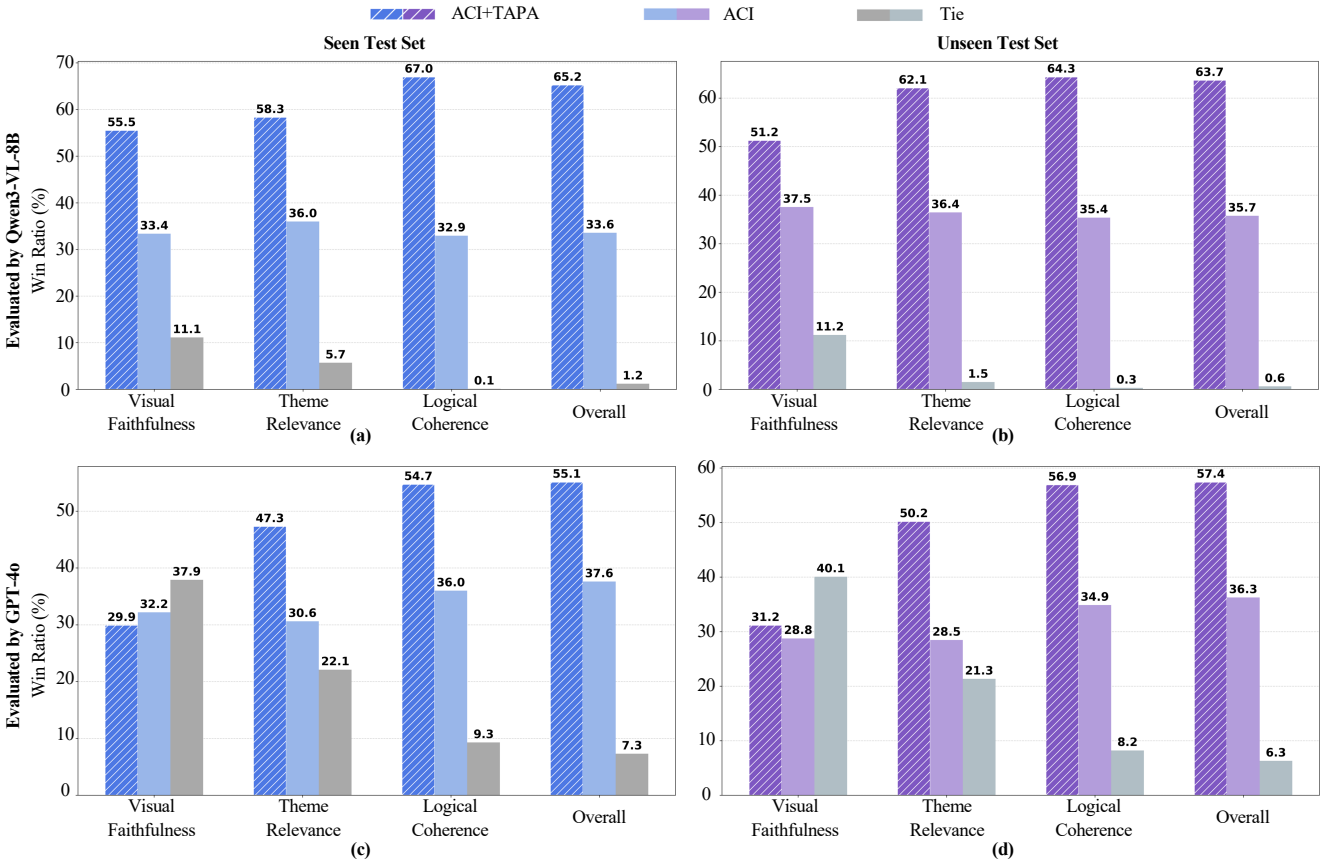


Figure 1. **Reasoning Quality Comparison of ACI versus ACI+TAPA, as Evaluated by Qwen-VL-8B and GPT-4o.** The charts illustrate the win ratio of the methods on (a, c) the seen set and (b, d) the unseen set. The top row (a, b) presents the evaluation results from the Qwen-VL-8B model, while the bottom row (c, d) shows the results from the GPT-4o model. The win ratio indicates the percentage of times a method was judged superior, and 'Tie' denotes instances of equal quality. The evaluation covers four dimensions: Visual Faithfulness, Theme Relevance, Logical Coherence, and Overall Quality.

Overview

This supplementary material supports the main paper with additional details and analyses, organized as follows:

- **Section A: Detailed analysis of the Synergistic Lift mechanism**, validating performance improvements on both seen and unseen test sets.
- **Section B: Extended ablation studies** focusing on pseudo label quality, crowd simulation parameters, and the robustness of the ACD metric.
- **Section C:** Qualitative visualizations **comparing the aesthetic reasoning capabilities** of the foundational policy (Stage 1) versus the adapted policy (Stage 2).
- **Section D:** Exact **prompt templates** used for theme annotation, pairwise CoT construction, and model training (Fig. 3–5).

A. Analysis of Synergistic Lift

To investigate the mechanism underlying the Synergistic Lift, we conducted a comprehensive pairwise evaluation of reasoning quality on both Unseen and Seen Test Sets. As illustrated in Fig. 1, the adapted model (ACI+TAPA) not only demonstrates superior reasoning capability on the Unseen Test Set (e.g., 63.7% Overall Win Ratio in Fig. 1 (b)), but crucially, this capability transfers to the Seen Test Set (Fig. 1 (a) and (c)), where it significantly outperforms the foundational policy (ACI) in both Theme Relevance (e.g., 58.3% vs. 36.0% win ratio in Fig. 1 (a)) and Logical Coherence (e.g., 54.7% vs. 36.0% in Fig. 1 (c)). These observations support the hypothesis for the Synergistic Lift: the adaptation process forces the model to escape "shortcuts" and activate generalizable aesthetic reasoning priors, result-

Table 1. Influence of Number of Rollout Samples N on the Quality of Pseudo Labels (Measured by MAE). Aggregating multiple samples ($N = 64$) effectively reduces prediction noise.

N	Low	Medium	High
1	0.389	0.366	0.436
4	0.373	0.357	0.405
16	0.360	0.348	0.398
64	**0.360**	**0.347**	**0.395**
128	0.362	0.348	0.397

Table 2. Influence of IAF Components on the Quality of Pseudo Labels (Measured by MAE). ACI is essential for initialization, while TAPA further refines the supervision signal.

ACI	TAPA	Low	Medium	High
×	×	3.817	3.753	3.715
✓	×	0.360	0.347	0.395
✓	✓	**0.358**	**0.336**	**0.372**

ing in evaluations that are more logically sound and strictly aligned with theme definitions, thereby explaining the concurrent performance enhancement on seen themes.

B. Additional Ablation Studies

B.1. Quality of Pseudo Labels

To ensure an effective adaptation process, the generated pseudo labels must be of high quality. We quantify this quality using the Mean Absolute Error (MAE) between the aggregated pseudo labels and the ground-truth scores.

Effectiveness of Crowd Simulation (N). We first validate the hypothesis that aggregating multiple stochastic inferences (i.e., crowd simulation) improves pseudo label quality. As shown in Table 1, the quality of pseudo labels improves consistently as the number of rollout samples N increases from 1 to 64. Specifically, the MAE decreases from 0.436 to 0.395 on the challenging High-deviation themes. This confirms that our crowd simulation strategy effectively captures a stable aesthetic consensus from the model's probabilistic outputs. Performance saturates beyond $N = 64$, justifying our choice of this value as a balance between accuracy and computational cost.

Necessity of IAF Components. We further analyze how each stage of our framework contributes to the generation of pseudo labels, as detailed in Table 2.

- *The Cold Start Problem (w/o ACI):* The vanilla backbone exhibits a catastrophic MAE (> 3.7), indicating that without the foundational aesthetic criterion induction from Stage 1, the model lacks a basic understanding of the scoring scale, rendering direct crowd simulation infeasible.
- *Foundational Policy (+ ACI):* Introducing ACI dramatically reduces the MAE to ~ 0.36. This proves that Stage 1 successfully installs a generalizable aesthetic policy, providing the necessary high-quality initial estimates for unsupervised adaptation.
- *Positive Feedback Loop (+ TAPA):* Finally, applying TAPA further decreases the MAE across all deviation levels (e.g., High: $0.395 \rightarrow 0.372$). This demonstrates a virtuous cycle: as the policy adapts to unseen themes, its self-generated pseudo labels become more precise, which in turn provides cleaner supervision for handling challenging aesthetic criteria.

Table 3. Robustness Validation of ACD across Backbones.

Backbone	Architecture	SRCC	PLCC
Siglip2 [4]	Transformer	1	1
DINOv2 [2]	Transformer	0.765	0.804
ConvNeXt V2 [3]	CNN	0.807	0.843

Table 4. Influence of k to Model Performance. (Measured by PLCC)

k	Low	Medium	High
1	0.852	0.832	0.746
5	0.853	0.833	0.749
10	0.852	0.833	0.749

B.2. Robustness of ACD.

To ensure that ACD captures the intrinsic aesthetic distinctiveness of themes rather than reflecting model-specific biases, we conduct a sensitivity analysis using distinct visual backbones: DINOv2 Large [2] (to test robustness against non-language supervision) and ConvNeXt V2 Large [3] (to test robustness against CNN inductive biases). As presented in Table 3, despite fundamental differences in architecture (ViT vs. CNN) and pre-training objectives (Self-Supervised vs. Supervised), the ACD rankings produced by these models exhibit high consistency with our Siglip2 baseline (SRCC > 0.76). This consistency demonstrates that ACD is a robust metric capable of quantifying thematic aesthetic deviations across diverse visual perception paradigms.

B.3. Sensitivity of Reward Function

Table 4 demonstrates the robustness of IAF to the hyperparameter k in the reward function. As the value of k varies across $\{1, 5, 10\}$, the method's performance, measured by PLCC, remains highly stable across all deviation levels.

C. Qualitative Analysis of Aesthetic Reasoning

To intuitively demonstrate the effectiveness of our Induce-and-Adapt Framework, we visualize the aesthetic reason-

ing generated by the foundational policy (Stage 1) and the adapted policy (Stage 2) in Fig. 2.

The results reveal a distinct contrast in reasoning quality. The foundational policy from Stage 1 (ACI), constrained by its training on seen themes, often succumbs to learning shortcuts. As highlighted in **red**, it tends to produce generic descriptions that lack thematic depth or even exhibit logical inconsistencies (e.g., criticizing "static" elements in a "Zen" theme where stillness is desirable). This indicates that the foundational policy has not fully decoupled generic aesthetic rules from theme-specific requirements.

In contrast, Stage 2 (TAPA) demonstrates a significant improvement. By adapting to the unseen themes via crowd-simulated pseudo-labels, the model breaks away from these shortcuts. As highlighted in **green**, the adapted policy generates more detailed and theme-aligned analyses, correctly grounding aesthetic attributes (such as lighting and composition) to the specific theme. This qualitative improvement empirically supports our hypothesis that test-time adaptation enables the model to escape the "shortcuts" of generic criteria learned in Stage 1, thereby achieving robust generalization in the open world.

D. Details of Prompts

This section presents the specific prompt templates employed across the different stages, covering Theme Annotation, pairwise CoT Construction, and model training.

Prompt Template for Theme Annotation (Fig. 3). This prompt is utilized during the theme annotation of OA-Bench. It instructs the language model to process raw challenge descriptions from DPChallenge [1] by filtering out conversational noise and encouraging language. The goal is to generate standardized theme labels and objective, factual descriptions that serve as the grounding context for aesthetic assessment.

Prompt Template for Pairwise CoT Construction (Fig. 4). Used in the data generation phase, this prompt directs the teacher model (GPT-4o) to first provide independent analyses for each image, followed by a comparative aesthetic evaluation under the specific theme. The output is strictly structured in JSON format, containing independent analyses of both images, a comparative justification, and a final quality ranking. This structured output facilitates the verification.

Prompt Template for Model Training (Fig. 5). This template serves as the standard input format for the Multimodal Large Language Model during both the ACI and TAPA stages. It conditions the model on the specific theme and description and enforces a Chain-of-Thought format, requiring the model to generate a reasoning process within `<think>` tags before predicting the final aesthetic score within `<answer>` tags.

References

[1] Challenging Technologies, LLC. DPChallenge - a digital photography contest. https://www.dpchallenge.com/, 2001–2025. Accessed: 2025-11-04. 3

[2] Maxime Oquab, Timothée Darcet, Theo Moutakanni, Huy V. Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Russell Howes, Po-Yao Huang, Hu Xu, Vasu Sharma, Shang-Wen Li, Wojciech Galuba, Mike Rabbat, Mido Assran, Nicolas Ballas, Gabriel Synnaeve, Ishan Misra, Herve Jegou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. Dinov2: Learning robust visual features without supervision, 2023. 2

[3] Ronghang Hu Xinlei Chen Zhuang Liu In So Kweon Sanghyun Woo, Shoubhik Debnath and Saining Xie. Convnext v2: Co-designing and scaling convnets with masked autoencoders. *arXiv preprint arXiv:2301.00808*, 2023. 2

[4] Michael Tschannen, Alexey Gritsenko, Xiao Wang, Muhammad Ferjad Naeem, Ibrahim Alabdulmohsin, Nikhil Parthasarathy, Talfan Evans, Lucas Beyer, Ye Xia, Basil Mustafa, Olivier Hénaff, Jeremiah Harmsen, Andreas Steiner, and Xiaohua Zhai. Siglip 2: Multilingual vision-language encoders with improved semantic understanding, localization, and dense features, 2025. 2

Theme:
Zen and Serene Photography
Score:
6.40

Stage 1 (ACI)
Reasoning: This image captures a contemplative mood with its cool blue tones and soft, misty water effect. The composition is balanced, with the figure adding a sense of scale and introspection. However, **the image feels slightly static**, and the darker tones may detract from the theme's emphasis on simplicity and serenity.
Score: 7.20

Stage 2 (TAPA)
Reasoning: This image features a lone figure standing by the ocean, with smooth, mist-like water and muted blue tones. **The long exposure creates a sense of tranquility and fluidity**, while the figure's solitary presence adds emotional depth. However, the **composition feels slightly unbalanced due to the placement of the figure**, and the rocky foreground introduces a sense of ruggedness that slightly detracts from the serene theme.
Score: 6.40

Theme:
Right Angles Composition
Score:
4.56

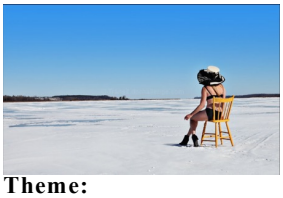
Stage 1 (ACI)
Reasoning: This image features intersecting power lines and a tree with angular branches, which **loosely align with the theme of right angles**. The composition is dynamic, with the lines creating a sense of depth and movement. However, the lighting is uneven, and the colors feel muted, which detracts from the overall impact. The right angles are present but not dominant.
Score: 3.9

Stage 2 (TAPA)
This image features a telephone pole and wires, which incorporate right angles into the composition. The warm lighting on the tree contrasts nicely with the cool tones of the sky, creating a pleasing color palette. However, the right angles are less dominant and **feel secondary to the organic shapes of the tree** and the overall balance of the composition.
Score: 4.4

Theme:
Coffee Themed Photography
Score:
4.77

Stage 1 (ACI)
Reasoning: This image features a cup of coffee alongside coffee beans and a jar of peppers, creating a **visually rich composition**. The warm tones and textures are appealing, and the coffee is clearly the focal point, aligning well with the theme. However, the lighting could be more refined to enhance the overall clarity and vibrancy.
Score: 4.2

Stage 2 (TAPA)
Reasoning: This image showcases a coffee cup alongside coffee beans and a jar of chili peppers, emphasizing the coffee directly. The composition is balanced, but the **lighting is uneven**, with some glare on the glass. The colors are rich but lack vibrancy, and the **background feels cluttered**, detracting from the focus on the coffee.
Score: 4.7



Theme:
Seating Furniture Photography
Score:
5.61

Stage 1 (ACI)
Reasoning: This image features a vibrant composition with a striking contrast between the bright snow and the vivid blue sky. The placement of the chair and the figure creates a sense of isolation and whimsy, **aligning well with the theme**. However, **the image could benefit from more dynamic lighting to enhance depth.**
Score: 6.5

Stage 2 (TAPA)
Reasoning: This image features a striking juxtaposition of a yellow chair against a vast snowy landscape and vibrant blue sky. The composition is clean and minimal, with the chair and seated figure positioned off-center, creating visual interest. The lighting is natural and well-balanced, emphasizing the textures of the snow and the chair. However, **the thematic focus on seating furniture is somewhat diluted by the prominence of the human subject**.
Score: 5.9

Figure 2. Qualitative comparison of aesthetic reasoning generated by the foundational policy (Stage 1 ACI) and the adapted policy (Stage 2 TAPA).

````
# ROLE
You are an expert text analyst specializing in information summarization and categorization.
# TASK
Your primary task is to read each piece of challenge data provided below (containing a "filename", "title", and
a "description"). For each entry, you must perform the following steps:
1. **Analyze the Core Theme**: Understand the central subject or requirement of the challenge title and
description.
2. **Generate a Theme**: Summarize the core theme into a concise "theme" using 4 English words or fewer.
3. **Ensure Consistency**: If multiple challenge descriptions are fundamentally about the same theme (even if
worded differently), you **must** assign them the exact same theme string. This promotes standardization across
your output.
4. **Generate Objective Description**: Create a neutral, factual summary of the challenge's requirements,
stripped of any conversational or encouraging language.
# OUTPUT FORMAT
Strictly return a single JSON array of objects. Each object must contain three key-value pairs:
1. `"filename"`: The filename from the input data (e.g., `"challenge_1591.html"`).
2. `"theme"`: The generated summary of 4 words or fewer.
3. `"objective_description"`: The newly generated neutral and factual summary of the challenge requirements.

The order of objects in the array must match the order of the input entries. Do not include any extra
explanations or text outside of the JSON array in your response.
# EXAMPLE
## Example Input:
filename,title,description
challenge_10.html,Curves,Just like the title says, use your creativity to photograph curvature. The photograph
must be taken this week (April 1st - April 7th). Be creative, and good luck!
challenge_11.html,Minimalism,Capture a scene with as few elements as possible. Focus on simplicity, clean lines,
and negative space.
challenge_12.html,Minimalist Photo,This week's challenge is about minimalism. Create a simple photo with
minimal elements.
## Example Output:
```json
[
{
"filename": "challenge_10.html",
"theme": "Creative curvature photography",
"objective_description": "Photograph a subject featuring curvature. The photo must be taken between April 1st
and April 7th."
},
{
"filename": "challenge_11.html",
"theme": "Minimalist photography",
"objective_description": "Capture a scene with minimal elements, focusing on simplicity, clean lines, and
negative space."
},
{
"filename": "challenge_12.html",
"theme": "Minimalist photography",
"objective_description": "Create a simple photo with minimal elements."
}
]
### TASK EXECUTION
Now, generate the corresponding theme and objective description for each of the following data entries and
organize them into a JSON array of objects as requested.
````

Figure 3. Prompt Template for Theme Annotation with Gemini2.5-Pro.
````

```
You are a world-class art director and critic.
[EVALUATION CONTEXT]
* Theme: {theme}
* Theme Description: {description}
[TASK INSTRUCTIONS]
You will be presented with two images: the first is [Image A], and the second is [Image B]. Please conduct a
concise evaluation of both images by leveraging your professional knowledge of aesthetics (e.g., composition,
lighting, color theory) and considering the provided theme context.
You must strictly output your analysis in the following JSON format. Do not add any text or explanations
before or after the JSON object.
{
"parsed_theme": "{theme}",
"analysis_A": "[Provide a concise analysis of Image A's key strengths and weaknesses. Focus on its aesthetic
qualities (composition, color, etc.) and its relevance to the theme '{theme}'. Keep the analysis brief and
to the point.]",
"grade_A": "[bad, poor, fair, good, or excellent]",
"analysis_B": "[Provide a concise analysis of Image B's key strengths and weaknesses, similar to the
evaluation of Image A. Keep it brief.]",
"grade_B": "[bad, poor, fair, good, or excellent]",
"comparison_analysis": "[Briefly compare Image A and Image B. From the combined perspective of overall
aesthetics and theme relevance, which image is more successful and why? Summarize the key deciding
factors.]",
"sort_order": "[A > B, B > A, or A = B]"
}
```

Figure 4. Prompt Template for Pairwise CoT Construction Strategy.

```
<image>\nYou are an expert art critic and photo analyst. Your task is to rate the aesthetic quality of the
given image on a scale from 1 to 10.

The image's theme is '{theme}', with the objective description: '{description}'.

Please provide your reasoning process within <think></think> tags and your final answer (the numerical score)
within <answer></answer> tags.
```

Figure 5. Prompt Template for Model Training.