Supplementary Materials

January 26, 2025

This file provides supplementary information of the paper *"Harnessing the Power of LLMs for Image Aesthetics Assessment through Semantic and Contextual Understanding"*.

1 Aesthetic Evaluation Factors

The proposed list of evaluation factors used in Exp. 1 and 2 in the main paper is shown in Fig. 1.

		Category		Item	Description
0	1	Composition	1	Arrangement	The arrangement of elements in the image (balance, symmetry, golden ratio, or rule of thirds).
1	1	Composition	2	Pattern and Texture	The arrangement of repeating elements and the surface quality.
2	1	Composition	3	Depth	The arrangement of foreground, middle ground, and background layers.
3	1	Composition	4	Background Simplicity	The simplicity of the background.
4	2	Color	1	Color Harmony	The harmony of colors.
5	2	Color	2	Color Contrast	The effectiveness of color contrast.
6	3	Lighting and Shadow	1	Effect of Lighting	The appropriate use of natural or artificial light.
7	3	Lighting and Shadow	2	Effect of Shadows	The appropriate use of shadows.
8	3	Lighting and Shadow	3	Dynamic Range	The range of contrast between the brightest and darkest areas.
9	4	Image Quality	1	Resolution and Sharpness	The clarity, detail, and sharpness of the image.
10	4	Image Quality	2	Focus and Blur	The appropriateness of focus and blur.
11	5	Subject	1	Clarity of Theme	The clarity and recognizability of the main subject or theme.
12	5	Subject	2	Content	The appeal and interest of the subject matter.
13	5	Subject	3	Shape	The appropriateness and attractiveness of the shape of the main subject.
14	5	Subject	4	Perspective	The appropriateness of the viewpoint.
15	6	Creativity	1	Creativity	The novelty, originality, or inventiveness displayed.
16	7	Emotional and Physical Reaction	1	Emotional Reaction	The emotional impact and response elicited by the image.
17	7	Emotional and Physical Reaction	2	Physical Reaction	The physical sensations or reactions provoked by the image.
18	8	Context and Knowledge	1	Cultural/Social/Historical Context	The cultural, social, or historical significance and context of the image.
19	8	Context and Knowledge	2	Expertise and Artistic Techniques	The demonstration of skill and artistry in the image.

Figure 1: Aesthetic evaluation factors

2 Prompting Methods

2.1 Q prompt

2.1.1 Default Q prompt

Default question prompt (no-factors Q prompt) is as follows. First, a system attribute message briefly explains the task.

```
You will be asked to provide an aesthetic evaluation of the
image that will be presented to you.  The aesthetic evaluation
of the image is conducted in three levels (0:  low, 1:  middle,
2:  high).  When answering, output only in JSON format.  Refrain
from outputting anything other than JSON format.
```

Next, a test image is presented, followed by a user-type message that asks an LLM to answer the prediction of aesthetic evaluation.

```
Predict the aesthetic evaluation of the image by choosing
from three classes (0:  low, 1:  middle, 2:  high) as an integer
value (0,1,2), and explain the reason and your confidence
level (as a floating-point number ranging from 0 to 1).

Note:
- Answer in English.
- Answer from your own perspective.
- Ensure that your output is strictly in JSON format without
any additional characters or formatting symbols such as ‘‘‘
before and after the JSON object.
- Strictly adhere to the example format provided below:
{"reason":  "~~~", "prediction":  0, "confidence":  0.5}
```

2.1.2 COnly_C1-4_text Q prompt

In the case where the evaluation factors are used, the Q prompt is as follows. Especially, we show an example in the case of COnly_C1-4_text (See Section 4.1 in this supplementary material for the detailed description of this setting). Only the user-type message that asks the question is changed.

```
Evaluate the following aspects of the image by considering
the listed categories and provide detailed descriptions for
each.
Q1:  Composition
Q2:  Color
Q3:  Lighting and Shadow
Q4:  Image Quality
Then, predict the aesthetic evaluation of the image by choosing
from three classes (0:  low, 1:  middle, 2:  high) as an integer
value (0,1,2), and explain the reason and your confidence
level (as a floating-point number ranging from 0 to 1).
Note:
- Answer in English.
- Ensure that your output is strictly in JSON format without
any additional characters or formatting symbols such as ‘‘‘
before and after the JSON object.
- Strictly adhere to the example format provided below:
{"Q1":  "...", "Q2":  "...", "Q3":  "...", "Q4":  "...", "reason":
"...", "prediction":  1, "confidence":  0.5}
```

2.1.3 C&I_C1-8_3-class Q prompt

We show another example in the case of C&I_C1-8_3-class (See Section 4.1 in this supplementary material for the detailed description of this setting). Only the user-type message that asks a question is changed.

```
Evaluate the following aspects of the image by considering
the listed categories and rate each aspect as one of the following:
low, middle, high, where low is bad and high is good.
~Composition~
Q1.1:  Arrangement --- The arrangement of elements in the
image (balance, symmetry, golden ratio, or rule of thirds).
Q1.2:  Pattern and Texture --- The arrangement of repeating
elements and the surface quality.
Q1.3:  Depth --- The arrangement of foreground, middle ground,
and background layers.
Q1.4:  Background Simplicity --- The simplicity of the background.
~Color~
Q2.1:  Color Harmony --- The harmony of colors.
Q2.2:  Color Contrast --- The effectiveness of color contrast.
~Lighting and Shadow~
Q3.1:  Effect of Lighting --- The appropriate use of natural
or artificial light.
Q3.2:  Effect of Shadows --- The appropriate use of shadows.
Q3.3:  Dynamic Range --- The range of contrast between the
brightest and darkest areas.
~Image Quality~
Q4.1:  Resolution and Sharpness --- The clarity, detail, and
sharpness of the image.
Q4.2:  Focus and Blur --- The appropriateness of focus and
blur.
~Subject~
Q5.1:  Clarity of Theme --- The clarity and recognizability
of the main subject or theme.
Q5.2:  Content --- The appeal and interest of the subject
matter.
Q5.3:  Shape --- The appropriateness and attractiveness of
the shape of the main subject.
Q5.4:  Perspective --- The appropriateness of the viewpoint.
~Creativity~
Q6.1:  Creativity --- The novelty, originality, or inventiveness
displayed.
~Emotional and Physical Reaction~
Q7.1:  Emotional Reaction --- The emotional impact and response
elicited by the image.
Q7.2:  Physical Reaction --- The physical sensations or reactions
```

provoked by the image.
~Context and Knowledge~
Q8.1: Cultural/Social/Historical Context --- The cultural, social, or historical significance and context of the image.
Q8.2: Expertise and Artistic Techniques --- The demonstration of skill and artistry in the image.

Then, predict the aesthetic evaluation of the image by choosing from three classes (0: low, 1: middle, 2: high) as an integer value (0,1,2), and explain the reason and your confidence level (as a floating-point number ranging from 0 to 1).
Note:
- Answer in English.
- Ensure that your output is strictly in JSON format without any additional characters or formatting symbols such as ‘‘‘ before and after the JSON object.
- Strictly adhere to the example format provided below:
{"Q1.1": "middle", "Q1.2": "middle", "Q1.3": "middle", "Q1.4": "middle", "Q2.1": "middle", "Q2.2": "middle", "Q3.1": "middle", "Q3.2": "middle", "Q3.3": "middle", "Q4.1": "middle", "Q4.2": "middle", "Q5.1": "middle", "Q5.2": "middle", "Q5.3": "middle", "Q5.4": "middle", "Q6.1": "middle", "Q7.1": "middle", "Q7.2": "middle", "Q8.1": "middle", "Q8.2": "middle", "reason": "...", "prediction": 1, "confidence": 0.5}

2.2 TU prompt

2.2.1 Default TU prompt

In the PIAA task, LLMs are asked to capture an user's tendency of aesthetic evaluation from few-shot example images using the tendency understanding prompt (TU prompt). An example of Default TU prompt (no-factors TU prompt) with the "Class-Balanced" selecting method is as follows.

First, how the few-shot examples are presented is explained.

```
You will be tasked with understanding a user's aesthetic evaluation
tendencies towards images.
A total of f images will be presented for this purpose, all
of which are themed around ~~~.
In addition, for each image, you will be provided with the
aesthetic evaluation given by the current user (PIAA_Score,
a discrete value).
The PIAA_Score ranges from 1 to 5 and is categorized into
three aesthetic evaluation classes:  scores of 1.0, 1.5, 2.0
are low; 2.5, 3.0, 3.5 are middle; and 4.0, 4.5, 5.0 are high.
An equal number of images will be presented from each aesthetic
evaluation class.
##############################################
```

Then, for each of the f few-shot images, the image and its rating are presented.

```
The aesthetic evaluation of the above image is:
PIAA_Score = ~~~(PIAA_Score class:  ~~~)
##############################################
```

Lastly, the following statement is presented to ask the user's tendency of aesthetic evaluation.

```
Based on the examples provided above, you will be asked to
predict the user's aesthetic evaluation of the images that
will be presented.

First, please describe the user's 'aesthetic evaluation' tendencies
towards images based on the f images presented above.
```

2.2.2 TU_C5-8 TU prompt

In the case where the evaluation factors are used, the TU prompt is as follows. Especially, we show an example in the case of TU_C5-8. Only the part that asks about tendency understanding is changed.

```
First, please describe the user's aesthetic evaluation tendencies
towards images based on the f images presented above.
Follow the items listed below to understand these tendencies.
~Subject~
Q5.1:  Clarity of Theme --- The clarity and recognizability
of the main subject or theme.
Q5.2:  Content --- The appeal and interest of the subject
matter.
Q5.3:  Shape --- The appropriateness and attractiveness of
the shape of the main subject.
Q5.4:  Perspective --- The appropriateness of the viewpoint.
~Creativity~
Q6.1:  Creativity --- The novelty, originality, or inventiveness
displayed.
~Emotional and Physical Reaction~
Q7.1:  Emotional Reaction --- The emotional impact and response
elicited by the image.
Q7.2:  Physical Reaction --- The physical sensations or reactions
provoked by the image.
~Context and Knowledge~
Q8.1:  Cultural/Social/Historical Context --- The cultural,
social, or historical significance and context of the image.
Q8.2:  Expertise and Artistic Techniques --- The demonstration
of skill and artistry in the image.
Note:
- Answer in English.
- Ensure that your output is strictly in JSON format without
any additional characters or formatting symbols such as ‘‘‘
before and after the JSON object.
- Strictly adhere to the example format provided below:
{"Q5.1":  "...", "Q5.2":  "...", "Q5.3":  "...", "Q5.4":  "...",
"Q6.1":  "...", "Q7.1":  "...", "Q7.2":  "...", "Q8.1":  "...",
"Q8.2":  "..."}
```

3 Parameter Settings

We use the "gpt-4o-2024-05-13" version of GPT-4o. Temperature is set to 0. The setting of "detail" for processing image is set to "low" mode.

4 Exp. 1-S: GIAA

In this section, we present supplementary information of additional experiments for the GIAA task.

4.1 Exp. 1-S1: Level of Question Detail

In Exp. 1-S1, we examine whether the way these evaluation factors are presented for an intermediate evaluation step affects the performance of LLMs in aesthetic evaluation. The intermediate evaluations are described in a free-text format.

COnly vs. C&I: The former (Category Only) refers to cases where only the category names are provided, while the latter (Category and Item) refers to cases where detailed items, including explanatory descriptions, are provied as evaluation factors in prompts.

C1-4 vs. C1-8 vs. C5-8: These variations represent differences in which categories are used as evaluation factors. In the case of C1-4, Categories 1 through 4 in Fig. 1 are used. Similarly, in C1-8 and C5-8, Categories 1 through 8 and Categories 5 through 8 are used, respectively.

We compare a total of seven Q prompts. The result of the distribution of accuracy across 10 seeds for the 180 test images under each prompt condition is shown in Table 1.

Looking at the differences between COnly and C&I, it is clear that C&I outperforms COnly in all cases, whether for C1-4, C1-8, or C5-8. Next, examining the differences between C1-4, C1-8, and C5-8, we observe that performance increases in the order of C5-8 < C1-8 < C1-4.

Table 1: Comparison of different conditions in question details in terms of accuracy in the GIAA task. Mean and standard deviation over 10 trials are reported. **Bold** indicates the best performance, while underlined values denote conditions better than the no-factors baseline.

Condition	Accuracy
no-factors	0.7200 ± 0.0284
COnly_C1-4	0.6822 ± 0.0327
COnly_C1-8	0.6783 ± 0.0270
COnly_C5-8	0.5889 ± 0.0257
C&I_C1-4	$\underline{\mathbf{0.7261}} \pm 0.0358$
C&I_C1-8	0.7144 ± 0.0246
C&I_C5-8	0.6544 ± 0.0297

4.2 Exp. 1-S2: Intermediate Evaluation Format (text vs. 3-class)

In Exp. 1-S2, we compare two conditions under C&I setting, which uses categories and detailed items as evaluation factors:
text vs. 3-class: The former is one where the intermediate evaluation is provided as free-text, and the latter is where it is given as a three-class evaluation. We use three patterns for the categories: C1-4, C1-8, and C5-8.

The result of the distribution of accuracy across 10 seeds for the 180 test images under each prompt condition is shown in Table 2.

For both the text and 3-class intermediate evaluation patterns, it is evident that performance increases in the order of C5-8 < C1-8 < C1-4. When comparing text and 3-class, the text pattern outperforms 3-class in the case of C5-8, while 3-class slightly outperforms text in the case of C1-8. In the case of C1-4, 3-class considerably outperforms text. The C1-4 with 3-class condition performs substantially better than the Default prompt, achieving an accuracy of 75.7%, which is considerably higher than the chance level of 33.3% for the three-class classification task.

Table 2: Comparison of intermediate evaluation format (free-text vs. three-classes) in terms of accuracy in the GIAA task. Mean and standard deviation over 10 trials are reported. **Bold** indicates the best performance, while underlined values denote conditions better than the no-factors baseline.

Condition	Accuracy
no-factors	0.7200 ± 0.0284
C1-4_txt	0.7261 ± 0.0358
C1-8_txt	0.7144 ± 0.0246
C5-8_txt	0.6544 ± 0.0297
C1-4_3cls	**0.7567** ± 0.0288
C1-8_3cls	0.7189 ± 0.0335
C5-8_3cls	0.6461 ± 0.0440

4.3 Exp. 1-S3: Intermediate Evaluation Format (0-1 vs. 3-class)

In this Exp. 1-S3, we examine a new type of format, "0-1". In the case of "0-1", the intermediate evaluation for each evaluation factor is answered in the format of a continuous value between 0 and 1. We use C&I_C1-4 or C&I_C1-8 types for the categories used and 5 different random seeds to take variability in image selection and LLM's outputs into account.

The result of the comparison is shown in Table 3. It shows that the cases using "0-1" intermediate evaluation underperform those using "3-class" format.

Table 3: Comparison of intermediate evaluation format (0-1 vs. 3-class) in terms of accuracy in the GIAA task. Mean and standard deviation over 5 trials are reported. **Bold** indicates the best performance.

Condition	Accuracy
C1-4_0-1	0.6833 ± 0.0654
C1-4_3-class	$\mathbf{0.7489} \pm 0.0186$
C1-8_0-1	0.6500 ± 0.0362
C1-8_3-class	0.7011 ± 0.0577

4.4 Exp. 1-S4: Focusing on Selected Evaluation Factors

In this Exp. 1-S4, we investigate the case where an LLM is asked to select some factors to focus on from the presented evaluation factors. We use C&I_C1-4_3-class or C&I_C1-8_3-class settings, varying the number of focus (3 or 5). We use 5 different random seeds to take variability in image selection and LLM's outputs into account.

The result of the comparison is shown in Table 4. It shows that the difference is small among the cases with different number of factors focused on. However, compared to the results in Table 2, it can be observed that focusing within C1-4 leads to a decrease in performance, whereas focusing within C1-8 contributes to performance improvement.

Table 4: Comparison of the number of factors focused on in terms of accuracy in the GIAA task. Mean and standard deviation over 5 trials are reported. **Bold** indicates the best performance.

Condition	Accuracy
C1-4_3Foc	0.7344 ± 0.0348
C1-4_5Foc	0.7333 ± 0.0219
C1-8_3Foc	$\mathbf{0.7356} \pm 0.0313$
C1-8_5Foc	0.7300 ± 0.0203

4.5 Exp. 1-S5: Perspective

In this Exp. 1-S5, we investigate the cases where the following sentence is added to Default Q prompt: "- Answer from your own perspective" or "- Answer from the perspective of the general public". We use 10 different random seeds to take variability in image selection and LLM's outputs into account.

The result of the comparison is shown in Table 5. This shows that it is better not to specify a perspective from which an LLM predicts the aesthetic evaluation.

Table 5: Comparison of perspectives in terms of accuracy in the GIAA task. Mean and standard deviation over 10 trials are reported. **Bold** indicates the best performance.

Condition	Accuracy
None	**0.7200** $\pm$ 0.0284
You	0.7072 $\pm$ 0.0285
Public	0.6950 $\pm$ 0.0199

5 Exp. 2-S: PIAA

In this section, we present supplementary information of additional experiments for the PIAA task.

5.1 Exp. 2-S1: Number of Few-Shot Examples

In this Exp. 2-S1, we examine the influence of the number of few-shot examples. We use Default Q prompt and TU_C1-4 prompt for the 12 target users.

The result of the comparison is shown in Table 6. This shows that the cases where few-shot examples are used outperform the case with no few-shot examples in terms of accuracy, indicating that the importance of few-shot prompting.

Table 6: Comparison of the number of few-shot examples in terms of accuracy in the PIAA task. Mean and standard deviation over 12 target users are reported. **Bold** indicates the best performance, while underlined values denote conditions better than the $f = 0$ baseline.

Condition	Accuracy
0	0.5690 ± 0.0589
4	**0.6162** ± 0.0618
8	0.6106 ± 0.0464
12	0.6037 ± 0.0604
16	0.6120 ± 0.0506
20	0.6083 ± 0.0599

Additionally, Fig. 2 shows an example of the distribution of scores in the few-shot image set of a target-semantic category (animal photographs) for a target user ($f = 20$). A total of 60 images (20 images per each of the 3 aesthetic classes) are sorted by residual score. The color of the arrows indicates the sign of the residual score r (red for positive and blue for negative). The base of the arrow represents the GIAA score g, while the tip represents the PIAA score p. This shows that the images with large absolute value of residual score mainly belong to either the class High ($p = 4.0, 4.5, 5.0$) or the class Low ($p = 1.0, 1.5, 2.0$), rather than the class Middle ($p = 2.5, 3.0, 3.5$).

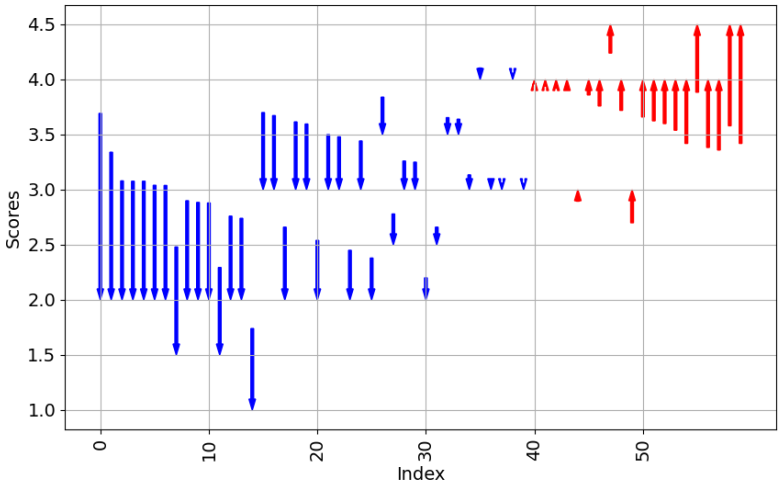

Figure 2: Distribution of scores in few-shot example images of animal scenes for a target user (PIAA)