

JoPPO: Hierarchical Photography Assessment via Contrastive Joint Conditional Probabilistic Reinforcement Learning

Supplementary Material

Appendix A Derivation of the Conditional Gaussian Preference Model

A.1 Conditional Distribution of Overall Scores Given Attribute Scores

Recall that we model the joint distribution of attribute scores and overall aesthetic scores as a multivariate Gaussian:

$$\begin{bmatrix} \mathbf{A} \\ S \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} \mu_A \\ \mu_S \end{bmatrix}, \begin{bmatrix} \Sigma_{AA} & \Sigma_{AS} \\ \Sigma_{SA} & \sigma_{SS} \end{bmatrix}\right) \quad (1)$$

where \mathbf{A} denotes the random vector of attribute scores, S is the random variable of the overall score, μ_A and μ_S are their means, and the covariance matrix is partitioned into the attribute covariance Σ_{AA} , cross-covariances Σ_{AS}, Σ_{SA} and overall-score variance σ_{SS} .

Our goal is to derive the conditional distribution of S given an observed attribute vector \mathbf{a} , i.e., $p(S | \mathbf{A} = \mathbf{a})$.

Step 1: Define a decorrelated residual of S . We construct the following linear transformation:

$$S' = S - \Sigma_{SA}\Sigma_{AA}^{-1}(\mathbf{A} - \mu_A) \quad (2)$$

Because (\mathbf{A}, S) is jointly Gaussian, any affine transformation of (\mathbf{A}, S) is also jointly Gaussian. In particular, (\mathbf{A}, S') is jointly Gaussian.

Step 2: Show that S' is uncorrelated with \mathbf{A} . First, compute the mean of S' :

$$\begin{aligned} \mathbb{E}[S'] &= \mathbb{E}[S - \Sigma_{SA}\Sigma_{AA}^{-1}(\mathbf{A} - \mu_A)] \\ &= \mathbb{E}[S] - \Sigma_{SA}\Sigma_{AA}^{-1}(\mathbb{E}[\mathbf{A}] - \mu_A) \\ &= \mu_S \end{aligned} \quad (3)$$

Then compute the covariance between \mathbf{A} and S' :

$$\begin{aligned} \text{Cov}(\mathbf{A}, S') &= \text{Cov}(\mathbf{A}, S - \Sigma_{SA}\Sigma_{AA}^{-1}(\mathbf{A} - \mu_A)) \\ &= \text{Cov}(\mathbf{A}, S) - \Sigma_{SA}\Sigma_{AA}^{-1}\text{Cov}(\mathbf{A}, \mathbf{A}) \\ &= \Sigma_{AS} - \Sigma_{SA}\Sigma_{AA}^{-1}\Sigma_{AA} \\ &= \Sigma_{AS} - \Sigma_{AS} \\ &= 0 \end{aligned} \quad (4)$$

Thus, \mathbf{A} and S' are uncorrelated. Because (\mathbf{A}, S') is jointly Gaussian, uncorrelatedness implies independence, i.e., S' is independent of \mathbf{A} .

Step 3: Compute the variance of S' . Using the definition in (2), we have

$$\begin{aligned} \text{Var}(S') &= \text{Var}(S - \Sigma_{SA}\Sigma_{AA}^{-1}(\mathbf{A} - \mu_A)) \\ &= \text{Var}(S) + \Sigma_{SA}\Sigma_{AA}^{-1}\text{Var}(\mathbf{A})\Sigma_{AA}^{-1}\Sigma_{AS} \\ &\quad - 2\text{Cov}(S, \Sigma_{SA}\Sigma_{AA}^{-1}(\mathbf{A} - \mu_A)) \\ &= \sigma_{SS} + \Sigma_{SA}\Sigma_{AA}^{-1}\Sigma_{AA}\Sigma_{AA}^{-1}\Sigma_{AS} \\ &\quad - 2\Sigma_{SA}\Sigma_{AA}^{-1}\Sigma_{AS} \\ &= \sigma_{SS} - \Sigma_{SA}\Sigma_{AA}^{-1}\Sigma_{AS} \end{aligned} \quad (5)$$

Therefore,

$$S' \sim \mathcal{N}(\mu_S, \sigma_{SS} - \Sigma_{SA}\Sigma_{AA}^{-1}\Sigma_{AS}) \quad (6)$$

Step 4: Obtain the conditional mean and variance.

Since S' is independent of \mathbf{A} , its distribution does not change when conditioning on $\mathbf{A} = \mathbf{a}$. From (2) we can write

$$S = S' + \Sigma_{SA}\Sigma_{AA}^{-1}(\mathbf{A} - \mu_A) \quad (7)$$

Conditioning on $\mathbf{A} = \mathbf{a}$ yields

$$S | \mathbf{A} = \mathbf{a} = S' + \Sigma_{SA}\Sigma_{AA}^{-1}(\mathbf{a} - \mu_A) \quad (8)$$

Taking expectation and variance on both sides, and using (6), we obtain

$$\mu_{S|\mathbf{a}} = \mathbb{E}[S | \mathbf{A} = \mathbf{a}] = \mu_S + \Sigma_{SA}\Sigma_{AA}^{-1}(\mathbf{a} - \mu_A) \quad (9)$$

$$\sigma_{S|\mathbf{a}}^2 = \text{Var}(S | \mathbf{A} = \mathbf{a}) = \sigma_{SS} - \Sigma_{SA}\Sigma_{AA}^{-1}\Sigma_{AS} \quad (10)$$

Equations (9) and (10) are exactly the conditional mean and variance used in the main paper.

A.2 Conditional Preference Probability Between Two Candidates

Consider two candidate items $(\mathbf{a}_i^{(m)}, s_i^{(m)})$ and $(\mathbf{a}_j^{(n)}, s_j^{(n)})$ drawn from candidate groups K_i and K_j , respectively. Using (9)–(10), we have the conditional distributions

$$s_i^{(m)} | \mathbf{a}_i^{(m)} \sim \mathcal{N}(\mu_{s_i^{(m)}|\mathbf{a}_i^{(m)}}, \sigma_{s_i^{(m)}|\mathbf{a}_i^{(m)}}^2) \quad (11)$$

$$s_j^{(n)} | \mathbf{a}_j^{(n)} \sim \mathcal{N}(\mu_{s_j^{(n)}|\mathbf{a}_j^{(n)}}, \sigma_{s_j^{(n)}|\mathbf{a}_j^{(n)}}^2) \quad (12)$$

where

$$\mu_{s_i^{(m)}|\mathbf{a}_i^{(m)}} = \mu_S + \Sigma_{SA}\Sigma_{AA}^{-1}(\mathbf{a}_i^{(m)} - \mu_A) \quad (13)$$

$$\mu_{s_j^{(n)}|\mathbf{a}_j^{(n)}} = \mu_S + \Sigma_{SA}\Sigma_{AA}^{-1}(\mathbf{a}_j^{(n)} - \mu_A) \quad (14)$$

$$\sigma_{s_i^{(m)}|\mathbf{a}_i^{(m)}}^2 = \sigma_{SS} - \Sigma_{SA}\Sigma_{AA}^{-1}\Sigma_{AS} \quad (15)$$

$$\sigma_{s_j^{(n)}|\mathbf{a}_j^{(n)}}^2 = \sigma_{SS} - \Sigma_{SA}\Sigma_{AA}^{-1}\Sigma_{AS} \quad (16)$$

Conditioned on the attribute vectors $\mathbf{a}_i^{(m)}$ and $\mathbf{a}_j^{(n)}$, we assume the overall scores $s_i^{(m)}$ and $s_j^{(n)}$ are independent. Define the difference

$$D_{i,j}^{(m,n)} = s_i^{(m)} - s_j^{(n)} \quad (17)$$

Then, given $(\mathbf{a}_i^{(m)}, \mathbf{a}_j^{(n)})$, $D_{i,j}^{(m,n)}$ is also Gaussian with

$$\mathbb{E}[D_{i,j}^{(m,n)} | \mathbf{a}_i^{(m)}, \mathbf{a}_j^{(n)}] = \mu_{s_i^{(m)}|\mathbf{a}_i^{(m)}} - \mu_{s_j^{(n)}|\mathbf{a}_j^{(n)}} \quad (18)$$

$$\text{Var}(D_{i,j}^{(m,n)} | \mathbf{a}_i^{(m)}, \mathbf{a}_j^{(n)}) = \sigma_{s_i^{(m)}|\mathbf{a}_i^{(m)}}^2 + \sigma_{s_j^{(n)}|\mathbf{a}_j^{(n)}}^2 \quad (19)$$

The conditional probability that candidate (i, m) is preferred over (j, n) is

$$\begin{aligned} p_\theta(s_i^{(m)} > s_j^{(n)} | \mathbf{a}_i^{(m)}, \mathbf{a}_j^{(n)}) \\ &= p(D_{i,j}^{(m,n)} > 0 | \mathbf{a}_i^{(m)}, \mathbf{a}_j^{(n)}) \\ &= \Phi\left(\frac{\mu_{s_i^{(m)}|\mathbf{a}_i^{(m)}} - \mu_{s_j^{(n)}|\mathbf{a}_j^{(n)}}}{\sqrt{\sigma_{s_i^{(m)}|\mathbf{a}_i^{(m)}}^2 + \sigma_{s_j^{(n)}|\mathbf{a}_j^{(n)}}^2}}\right) \end{aligned} \quad (20)$$

where $\Phi(\cdot)$ is the CDF of the standard normal distribution. In practice, we add a small constant γ inside the square root for numerical stability, which yields the conditional preference probability used in the main paper:

$$p_\theta(s_i^{(m)} > s_j^{(n)} | \mathbf{a}_i^{(m)}, \mathbf{a}_j^{(n)}) \quad (21)$$

$$= \Phi\left(\frac{\mu_{s_i^{(m)}|\mathbf{a}_i^{(m)}} - \mu_{s_j^{(n)}|\mathbf{a}_j^{(n)}}}{\sqrt{\sigma_{s_i^{(m)}|\mathbf{a}_i^{(m)}}^2 + \sigma_{s_j^{(n)}|\mathbf{a}_j^{(n)}}^2 + \gamma}}\right) \quad (22)$$

Equations (9), (10), and (21) thus provide a complete derivation of the conditional mean, conditional variance, and conditional preference probability under the joint Gaussian assumption.

Appendix B Attribute-Overall Correlation Analysis

To further validate the structural reliability of our model’s attribute scores, we conduct an experiment on the out-of-distribution dataset PARA to measure the correlation between attribute-level scores and overall scores. Specifically, we perform ridge regression from the five attribute

scores to the overall score and compute standard regression metrics including Mean Squared Error (MSE), Root Mean Squared Error (RMSE), Coefficient of Determination (R^2), and Cross-Validated R^2 (CV_R^2).

We evaluate this on four sources of score data: (1) Human-preference scores (manual annotation of the test set), (2) Our model’s predictions, (3) Q-Insight predictions, (4) Artimuse predictions.

Table 1. Ridge regression metrics between predicted attribute scores and overall scores on the PARA test set, the best result are highlighted in bold.

Model	MSE↓	RMSE↓	R ² ↑	CV_R ² ↑
Q-Insight	0.102	0.319	0.432	0.515
Artimuse	0.135	0.367	0.572	0.537
<i>Human-preference</i>	<i>0.009</i>	<i>0.096</i>	<i>0.965</i>	<i>0.966</i>
Ours	0.083	0.288	0.713	0.684

As shown in Fig. 1 and Tab. 1, our model achieves a significantly stronger attribute-to-overall alignment than the baselines, as indicated by a higher R^2 score and lower error. Compared to Q-Insight and Artimuse, our model’s attribute scores exhibit better linear predictability of overall aesthetics, confirming that our model successfully captures human-consistent aesthetic reasoning.

Moreover, the regression performance of our model approaches that of the human scores, demonstrating that our attribute outputs are not only interpretable but also semantically aligned with the human aesthetic judgment process.

Appendix C Further Analysis

C.1 Case Study on Composition Classification

To demonstrate our model’s ability on performing interpretable composition classification, we showcase six examples from the PICD dataset in Fig. 2, where the model is asked to identify the underlying composition type and provide a textual explanation for its decision.

The selected examples cover a variety of common compositional patterns. In each case, the model accurately classifies the image and offers a reasoning statement that reflects an understanding of spatial structure and visual balance. Most notably, in the final example, an image depicting a casual gathering without any intentional layout—the model correctly classifies it as “No composition.” Its explanation highlights the lack of perspective, symmetry, or rule-of-thirds usage, demonstrating that the model is not overconfident and can recognize unstructured scenes. This reinforces the model’s robustness and compositional discernment even when no explicit rule is present.

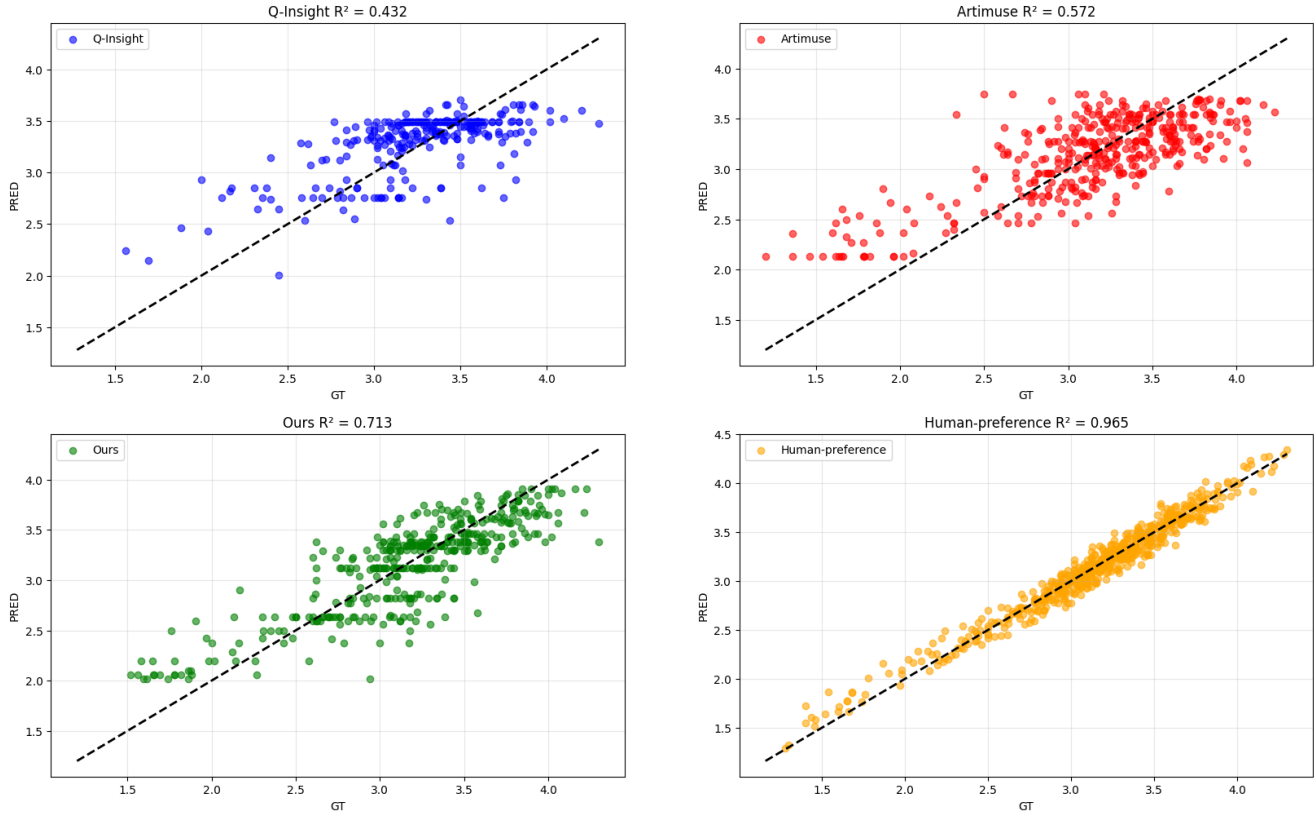


Figure 1. Visualization of ridge regression fit between attribute scores and overall scores.

C.2 Case Study on Aesthetic Assessment

To further demonstrate the effectiveness of our model in interpretable aesthetic assessment, we present a set of examples chosen on PARA in Fig. 3, covering different quality images. This selection allows us to examine whether the model can consistently align with human judgments across the quality spectrum. In all three cases, the model generates detailed, multi-attribute reasoning followed by corresponding sub-scores and an overall prediction.

Our model demonstrates strong alignment with human ratings on high aesthetic quality images. For the low aesthetic quality image, our model clearly articulates its critique—highlighting the lack of compositional intent, depth layering, and dynamic elements. The predicted scores across all dimensions are significantly lower than in other cases, and the final overall rating (2.0 vs. GT 2.24) again aligns closely with human perception.

These examples collectively demonstrate that our model not only provides semantically grounded explanations, but also exhibits high score alignment with human annotations at both attribute and overall levels. Moreover, the relative differences in predictions across cases show the model’s capacity to reason comparatively, assigning lower scores

when images lack structure or appeal, and rewarding well-composed scenes with appropriately higher ratings. This confirms the model’s robustness in aesthetic ranking, compositional analysis, and preference alignment across diverse visual contexts.

Besides, we also demonstrate a failure case in Fig. 4. In this case, although the attribute and overall scores predicted by the model are close to the ground truth, the textual reasoning reveals notable inaccuracies. Specifically, when the input image is blurry or lacks a clearly defined subject, the model struggles to accurately recognize the image content, leading to incorrect reasoning regarding elements such as lighting or semantic content. While our framework ensures structural consistency between attribute scores and overall assessment, this case highlights that its performance can degrade when the base model’s perception capabilities are compromised by poor image quality.



Does these images follow composition rules? If yes, which composition rule does it follow?

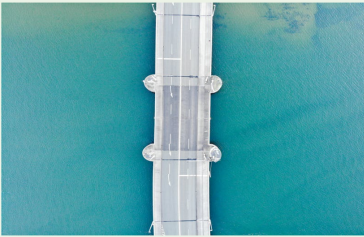


<think> The heron is positioned on the right third, creating a dynamic composition. Its long neck and open beak draw attention, while the water and vegetation in the background provide balance. This placement guides the viewer's gaze, emphasizing the heron as the focal point.</think>

<answer> Composition Category: Vertical Third Line </answer>



Vertical Third Line



<think> The image shows a long, straight bridge spanning the width of the frame. The bridge's central alignment creates a clear vertical line, dividing the image into two symmetrical halves. This central axis guides the viewer's eye along the bridge, emphasizing its length and structure. The uniform blue water on both sides enhances the sense of order and balance, typical of a vertical middle line composition.</think>

<answer> Composition Category: Vertical Middle Line </answer>



Vertical Middle Line

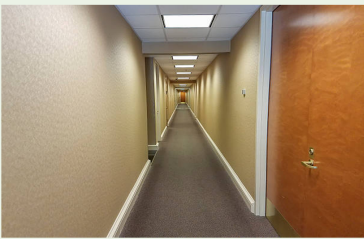


<think> The image shows a long wooden dock extending into a calm lake, with the horizon line dividing the scene into two distinct zones. The dock aligns with the left third, guiding the eye toward the center. The clear sky and reflective water create a balanced, serene composition. This setup emphasizes the dock's length and the lake's expanse, adhering to the Horizontal Third Part rule.</think>

<answer> Composition Category: Horizontal Third Part </answer>



Horizontal Third Part



<think> The image shows a long corridor with a central vanishing point at the end. The perspective lines of the floor and walls lead to this point, creating a sense of depth. The doors on the right and the wall on the left align with the vanishing point, guiding the viewer's eye. This setup exemplifies the Perspective composition rule, enhancing the corridor's length and depth.</think>

<answer> Composition Category: Perspective </answer>



Perspective



<think> The train is positioned on the right side of the image, creating a balanced composition. The railing and tracks guide the eye toward the train, emphasizing its presence. The sea in the background provides a natural horizon, enhancing the visual flow. This arrangement follows the rule of placing the train as the focal point, with the sea as a secondary element, creating a harmonious and dynamic scene.</think>

<answer> Composition Category: Single Shape Rule of Thirds </answer>



Single Shape Rule of Thirds



<think> The image shows a group of people seated at a table, with the focus on the foreground. The lighting and arrangement of the subjects create a balanced visual flow, but there's no evident use of perspective, symmetry, or rule of thirds. The composition appears to be a casual gathering rather than a structured artistic arrangement.</think>

<answer> Composition Category: No composition </answer>



No composition

Figure 2. Some cases on our model's composition classification.



Please evaluate these images from aesthetic perspective. Provide reason and score.



Reason:

Composition: The image shows a clear linear composition with the buildings aligning vertically. The Handelsbanken sign is prominent, guiding the eye. The curved building on the right contrasts with the straight lines, adding dynamism. The composition uses depth and contrast effectively.

Color: The image lacks vibrant colors, focusing on architectural elements. The blue sky and glass facades create a balanced, somewhat monochromatic palette. This enhances the structural focus, avoiding distractions.

Depth of Field: The image uses shallow depth of field, focusing on the buildings' upper sections. The blurred background emphasizes the foreground, creating a sense of depth. This directs attention to the architectural details, enhancing visual flow.

Lighting: The lighting appears natural, with the sun casting soft shadows. The reflective glass surfaces and clear sky suggest a daytime setting. This natural light enhances the image's clarity and balance, highlighting the buildings' textures.

Content: The image depicts a modern urban landscape, emphasizing architectural design. The Handelsbanken sign and curved building draw focus, while the straight-lined buildings provide a balanced contrast. This content appeals to those interested in urban planning and architecture.

Overall: The image's composition, color, and depth of field work together to create a visually appealing urban scene. The balanced visual flow and clear focus on architectural elements enhance its aesthetic quality.

Scores:

Composition: 4.25; Color: 2.5; Depth of Field: 3.75; Lighting: 4.0; Content: 3.75; Overall: 3.9



Composition: 4.04; Color: 3.52; Depth of Field: 3.56; Lighting: 3.68; Content: 3.52; Overall: 3.98



Reason:

Composition: The image shows a vast open field with flowers and mountains in the background. The horizon line is clear, dividing the sky and field. The flowers are scattered, creating a natural flow. The composition guides the eye across the scene, emphasizing the expansive landscape.

Color: The image features a vibrant mix of yellow, red, and white flowers against the green field and blue sky. The sunset adds warm tones, enhancing the visual appeal. The color contrast between the flowers and the sky creates a balanced and harmonious composition.

Depth of Field: The flowers in the foreground are in sharp focus, while the mountains in the background are slightly blurred. This creates a sense of depth, drawing attention to the foreground elements. The shallow depth of field enhances the natural beauty and focus on the flowers.

Lighting: The soft, diffused light of the sunset highlights the flowers and field, creating a warm, inviting glow. The contrast between the bright foreground and darker background sky adds depth. This lighting enhances the image's natural beauty and emotional resonance.

Content: The content is rich in visual elements, creating a peaceful and inviting scene. The composition and color palette suggest a focus on nature's beauty and tranquility.

Overall: The image showcases a harmonious blend of natural elements, with the flowers and mountains creating a balanced and visually appealing composition. The lighting and depth of field enhance the natural beauty, making it a captivating and aesthetically pleasing scene.

Scores:

Composition: 4.4; Color: 4.6; Depth of Field: 3.9; Lighting: 4.25; Content: 4.4; Overall: 4.45



Composition: 4.12; Color: 4.2; Depth of Field: 4.0; Lighting: 4.28; Content: 4.04; Overall: 4.28



Reason:

Composition: The image shows a street scene with a cart and a dog, but the composition lacks a clear focal point. The arrangement of elements appears random, lacking intentional spatial guidance. There is no evident use of perspective or symmetry.

Color: The color palette is muted and somewhat monochromatic, with the buildings and ground dominating the scene. There is no vibrant contrast or harmonious color arrangement. The lighting appears natural but lacks depth, contributing to the overall subdued tone.

Depth of Field: The image lacks a clear depth of field, with all elements appearing equally sharp or blurred. There is no evident use of foreground, middle ground, or background to create depth, lacking spatial layering.

Lighting: The lighting appears natural, possibly overcast, as there are no strong shadows or highlights. This even lighting contributes to the image's calm and somewhat a plain atmosphere.

Content: The content is a street scene with a cart and a dog, but the composition and focus are not clear. There is no evident narrative or central theme. The subjects are static, and the image lacks dynamic visual flow.

Overall: The overall aesthetic is functional rather than artistic. The image lacks visual impact due to its lack of composition, color harmony, and depth. The content is straightforward but not engaging, and the lighting is balanced but not dramatic.

Scores:

Composition: 2.5; Color: 2.0; Depth of Field: 1.5; Lighting: 1.5; Content: 2.0; Overall: 2.0



Composition: 2.36; Color: 2.32; Depth of Field: 2.24; Lighting: 2.16; Content: 2.28; Overall: 2.24

Figure 3. Some cases on our model's aesthetic assessment.



Please evaluate these images from aesthetic perspective. Provide reason and score.



Composition: 1.86; Color: 1.76; Depth of Field: 2.03;
Lighting: 1.9; Content: 1.93; Overall: 1.67



Reason:

Composition: The image appears to lack a clear compositional structure. There is no evident use of perspective, symmetry, rule of thirds, or leading lines. The arrangement of visual elements seems random or disorganized, lacking intentional spatial guidance.
Color: The color palette is muted and lacks vibrancy. The dominant colors are white and gray, with no vivid or contrasting hues. This creates a subdued and somewhat monotonous visual experience, emphasizing the image's lack of aesthetic appeal.

Depth of Field: The image shows a shallow depth of field, with the foreground elements (curtain) in sharp focus and the background (shower area) blurred. This effect draws attention to the foreground while creating a sense of depth and guiding the viewer's gaze.

Lighting: The lighting appears to be natural from a window, but it is not particularly directional or dramatic. The shadows and highlights are soft and even, contributing to the image's calm and balanced aesthetic.

Content: The content is a close-up of a curtain with a face pattern. The focus is on the curtain's texture and pattern, but the composition lacks narrative or symbolic depth. The subject is static, and there is no evident action or interaction.

Overall: The image lacks a clear artistic vision. The composition is disorganized, the color palette is muted, and the depth of field is shallow. The lighting is even, and the content is static. The overall aesthetic is calm but unengaging, lacking visual interest or narrative clarity.

Scores:

Composition: 1.0; Color: 1.5; Depth of Field: 2.5; Lighting: 2.0; Content: 1.0;
Overall: 1.5

Figure 4. A bad case on our model's aesthetic assessment.