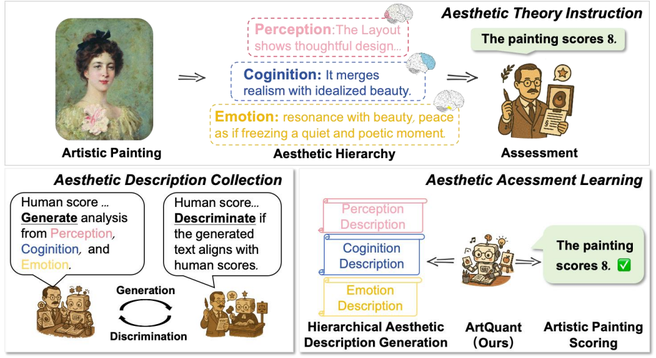
Bridging Cognitive Gap: Hierarchical Description Learning for Artistic Image Aesthetics Assessment

Henglin Liu[1,2], Nisha Huang[2,3], Chang Liu[1‡], Jiangpeng Yan[1], Huijuan Huang[2], jixuan Ying[1], Tong-Yee Lee[4], Pengfei Wan[2], Xiangyang Ji[1‡]

1. Tsinghua University 2. Kling Team, Kuaishou Technology 3. Pengcheng Laboratory 4. National Cheng Kung University

AAAI-26 / IAAI-26 / EAAI-26
JANUARY 20-27, 2026 | SINGAPORE

Teaser



- ArtQuant enhances the alignment between models and human aesthetic judgment through auxiliary hierarchical description generation tasks.

Motivation & Contribution

Motivations

- **Data aspect:** Manual aesthetic annotation is costly and scarce. Current datasets, like APDD, are *limited to surface-level technical analysis,* and their descriptions are often *too general and not well-grounded in visual details*.
- **Method aspect:** Existing models are fragmented: visual networks use isolated multi-branch encoders for aesthetic attributes, and multimodal contrastive learning methods struggle with long-text descriptions.

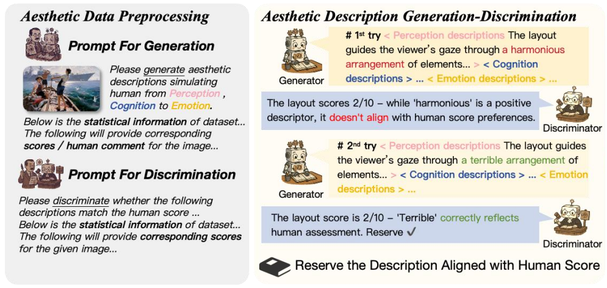
APDD (Human annotation): The composition is rigorous and complete the shape is exquisite, accurate and realistic, the character landscape is delicately portrayed, and the details are diverse. *(1. Overly generalized with insufficient visual evidence 2. Limited to technical analysis)*

Ours: The painting presents a harmonious blend of classicism and landscape, showcasing a serene scene that invites contemplation... The placement of the castle and the figures by the water creates a sense of depth, drawing the viewer into the scene *(Specifically analyzed how the positions of the castle and the characters create depth of field)* ... The warm tones of the sunset contrast beautifully with the cooler hues of the landscape, creating a harmonious balance *(It describes in detail the contrast effect of warm and cool colors)* ... The theme conveys a connection between humanity and nature, resonating well within the classical framework. *(Point out the theme of "Man and Nature" clearly.)*... The painting captures a moment in time, inviting viewers to ponder the beauty and quietness of nature. *(Deeply interpreted the emotional connotation of the painting and the viewer's experience)* ...

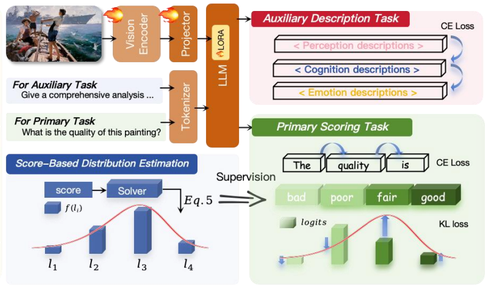
Contributions

- **Data level**: A scalable, hierarchical aesthetic description generation framework is proposed.
- **Method level**: The ArtQuant framework couples isolated aesthetic dimensions through dual-task training (description generation + score prediction). A score-based distribution estimation method optimizes continuous score prediction.
- **Theoretical level**: Information theory proves that description sufficiency and representation quality bound score prediction entropy, establishing a mathematical basis for auxiliary descriptions enhancing aesthetic assessment.

Method



Aesthetic Description Collection



Multi-Task Aesthetic Training (MAT)

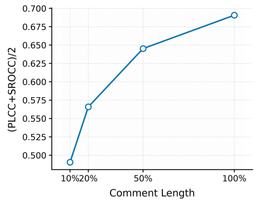
- Hierarchical aesthetic template.
- Correct biases by combining dataset statistical information.
- Discriminative quality control: An iterative framework of generator + discriminator to ensure the descriptions are consistent with scores.

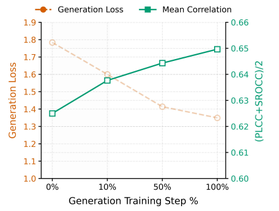
- Jointly optimize the generation and scoring tasks.
- Convert continuous ratings into the expected value of discrete token probabilities by following equation:

$$\mu^*, \sigma^* = \arg\min \left\| \sum f(l_i) l_i - x \right\|_2 \quad \text{s.t.} \sum f(l_i) = 1$$

Analysis

Score prediction entropy $H(Y|Z) \leq H(Y|D) + H(D|Z)$





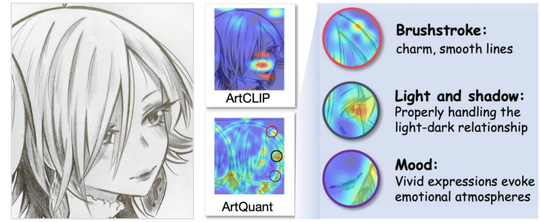
Description sufficiency $H(Y|D)$ **Description generation ability $H(D|Z)$**

Components	SROCC Value	SROCC Δ_{human}	PLCC Value	PLCC Δ_{human}
Human	0.837	–	0.864	–
Perception	0.865	+3.35%	0.888	+2.78%
+ Cognition	0.866	+3.46%	0.890	+3.01%
++ Emotion	**0.871**	**+4.06%**	**0.894**	**+3.47%**

Dataset	w/o MAT	w/ MAT	Gain (%)
APDD	0.863/0.889	0.871/0.894	+0.9%/+0.5%
BAID	0.499/0.580	0.543/0.589	+8.82%/+1.55%
VAPS	0.545/0.634	0.625/0.681	+14.68%/+7.41%

Insight: Description sufficiency and generation ability are both important for quality assessment estimation.

Feature Visualization



After auxiliary description learning, the heatmap and semantic descriptions have better alignment than ArtCLIP.

Henglin Liu 刘恒霖
(liu-hl24@mails.tsinghua.edu.cn)
Seeking **PhD** programs (Fall 2027) and **internships** on multimodal understanding and generation.
Wechat