CPA-IQA: A Contextual Preference-Aligned Framework for AIGC Quality Assessment

Authors: Zhang Hao, Lei Xiang, Zhang Hao

Date: 2026-02-05T22:48:04+00:00

Abstract

With the rapid advancement of AI-Generated Content (AIGC) technologies, evaluating the quality of generated images has become increasingly critical, particularly in specialized domains such as cultural heritage preservation. This paper presents CPA-IQA (Contextual Preference-Aligned Image Quality Assessment), a novel no-reference multi-modal evaluation framework that addresses the limitations of existing quality assessment methods. Unlike traditional approaches that focus solely on low-level perceptual quality or semantic alignment, CPA-IQA integrates hierarchical multi-modal alignment, distortion identification, and scenario-adaptive preference modeling to provide comprehensive quality assessment tailored to specific application contexts. Our experimental validation on a dataset of 192 heritage craft images generated by three state-of-the-art models (Doubao, Kwan 2.6, and JiMengAI 3.0) demonstrates the framework' s effectiveness in distinguishing quality differences across diverse scenarios and identifying AIGC-specific artifacts. The results show that Kwan 2.6 achieves the highest average quality score (43.75), followed by JiMengAI 3.0 (43.37) and Doubao (41.60), with significant variations across different cultural themes and artistic styles.

Full Text

Preamble

CPA-IQA: A Contextual Preference-Aligned Framework for AIGC Quality Assessment Hao Zhanga,b*, Xiang Leia a Chongqing College of Mobile Communication, Chongqing 401420, China b Chongqing Key Laboratory of Public Big Data Security Technology, Chongqing 401420, China Abstract: With the rapid advancement of AI-Generated Content (AIGC) technologies, evaluating the quality of generated images has become increasingly critical, particularly in specialized domains such as cultural heritage preservation. This paper presents CPA-IQA (Contextual Preference-Aligned Image Quality Assessment), a novel

no-reference multi-modal evaluation framework that addresses the limitations of existing quality assessment methods. Unlike traditional approaches that focus solely on low-level perceptual quality or semantic alignment, CPA-IQA integrates hierarchical multi-modal alignment, distortion to provide comprehensive quality assessment tailored to specific application contexts. Our experimental validation on a dataset of 192 heritage craft images generated by three state-of-the-art models (Doubao, Kwan 2.6, and JiMengAI 3.0) demonstrates the framework' s effectiveness in distinguishing quality differences across diverse scenarios and identifying AIGC-specific artifacts. The results show that Kwan 2.6 achieves the highest average quality score (43.75), followed by JiMengAI 3.0 (43.37) and Doubao (41.60), with significant variations across different cultural themes and artistic styles.

Keyword: AIGC Quality Assessment; CPA-IQA; Contextual Preference Alignment; No-Reference Image Quality Assessment; Distortion Identification

Introduction

identification, and scenario-adaptive preference modeling image misalignment, and context-inappropriate stylistic choices.This paper makes the following contributions: proliferation text-to-image generation models has revolutionized content creation across industries[1],[2], from digital art and advertising to cultural heritage preservation and educational materials. However, the quality of AI-generated images varies significantly depending on the model architecture, training data, parameters. More importantly, quality itself is context-dependent: an image suitable for abstract artistic expression may be inadequate for technical documentation or heritage craft representation. generation Traditional Image Quality Assessment (IQA) methods, including both full-reference (FR) and no-reference (NR) approaches, were primarily designed for natural images and focus on low-level distortions such as blur, noise, and compression artifacts. These methods fail to capture AIGC-specific issues such as anatomical inconsistencies, physical law violations, prompt- that evaluates global Hierarchical Multi-modal Alignment three-tier semantic verification (HMA): A system scene-level alignment, local entity-region correspondence, and implicit attribute consistency between text prompts and generated images[6]. Distortion Identification and Localization Network: A specialized module for detecting and quantifying 12 categories of AIGC-specific distortions[7], including anatomical errors, physical violations, rendering artifacts. Scenario-Adaptive Preference Module: A context-aware weighting mechanism that dynamically adjusts quality criteria based on scenarios.

Comprehensive Experimental Validation: Evaluation on 192 heritage craft images across multiple themes (lion dance, embroidery, paper- cutting) and styles (formal, ink-wash), with application detailed distortion analysis and classification. scenario Classical IQA methods can be categorized into full-reference (FR) methods like PSNR and SSIM[3], which require reference images, and no- reference (NR) methods like BRISQUE and NIQE, which assess quality based on

natural scene photographic

methods

inadequately address semantic consistency and context-specific requirements. statistics[4],[5]. While effective images, these Recent work has begun addressing AIGC quality assessment[8],[9]. Methods like CLIP-IQA leverage vision-language models for semantic alignment evaluation, while AGIQA-3K provides a benchmark dataset for AIGC quality assessment. However, these approaches lack adaptation scenario-specific systematically identify AIGC-specific distortion patterns.

Vision-language models such as CLIP, BLIP, and Grounding DINO have demonstrated strong capabilities in image-text matching[10]-[12].

Our framework extends these capabilities by incorporating hierarchical alignment at multiple semantic levels and integrating context-aware preference modeling[13].

Methodology

The CPA-IQA framework is designed as a no-reference multi-modal evaluation system. It and a contextual maps an image-text pair scenario to a scalar quality score shown in Figure 1 [Figure 1: see original paper], its architecture consists of three core modules: Hierarchical Multimodal Alignment, Distortion Recognition Localization Network, and Scene Adaptive Preference Module.
(,) Alignment comprehensive score 100 | Comprehensive score for image-text alignment Distortion Penalty Score 100 | Image distortion penalty score Scene-adaptive preferences 100 | Scene-adaptive preference weights Final quality score 100 | Normalized score S_{align} P_{dist} Q_final Figure 1: The workflow diagram of the CPA-IQA framework To ensure semantic fidelity across scales, we define the alignment objective as a weighted sum of global and local correspondences.

We compute the global cosine similarity between the image embedding the text embedding a pre-trained CLIP-based encoder: = ( ) extracted from Given a set of = entities | || | via Grounding DINO, { } =1 extracted = ( ) corresponding visual regions by SAM, the local alignment is: {ri}i=1 segmented C. Implicit Attribute Verification = ( ), ( ) For non-explicit descriptor, we utilize a Large Language Model (LLM) to decompose into a set of latent attributes lightweight MLP-based estimates the probability classifier = { 1, ···, } = log
III. Distortion Identification & Localization (DILN) To quantify AIGC-specific artifacts, we employ a multi-label classification head and a spatial activation mapping.

The distortion penalty is calculated = ω represents the predicted -th distortion type, and [0, 1] where severity of the its impact factor.

IV. Scenario-Adaptive Preference Module (SAPM) This module dynamically modulates the final score based on the application context.

Contextual Weighting The scenario classifier outputs a context softmax vector . This vector via a determines the dynamic weights ( ) Scenario-Adaptive Weighting Module (SAWM): α, β, γ where is a pre-calibrated preference human-centric $= [\alpha, \beta, \gamma] =$ matrix derived psychological experiments.

B. Preference Ranking The module is optimized using a pairwise ranking loss on human preference pairs for the same prompt (1, 2) where is preferred over $= \max 0,1 -$ (1) $-$ (2) otherwise. $= 1$ V. Final Quality Scoring $= -1$ The integrated quality score synthesized through a scenario-modulated fusion of the aforementioned metrics: $= \alpha$ $+$ $+$ where is the low-level fidelity scor and all terms are normalized to the range $+ \beta$ $- \gamma$ The comprehensive evaluation across 192 [0,100] heritage distinct images performance characteristics, as shown in table 1 : reveals craft Table 1: Comparison of Different Models Avg Quality Std Dev Min Max Median Model Kwan 2.6 Jimeng 3.0 Doubao Key Findings:

Kwan 2.6 leads overall: Achieves highest balanced (43.75) with average quality performance across scenarios JiMengAI 3.0 shows consistency: Lowest standard deviation (7.57) indicates more reliable output quality, though slightly lower average than Kwan Doubao exhibits high variance: Largest standard deviation (9.22) and lowest average (41.60), but achieves the highest single image inconsistent but Breaking down quality scores by artistic (56.24), suggesting score occasionally excellent results Quality range: All models operate in the 29- 56 range, indicating moderate to good quality but room for improvement Analyzing the distortion patterns across all 192 images reveals model-specific weaknesses:

Most Common Distortions (averaged across all models):

Physical Violation: 0.96 average severity - nearly universal issue Anatomical Distortion: 0.78 average severity - particularly problematic for character elements Color Overflow: 0.19 average severity Text Rendering Error: 0.18 average severity average Perspective Distortion: 0.18 severity Model-Specific Patterns:

Doubao: Higher texture repetition and edge blur, particularly in ink-wash style images incidence of Kwan 2.6: Better handling of anatomical light-shadow occasional accuracy contradictions JiMengAI 3.0: Most consistent across distortion types, with balanced performance Scenario Distribution scenario classification reveals interesting patterns in how models handle different context types, as shown in table 2 :

Table 2:Average Scenario Probabilities: Average Scenario Realistic Photography Heritage Craft Character Design Product Design Technical Diagram This distribution Probabilities: 31.2% 28.4% 18.7% 12.5% the cultural heritage focus of our dataset, with most images classified as realistic photography or heritage

craft contexts. reflects Quality by Theme and Style style, as shown in table 3 :

Table 3: Comparison of AI models for scoring the image generation effects of Formal Style and Ink-Wash Style AI model Kwan 2.6 Image Style Effect Score 45.2 average JiMengAI 3.0 44.8 average 42.3 average JiMengAI 3.0 42.1 average 41.9 average 38.7 average Kwan 2.6 Doubao Doubao Formal Style Images Ink-Wash Style Images This suggests that JiMengAI 3.0 handles traditional Chinese artistic styles (ink-wash) more effectively, while Kwan 2.6 excels at formal, realistic representations.

This 27-point gap illustrates the significant variability in Doubao' s output quality and highlights the importance of comprehensive quality assessment beyond simple averaging.

Discussion

The CPA-IQA framework successfully addresses key limitations of existing AIGC quality assessment methods:

Context Sensitivity: The scenario-adaptive module enables appropriate evaluation criteria for different use cases, as evidenced by the varying performance across heritage craft vs. general purpose scenarios.

Comprehensive Distortion Detection: The 12-category distortion captures AIGC-specific metrics miss, particularly anatomical and physical violations. traditional taxonomy issues Multi-level Alignment: The hierarchical approach to semantic verification (global, local, implicit) provides nuanced evaluation beyond simple CLIP similarity scores. evaluation reveals important characteristics of current AIGC models:

Physical consistency remains challenging: All models struggle with physical law violations (0.96 average severity), suggesting that current generation methods inadequately model real- world physics and spatial relationships. modeling, the framework addresses critical gaps in existing AIGC evaluation methods.

Style-dependent performance: Models show differential strengths across artistic styles, with JiMengAI 3.0 excelling at traditional Chinese aesthetics while Kwan 2.6 leads in formal realistic rendering.

Consistency vs. peak performance trade-off: Doubao' s high variance indicates a model lacking capable of excellent consistent quality control results but Several investigation: limitations warrant further Dataset Scope: Our evaluation focuses on cultural heritage themes; broader domain coverage would strengthen generalizability claims.

Preference Calibration: The preference matrix M_{pref} currently derives from limited expert evaluations; larger-scale human studies could improve scenario weighting accuracy.

Computational Efficiency: The multi- requires significant component architecture computational resources; model distillation or efficient architecture search could enable practical deployment.

Temporal video Consistency: generation, extending the framework to assess temporal coherence and motion quality necessary. quality Interactive Refinement: Integrating framework into iterative generation workflows guides model feedback where improvements important represents application direction.

Conclusion

contextual framework comprehensive preference-aligned quality assessment of AI- generated images. By integrating hierarchical multi-modal alignment, systematic distortion identification, and scenario-adaptive preference presents CPA-IQA, paper As AIGC technologies continue advancing, robust quality assessment frameworks like CPA- IQA become increasingly essential for ensuring generated content meets user expectations and domain-specific requirements. Future work will expand the framework to additional domains, improve computational efficiency, and explore integration with iterative generation workflows. [1] Ramesh A, Pavlov M, Goh G, et al. Zero- shot text-to-image generation[C]//Internati- onalconference on machine learning. Pmlr, 2021: 8821-8831. [2] Rombach R, Blattmann A, Lorenz D, et al.

High-resolution image synthesis with latent diffusion models[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2022: 10684-10695. [3] Wang A C, Bovik H R. Sheikh, and EP assessment:

Simoncelli, "Image From error visibility to structural similari- ty," [J]. IEEE Trans. Image Process, 2004, 13(4): 600-612. quality [4] Mittal A, Moorthy A K, Bovik A C. No- reference image quality assessment in the spatial domain[J]. IEEE Transactions on image processing, 2012, 21(12): 4695-4708. [5] Mittal A, Soundararajan R, Bovik A C.

Making a "completely blind" image quality analyzer[J]. IEEE Signal processing letters, 2012, 20(3): 209-212. [6] Ren T, Jiang Q, Liu S, et al. Grounding dino 1.5: Advance the" edge" of open-set object detection[J]. arXiv preprint arXiv:2405. 10300, 2024. [7] Ryu J. Learning Gene Regulation With Cross-Modal Integration of Observations and Perturbations[D]. Harvard University, [8] Li C, Zhang Z, Wu H, et al. Agiqa-3k: An open database for ai-generated image quality assessment[J]. IEEE Transactions on Circuits and Systems for Video Technology, 2023, 34(8): 6833-6846. [9] Wang J, Chan K C K, Loy C C. Exploring clip for assessing the look and feel of images[C]//Proceedings of the AAAI conference on artificial intelligence. 2023, 37(2): 2555-2563. [10] Radford A, Kim J W, Hallacy C, et al.

Learning transferable visual models from natural language superviseon[C]//International conference on machine learning. PmLR, 2021: 8748-8763. [11] Li J, Li D, Xiong C, et al. Blip: Bootstrap- language-image pre-training unified

vision-language understanding and generation[C]//International conference on machine learning. PMLR, 2022: 12888- [12] Liu S, Zeng Z, Ren T, et al. Grounding dino:

Marrying dino with grounded pre-training for open-set object detection[C]//European conference on computer vision. Cham:

Springer Nature Switzerland, 2024: 38-55. [13] Lee J Y, Cha B, Kim J, et al. Aligning text to image in diffusion models is easier than you think[J]. arXiv preprint arXiv:2503. 08250, 2025.

Note: Figure translations are in progress. See original paper for figures.