HEIE: MLLM-Based Hierarchical Explainable AIGC Image Implausibility Evaluator
Supplementary Material

Fan Yang[1,2,5*] Ru Zhen[3*] Jianing Wang[4] Yanhao Zhang[3†] Haoxiang Chen[3]
Haonan Lu[3] Sicheng Zhao[1,2 ‡] Guiguang Ding[1,2 ‡]
[1]Tsinghua University [2]BNRist [3]OPPO AI Center
[4]Peking University [5]Hangzhou Zhuoxi Institute of Brain and Intelligence
yfthu@outlook.com, {zhenru1,zhangyanhao,luhaonan}@oppo.com,
schzhao@gmail.com, dinggg@tsinghua.edu.cn

Contents

1. Experimental Environment

Environment	Details
Operating System	Ubuntu 22.04.4 LTS (Linux 3.10.0-957.27.2.el7.x86_64)
CPU	Intel(R) Xeon(R) Platinum 8362 CPU @ 2.80GHz
GPU	4 × NVIDIA A100-SXM4-80GB
CUDA	Version 12.2
Python	Version 3.9.19
PyTorch	Version 2.0.1
Torchvision	Version 0.15.2
Transformers	Version 4.37.2
DeepSpeed	Version 0.13.5
Opencv-Python	Version 4.10.0.84

Table 1. Details of the experiment environment.

Our experimental setup is summarized in Table 1.

*Equal contribution. † Project lead. ‡ Corresponding authors.

2. More Related Works

2.1. Early Research

Early research utilized neural networks to model scoring systems. In evaluating the authenticity of AIGC images, [1] employed visual question answering (VQA) to assess image-text consistency. Subsequent studies refined and scientifically decomposed evaluation dimensions. [6] released an open dataset annotated in aspects like alignment, fidelity, and harmlessness, providing a foundation for subsequent preference model training. [5] and [2] fine-tuned CLIP models using human-annotated datasets to fit the probability distribution of user preferences based on image-text similarity.

2.2. Evaluation with MLLMs

With the development of large-scale visual understanding models, their superior semantic understanding capabilities have been widely applied in evaluating generative quality and recognized as reliable evaluation metrics. To address the challenge of aligning automated evaluations with human perception, [3] adeptly programmed GPT-4o to circumvent the limitations of direct comparisons, avoiding sequence bias and time-consuming paired assessments. The carefully designed prompt framework not only clarified the core evaluation tasks but also refined the scoring criteria, defined quantifiable scoring intervals, and standardized the format. This guided GPT-4o to accurately capture individualized image nuances across dimensions such as shape, color, texture, and biological features. Q-benchmark [4] systematically demonstrated the potential of MLLMs in three areas: low-level vision perception, low-level vision description, and overall visual quality assessment.

3. More Visualization Results

Additional visual analyses are provided in the supplementary material to fully demonstrate the performance of our HEIE.

3.1. More Visualizations of Implausibility Heatmap Predictions

As illustrated in Figure 1, we provide additional visualizations of implausibility heatmap predictions. Our model demonstrates highly accurate predictions for implausibility in AIGC images. Notably, for AIGC images without evident implausibility, the model does not produce false positive results.

3.2. Additional Heatmap, Score, and Explanation Evaluation Results of HEIE

As shown in Figure 2 and Figure 3, we present further model output results. Our model outputs not only heatmaps and scores but also textual explanations, including Image Description, Problematic Regions Identification, and Analysis of Issues. These outputs leverage the CoT to mutually enhance each other, enabling a comprehensive assessment of implausibility in AIGC images.

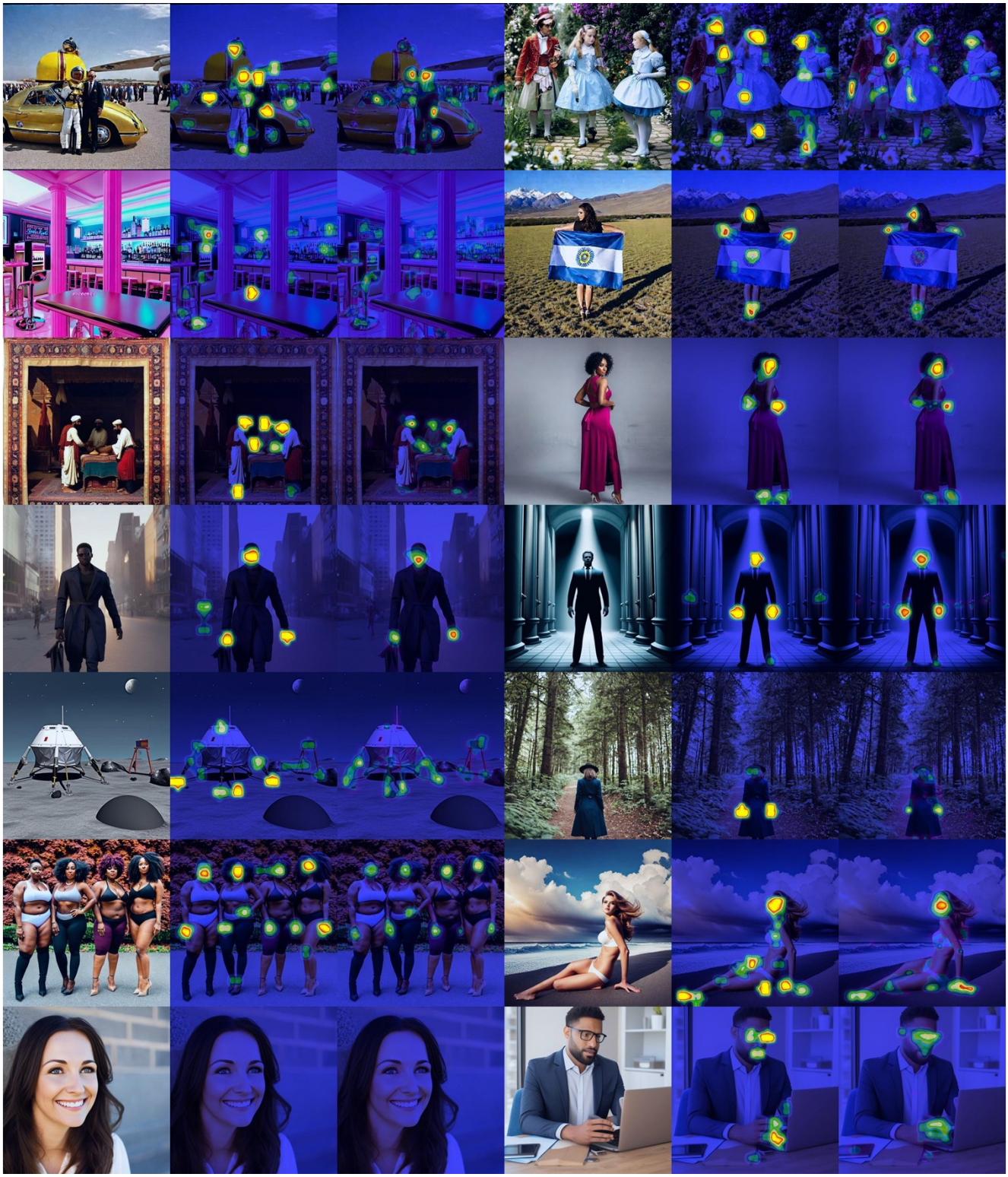

Figure 1. **Heatmap output of HEIE.** Each set of images, from left to right, includes: (a) Input Image, (b) Ground Truth, (c) Output of HEIE. Note that for the last AIGC image on the left of the bottom row, which has no evident defects, our model avoids false positives.

Image Description:
A man in a black suit and hat stands in a surreal landscape with large mushrooms, in black and white. He holds a camera, suggesting he might be a photographer or explorer.

Problematic Regions Identification:
Issues with the man's face and hat, the camera, and his feet.

Analysis of Issues:
Face and Hat: Distorted, lack clear definition.
Camera: Poorly defined, unrealistic shape, lacks depth.
Feet: Disproportionately small, lack definition, abrupt transition from legs.

Our score: 0.5298
Gt score: 0.5

Image Description:
A digital recreation of the Mona Lisa, featuring a woman with long dark hair in a Renaissance dress, with an enigmatic smile in a three-quarter view, against a muted, hazy landscape.

Problematic Regions Identification:
Issues in the right side of the subject's hair and the left hand.

Analysis of Issues:
Hair: Unnatural texture break and abrupt change.
Hand: Distorted, elongated fingers with unclear joints.

Our score: 0.7701
Gt score: 0.75

Image Description:
An anime-style portrait of a young woman with short, light purple hair against a solid purple background, wearing a black headband and a white top with a dark vest or jacket. She has large, expressive eyes and a serious or concerned expression.

Problematic Regions Identification:
Issues located on the headband and near the woman's right eye (from the viewer's perspective).

Analysis of Issues:
Headband: Inconsistent rendering, unrealistic structure, abrupt end or direction change.
Right Eye: Anatomical inaccuracy, misaligned position and angle, inconsistent shading and detailing.

Our score: 0.9348
Gt score: 0.91796875

Image Description:
A woman in a black latex or leather outfit poses confidently in an urban setting with concrete columns and a smooth floor.

Problematic Regions Identification:
Issues with the woman's face, upper right torso, and right hip and thigh area.

Analysis of Issues:
Face: Unnaturally smooth, lacks realistic texture.
Upper Right Torso: Inconsistent sheen, abrupt transition between skin and clothing.
Right Hip and Thigh: Unnatural warping, inaccurate contours and reflections.

Our score: 0.5328
Gt score: 0.5

Image Description:
A cat in a full white astronaut suit, including a helmet showing its face, stands on the rocky, uneven surface of the moon. The background features a starry sky, a bright sun or star, and the Earth in the distance.

Problematic Regions Identification:
Issues with the cat's face, the NASA logo on the chest, the American flag on the arm, the hands, the leg joints, and the boots of the spacesuit.

Analysis of Issues:
Cat's Face: Not properly integrated with the helmet.
NASA Logo: Blurry and lacks detail.
American Flag: Poorly defined, incorrect proportions.
Hands: Disproportionately small, lack gloved appearance.
Leg Joints: Unrealistic bending, lack proper articulation.
Boots: Oversimplified, lack detail and proper proportions.

Our score: 0.58984375
Gt score: 0.58203125

Image Description:
A close-up portrait of a man with long, wavy brown hair and a full beard, wearing a colorful plaid shirt over a white t-shirt. He has a warm, friendly expression with a slight smile, with a background featuring palm trees. The image appears painted or digitally rendered.

Problematic Regions Identification:
Issues around the man's eyes, near his left ear, and on his shirt.

Analysis of Issues:
Man's Eyes: Inconsistent rendering, smooth skin, lacks texture.
Near Left Ear: Hair blends unnaturally with skin, unclear boundary.
Shirt: Distorted plaid pattern, issues with shading and fabric contours.

Our score: 0.9397
Gt score: 0.91796875

Image Description:
A tabby cat is intently looking at a large pepperoni pizza on a black plate on a wooden surface, with a red surface in the background indicating a domestic setting.

Problematic Regions Identification:
Issues with the cat's nose and mouth area, the top of the cat's head, part of the pizza crust, and the top right corner of the image.

Analysis of Issues:
Cat's Nose: Oddly shaped, lacking typical structure.
Cat's Mouth: Blurred, lacking definition.
Top of Cat's Head: Unnatural color transition, red patch.
Pizza Crust: Distorted, irregular bumps and indentations.
Top Right Corner: Distorted unidentified object.

Our score: 0.8514
Gt score: 0.83203125

Figure 2. **Model outputs of HEIE.** HEIE not only predicts implausibility heatmaps but also provides image descriptions, problematic regions identification, analysis of issues, and score, achieving reliable and explainable implausibility evaluation.

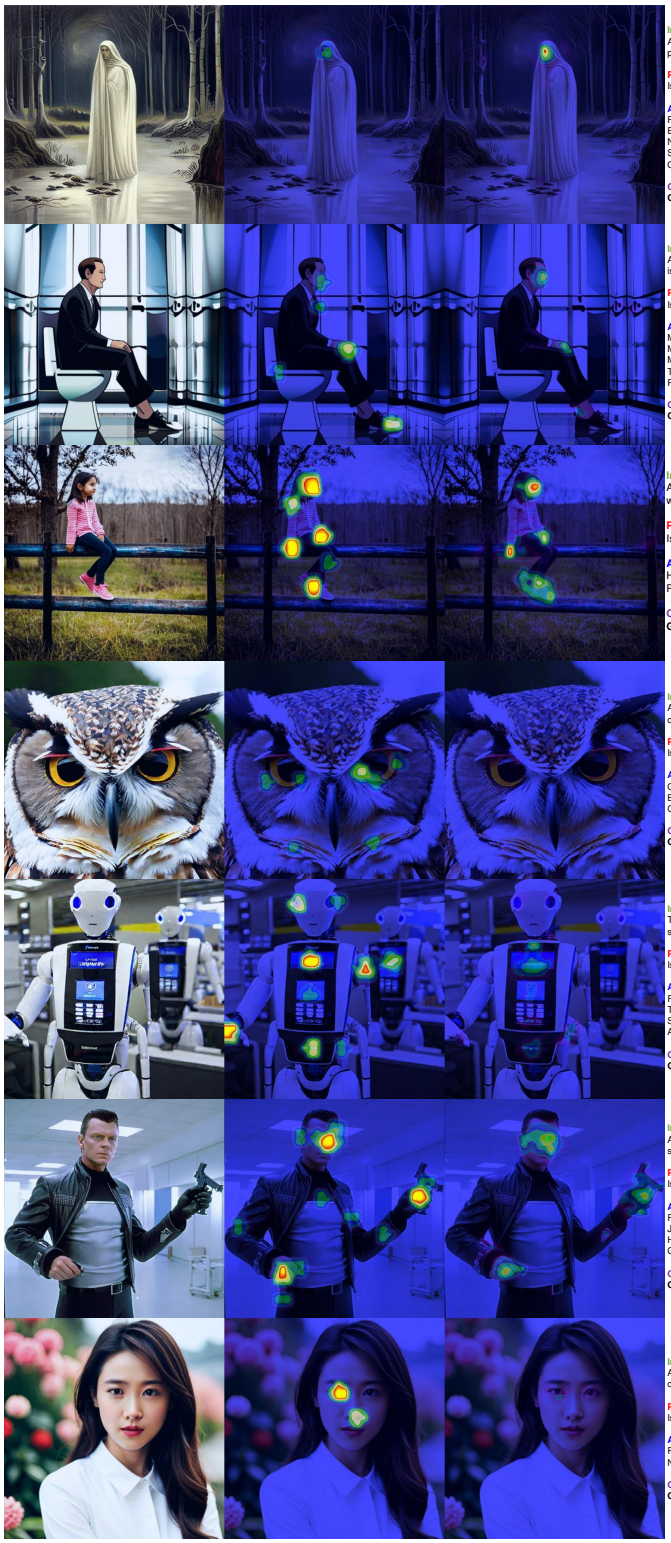

Image Description:
A mysterious, cloaked figure stands in a dark forest, wearing a long white cloak. Bare trees form a tunnel-like scene with fallen leaves on the ground and a puddle reflecting the figure. The mood is somber and ominous.

Problematic Regions Identification:
Issue with the face of the cloaked figure.

Analysis of Issues:
Face: Unnatural, disproportionate features, poor shading and detail.
Eyes: Too large, misaligned.
Nose and Mouth: Poorly defined.
Skin Texture: Artificial, mask-like appearance.
Overall: Inconsistency in style and quality with the rest of the image.

Our score: 0.8875
Gt score: 0.91796875

Image Description:
A stylized illustration of a man in a black suit sitting on a modern toilet or chair in a futuristic bathroom with large windows or glass panels. The color scheme is blue, white, and black.

Problematic Regions Identification:
Issues on the man's face, hands, legs, feet, and the toilet or chair he is sitting on.

Analysis of Issues:
Man's Face: Distorted and pixelated.
Man's Hands: Poorly defined, lack anatomical structure.
Man's Legs and Feet: Abrupt leg ends, poorly formed feet.
Toilet/Chair: Structural inconsistencies, seat disconnected from base.
Pixelated, low-resolution quality in problematic areas.

Our score: 0.6976
Gt score: 0.75

Image Description:
An anime-style portrait of a young woman with short, light purple hair against a solid purple background, wearing a black headband and a white top with a dark vest or jacket. She has large, expressive eyes and a serious or concerned expression.

Problematic Regions Identification:
Issues located on the headband and near the woman's right eye (from the viewer's perspective).

Analysis of Issues:
Headband: Inconsistent rendering, unrealistic structure, abrupt end or direction change.
Right Eye: Anatomical inaccuracy, misaligned position and angle, inconsistent shading and detailing.

Our score: 0.8945
Gt score: 0.91796875

Image Description:
A close-up portrait of an owl with large yellow eyes, intricate feather patterns in brown, white, and black, a prominent facial disk, and a dark hooked beak. The owl is looking directly at the camera with an intense expression.

Problematic Regions Identification:
Issues with the owl's eyes, particularly the irises, and the area below the beak.

Analysis of Issues:
Owl's Eyes: Unnatural pixelation and geometric shapes in the irises.
Below Beak: Misplaced yellow shapes not anatomically correct.
General: AI generation errors with feature misplacement, lack of realistic detail rendering.

Our score: 0.7953
Gt score: 0.83203125

Image Description:
Two humanoid robots in an indoor setting, possibly a laboratory or warehouse. Foreground robot has a white body, large blue eyes, and a black torso with a screen; background robot is partially visible. The setting has fluorescent lighting and an industrial feel.

Problematic Regions Identification:
Issues with the robot's eyes, torso screen, side of torso, arm, and waist area.

Analysis of Issues:
Robot's Eyes: Unnaturally large and perfectly circular, lacking realistic detail.
Torso Screen: Displays illegible, inconsistent text.
Side of Torso/Waist: Inconsistent construction, poor alignment.
Arm: Unnatural bending near the joint, unrealistic structure.

Our score: 0.5198
Gt score: 0.5

Image Description:
A person in a sterile, white office or laboratory environment, wearing a black leather jacket over a light-colored shirt, holding a gun in their right hand with a stern expression, and standing in a confrontational posture.

Problematic Regions Identification:
Issues on the person's face, jacket's chest area, both hands, and the gun.

Analysis of Issues:
Face: Inconsistent skin texture and facial feature proportions, uncanny valley effect.
Jacket: Lacks realistic detail and depth, flat and artificial appearance.
Hands: Unnatural proportions and positioning, improper detail in fingers, anatomical structure issues.
Gun: Poorly defined, blurred edges, unrealistic shape.

Our score: 0.5337
Gt score: 0.5

Image Description:
A young woman with long dark hair in a white collared shirt is in front of a backdrop of pink flowers. She has a serene expression, looking directly at the camera, with a soft, dreamy background.

Problematic Regions Identification:
Issues around the woman's right eye and the bridge of her nose.

Analysis of Issues:
Right Eye: Unnatural pixelation, jagged edges.
Nose Bridge: Flattened appearance, improper shading and depth.

Our score: 0.7680
Gt score: 0.75

Figure 3. **Model outputs of HEIE.** HEIE not only predicts implausibility heatmaps but also provides image descriptions, problematic regions identification, analysis of issues, and score, achieving reliable and explainable implausibility evaluation.

References

[1] Yushi Hu, Benlin Liu, Jungo Kasai, Yizhong Wang, Mari Ostendorf, Ranjay Krishna, and Noah A Smith. Tifa: Accurate and interpretable text-to-image faithfulness evaluation with question answering. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 20406–20417, 2023. 1

[2] Yuval Kirstain, Adam Polyak, Uriel Singer, Shahbuland Matiana, Joe Penna, and Omer Levy. Pick-a-pic: An open dataset of user preferences for text-to-image generation. *Advances in Neural Information Processing Systems*, 36:36652–36663, 2023. 1

[3] Yuang Peng, Yuxin Cui, Haomiao Tang, Zekun Qi, Runpei Dong, Jing Bai, Chunrui Han, Zheng Ge, Xiangyu Zhang, and Shu-Tao Xia. Dreambench++: A human-aligned benchmark for personalized image generation. *arXiv preprint arXiv:2406.16855*, 2024. 1

[4] Haoning Wu, Zicheng Zhang, Erli Zhang, Chaofeng Chen, Liang Liao, Annan Wang, Chunyi Li, Wenxiu Sun, Qiong Yan, Guangtao Zhai, et al. Q-bench: A benchmark for general-purpose foundation models on low-level vision. *arXiv preprint arXiv:2309.14181*, 2023. 1

[5] Xiaoshi Wu, Keqiang Sun, Feng Zhu, Rui Zhao, and Hongsheng Li. Human preference score: Better aligning text-to-image models with human preference. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2096–2105, 2023. 1

[6] Jiazheng Xu, Xiao Liu, Yuchen Wu, Yuxuan Tong, Qinkai Li, Ming Ding, Jie Tang, and Yuxiao Dong. Imagereward: Learning and evaluating human preferences for text-to-image generation. *Advances in Neural Information Processing Systems*, 36, 2024. 1