HEIE: MLLM-Based Hierarchical Explainable AIGC Image Implausibility Evaluator

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

Abstract

*AIGC images are prevalent across various fields, yet they frequently suffer from quality issues like artifacts and unnatural textures. Specialized models aim to predict defect region heatmaps but face two primary challenges: (1) lack of explainability, failing to provide reasons and analyses for subtle defects, and (2) inability to leverage common sense and logical reasoning, leading to poor generalization. Multimodal large language models (MLLMs) promise better comprehension and reasoning but face their own challenges: (1) difficulty in fine-grained defect localization due to the limitations in capturing tiny details, and (2) constraints in providing pixel-wise outputs necessary for precise heatmap generation. To address these challenges, we propose HEIE: a novel MLLM-Based **H**ierarchical **E**xplainable Image **I**mplausibility **E**valuator. We introduce the CoT-Driven Explainable Trinity Evaluator, which integrates heatmaps, scores, and explanation outputs, using CoT to decompose complex tasks into subtasks of increasing difficulty and enhance interpretability. Our Adaptive Hierarchical Implausibility Mapper synergizes low-level image features with high-level mapper tokens from LLMs, enabling precise local-to-global hierarchical heatmap predictions through an uncertainty-based adaptive token approach. Moreover, we propose a new dataset: Expl-AIGI-Eval, designed to facilitate interpretable implausibility evaluation of AIGC images. Our method demonstrates state-of-the-art performance through extensive experiments. Our project is at https://yfthu.github.io/HEIE/.*

1. Introduction

Recently, Artificial Intelligence Generated Content (AIGC) has rapidly advanced, leading to the widespread use of

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

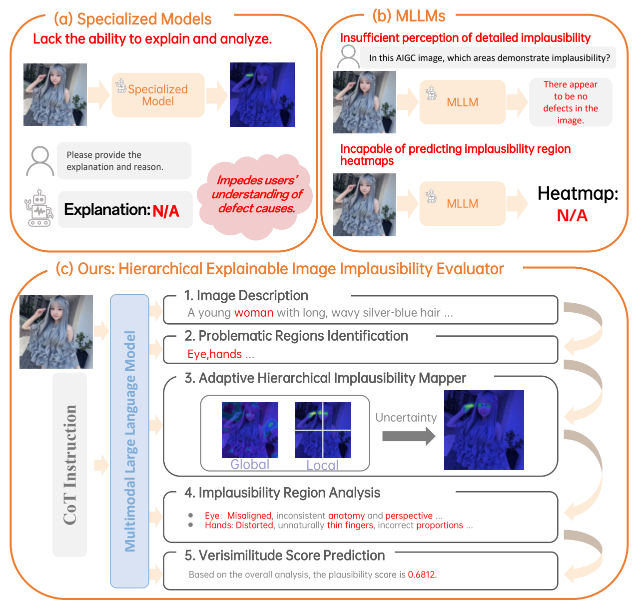

Figure 1. (a) Specialized models lack the ability to explain and analyze subtle implausibility regions, hindering understanding for general users. (b) MLLMs struggle with precise localization of local defects and cannot directly output pixel-level implausibility areas. (c) Our CoT-Driven Explainable Trinity Evaluator can generate heatmaps, analyses, and scores. In our Adaptive Hierarchical Implausibility Mapper, local and global heatmaps are predicted separately, improving the localization of tiny implausibilities.

AIGC images across various fields. However, they often face quality issues such as artifacts, unnatural textures, and problematic regions. Accordingly, many studies focus on the evaluation of AIGC images [22, 47, 48]. Current evaluations primarily yield scalar scores, simple 0-1 values, failing to identify specific defect regions, which hinders targeted improvements in generative models.

A few months ago, RichHF [26] introduced the task of predicting defect region heatmaps and the RAHF model, which locates specific problematic areas. However, specialized models like RAHF have several limitations in the implausibility evaluation task:

(1) Lack of explainability. As shown in Fig. 1 (a), implausibility regions in AIGC images can be complex, such as twisted fingers or overly large eye corners. Specialized small models fail to provide detailed explanations for these subtle defects, making it difficult for general users to understand and optimize the AIGC models.

(2) Lack of general knowledge. These small specialized models are trained on limited datasets and lack comprehensive world knowledge, common sense, and logical reasoning skills. As a result, they struggle to generalize to out-of-domain samples during practical usage.

One approach to tackle the above issues is to use MLLMs [4, 14, 27]. MLLMs have rapidly advanced through large training data and excel in understanding, generating, and possessing extensive knowledge and common. However, directly applying vanilla MLLM to image implausibility evaluation remains challenging:

(1) Fine-Grained Defect Localization: While MLLMs excel in capturing global image structure, they struggle with pinpointing fine-grained defects. Tiny defect areas like the corners of eyes or eyebrows (Fig. 1 (b)) are often neglected by strong MLLMs such as GPT-4o.

(2) Pixel-Wise Output Limitations: General-purpose MLLMs handle various modalities (text, images, videos, etc.) but usually output text only, lacking the capability to produce pixel-level defect heatmaps (Fig. 1 (b)).

To overcome the issues of specialized small models and MLLMs in image implausibility evaluation, we propose HEIE: an MLLM-based **H**ierarchical **E**xplainable image **I**mplausibility **E**valuator, which overcomes various problems as follows:

Firstly, to address the issues of insufficient interpretability, common sense, and logical reasoning in previous methods, we propose the CoT-Driven Explainable Trinity Evaluator. By decomposing the complex task into a sequential task structure, progressing from simple to difficult, our Chain of Thought (CoT) framework harnesses the full potential of LLMs to conduct step-by-step reasoning for intricate implausibility assessments. Our trinity evaluator not only generates scores and heatmaps but also incorporates the analysis and explanation of the implausibility. Crucially, through the integration of CoT [23, 41, 54] and structural design, our heatmaps, scores, and explanations are synergistic and mutually reinforcing.

Secondly, to overcome limitations in precise implausibility localization and pixel-level heatmap generation, we introduce the Adaptive Hierarchical Implausibility Mapper. Our Implausibility Mapper, based on two-way cross-attention, combines low-level image features from ViT [1] with high-level special map tokens from LLMs to predict implausibility heatmaps. AIGC image defects manifest as either global anomalies, such as extra limbs on an animal, or local irregularities, like issues with individual fingers. So, we propose an uncertainty-driven adaptive algorithm that integrates both global and local tokens. Furthermore, the adaptive map token dynamically adjusts according to image resolution and aspect ratio, thereby ensuring both flexibility and optimal performance.

Lastly, current image implausibility datasets do not meet our explainable evaluation needs. We propose a new dataset: Expl-AIGI-Eval (Explainable AIGC Image Implausibility Evaluation Dataset). Utilizing visual prompting, LLM free-form outputting, and in-context learning-based formatting, this dataset provides interpretable implausibility evaluations, including image descriptions, problematic region identifications, and issue analyses, aiding CoT-based reasoning for LLMs.

Our contributions are as follows:

1. We are the first to achieve explainable image implausibility evaluation. Our CoT-Driven Explainable System guides the LLM to decompose complex problems into subproblems of increasing difficulty and mutually reinforcing, improving heatmaps, scores, and explanations.
2. We are the first to use MLLM for image implausibility heatmap prediction, leveraging MLLM's commonsense, logical reasoning, and generalization capabilities.
3. We propose the Adaptive Hierarchical Implausibility Mapper, employing dynamic special map tokens for uncertainty-based global-local heatmap predictions, enhancing the localization of subtle defects.
4. We construct and publish a high-quality explainable AIGC image implausibility evaluation dataset. Our method achieves state-of-the-art performance on various datasets and tasks. Extensive comparisons and ablation studies validate the effectiveness of our approach.

2. Related Work

2.1. AI-Generated Images

Advancements in text-to-image generation models have significantly impacted creative industries like art, design, and advertising. Stable Diffusion [32]uses denoising processes to produce high-quality images. It is open-source, encouraging academic and industrial applications. DALL·E 3 [5] refines the generation of high-resolution, realistic images with detailed text integration. MidJourney [29], by an independent lab, focuses on creating artistic and imaginative images, prioritizing aesthetics over realism. Recent works [25] explore combining diffusion models with GANs, showing

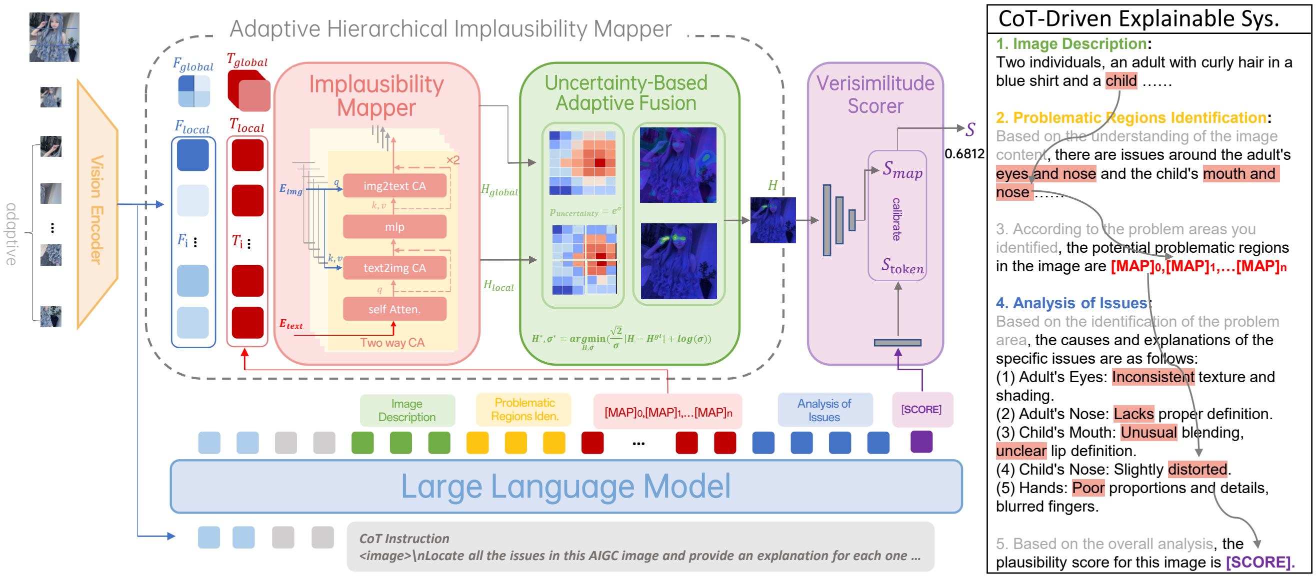

Figure 2. The Implausibility Mapper processes a dynamic number of special **[MAP]** tokens and the image features to enhance detailed implausibility localization. We implement the Adaptive Hierarchical Implausibility Mapper through the local and global heatmaps and the uncertainty-based adaptive fusion. In the verisimilitude scorer, features from the heatmap and the special **[SCORE]** token are integrated for score prediction. Furthermore, our Cot-Driven Explainable System guides the LLM to decompose complex issues into progressive subproblems, facilitating the mutual enhancement of heatmap, analysis, and score prediction, thus improving explainability.

potential for enhancing image quality and diversity. However, AIGC images often contain many implausible elements that require careful evaluation.

2.2. AIGC Image Evaluation

Effective evaluation are required to optimize generative models. Many studies have released datasets to emulate human assessment [17, 22, 24, 47, 48]. Large-scale visual understanding models are now commonly used for reliable evaluation due to their strong semantic capabilities [30, 46, 53]. Prior research has shown that MLLMs are proficient in both low-level visual evaluations [42, 44] and high-level semantic comprehension [52, 55]. Some studies [20, 36, 43, 45] activate the visual evaluation capabilities of multimodal MLLMs through data fine-tuning.

Existing models primarily aim to align their scores with human preferences [43, 47] or aesthetic evaluation [18, 19]. These approaches fail to identify specific fine-grained problem areas within images, which hinders users and researchers from making targeted improvements.

Recently, HumanRefiner [13] and RichHF [26] have introduced tasks focused on detecting fine-grained problematic regions in images. HumanRefiner targets defects in AIGC-generated human images, while RichHF encompasses defects in various objects. However, these approaches rely on specialized small models to detect defect regions. These models are limited in common sense and knowledge, resulting in insufficient generalization capabilities in complex scenarios. Additionally, specialized small models lack interpretability and cannot provide detailed reasons or explanations [2, 12] for the detected issues.

3. Method

3.1. Overview

As shown in Fig. 2, HEIE first employs an Adaptive Hierarchical Implausibility Mapper to generate an image implausibility region heatmap. Subsequently, a verisimilitude scorer is used to predict the image score. Furthermore, with the aid of a Co-driven Explainable System, we stimulated the potential of the LLM to achieve mutual enhancement and joint output of the heatmap, score, and explanation.

3.2. Adaptive Hierarchical Implausibility Mapper

3.2.1. MLLM-Based Implausibility Mapper

To obtain fine-grained, precise image implausibility regions, pixel-level mask outputs are necessary. However, most MLLMs only provide text information and cannot produce heatmap predictions. To address this, we design the Image Implausibility Mapper module for heatmap prediction. We introduced a special map token **[MAP]** to extract implausibility region information from LLM. In LLM, the **[MAP]** token generates many feature layers, and its final linear layer feature is used to predict the next token for LLM. We select this final LLM layer's hidden state of the **[MAP]** token as the features T. Then we use image features F from the image encoder and **[MAP]** token features T from the LLM for a two-way cross-attention calculation [1, 21, 37], as shown in Eq. 1.

$$
\begin{aligned}
C_{\mathrm{T}\to\mathrm{F}} &= \mathrm{CrossAtten}(\mathrm{SelfAtten}(T), F, F) \\
T' &= \mathrm{mlp}(\mathrm{norm}(C_{\mathrm{T}\to\mathrm{F}})) \\
C_{\mathrm{F}\to\mathrm{T}} &= \mathrm{CrossAtten}(F, T', T') \\
O_{2way\_CA} &= \mathrm{norm}(C_{\mathrm{F}\to\mathrm{T}})
\end{aligned}
\tag{1}
$$

To fuse the **[MAP]** token features T with image features F, each input alternates as the query and value during cross-attention operations. By stacking the above two-way cross-attention Eq. (1) twice, we get the predicted heatmap H.

Data imbalance poses a challenge in heatmap prediction, with mean squared error (MSE) loss often leading to conservative and low-value predictions. And we simply adopt focal loss to overcome the data imbalance issue [33].

3.2.2. Hierarchical Implausibility Mapper

Implausibility detection is not the primary objective of mainstream general-purpose MLLMs. Defect localization often requires detailed information, including small defects like fingers and eyes, making it crucial to detect these nuances. We developed a method targeting both global and local defects by segmenting images into adaptive patches according to resolution [8]. The image encoder processes the thumbnail and N segmented patches, generating global features F_g and local features F_i. The LLM then outputs N local information-enhanced tokens, $[\mathbf{MAP}]_\mathbf{i}$, for each patch. Tokens T_i combine with local features F_i and input into the Implausibility Mapper (see Sec. 3.2.1), where cross-attention mechanisms generate local implausibility heatmaps L_i. These heatmaps are concatenated to form a complete heatmap:

$$
H_l = F_{concat}(L_1, L_2, ..., L_N) \tag{2}
$$

To handle large implausibility regions and preserve semantics, we predict a global map token T_g. We utilize global features F_g and the LLM token T_g as inputs to the Implausibility Mapper to generate the heatmap H_g.

3.2.3. Adaptive Implausibility Mapper

AIGC images vary in resolution, aspect ratio, and content, necessitating adaptive strategies for handling implausibility.
Adaptive special map token: We propose an adaptive special map token to manage local information for AIGC images. Our hierarchical design adjusts the local token count based on image scale: more tokens for higher resolution, fewer for lower.
Adaptive Heatmap Fuser: Images can have both global and local defects. Global defects, like a person with three legs, require global features for implausibility prediction, while local errors, like incorrect hand or eye details, need local features.

We introduce an uncertainty-based local-global fusion strategy. We model Local Heatmap H_l and Global Heatmap H_g as Laplace distributions[1] [28], with standard deviations σ_l and σ_g. During training, we minimize the terms H and σ, where H represents either H_l or H_g, and σ represents σ_l or σ_g, respectively:

$$
H^*, \sigma^* = \arg\min_{H,\sigma} \left(\frac{\sqrt{2}}{\sigma} |H - H^{gt}| + \log(\sigma) \right) \tag{3}
$$

During inference, we calculate confidence with Laplace-derived uncertainty and perform weighted summation using $p_{uncertainty} = e^{-\sigma}$ to obtain the final heatmap.

3.3. CoT-Driven Explainable Trinity Evaluator

3.3.1. Verisimilitude Scorer

Our model generates a heatmap of implausible regions and an overall verisimilitude score for comprehensive evaluation. LLMs struggle with numerical outputs, so we propose the Verisimilitude Scorer method to regress the score.

We define a special score token **[SCORE]** within the LLM, capturing the image's verisimilitude. The hidden state of the **[SCORE]** token is processed by a Feed-Forward Network (FFN) to regress an initial score, S_{token}.

Our experiments (Tab. 6) show a strong correlation between the heatmap and verisimilitude score. So we employ several convolution and FFN layers to get the heatmap score S_{map} from the predicted heatmap. The final score S is derived by calibrating S_{token} and S_{map}: $S = Calib(S_{token}, S_{map})$. In our implementation, we explored various calibration functions, such as weighted summation and dynamic fusion. They exhibited similar performance.

3.3.2. CoT-Driven Explainable System

In addition to generating the implausibility region heatmap and verisimilitude score, it is essential to analyze and explain these outputs for AIGC system improvement. We introduce the CoT-Driven Explainable Trinity Evaluator, depicted in Fig. 2. The LLM is guided through a Chain of Thought (CoT) prompting process in five steps:

(1) Image Description: The LLM describes the image, capturing key elements in natural language.

(2) Problematic Regions Identification: Based on the description and the understanding of the image, the LLM identifies potential problematic areas for further analysis.

(3) Special map token: Based on the preceding Problematic Regions Identification, the LLM injects defect re-

[1] The Laplace distribution is $f_X(x) = \frac{1}{2\lambda} \exp\left(-\frac{|x-\mu|}{\lambda}\right)$, where μ and λ are parameters. The standard deviation σ is $\sigma = \sqrt{2}\lambda$.

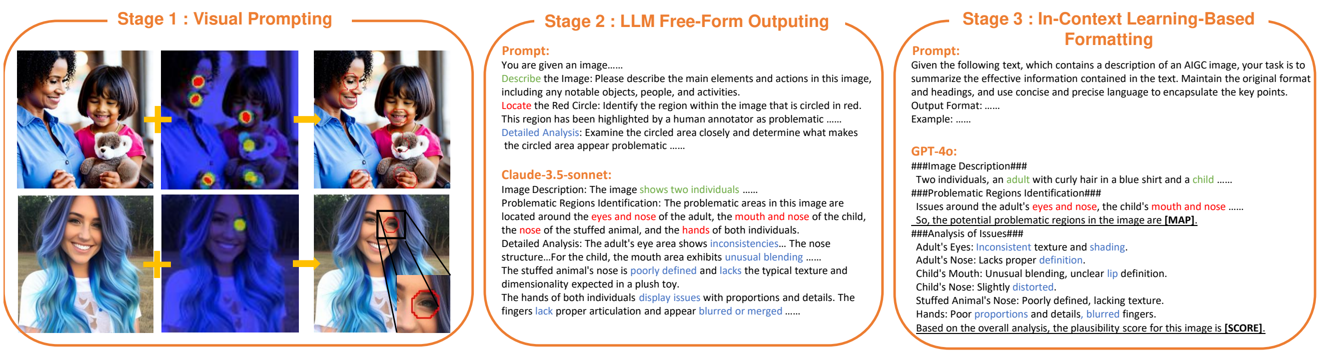

Figure 3. **Three stages in Expl-AIGI-Eval dataset construction**: **Stage 1** - Visual Prompting: Defect regions are circled on images to aid Claude-3.5-sonnet in accurately locating problem areas. **Stage 2** - LLM Free-Form Output: Claude-3.5-sonnet generates free-form defect location and analysis. **Stage 3** - In-Context Learning-Based Formatting: GPT-4o is used for format standardization.

gion information into the special map token for use by the implausibility mapper.

(4) Analysis of Issues: Based on the preceding implausibility localization, the LLM provides detailed text explanations for each region, including type, causes, and so on.

(5) Special Score Token: Based on the comprehensive understanding, the LLM incorporates overall information into a special score token for use by our Verisimilitude Scorer.

This CoT-driven process breaks down the complex evaluation task into five progressively challenging and interrelated tasks, fully leveraging the LLM's potential.

Trinity Implausibility Evaluator. Textual explanations, heatmaps, and scores synergize in image assessment. Texts describe image content, offering semantic context and guiding heatmap generation to highlight defect-prone areas. Heatmaps, using visual saliency, pinpoint critical regions, aiding scores in accurately quantifying defect levels. Conversely, scores refine heatmap focus, enhancing assessment precision. This interaction enables models to better understand and evaluate image issues.

3.4. Expl-AIGI-Eval Dataset

Existing AIGC image implausibility datasets, such as RichHF-18K [26] and AbHuman [13], include AIGC images and annotated defect regions but lack interpretability explanations. To address this, we introduce Expl-AIGI-Eval: Explainable AIGC Image Implausibility Evaluation. We provide detailed analyses and explanations for each implausibility region. The annotation process is conducted in the following three stages, shown in Fig. 3.

(1) Visual Prompting: We follow [7] to add visual prompts. Specifically, the original dataset provides masks of the defective regions for each image. We draw a red contour for each defective region, clearly highlighting these areas, as shown in Stage 1 in Fig. 3.

(2) LLM Free-Form Outputting: Using Claude-3.5-sonnet [3], we generate detailed analyses of the defect regions. Previous study [35] indicates that enforcing fixed-format outputs impairs the performance of LLM, so we allow the LLM to produce free-form outputs for analyzing defect causes.

(3) In-Context Learning-Based Formatting: As shown in Stage-3 (Fig.3), we employ GPT-4o and in-context learning to format the defect analysis texts obtained in Stage 2.

This hierarchical approach ensures comprehensive annotations. Each explanation includes image description, problematic regions identification, and issue analysis, supporting our CoT process in Sec. 3.3.2 and aiding CoT training.

Sensitive images related to pornography, privacy, animal protection, and political topics are filtered out to avoid ethical issues.

4. Experiments

4.1. Experimental Setup

4.1.1. Implementation Details

Our model is based on the InternVL-8B [9]. Our Expl-AIGI-Eval dataset is constructed by expanding upon two existing datasets, RichHF-18K and AbHuman, by adding explanatory and analytical annotations. We follow the original train-validation split. We use Deepspeed to finetune the MLLM, with a learning rate of 3×10^{-4}, a warmup ratio of 0.03, and a batch size of 16.

4.1.2. Evaluation Metrics

We follow the method in [26] for evaluation. We report the MSE for all samples, including images with no problematic region. For images with implausibility regions, we provide evaluation metrics for the saliency heatmaps, such as KLD, AUC-Judd, SIM, and CC, as referenced in [6]. For the score

Figure 4. **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. Note that for the last AIGC image, which has no evident defects, our model avoids false positives.

Figure 5. **Comparison with Baselines.** Each set of images, from left to right, includes: (a) Input Image, (b) Output of InternViT-300M-448px, (c) Output of CLIP-ViT-Base-Patch16, (d) Output of ours HEIE, (e) Ground Truth.

prediction tasks, we report the Pearson linear correlation coefficient (PLCC) and Spearman rank correlation coefficient (SRCC) [6]. Additionally, since the RAHF model [26] is not open source, we provide its results reported in its paper.

4.2. Comparison with State-of-the-art

We evaluate our method on the RichHF-18K benchmark [26], as presented in Tab. 1. Conventional vision models typically lack physical common sense and struggle with high-level semantic understanding, complicating their ability to perceive and recognize content that goes against human intuition. In contrast, our method excels in accurately predicting heatmaps. Given the strong correlation between heatmap accuracy and image plausibility scores, our approach provides quantitative metrics that align with human evaluations. In comparison, our model achieved SOTA performance.

As shown in Tab. 2, we evaluate the performance of image implausibility explanations on our Expl-AIGI-Eval Dataset. We employ three evaluation methods to eval-

	Heatmap Prediction						Plausibility Score	
	MSE (All Data)↓	MSE (GT=0)↓	KLD↓	CC↑	SIM↑	AUC-Judd↑	PLCC↑	SRCC↑
PickScore (off-the-shelf)	-	-	-	-	-	-	0.010	0.028
EVA-CLIP encoder (fine-tuned)	0.01614	0.00512	2.835	0.350	0.082	0.549	0.157	0.143
CLIP encoder (fine-tuned)	0.01437	0.00425	2.462	0.251	0.122	0.747	0.390	0.378
RAHF (multi-head)	0.01216	0.00141	1.971	0.425	0.302	0.877	0.666	0.654
RAHF (augmented prompt)	0.00920	0.00095	1.652	0.556	0.409	0.913	0.693	0.681
HEIE (ours)	**0.00825**	**0.00014**	**1.634**	**0.574**	**0.417**	**0.915**	**0.697**	**0.683**

Table 1. Comparison with State-of-the-Art on the RichHF-18K Dataset.

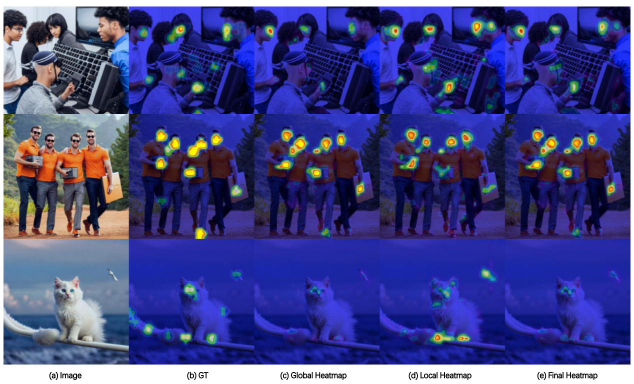

Figure 6. **Results of Hierarchical Implausibility Mapper.** Columns (c) and (d) are the global and local heatmaps from our Hierarchical Implausibility Mapper. Column (e) illustrates the final heatmap after adaptively fusing the global and local heatmaps.

	Perplexity ↓	GPT-4o Eval ↑	Human Eval ↑
Qwen2-VL-7B-Instruct [4]	1.924209	1.910995	1.979058
DeepSeek-VL-7B-chat [27]	1.794179	1.952880	1.883770
InternVL2-8B [9]	1.456884	2.695288	2.603141
GLM-4V-9B [14]	1.320043	2.486911	2.653403
GPT-4o	-	3.828272	3.998953
Claude-3.5-Sonnet	-	3.938220	4.080628
HEIE (ours)	**1.031390**	**4.582199**	**4.352880**

Table 2. Performance of image implausibility explanations on our Expl-AIGI-Eval Dataset.

	All Data	$GT = 0$	$GT > 0$			
	MSE ↓	MSE ↓	KLD↓	CC↑	SIM↑	AUC-Judd↑
InternViT[9]	0.07318	0.07248	3.515	0.019	0.091	0.524
EVA-CLIP[34]	0.00924	0.00207	3.226	0.582	0.095	0.607
CLIP[31]	0.00916	0.00920	1.953	0.244	0.154	0.636
HEIE (ours)	**0.00510**	**0.00076**	**1.629**	**0.684**	**0.423**	**0.938**

Table 3. Results on AbHuman. All models are finetuned.

uate the models' text explanation capability: (1) Compare the perplexity (ppl) of model outputs to ground truth; (2) Employ GPT-4o to compare the content similarity between the model outputs and the ground truth. (3) Conduct human evaluations of the model-generated explanations. Our carefully designed CoT approach significantly enhances our analysis and explanation capabilities for AIGC image implausibility. Our textual analysis outperforms both other open-source MLLMs and surpasses the general-purpose SOTA models like GPT-4o and Claude-3.5-sonnet. Across various textual explanation metrics, our model achieves SOTA performance.

Tab. 3 shows the results on the AbHuman dataset [13], which focuses more on the flaws in AI-generated images of humans. Baseline models underperform due to limited common sense and reasoning skills. In contrast, our model achieves SOTA results.

4.3. Zero-shot Domain Generalization Results

Tab. 4 presents zero-shot cross-domain generalization experiments. We train on either the RichHF-18K or AbHuman datasets and perform zero-shot inference on the other. Cross-domain tasks [16, 51], common in real-world applications [10, 11, 38, 39, 49, 50], test a model's fundamental world knowledge beyond domain-specific fitting [15, 40]. As shown, smaller models suffer significant accuracy drops, losing defect assessment ability. In contrast, our model excels, leveraging MLLMs' strong common-sense knowledge and zero-shot capabilities.

4.4. Ablation Study

Ablation Study for Adaptive Hierarchical Implausibility Mapper. As shown in Tab. 5, Exp. (a) uses only a special global map token and global image features for implausibility prediction, resulting in poor accuracy due to lost local details. Exp. (b) uses only a special local map token, preserving fine-grained image information and improving accuracy over Exp. (a). Exp. (c) combines global and local tokens with fixed weights, outperforming both singular approaches. Exp. (d) employs learnable weights, further improving results over Exp. (c). Exp. (e) uses an uncertainty-based adaptive strategy, dynamically weighting local and global heatmaps and achieving optimal performance.

Ablation for Verisimilitude Scorer. Tab. 6 presents the results: Exp. (a) outputs scores in plain text via LLM, showing low accuracy due to LLM's numerical insensitivity. Exp. (b) uses the LLM's special score token **[SCORE]**

Train: RichHF-18K	All Data	$GT = 0$	$GT > 0$		Train: Abhuman	All Data	$GT = 0$	$GT > 0$	
Test: AbHuman [13]	MSE (All Data)↓	MSE (GT=0)↓	KLD↓	CC↑	Test: RichHF-18K	MSE (All Data)↓	MSE (GT=0)↓	KLD↓	CC↑
InternViT-300MB [9]	0.12610	0.12384	3.263	-0.046	InternViT-300MB	0.25362	0.25170	10.547	-0.017
CLIP [31] encoder	0.01181	0.00492	2.730	0.212	CLIP encoder	0.01510	0.00384	3.448	0.015
EVA-CLIP[34]	0.01170	0.00562	3.225	0.482	EVA-CLIP	0.01432	0.00148	2.835	0.414
HEIE (ours)	**0.00858**	**0.00298**	**1.796**	**0.506**	HEIE (ours)	**0.01091**	**0.00021**	**1.996**	**0.433**

Table 4. Performance of zero-shot domain generalization on the RichHF-18K and AbHuman dataset.

Exps	Global	Local	FW	LW	UncW	MSE↓	KLD↓	CC↑
(a)	✓					0.01071	1.950	0.502
(b)		✓				0.00980	1.921	0.504
(c)	✓	✓	✓			0.00954	1.874	0.511
(d)	✓	✓		✓		0.00873	1.680	0.557
(e)	✓	✓			✓	**0.00825**	**1.634**	**0.574**

Table 5. Ablation study for Adaptive Hierarchical Implausibility Mapper on RichHF-18K dataset.

Exps	Token	Heatmap	GT Heatmap	PLCC↑	SRCC↑
(a)				0.216	0.197
(b)	✓			0.653	0.640
(c)		✓		0.597	0.575
(d)	✓	✓		0.697	0.683
(e)	✓		✓	0.743	0.710

Table 6. Ablation for Verisimilitude Scorer on RichHF-18K dataset.

for score prediction, while Exp. (c) uses only heatmap data. Both improve significantly over plain text. Exp. (d), our Verisimilitude Scorer, combines heatmap and **[SCORE]** token features with a calibration function for superior accuracy, validating its effectiveness. Exp. (e) further validates using ground truth heatmaps and **[SCORE]** token, achieving excellent score predictions, demonstrating the importance of accurate heatmaps.

	Heatmap			Score	
	MSE↓	KLD↓	CC↑	PLCC↑	SRCC↑
w/o CoT Text	0.00913	1.689	0.553	0.669	0.658
w/ CoT Text	0.00825	1.634	0.574	0.697	0.683
w/ GT CoT Text	0.00792	1.601	0.580	0.701	0.685

Table 7. Ablation Study for CoT-Driven Explainable System.

Ablation Study for CoT-Driven Explainable System. As shown in Tab. 7, we evaluate our CoT Instruction. The first experiment without CoT Text (w/o CoT Text) reports results lacking our CoT Instruction, while the second (w/ CoT Text) shows outcomes with it. Comparing these, CoT Prompt significantly enhances heatmap and score predictions, demonstrating its effectiveness in guiding logical reasoning in LLMs. The third experiment (w/ GT CoT Text) uses ground truth CoT text labels, including image descriptions and problematic region identification, as instruction. These further boost prediction accuracy, affirming CoT's value in enhancing LLM reasoning.

4.5. Visualization Analysis

Evaluation Results of HEIE: As shown in Fig. 4, by using CoT, HEIE implements comprehensive and explainable AIGC image implausibility evaluation. Our model does not produce false positive predictions for defect-free images.

Comparison with Baseline: As Fig. 5 illustrates, predicting image implausibility heatmaps is challenging, and baselines perform poorly. Our model, however, shows excellent predictive performance. As well, we cannot obtain RAHF's prediction results due to the unavailability of RAHF's open model.

Results of Hierarchical Implausibility Mapper: As shown in Fig. 6, global heatmaps tend to predict obvious, coarse-grained regions, while local heatmaps focus on subtler, finer details. Their adaptive combination yields the final heatmap.

5. Conclusion

In summary, we introduce the CoT-Driven Explainable Trinity Evaluator, which guides LLM to generate the heatmap, score, and explanation for image implausibility. We propose an Adaptive Hierarchical Implausibility Mapper capable of enhancing the prediction of subtle implausibilities. Additionally, we present the Expl-AIGI-Eval dataset, which offers high-quality, explainable implausibility annotations. Our work facilitates users in understanding the image implausibility and improving the images. In the future, we aim to improve the reasoning efficiency and performance of the MLLM-based framework.

6. Acknowledgment

This research was supported by Zhejiang Provincial Natural Science Foundation of China (No.LDT23F01013F01), National Natural Science Foundation of China (No.62021002, No.62441614), and CCF-DiDi GAIA Collaborative Research Funds.

References

[1] Dosovitskiy Alexey. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv: 2010.11929*, 2020. 2, 4

[2] Plamen P Angelov, Eduardo A Soares, Richard Jiang, Nicholas I Arnold, and Peter M Atkinson. Explainable artificial intelligence: an analytical review. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 11(5):e1424, 2021. 3

[3] Anthropic. Claude 3.5 sonnet: A next-generation language model. Online, 2024. Version 3.5, Accessed: 2025-03-12. 5

[4] Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. Qwen-vl: A frontier large vision-language model with versatile abilities. *arXiv preprint arXiv:2308.12966*, 2023. 2, 7

[5] James Betker, Gabriel Goh, Li Jing, Tim Brooks, Jianfeng Wang, Linjie Li, Long Ouyang, Juntang Zhuang, Joyce Lee, Yufei Guo, et al. Improving image generation with better captions. *Computer Science. https://cdn. openai. com/papers/dall-e-3. pdf*, 2(3):8, 2023. 2

[6] Zoya Bylinskii, Tilke Judd, Aude Oliva, Antonio Torralba, and Frédo Durand. What do different evaluation metrics tell us about saliency models? *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41(3):740–757, 2018. 5, 6

[7] Jun Chen, Han Guo, Kai Yi, Boyang Li, and Mohamed Elhoseiny. Visualgpt: Data-efficient adaptation of pretrained language models for image captioning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18030–18040, 2022. 5

[8] Zhe Chen, Weiyun Wang, Hao Tian, Shenglong Ye, Zhangwei Gao, Erfei Cui, Wenwen Tong, Kongzhi Hu, Jiapeng Luo, Zheng Ma, et al. How far are we to gpt-4v? closing the gap to commercial multimodal models with open-source suites. *arXiv preprint arXiv:2404.16821*, 2024. 4

[9] Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, et al. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 24185–24198, 2024. 5, 7, 8

[10] Xiaohan Ding, Yuchen Guo, Guiguang Ding, and Jungong Han. Acnet: Strengthening the kernel skeletons for powerful cnn via asymmetric convolution blocks. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 1911–1920, 2019. 7

[11] Xiaohan Ding, Xiangyu Zhang, Ningning Ma, Jungong Han, Guiguang Ding, and Jian Sun. Repvgg: Making vgg-style convnets great again. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 13733–13742, 2021. 7

[12] Filip Karlo Došilović, Mario Brčić, and Nikica Hlupić. Explainable artificial intelligence: A survey. In *2018 41st International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO)*, pages 0210–0215. IEEE, 2018. 3

[13] Guian Fang, Wenbiao Yan, Yuanfan Guo, Jianhua Han, Zutao Jiang, Hang Xu, Shengcai Liao, and Xiaodan Liang. Humanrefiner: Benchmarking abnormal human generation and refining with coarse-to-fine pose-reversible guidance. *arXiv preprint arXiv:2407.06937*, 2024. 3, 5, 7, 8

[14] Team GLM, Aohan Zeng, Bin Xu, Bowen Wang, Chenhui Zhang, Da Yin, Dan Zhang, Diego Rojas, Guanyu Feng, Hanlin Zhao, et al. Chatglm: A family of large language models from glm-130b to glm-4 all tools. *arXiv preprint arXiv:2406.12793*, 2024. 2, 7

[15] Tianxiang Hao, Xiaohan Ding, Juexiao Feng, Yuhong Yang, Hui Chen, and Guiguang Ding. Quantized prompt for efficient generalization of vision-language models. *arXiv preprint arXiv:2407.10704*, 2024. 7

[16] Tao He, Leqi Shen, Yuchen Guo, Guiguang Ding, and Zhenhua Guo. Secret: Self-consistent pseudo label refinement for unsupervised domain adaptive person re-identification. In *Proceedings of the AAAI conference on artificial intelligence*, pages 879–887, 2022. 7

[17] 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. 3

[18] Yipo Huang, Xiangfei Sheng, Zhichao Yang, Quan Yuan, Zhichao Duan, Pengfei Chen, Leida Li, Weisi Lin, and Guangming Shi. Aesexpert: Towards multi-modality foundation model for image aesthetics perception. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 5911–5920, 2024. 3

[19] Yipo Huang, Quan Yuan, Xiangfei Sheng, Zhichao Yang, Haoning Wu, Pengfei Chen, Yuzhe Yang, Leida Li, and Weisi Lin. Aesbench: An expert benchmark for multimodal large language models on image aesthetics perception. *arXiv preprint arXiv:2401.08276*, 2024. 3

[20] Zhipeng Huang, Zhizheng Zhang, Yiting Lu, Zheng-Jun Zha, Zhibo Chen, and Baining Guo. Visualcritic: Making lmms perceive visual quality like humans. *arXiv preprint arXiv:2403.12806*, 2024. 3

[21] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4015–4026, 2023. 4

[22] 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

[23] Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. Large language models are zero-shot reasoners. *Advances in Neural Information Processing Systems*, 35:22199–22213, 2022. 2

[24] Chunyi Li, Zicheng Zhang, Haoning Wu, Wei Sun, Xiongkuo Min, Xiaohong Liu, Guangtao Zhai, and Weisi Lin. Agiqa-3k: An open database for ai-generated image

quality assessment. *IEEE Transactions on Circuits and Systems for Video Technology*, 2023. 3

[25] Yanyu Li, Huan Wang, Qing Jin, Ju Hu, Pavlo Chemerys, Yun Fu, Yanzhi Wang, Sergey Tulyakov, and Jian Ren. Snapfusion: Text-to-image diffusion model on mobile devices within two seconds. *Advances in Neural Information Processing Systems*, 36, 2024. 2

[26] Youwei Liang, Junfeng He, Gang Li, Peizhao Li, Arseniy Klimovskiy, Nicholas Carolan, Jiao Sun, Jordi Pont-Tuset, Sarah Young, Feng Yang, et al. Rich human feedback for text-to-image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19401–19411, 2024. 2, 3, 5, 6

[27] Haoyu Lu, Wen Liu, Bo Zhang, Bingxuan Wang, Kai Dong, Bo Liu, Jingxiang Sun, Tongzheng Ren, Zhuoshu Li, Hao Yang, et al. Deepseek-vl: towards real-world vision-language understanding. *arXiv preprint arXiv:2403.05525*, 2024. 2, 7

[28] Yan Lu, Xinzhu Ma, Lei Yang, Tianzhu Zhang, Yating Liu, Qi Chu, Junjie Yan, and Wanli Ouyang. Geometry uncertainty projection network for monocular 3d object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3111–3121, 2021. 4

[29] MidJourney. Midjourney: Ai-based image synthesis tool. https://www.midjourney.com, 2023. Accessed: 2023-10-01. 2

[30] 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. 3

[31] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, pages 8748–8763. PMLR, 2021. 7, 8

[32] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10684–10695, 2022. 2

[33] Tsung-Yi Lin Priya Goyal Ross and Girshick Kaiming He Piotr Dollár. Focal loss for dense object detection. In *Proceedings of the IEEE International Conference on Computer Vision*, page 2980–2988, 2017. 4

[34] Quan Sun, Jinsheng Wang, Qiying Yu, Yufeng Cui, Fan Zhang, Xiaosong Zhang, and Xinlong Wang. Eva-clip-18b: Scaling clip to 18 billion parameters. *arXiv preprint arXiv:2402.04252*, 2024. 7, 8

[35] Zhi Rui Tam, Cheng-Kuang Wu, Yi-Lin Tsai, Chieh-Yen Lin, Hung-yi Lee, and Yun-Nung Chen. Let me speak freely? a study on the impact of format restrictions on performance of large language models. *arXiv preprint arXiv:2408.02442*, 2024. 5

[36] Zhiyu Tan, Xiaomeng Yang, Luozheng Qin, Mengping Yang, Cheng Zhang, and Hao Li. Evalalign: Evaluating text-to-image models through precision alignment of multimodal large models with supervised fine-tuning to human annotations. *arXiv preprint arXiv:2406.16562*, 2024. 3

[37] A Vaswani. Attention is all you need. *Advances in Neural Information Processing Systems*, 2017. 4

[38] Ao Wang, Hui Chen, Zijia Lin, Jungong Han, and Guiguang Ding. Repvit: Revisiting mobile cnn from vit perspective. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15909–15920, 2024. 7

[39] Ao Wang, Hui Chen, Lihao Liu, Kai Chen, Zijia Lin, Jungong Han, et al. Yolov10: Real-time end-to-end object detection. *Advances in Neural Information Processing Systems*, 37:107984–108011, 2024. 7

[40] Zerun Wang, Liuyu Xiang, Fan Yang, Jinzhao Qian, Jie Hu, Haidong Huang, Jungong Han, Yuchen Guo, and Guiguang Ding. Lode: Deep local deblurring and a new benchmark. *arXiv preprint arXiv:2109.09149*, 2021. 7

[41] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems*, 35:24824–24837, 2022. 2

[42] 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. 3

[43] Haoning Wu, Zicheng Zhang, Weixia Zhang, Chaofeng Chen, Liang Liao, Chunyi Li, Yixuan Gao, Annan Wang, Erli Zhang, Wenxiu Sun, et al. Q-align: Teaching lmms for visual scoring via discrete text-defined levels. *arXiv preprint arXiv:2312.17090*, 2023. 3

[44] Haoning Wu, Zicheng Zhang, Erli Zhang, Chaofeng Chen, Liang Liao, Annan Wang, Kaixin Xu, Chunyi Li, Jingwen Hou, Guangtao Zhai, et al. Q-instruct: Improving low-level visual abilities for multi-modality foundation models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 25490–25500, 2024. 3

[45] Haoning Wu, Hanwei Zhu, Zicheng Zhang, Erli Zhang, Chaofeng Chen, Liang Liao, Chunyi Li, Annan Wang, Wenxiu Sun, Qiong Yan, et al. Towards open-ended visual quality comparison. In *European Conference on Computer Vision*, pages 360–377. Springer, 2025. 3

[46] Tianhe Wu, Kede Ma, Jie Liang, Yujiu Yang, and Lei Zhang. A comprehensive study of multimodal large language models for image quality assessment. *arXiv preprint arXiv:2403.10854*, 2024. 3

[47] 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, 3

[48] 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, 3

[49] Fan Yang, Xinhao Xu, Hui Chen, Yuchen Guo, Jungong Han, Kai Ni, and Guiguang Ding. Ground plane matters: Picking up ground plane prior in monocular 3d object detection. *arXiv preprint arXiv:2211.01556*, 2022. 7

[50] Fan Yang, Xinhao Xu, Hui Chen, Yuchen Guo, Yuwei He, Kai Ni, and Guiguang Ding. Gpro3d: Deriving 3d bbox from ground plane in monocular 3d object detection. *Neurocomputing*, 562:126894, 2023. 7

[51] Fan Yang, Hui Chen, Yuwei He, Sicheng Zhao, Chenghao Zhang, Kai Ni, and Guiguang Ding. Geometry-guided domain generalization for monocular 3d object detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 6467–6476, 2024. 7

[52] Fan Yang, Sicheng Zhao, Yanhao Zhang, Haoxiang Chen, Hui Chen, Wenbo Tang, Haonan Lu, Pengfei Xu, Zhenyu Yang, Jungong Han, et al. Llmi3d: Empowering llm with 3d perception from a single 2d image. *arXiv preprint arXiv:2408.07422*, 2024. 3

[53] Zhiyuan You, Zheyuan Li, Jinjin Gu, Zhenfei Yin, Tianfan Xue, and Chao Dong. Depicting beyond scores: Advancing image quality assessment through multi-modal language models. *arXiv preprint arXiv:2312.08962*, 2023. 3

[54] Zhuosheng Zhang, Aston Zhang, Mu Li, and Alex Smola. Automatic chain of thought prompting in large language models. *arXiv preprint arXiv:2210.03493*, 2022. 2

[55] Zicheng Zhang, Haoning Wu, Chunyi Li, Yingjie Zhou, Wei Sun, Xiongkuo Min, Zijian Chen, Xiaohong Liu, Weisi Lin, and Guangtao Zhai. A-bench: Are lmms masters at evaluating ai-generated images? *arXiv preprint arXiv:2406.03070*, 2024. 3