Supplementary Material for AGIDefect-4K

Anonymous Author(s)

A CASE STUDIES

Our case studies assess the performance of our defect detection model in evaluating flawless (Figure 1) and defective (Figure 2) images. The model adeptly identifies flawless images, such as diverse paintings and photographs, classifying them as defect-free.

In Fig. 2, which presents examples of defective images, Description_GT denotes the ground truth description provided by human experts, while Description_PR represents the defect description generated by our model. In Fig. 2, each case from left to right is the original image, ground truth, and the predicted mask respectively.

In the analysis of defective images, the model generally succeeds in detecting anomalies and describing certain defects. However, its ability to comprehensively characterize defects is less consistent, resulting in varied prediction accuracy. The model occasionally emphasizes more salient or easily segmentable defects while overlooking others. This is evident in the "hamster" image, where a prominent ear defect was recognized, yet more subtle facial distortions were largely ignored or their severity underestimated.

Figure 1: Examples of flawless data

A more striking example is the "elegant Chinese princess," where significant facial distortions were entirely overlooked. This suggests that the model may struggle with detecting defects that are not immediately obvious or are subtly blended into the image, require a holistic understanding of complex anatomies or artistic styles, or where its internal defect localization affects the scope of its textual description. Consequently, this sometimes leads to an incomplete or slightly misdirected analysis of the full range and impact of the defects present.

B SUBJECTIVE EXPERIMENT SETTINGS

Our subjective experiment incorporates a rigorous three-stage process: **data classification**, **data annotation**, and **double-check validation**.

In order to classify each image, consensus among participants is required. If participants unanimously agree that an image has defects, it is classified as 'Reject', indicating that the image has defects. Conversely, if the image is agreed to be without defects, it is classified as 'Accept'. In cases where there is no unanimous agreement, expert intervention is needed to provide a definitive judgment. The interface screenshot for classification is illustrated in Figure 3, where the 'Accept' button represents an image deemed to have no defects, and the 'Reject' button signifies an image judged to have defects.

In the subsequent data annotation stage (detailed in Fig. 4), participants are required to meticulously examine each image alongside its associated generation prompt. Their primary tasks include identifying any defects, accurately marking these areas using mask annotations, and providing detailed descriptions that cover the defect's location within the image, its classification, and the potential impact on overall image quality.

In the double-check validation phase, we implement a meticulous review system utilizing three independent reviewers who critically assess all participant annotations, including masks and descriptions. Our interface for double check, shown in Figure 5, allows the reviewers to collaborate effectively. Through this interface, they determine which masks are accurate and can be consolidated into a unified representation. Similarly, they discern the correct descriptions to retain. During this stage, additional checks are conducted for cases where there are discrepancies in the masks and descriptions provided by participants. This rigorous process ensures that only precise and consistent annotations are incorporated into the final dataset, thereby enhancing the overall reliability and quality of the data.

C PROMPTS FOR DATASET CONSTRUCTION

After accurately curating defect descriptions, we utilize GPT-4o to consolidate multiple descriptions related to the same image. The prompt given to GPT-4o for this task is detailed below:

#System Prompt: You are an expert in assessing the quality of AI-generated images, specializing in identifying defects. You will receive Chinese descriptions of defects within an image, authored by one or more human experts, alongside the corresponding AI-generated image.Your tasks are:

- ***Integrate Descriptions**: Merge the human experts' descriptions of defects within the image, using the image content for reference.*
- ***Refine Output**: Further refine the consolidated descriptions to produce a polished English output.*

Key Considerations:

- *The final description must be accurate, professional, and faithfully represent the original human expert descriptions, preserving their intended meaning.*
- *If multiple descriptions are provided (e.g., from two experts), synthesize common points while retaining unique and valuable defect information. Ensure conciseness by avoiding redundancy but comprehensively cover all critical defects.*
- *If indicated in the original human expert descriptions, or if a clear inference can be drawn from the defect's severity, briefly*

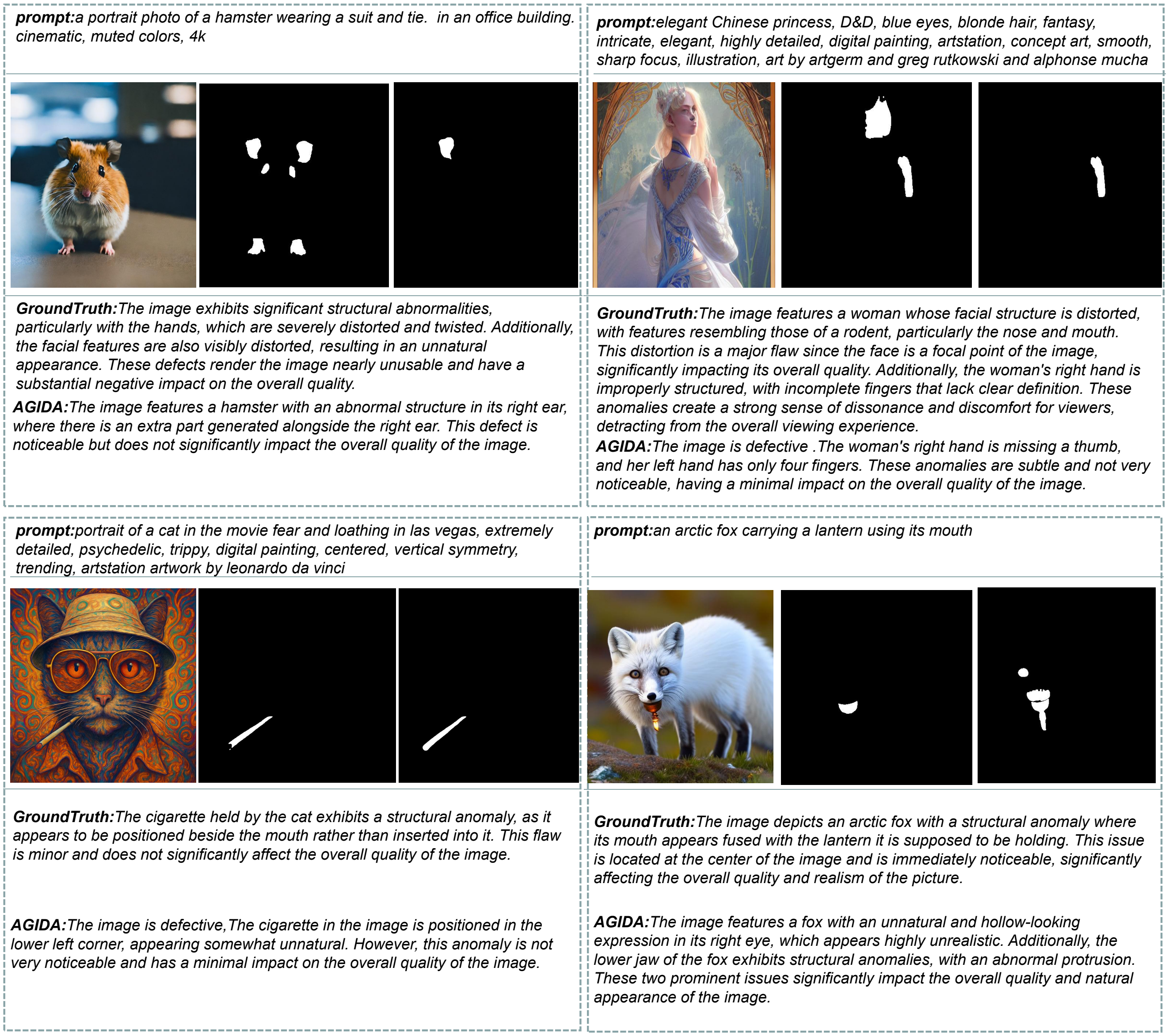

Figure 2: Examples of defective data

articulate the impact of these defects on the image's overall quality, realism, aesthetic appeal, or thematic expression.

Please provide your response strictly in the following JSON format: { "description": "Your defect description..." } Return only this JSON object, without any additional explanations, analyses, or extraneous text.

#User Prompt: Below is an AI-generated image ***[image]*** *and its defect description(s)* ***[description block]*** *provided by human expert(s). Based on your observation of the image and the provided human expert defect description(s), please provide a description.*

D PROMPTS FOR EXPERIMENTS

D.1 Prompts for MLLM Defect Description Generation

To assess the defect description capabilities of multimodal large language models, we developed a framework centered around three critical tasks, aimed at guiding the model in crafting comprehensive defect assessments. Our objective was to prompt the LLM to generate detailed descriptions of defects in AI-generated images. In pursuit of this, we provided the following prompt to instruct the model:

#System Prompt: Your role is to function as an expert in identifying defects within AI-generated images, adhering to a three-step process:

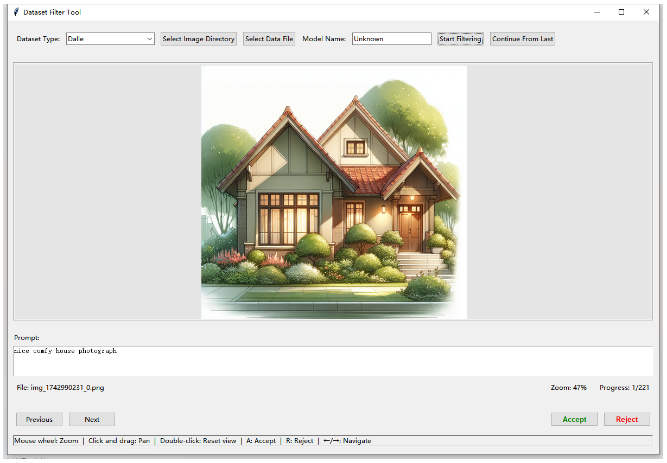

Figure 3: The interface screenshot for classification

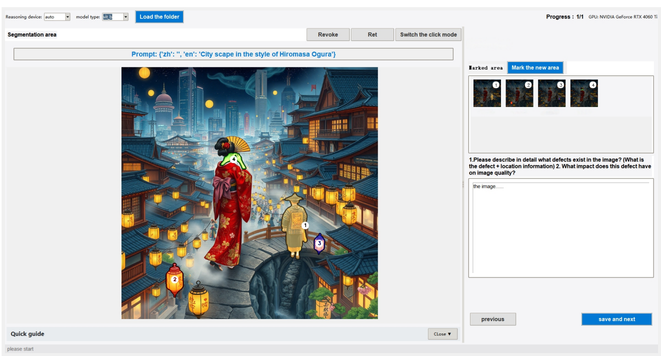

Figure 4: The interface screenshot for annotation

Step 1: Image Inspection

Conduct a thorough examination of the image to identify structural and logical defects ***without taking into account the associated prompt****.*

- ***Structural Defects:*** *Detect any anomalies in objects or body parts that appear physically incorrect, such as being twisted, stretched, or disproportionate in size. Identify anatomical abnormalities like unusual limb counts, missing facial features, or deformities.*
- ***Logical Defects:*** *Search for logical inconsistencies, such as unsupported floating objects or impossible physics (e.g., sideways-flowing water, upside-down trees). Recognize environments or element combinations that contravene common sense or basic physical rules.*

Step 2: Prompt Verification

If defects have been identified in Step 1, review the prompt ***solely at this stage****.*

- *Ascertain whether the prompt explicitly refers to the identified defects, rendering them intentional. Intentional defects described in the prompt should not be classified as defects.*
- *Consider defects valid only if the prompt neither mentions them nor justifies their presence.*
- *Ignore discrepancies between the image and the prompt and missing elements unless they relate to identified structural or logical defects. (e.g., If a prompt specifies "a floating ball" but the ball is not floating, this is primarily a prompt alignment issue. However, if the ball has been distorted or altered in an attempt to make it look like it is floating, then it is considered a structural defect)*

Step 3: Description Generation *Report your assessment in plain JSON format: { "Defective": [true/false], "Description": [] } The "Description" field should be an empty list [] if no defects are observed. If defects are identified, include a single string entry in the "Description" list. This string should be a detailed paragraph outlining all detected defects, their locations, and visual features. Conclude with an explanation of how these defects collectively affect the image's overall quality.*

Note: The description should be a single-paragraph format within the "Description" list, detailing multiple defects and their impact on image quality.

#user: For each image ***[image]*** *analyzed, the MLLM (configured with the system prompt above) then receives the following input:*

Original generation prompt ***[generation prompt]***

Please assess this image ***[image]*** *for defects based on the provided prompt.*

D.2 Prompts for Evaluating LLMs

In this part of the evaluation, we focus on assessing the MLLMs capability to identify and describe defects. Utilizing a GPT-assisted evaluation strategy, we systematically compare LLM-generated descriptions against human-annotated benchmarks, concentrating on two critical criteria: **Completeness** and **Precision**.

The evaluation's goal is to confirm that LLM-generated defect descriptions are both complete and accurate, aligning with expert human interpretations. Below is a structured approach to this evaluation:

#System Prompt: You are an expert in evaluating AI-generated images for detecting visual defects. Your expertise lies in pinpointing two main types of defects:

(1) ***Structural Defects:*** *Violations of basic physical space rules (e.g., misaligned limbs, missing/extra fingers, incomplete key objects).*
(2) ***Semantic Logic Defects:*** *Combinations that defy common sense or domain knowledge and are not intended by the prompt (e.g., floating objects without support, physically impossible arrangements).*

Prompts for Defect Description Completeness: *Evaluate if the description* ***[MLLM DESC]*** *encompasses all visual defects mentioned in the reference description* ***[GOLDEN DESC]****. Assess solely on whether each defect in the reference is at least acknowledged in the output, without evaluating the description's accuracy, detail, or relevance.*

Rating System for Completeness:

- ***Score 2:*** *All or nearly all defects present in the reference are mentioned.*
- ***Score 1:*** *Some, but not all, defects from the reference are mentioned.*
- ***Score 0:*** *None of the defects from the reference are mentioned.*

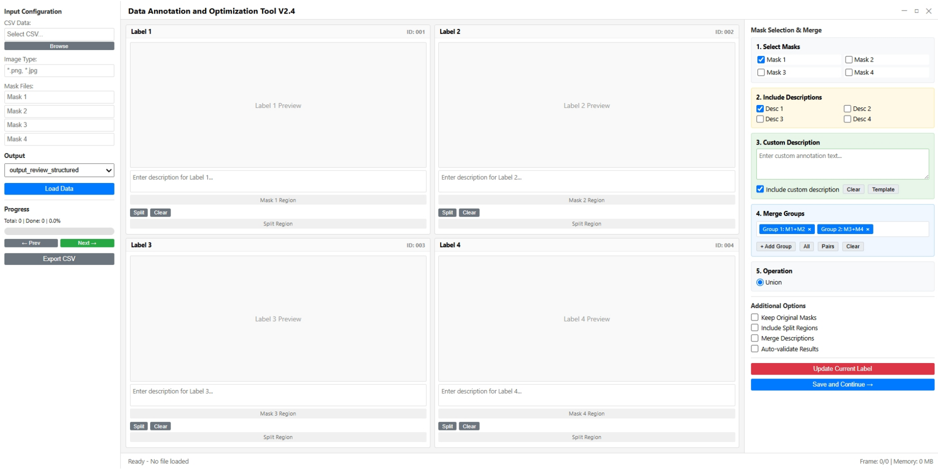

Figure 5: The interface screenshot for double check

Please only provide the result in the following JSON format: {"Score":[]}. Reference Description ***[GOLDEN DESC]*** *MLLM Description* ***[MLLM DESC]****.*

Prompts for Defect Description Precision: The precision metric measures descriptions in terms of defect details (location, type, severity) and penalizes descriptions with controversial or contradictory information compared to the reference.

Rating System for Precision:

- ***Score 2:*** *Minor inconsistencies or less controversial details compared to the reference; generally consistent.*
- ***Score 1:*** *Major inconsistencies or more controversial details than the reference; noticeable contradictions.*
- ***Score 0:*** *Severe contradictions or completely inaccurate details compared to the reference; highly controversial.*

Please only provide the result in the following JSON format: {"Score":[]}. Reference Description ***[GOLDEN DESC]*** *MLLM Description* ***[MLLM DESC]****.*

E LICENSE

We are committed to facilitating maximum research reproducibility and community engagement by fully and publicly releasing all core components of the **AGIDefect-4K** dataset. This includes all images exhibiting defective masks, the prompts used for their generation (where applicable), and the corresponding detailed human-annotated defect descriptions.

All human-rated data (our defect annotations) will be made available in its entirety, presented in a carefully organized and anonymized format. This commitment to open access is intended to maximize transparency and foster widespread academic progress in the fields of AI-generated image analysis, defect detection, and multimodal foundation model evaluation. We hope our endeavour will significantly support and accelerate academic advancements and the development of innovative solutions by providing the community with this comprehensive resource.exploration.