LEGION: Learning to Ground and Explain for Synthetic Image Detection

Supplementary Material

Contents of the Appendices:

A. SynthScars Dataset

A.1. Artifact Definition

Inspired by [32], we categorize artifacts in synthetic images into three types: physics, distortion, and structure. To eliminate subjective differences among annotators and to clarify and standardize the criteria for artifact classification during the annotation process, we established a guideline that explicitly defines the nature and scope of various artifacts, as shown in Table 3.

A.2. Annotation Details

We recruited 12 experienced annotators with high-education backgrounds and a unified guideline was provided along with dedicated training. They were required to strictly follow the guideline and discard samples where artifacts were entirely imperceptible to human eyes. The annotation process for 12,236 samples in SynthScars took ∼12×240 hours and underwent multiple rounds of quality inspection to ensure label consistency and standardization.

A.3. Data Curation

To obtain high-quality, deceptive, and challenging synthetic images, we carry out a multistage filtering process using Qwen2-VL-72B-Instruct [52], which removes low-quality samples (*e.g.*, blurred or compressed artifacts), non-photorealistic content (*e.g.*, cartoonish or watercolor-style images), and samples exhibiting conspicuous synthetic patterns. Specifically, we designed a prompt, as shown in the Table 4, for the model to sequentially inspect each data sample against the given criteria. Only samples that meet all the standards are retained.

Image Content	Human	Object	Animal	Scene	Total
Train	6253	1940	1183	1860	11236
Test	587	162	134	117	1000
Total	6840	2102	1317	1977	12236

Table 1. **Statistics on Image Content.** SynthScars encompasses a diverse range of real-world scenarios, including 12,236 fully synthesized images from different generators.

Artifact Type	Physics	Distortion	Structure	Total
Train	1431	1249	21233	23913
Test	111	136	2406	2653
Total	1542	1385	23639	26566

Table 2. **Statistics on Artifact Types.** SynthScars classifies artifacts into three fine-grained anomaly types, and contains a total of 26,566 artifact instances.

A.4. Dataset Statistics

As shown in Table 1, SynthScars includes 12,236 fully synthesized images across diverse real-world scenarios, with 11,236 training and 1,000 test samples categorized into human, object, animal, and scene. The dataset features 26,566 artifact instances (Table 2), annotated with irregular polygon masks and classified into three types: physics-related (6%), distortion (5%), and structural anomalies (89%).

B. Experimental Details

B.1. Prompt Design

When designing the prompt, in order to fully unleash the LLM's broad reasoning ability, we incorporated prior knowledge of different artifacts (denoted as `<Diverse Artifact Prior>`). Specifically, it consists of common cases from the three types of artifacts we defined, guiding the model to examine the image from the corresponding perspectives. To provide a concrete example, we define it as follows:

> **Physics artifacts** (*e.g., optical display issues, violations of physical laws, and spatial/perspective errors*), **Structure artifacts** (*e.g., deformed objects, asymmetry, or distorted text*), and **Distortion artifacts** (*e.g., color/texture distortion, noise/blur, artistic style errors, and material misrepresentation*)

B.2. Explanation Evaluation

Following Fakeshield [58], we use paraphrase-MiniLM-L6-v2[4] from HuggingFace as our text embedding model to transform the outputs into semantic feature space.

C. Robustness Study

We compare the artifact localization performance between LEGION and PAL4VST (the strongest expert model from

[4] https://huggingface.co/sentence-transformers/paraphrase-MiniLM-L6-v2

Artifact Definition

1. **Physics**
   (a) **Optical Display**: These artifacts arise from inconsistencies in the propagation and reflection of light, violating fundamental optical principles. They can occur across different objects and scenes, leading to unrealistic visual effects. Common cases include incorrect reflections, shadows, and light source positioning errors, causing synthetic images to deviate from real-world optical phenomena.
   (b) **Physical Law Violations**: These artifacts result from the failure to adhere to fundamental physical laws during image synthesis. They typically manifest as illogical scenes, such as water flowing upward or objects floating in mid-air, which contradict natural laws.
   (c) **Space and Perspective**: These artifacts stem from inaccuracies in object proportions and spatial relationships during image generation, leading to inconsistencies with real-world perspective rules. Examples include incorrect depth perception, mismatched object sizes, or spatial distortions that prevent accurate perspective alignment.
2. **Structure**
   (a) **Deformed Objects**: These artifacts arise when the shape or structure of objects is distorted due to errors in the generative model. Contributing factors include geometric inconsistencies, texture mapping errors, and rendering issues.
   (b) **Asymmetrical Objects**: These artifacts occur when an object exhibits unnatural asymmetry, deviating from expected structural balance.
   (c) **Incomplete/Redundant Structures**: These artifacts appear as missing or excessive structural components, leading to unrealistic representations of objects.
   (d) **Illogical Structures**: These artifacts involve the generation of unrecognizable or non-existent objects, as well as the appearance of elements that should not logically exist within the given context.
   (e) **Text Distortion and Illegibility**: These artifacts include warped, irregular, or unrecognizable text, affecting the readability and coherence of textual content within the generated image.
3. **Distortion**
   (a) **Color and Texture**: These artifacts result from errors in color rendering or color space conversion, leading to unnatural hues, inappropriate saturation, or other inconsistencies in color perception.
   (b) **Noise and Blurring**: These artifacts are associated with image noise reduction and clarity enhancement processes. They may arise when algorithms fail to effectively remove noise or introduce excessive blurring, causing local details to appear distorted or unnatural.
   (c) **Artistic Style**: These artifacts occur when synthetic images exhibit unintended stylization, such as cartoonish or painterly appearances that deviate from realistic textures. Such distortions are often caused by errors in style transfer or texture generation algorithms.

Table 3. **Artifact Definition**. We clearly define three types of artifacts and require annotators to strictly follow this guideline for annotation.

Table 2) on SynthScars under three types of distortion. Table 5 reveal that Gaussian noise induces the most severe performance degradation, followed by Gaussian blur, while JPEG compression exhibits the least negative effects. Notably, as intensity increases, LEGION remains stable, while PAL4VST degrades sharply, highlighting our model's superior robustness under strong interference—an unattainable ability for traditional expert models.

D. More Visual Examples

D.1. Localization Comparison

In this section, we provide more visual comparison cases of LEGION on the artifact localization task against other traditional experts (*e.g.*, HiFi-Net, TruFor, PAL4VST), object-grounding VLMs (*e.g.*, Ferret, Griffon), and general MLLMs (*e.g.*, InternVL2), as shown in Figure 1, 2.

It is evident that LEGION achieves the most accurate localization among all models, without failing to recognize artifacts entirely or mistakenly identifying the majority of the image as artifacts.

D.2. Explanation Comparison

In this section, we provide multiple cases to conduct a detailed comparison on the explanation generation task with the latest released open-source (*e.g.*, LLaVA-v1.6, InternVL2, Qwen2-VL, DeepSeek-VL2) and closed-source (*e.g.*, December,2024 updated GPT-4o) MLLMs with varying parameters, as shown in Figure 3.

Notably, LEGION achieves the highest CSS and ROUGE-L scores, indicating the highest alignment with ground truth in describing artifact locations and specific abnormal causes, demonstrating its strong interpretability. In contrast, other models exhibit various issues to some extent.

System Prompt

You are a helpful assistant. Analyze the given images based on the following three criteria and assign one label to each image. You only need to return the label for each image without providing any additional explanations:

Evaluation Criteria

1. **Clarity**
   (a) The image should be well-lit, sharp, and visually clear without blurriness, noise, or distortion.
   (b) The image must not show obvious signs of artificial manipulation, such as pixelated edges or unnatural distortions.
2. **Safety**
   (a) The image must not contain violence, blood, gore, explicit sexual content, hate symbols, discriminatory elements, or any harmful or inappropriate material.
   (b) Any content that could evoke strong negative emotions or discomfort should be classified as unsafe.
3. **Realism**
   (a) The image should look realistic and have a photo-like appearance.
   (b) It must not be cartoonish, animated, or heavily stylized in an artistic manner.

Labeling Task

Assign one of the following labels to each image:

1. **Acceptable**: If the image meets all three criteria;
2. **Rejected[Clarity]**: If the image is unclear, blurry, or distorted;
3. **Rejected[Safety]**: If the image contains unsafe or inappropriate content;
4. **Rejected[Realism]**: If the image is stylized, animated, or lacks realism.

User Prompt

Please strictly follow the instructions to label the input image: {*image*}

Table 4. **Curation Prompt**. Only samples that meet all the standards are retained.

Distortion	PAL4VST		LEGION (Ours)	
	mIoU	F1	mIoU	F1
No Distortion	56.10	29.21	59.41	36.96
JPEG Comp. (QF = 50)	55.95	28.85	57.78	33.97
JPEG Comp. (QF = 35)	55.55	27.60	58.04	34.08
JPEG Comp. (QF = 20)	55.01 (-1.9%)	26.36 (-9.8%)	57.91 (-2.5%)	34.28 (-7.3%)
Gaussian Noise (σ = 0.1)	56.01	28.96	57.31	33.00
Gaussian Noise (σ = 0.2)	54.42	25.16	56.77	32.52
Gaussian Noise (σ = 0.3)	52.91 (-5.7%)	21.11 (-27.7%)	56.49 (-4.9%)	32.12 (-13.1%)
Gaussian Blur (Ksize = 5)	55.62	27.76	57.75	33.78
Gaussian Blur (Ksize = 9)	54.58	25.23	57.27	32.63
Gaussian Blur (Ksize = 15)	53.24 (-5.1%)	22.30 (-23.7%)	57.50 (-3.2%)	33.52 (-9.3%)

Table 5. **Robustness Comparison Under Different Perturbations.** LEGION significantly outperforms the strongest existing expert model under severe JPEG compression (denoted as JPEG Comp.), Gaussian noise, and Gaussian blur (Ksize represents kernel size). Values in parentheses indicate degradation ratios, with the more robust method highlighted in **green**, otherwise in **red**.

For example, DeepSeek-VL2 often falls into meaningless repetition, while GPT-4o tends to provide overly lengthy responses with a large amount of distracting information.

D.3. More Cases of LEGION

In addition to comparing LEGION's predictions with other methods, including multi-modal large language models and expert models, this section provides an extended visualization of artifact segmentation masks and their corresponding explanations. As shown in Figure 4, LEGION excels in predicting artifacts on highly realistic synthetic images, achieving both positional and contour accuracy in segmentation. The accompanying explanations are insightful, highlighting not only the location of the artifact but also offering a plausible rationale for its artificial nature. These results highlight LEGION's ability to deliver precise artifact detection alongside interpretable insights, enhancing the transparency and trustworthiness of synthetic image generation.

E. Limitations and Analysis

While our model demonstrates promising results in detecting and segmenting artifacts on AI-generated images, there remain areas for improvement. A qualitative analysis of failure cases reveals two primary challenges. First, in scenarios with high scene complexity and a multitude of elements, our model sometimes tends to miss subtle artifact regions. As illustrated in Figure 5, the predicted masks may incompletely cover the areas affected by anomalies, particularly when the artifacts are intertwined with intricate background details. Second, the model struggles with detecting very subtle artifacts that occupy a small image area, especially in human portraits. These artifacts, often manifesting as minor distortions or unnatural textures, can be difficult to perceive even for the human eye. We argue that the model is being overwhelmed by the sheer volume of information, leading to a prioritization of more prominent anomalies at the expense of smaller, less conspicuous ones.

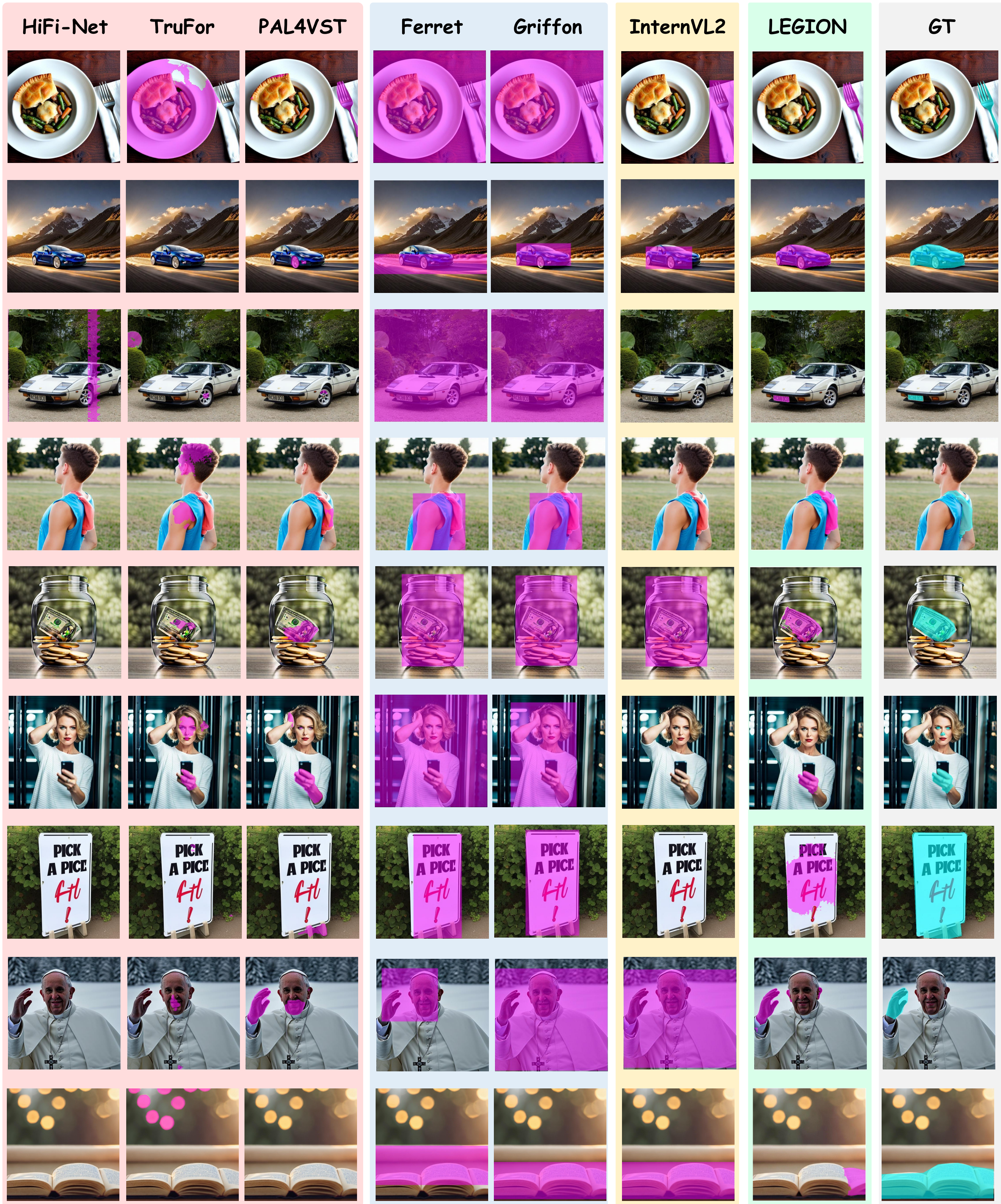

Figure 1. **More Visual Comparison Examples Between Existing Methods and LEGION on the Artifact Localization Task.** The rightmost column shows the ground truth.

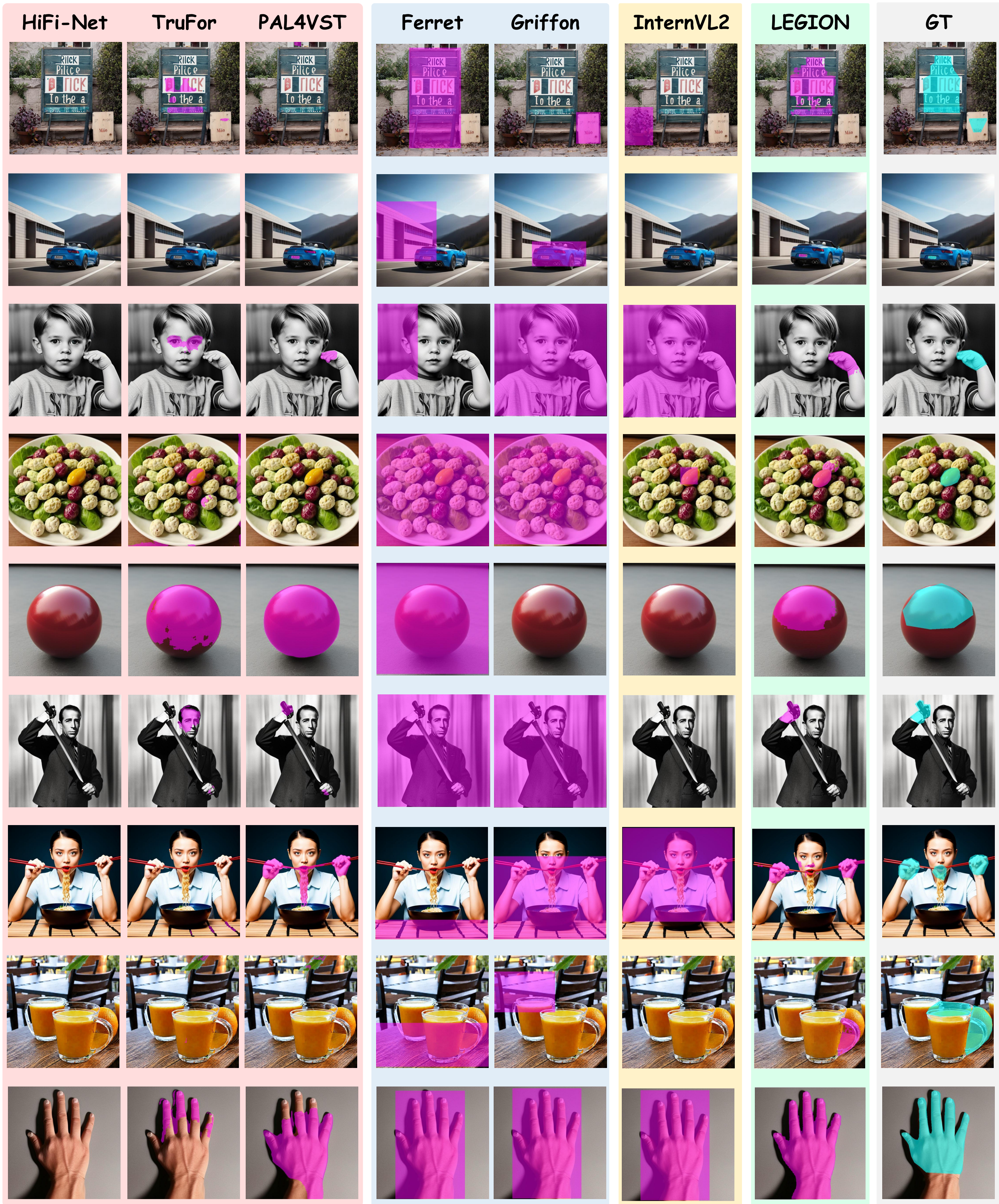

Figure 2. **More Visual Comparison Examples Between Existing Methods and LEGION on the Artifact Localization Task.** The rightmost column shows the ground truth.

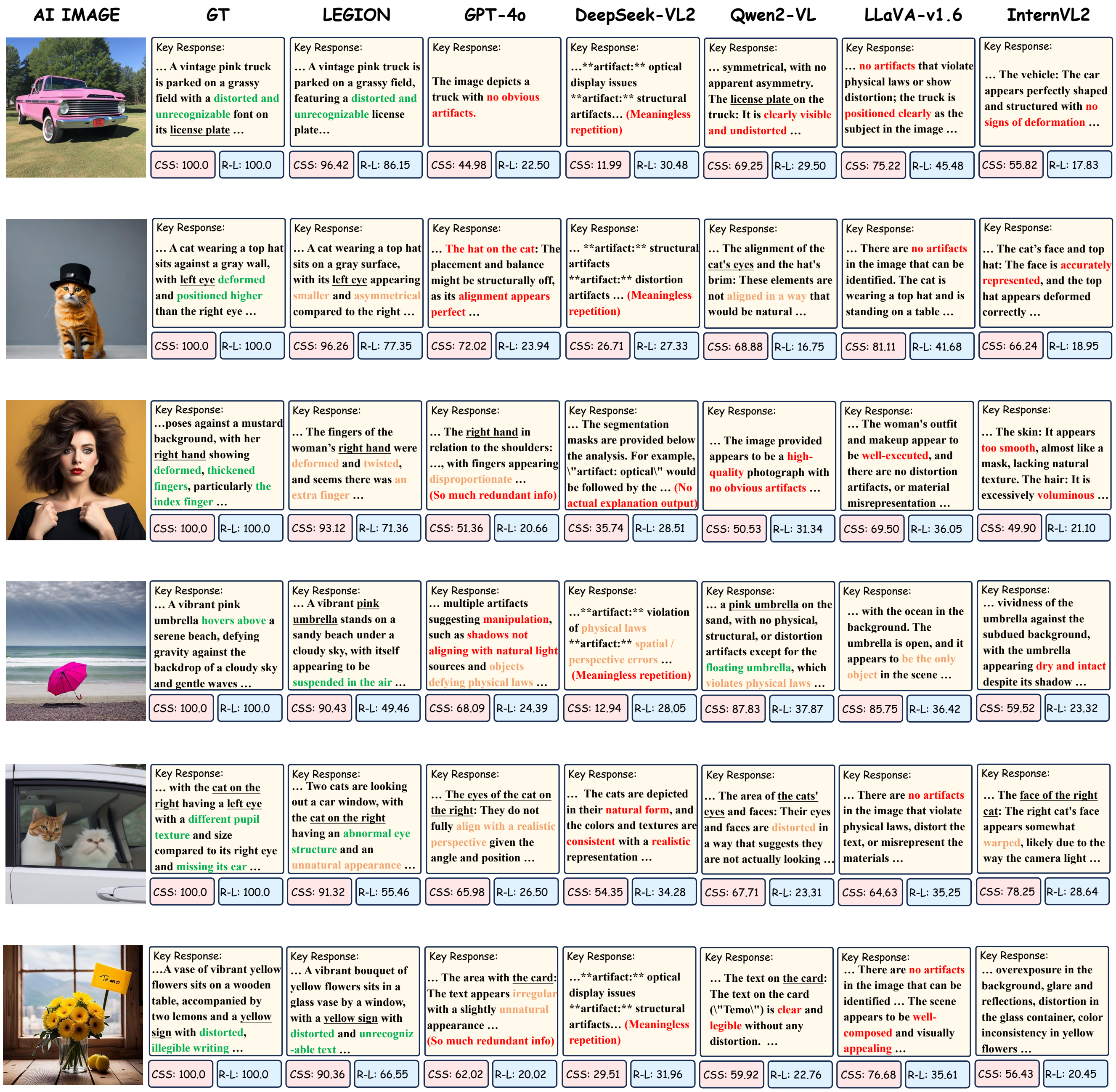

Figure 3. **More Visual Comparison Examples Between Existing MLLMs and LEGION on the Explanation Generation Task.** The second column shows the ground truth. For intuitive comparison, we highlight the model outputs that are semantically identical to the ground truth in **green**, approximately correct in **orange**, and incorrect in **red**, respectively.

Figure 4. **More Visual Cases of Artifact Segmentation Masks and Corresponding Explanations.** The figure illustrates a qualitative comparison between the ground truth (**Top row**) and the corresponding predictions obtained from our proposed model LEGION (**Bottom row**).

Figure 5. **Failures Occur in Complex Scenes and with Intricate Small Artifacts.** The figure illustrates a qualitative comparison between the ground truth (**Top row**) and the corresponding predictions obtained from our proposed model LEGION (**Bottom row**).