

# Rethinking VLMs for Image Forgery Detection and Localization

## Supplementary Material

### A. Appendix / Supplemental Material

#### A.1. Ablation on $\alpha$ for Region-aware Visual Feature Enhancement

We conduct an ablation study to evaluate the effect of  $\alpha$  in Eq. (4):

$$T_{\text{vis}} = \alpha \text{CLIP}(x) + (1 - \alpha) \text{CLIP}(x \odot M),$$

where  $\alpha$  controls the balance between the global semantics (from  $\text{CLIP}(x)$ ) and region-specific forensic cues (from  $\text{CLIP}(x \odot M)$ , where  $M$  is the mask). We evaluate  $\alpha \in \{0, 0.3, 0.5, 0.7, 1.0\}$  and report performance on the SID-Set using the Cosine Semantic Similarity (CSS) score.

The results show that  $\alpha = 0.5$  achieves the best trade-off between global semantics and region-specific forensic cues, balancing the need for general image context with the local detail required for accurate forgery detection. Specifically,  $\alpha = 0.5$  provides the highest CSS score, indicating a better alignment between the model’s predicted explanation and the ground-truth rationale.

Further analysis reveals that values of  $\alpha$  closer to 0 (favoring global semantics) or closer to 1 (favoring localized features) lead to suboptimal performance, as they either lose important contextual understanding or miss crucial localized details. The results are summarized in Table 1, which shows how varying  $\alpha$  influences the model’s overall performance.

Table 1. Ablation study on  $\alpha$  for region-aware visual feature enhancement technique with CSS scores.

$\alpha$ Value	Type	Areas	Tampered Content	Visual Inconsistencies	Summary
0.0	0.79	0.64	0.54	0.59	0.77
0.3	0.83	0.66	0.58	0.65	0.80
0.5	<b>0.87</b>	0.67	<b>0.60</b>	<b>0.70</b>	<b>0.84</b>
0.7	0.80	<b>0.68</b>	0.55	0.62	0.79
1.0	0.75	0.65	0.50	0.60	0.77

#### A.2. Effect of Unfreezing CLIP

While the vision-language alignment in pre-trained VLMs aids in generating language explanations, these models inherently lack forgery-specific priors, such as sensitivity to inconsistent low-level cues. Furthermore, CLIP’s shared image–text feature space is established through large-scale pre-training on up to 400M image–text pairs, acting as a crucial bridge for downstream reasoning tasks such as VQA. Directly fine-tuning the CLIP visual encoder on a relatively small image forgery dataset with localized segmentation losses disrupts this delicate cross-modal alignment. Consequently, this degradation severely impairs the model’s ability to generate language explanations. As shown in Table 2, unfreezing CLIP in the SIDA baseline leads to noticeable drops in both explanation quality (CSS and ROUGE-L) and localization performance (IoU), further validating the necessity of our decoupled architecture.

#### A.3. Attention Module Details

As illustrated in Figure 1, the Attention Module functions as a Cross-Modal Feature Enhancement mechanism, effectively bridging global classification semantics with local visual features. The process begins with the CLIP ViT backbone extracting a global class token and spatial patch tokens from the input image. The class token is processed to generate classification logits, which represent the model’s high-level semantic intent (e.g., the likelihood of tampering). To translate this semantic intent into spatial guidance for segmentation, the classification logits and the patch tokens are linearly projected into a shared 256-dimensional space. The projected logits serve as the Query ( $Q$ ), while the projected patch tokens serve as the Key ( $K$ ) and Value ( $V$ ). Through a Multi-Head Attention layer, the semantic query interacts with the local visual context, dynamically

Table 2. Comparison with SIDA (unfrozen CLIP).

Method	AUC	F1	IoU	CSS	ROUGE-L
SIDA-7B	0.87	0.74	0.44	0.66	0.38
SIDA-13B	0.94	0.72	0.54	0.80	0.41
SIDA-7B-Unfrozen-CLIP	0.94	0.63	0.40	0.65	0.39
SIDA-13B-Unfrozen-CLIP	0.95	0.72	0.50	0.73	0.40
<b>IFDL-VLM (Ours)</b>	<b>0.99</b>	<b>0.87</b>	<b>0.65</b>	<b>0.84</b>	<b>0.43</b>

highlighting regions relevant to the predicted class. The attention output is then added to the projected spatial features via a residual connection. Finally, an average pooling operation aggregates these enhanced features into a single 256-dimensional Global Prompt Embedding. This embedding acts as a substitute for text embeddings and is fed into the SAM Prompt Encoder to guide the mask decoder to focus specifically on the manipulated areas.

#### A.4. Efficiency comparison

We report inference-time efficiency metrics (parameter count, FLOPs, and peak memory usage) and compare with SIDA and FakeShield under the same setting.

Table 3. Efficiency comparison.

Metric	FakeShield	IFDL-VLM	SIDA
Params (B)	22.0	14.3	14.0
FLOPs (T)	6.3	6.2	6.1
Peak Mem. (GB)	27.1	27.5	29.5

#### A.5. Standard Text Similarity Metrics.

In addition to the model-based semantic similarity evaluations (GPT-5 and CSS), we also assess the generated explanations using standard natural language generation metrics, including BLEU-1, ROUGE-L, METEOR, and CIDEr. These metrics emphasize lexical overlap and n-gram matching. As shown in Table 4, our IFDL-VLM consistently outperforms the SIDA-13B baseline across all traditional text similarity metrics. This further demonstrates that our framework not only aligns better with human semantics but also generates text that is lexically more faithful to the ground-truth annotations.

Method	BLEU-1	ROUGE-L	METEOR	CIDEr
SIDA-13B	0.548	0.406	0.272	0.016
IFDL-VLM (Ours)	<b>0.580</b>	<b>0.434</b>	<b>0.290</b>	<b>0.030</b>

Table 4. Quantitative comparison on standard text similarity metrics.

#### A.6. Automated GPT-5 Evaluation Protocol for IFDL Explanation.

To assess the quality of generated language explanations, we employ an automated evaluation protocol using the multimodal large language model (GPT-5). The input to GPT-5 consists of: (i) the tampered image, (ii) the ground-truth (GT) localization mask, (iii) two predicted masks from Model A and Model B, and (iv) their corresponding textual explanations, alongside a human-written reference rationale (GT text).

To ensure the reproducibility of our automated evaluation, the GPT-5 model is applied across all methods with fixed decoding parameters: temperature = 0.7, top-p = 0.95, and maximum output tokens = 8000.

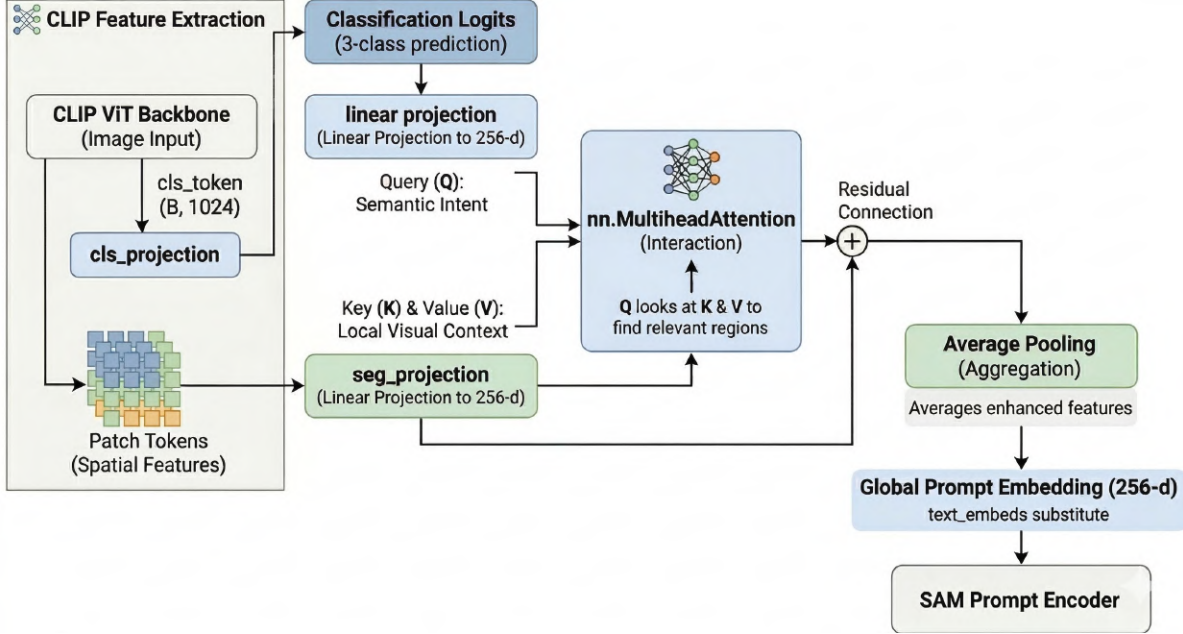


Figure 1. Attention details in the IFDL-VLM framework.

GPT-5 evaluates both models on six criteria, using a 0–5 scale: **mask** (localization fidelity), **type** (forgery type), **areas** (tampered regions), **tampered** (tampered objects/parts), **visual** (consistency of visual evidence), and **summary** (overall summary accuracy). The final score is computed by aggregating the five textual dimensions with weights 0.05, 0.35, 0.40, 0.15, 0.05 to obtain a weighted text score, which is then fused with the mask score as:

$$\text{Overall} = 0.5 \cdot \text{score}_{\text{Mask}} + 0.5 \cdot \text{Overall}_{\text{text}}.$$

All results reported in Table 4 are averaged over all of SIDA validation images, with the mean  $\pm$  standard deviation.

The complete GPT-5 prompt used for this evaluation is shown in Figure 2.

## A.7. Dataset Details

We used several publicly available datasets for both training and evaluation. Below are the details of the datasets, their composition, and how we used them.

**SID-Set:** The SID-Set consists of 300,000 images evenly split into three categories: Real, Full Synthetic, and Tampered images. We divide the dataset into training, validation, and test sets with a 7:1:2 ratio, ensuring balanced class distribution across all splits. The mask annotations are used as ground truth for evaluating the localization performance. This dataset was preprocessed using the setup detailed in the original paper.

**MMTD-Set:** The MMTD-Set is composed of 8 different evaluation benchmarks, including CASIA1+, IMD2020, Columbia, and others. It includes images of various forgery types, such as copy-move, splicing, and removal. All of these datasets provide masks, which we utilize for our evaluation of localization.

## A.8. Discussion: Error Propagation and Hallucination

A natural question for a decoupled pipeline is whether errors in Stage-1 localization could propagate to Stage-2 and cause the LLM to generate hallucinated forgery explanations. In particular, if the localization module produces an incorrect mask, such as a false positive over an authentic region, one may wonder whether the LLM would simply follow the mask and describe non-existent manipulation artifacts.

**Secondary Verification by the LLM.** Our IFDL-VLM is designed to mitigate this risk. During Stage-2 training, the LLM is exposed to real, fully synthetic, and tampered images, together with their corresponding descriptive annotations. As a result, the model learns to verify the visual evidence within the prompted region, rather than assuming that the highlighted area must be manipulated solely because it is specified by the mask. Therefore, when presented with an erroneous mask covering

You are a strict, reproducible evaluator for image forensics. Compare Model A and Model B on 'tampering localization + explanation' using the given tampered image, GT mask, and predicted masks, along with textual explanations. Score ONLY the requested sub-dimensions (0–5 each) and DO NOT compute totals. Output must include a #scores JSON block exactly as specified

[Scoring dimensions (0–5 each)]

- mask: Localization quality of predicted mask vs GT (coverage, precision/recall, shape alignment; do NOT hallucinate).
- type: Correctness of Type ({real/object/part tampered} etc.) against GT text.
- areas: Correctness of tampered area/region description vs GT text.
- tampered: Correctness of what objects/parts are tampered vs GT text.
- visual: Consistency of visual inconsistencies/evidence vs GT (generic phrases should NOT be over-rewarded).
- summary: Summary faithfulness to GT text.

[Important]

- 1) Use the images to assess mask quality visually; do not rely solely on text.
- 2) For 'visual', reward concrete, matching evidence; penalize generic template phrases.
- 3) Provide brief rationale in #analysis, then only the JSON in #scores.

[Output format — strictly follow]

#analysis

#scores

```
{
  "ModelA": { "mask": <0-5>, "type": <0-5>, "areas": <0-5>, "tampered": <0-5>, "visual": <0-5>, "summary": <0-5> },
  "ModelB": { "mask": <0-5>, "type": <0-5>, "areas": <0-5>, "tampered": <0-5>, "visual": <0-5>, "summary": <0-5> },
  "winner": "ModelA|ModelB|Tie",
  "notes": "brief notes"
}
```

Figure 2. The full multimodal prompt used to instruct GPT-5 for pairwise evaluation of IFDL models. The prompt specifies scoring criteria, output format, and input modalities (image, masks, texts).

Table 5. Quantitative evaluation of explanation robustness under boundary-perturbed masks. “Boundary-Pert.” denotes applying morphological operations to the predicted masks.

Setting	BLEU-1	ROUGE-L	METEOR	CIDEr	CSS
PM (Clean)	<b>0.579</b>	<b>0.434</b>	0.290	<b>0.030</b>	<u>0.842</u>
PM (Boundary-Pert.)	<u>0.574</u>	<u>0.431</u>	<b>0.292</b>	0.020	0.842
GT	0.565	0.428	<u>0.291</u>	<u>0.026</u>	<b>0.853</b>

a genuine region, the LLM can still produce authenticity-consistent explanations, such as noting that the lighting is natural and the boundaries are smooth without clear signs of manipulation. In this sense, the Stage-2 LLM serves as a secondary verification module that can reject incorrect localization cues and avoid hallucinating non-existent forensic traces.

**Robustness to Mask Imprecision.** To further quantify the effect of imperfect Stage-1 localization on explanation generation, we conduct a boundary perturbation experiment. Specifically, we apply morphological dilation and erosion to the predicted masks to simulate imprecise localization. The results are reported in Table 5.

Notably, the semantic similarity score (CSS) remains at 0.842 under boundary perturbation, which is very close to the performance obtained with ground-truth masks (0.853). The other text-generation metrics also exhibit only marginal changes. These results suggest that the Stage-2 LLM is not overly sensitive to small geometric deviations in the Stage-1 masks. Instead, it is able to focus on the salient visual inconsistencies within the indicated region while tolerating moderate boundary noise. Overall, these findings support that the proposed decoupled framework is robust to imperfect localization and effectively alleviates error propagation from Stage-1 to Stage-2.

## A.9. More Qualitative Comparison Examples

To demonstrate the advantages of our method, we show more qualitative comparison examples in Figures 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17.



GT





Can you identify whether this image is real, full synthetic, or tampered? If the image is tampered, **output a mask for the tampered region and explain why.**

Type: {object tampered}  
Areas:  
<tampered position>In the central foreground of the image, directly in front of the bar or counter area, positioned in the bottom center of the visual frame.  
Tampered Content:  
<Types of objects or parts> A human figure  
Visual Inconsistencies:  
<lighting> The lighting on the figure does not match the ambient lighting focused on the shelved items which diffuses more softly  
<edges> The edges around the figure are overly sharp and distinct compared to the naturally blending backgrounds  
<resolution> Differences in pixelation or resolution between the figure and surrounding items  
<shadows> Shadow angles and reflections do not match with those in the environment, suggesting inconsistencies  
<boundary> Strange pixel behavior or abrupt changes in texture at the boundaries indicating digital insertion  
Summary: The given image is {object tampered} based on the above analysis.

Type: {part tampered}  
Areas:  
<tampered position> On the person standing in front of the bar, **specifically affecting the attire and possibly the facial features.**  
Tampered Content:  
<Types of objects or parts> The person's attire, particularly the apron and possibly the facial features  
Visual Inconsistencies:  
<lighting> The lighting on the tampered areas does not match the surrounding environment  
<edges> The edges around the tampered areas are unnaturally sharp or blurred  
<resolution> The resolution of the tampered areas is either too high or too low compared to the rest of the image  
<shadows> Shadows and reflections in the tampered areas do not align with the light sources  
<boundary> The boundaries of the tampered areas are unnaturally sharp or blurred  
Summary: The given image is {part tampered} based on the above analysis.

Type: {object tampered}  
Areas:  
<tampered position> In the central part of the image, a person is standing in front of a shelf filled with bottles, positioned in the middle horizontally and slightly lower third vertically.  
Tampered Content:  
<Types of objects or parts> A person, appearing to be a bartender or staff, dressed in a white shirt with a dark stripe or sash, interacting with the background by possibly reaching for a bottle on the shelf.  
Visual Inconsistencies:  
<lighting> The lighting on the person does not match the ambient lighting of the scene, with discrepancies in shadow direction and intensity.  
<edges> The edges around the person are sharper and more defined than those of the background elements.  
<resolution> The resolution and texture details of the person's clothing differ from the surrounding environment.  
<shadows> The shadows cast by the person do not align with the light sources visible in the image.  
<boundary> The integration of the person into the scene shows unnatural sharpness and clarity compared to the slightly blurred background.  
Summary: The given image is {object tampered} based on the above analysis.



SIDA-13B





IFDL-VLM



Figure 3. Qualitative comparison example. We highlight the accurate descriptions from GT in red, while the areas where the model fails to identify the modified regions correctly are marked in blue.



GT





Can you identify whether this image is real, full synthetic, or tampered? If the image is tampered, **output a mask for the tampered region and explain why.**

Type: {object tampered}

Areas:<tampered position> Centrally and prominently positioned in the environment, standing above the tree line against a sparsely vegetated backdrop, occupying the central part of the image, stretching across almost the entirety of the mid to upper portion of the frame.

Tampered Content:

<Types of objects or parts> A zebra with exaggerated features; an extremely elongated neck and body that stretch unnaturally, depicted as standing upright with its head and part of its neck appearing out of the normal view above the tree canopy.

Visual Inconsistencies:<lighting> The overall lighting on the zebra's body does not match the ambient lighting of the background environment, suggesting a discrepancy in light source direction or intensity.

<edges> The edges around the zebra are slightly smoother and less integrated with the natural rough interplay of light and shadow seen on the natural elements of the background.<resolution> The resolution and texture details on the zebra are consistent throughout its body but slightly vary when compared to the detailing and graininess of the trees and ground.<shadows> Lack of consistent shadow casting from the zebra onto the ground below, which would be expected under natural light conditions.

<boundary> The zebra's proportions are severely distorted and do not align with the perspective and scale expected from the surrounding environment, making the integration appear unnatural.

Summary: The given image is object tampered based on the above analysis.

Type: {part tampered}

Areas:<tampered position> On the zebra's head, specifically affecting the stripes and the area around the mouth, located in the central part of the image.

Tampered Content:

<Types of objects or parts> Zebra's head, **particularly the stripes and mouth area**

Visual Inconsistencies:

<lighting> The lighting on the tampered area does not match the natural lighting of the rest of the image

<edges> The edges around the tampered area are unnaturally sharp and distinct

<resolution> The resolution within the tampered area is slightly different from the rest of the image

<shadows> Shadows and reflections in the tampered area do not align with the lighting conditions of the rest of the image

<boundary> The boundaries of the tampered area are overly defined and do not blend seamlessly with the surrounding area

Summary: The given image is {part tampered} based on the above analysis.

Type: {object tampered}

Areas:<tampered position> The tampered area is the zebra, centrally located in the image against a backdrop of a clear sky and sparse, leafless trees.

Tampered Content:

<Types of objects or parts> A single zebra standing with its head turned slightly to the left and body facing forward.

Visual Inconsistencies:

<lighting> The zebra's lighting is sharper and more defined compared to the softer, diffused lighting of the surrounding environment.

<edges> The edges around the zebra are overly sharp and distinct, lacking the natural blending seen in the rest of the image.

<resolution> The zebra exhibits a higher resolution and crisper detail than the softer, less detailed background.

<shadows> The shadowing under the zebra does not match the direction and softness of shadows in the rest of the environment.

<boundary> The integration of the zebra into the natural setting is imperfect, with clear demarcations where the zebra's image meets the background.

Summary: The given image is {object tampered} based on the above analysis.

Figure 4. Qualitative comparison example. We highlight the accurate descriptions from GT in red, while the areas where the model fails to identify the modified regions correctly are marked in blue.



GT



Can you identify whether this image is real, full synthetic, or tampered? If the image is tampered, **output a mask for the tampered region and explain why.**



SIDA-13B

Type: {part tampered}

Areas:<tampered position> On the racetrack, specifically focusing on a horse and its jockey in the lower middle quadrant of the image during a race.

Tampered Content:

<Types of objects or parts> Horse and jockey in dynamic motion, possibly altered to enhance detail or appearance.

Visual Inconsistencies:

<lighting> Mismatch in lighting on the tampered horse and jockey compared to the natural lighting of the scene

<edges> Slight blurring or mismatch in the edges around the tampered horse and jockey compared to the crisp boundaries of nearby objects

<resolution> Noticeable differences in resolution or quality in the tampered areas compared to the surroundings

<shadows> Discrepancies in how shadows align with light sources around the tampered horse and jockey

<boundary> Examined boundaries may reveal blended areas or slight overlaps inconsistent with natural images

Summary: The given image is {part tampered} based on above analysis.



IFDL-VLM

Type: object tampered

Areas:<tampered position> In the central part of the image, slightly towards the left side, featuring a jockey riding a horse in a horse race.

Tampered Content:

<Types of objects or parts> A jockey riding a horse, with the horse appearing to be running in a race.

Visual Inconsistencies:

<lighting> The lighting on the jockey and horse does not match the surrounding environment, suggesting different light sources.

<edges> The edges around the jockey and horse are unnaturally sharp, indicating potential digital manipulation.

<resolution> The resolution of the jockey and horse is higher than the rest of the image, indicating a different level of detail.

<shadows> The shadows cast by the jockey and horse do not align with the light sources in the rest of the image.

<boundary> The integration of the jockey and horse with the background is poor, with clear signs of insertion.

Summary: The given image is object tampered based on the above analysis.



Type: {part tampered}

Areas:<tampered position> On the horse and its rider, centrally located in the lower half of the image, slightly towards the left side.

Tampered Content:

<Types of objects or parts> The horse and its rider, including the jockey's attire and equipment

Visual Inconsistencies:

<lighting> The lighting on the horse and rider does not match the surrounding environment, indicating altered brightness or contrast

<edges> The edges around the horse and rider are unusually sharp or blurred compared to the rest of the image

<resolution> The resolution around the horse and rider differs from the surrounding area, showing a disparity in image quality

<shadows> Shadows cast by the horse and rider do not align with the light source affecting other parts of the image

<boundary> The boundaries around the horse and rider show signs of digital alteration, such as unnatural sharpness or blurring

Summary: The given image is {part tampered} based on the above analysis.

Figure 5. Qualitative comparison example. We highlight the accurate descriptions from GT in red, while the areas where the model fails to identify the modified regions correctly are marked in blue.

7



GT





Can you identify whether this image is real, full synthetic, or tampered? If the image is tampered, **output a mask for the tampered region and explain why.**

Type: {object tampered}

Areas:  
 <tampered position> **The full figure of a person centrally positioned on the road**, standing slightly to the right of the center of the image, facing slightly towards the camera. Occupies the lower-central part of the image.

Tampered Content:  
 <Types of objects or parts> **A full human figure wearing bright clothing and accessories**

Visual Inconsistencies:  
 <lighting> The colors of the clothing are overly vibrant and somewhat unnatural compared to the duller, more natural tones of the surrounding environment  
 <edges> The contour of the figure is smoother and less detailed compared to the naturally textured environment  
 <resolution> The figure appears marginally less sharp than the surrounding details  
 <shadows> The direction or absence of shadows beneath the person does not correspond with the natural lighting and shadows in the environment  
 <boundary> The perspective and proportions of the person and accessories (e.g., skateboard) do not align perfectly with the environment  
 Summary: The given image is object tampered based on above analysis.

SIDA-13B



Type: {part tampered}

Areas:<tampered position>  
**The tampered area is located in the central part of the image, specifically affecting the person on the skateboard, centrally positioned in the image.**

Tampered Content:  
 <Types of objects or parts> **The person on the skateboard, particularly the torso and upper body**, including the arms and head, have been modified.

Visual Inconsistencies:  
 <lighting> The lighting on the tampered parts of the person does not match the surrounding areas, showing inconsistencies in shadow and highlights.  
 <edges> The edges around the tampered parts of the person are unnaturally sharp or blurred, indicating digital manipulation.  
 <resolution> The resolution of the tampered parts is either overly smooth or unnaturally detailed, not matching the rest of the image.  
 <shadows> Shadows and reflections on the tampered parts do not align with the light sources and other objects in the image.  
 <boundary> The boundaries of the tampered parts show signs of digital alteration, such as abrupt changes in texture or color.  
 Summary: The given image is {part tampered} based on the above analysis.

IFDL-VLM



Type: {object tampered}

Areas:<tampered position>  
**The entire figure of a man standing on a sidewalk, centrally located in the image, slightly skewed towards the left side.**

Tampered Content:  
 <Types of objects or parts> **A human figure dressed in a sleeveless top, short shorts, pink shoes, and pink socks, wearing a green vest and sunglasses.**

Visual Inconsistencies:  
 <lighting> The lighting on the figure does not match the ambient lighting of the scene, with incorrect shadow and highlight directions.  
 <edges> The edges around the figure are unnaturally smooth and lack texture gradient, indicating digital insertion.  
 <resolution> The figure appears slightly blurrier and less detailed compared to the sharper background.  
 <shadows> The shadows cast by the figure do not align with other shadows in the image, suggesting incorrect lighting angles.  
 <boundary> The integration of the figure with the surrounding environment appears artificial, with clear boundaries.  
 Summary: The given image is {object tampered} based on the above analysis.

Figure 6. Qualitative comparison example. We highlight the accurate descriptions from GT in red, while the areas where the model fails to identify the modified regions correctly are marked in blue.





Can you identify whether this image is real, full synthetic, or tampered? If the image is tampered, **output a mask for the tampered region and explain why.**

GT



Type: **{object tampered}**

Areas: **<tampered position> The skier is centrally located in the image, slightly skewed towards the bottom but predominantly occupying the middle of the frame.**

Tampered Content: **<Types of objects or parts> A single human figure, a skier, dressed in full skiing attire, positioned as if skiing downhill.**

Visual Inconsistencies:

- <lighting> The lighting on the skier does not perfectly match the surrounding snowy landscape, with shadows and highlights appearing slightly off.
- <edges> The edges around the skier are inconsistently sharp or blurred compared to the natural sharpness of the surrounding snow and trees.
- <resolution> There is a noticeable difference in resolution or graininess between the skier and the background, with the skier appearing either too smooth or too detailed.
- <shadows> The direction and length of shadows cast by the skier do not align with those cast by natural elements like trees.
- <boundary> There might be misalignments in the skier's scale and perspective compared to surrounding elements, indicating potential digital insertion.

Summary: The given image is {object tampered} based on the above analysis.

SIDA-13B



Type: **{part tampered}**

Areas: **<tampered position> On the skier, specifically affecting the hand and arm, located in the central part of the image, slightly towards the left side.**

Tampered Content: **<Types of objects or parts> The hand and arm of the skier**

Visual Inconsistencies:

- <lighting> The lighting on the tampered hand does not match the natural lighting of the rest of the image
- <edges> The edges around the tampered hand are unnaturally sharp and distinct
- <resolution> The resolution of the tampered hand is slightly different from the rest of the image
- <shadows> Shadows and reflections on the tampered hand do not align with the light source affecting other parts of the image
- <boundary> The boundaries of the tampered hand are overly defined, lacking the natural blending seen in untouched parts of the image

Summary: The given image is {part tampered} based on the above analysis.

IFDL-VLM



Type: **{object tampered}**

Areas: **<tampered position> The skier is centrally located on the snow-covered slope, surrounded by sparse forests in the background.**


Tampered Content: **<Types of objects or parts> A single skier depicted in motion, wearing a bright red jacket, black pants, and equipped with a ski pole and skis.**

Visual Inconsistencies:

- <lighting> The skier is illuminated with frontal lighting, which contrasts with the diffuse sunlight in the background.
- <edges> The edges around the skier are overly smooth and lack the granularity present in the rest of the snowy landscape.
- <resolution> The skier appears sharper than the surrounding environment, indicating a disparity in image quality.
- <shadows> The shadow cast by the skier does not align with the natural light direction seen elsewhere in the image.
- <boundary> The integration of the skier into the landscape shows slight perspective and proportion misalignments.

Summary: The given image is {object tampered} based on the above analysis.

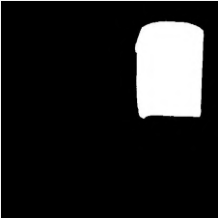
Figure 7. Qualitative comparison example. We highlight the accurate descriptions from GT in red, while the areas where the model fails to identify the modified regions correctly are marked in blue.





Can you identify whether this image is real, full synthetic, or tampered? If the image is tampered, **output a mask for the tampered region and explain why.**

GT



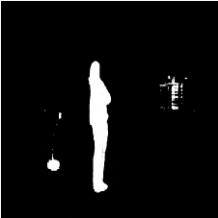
Type: {object tampered}

Areas:  
<tampered position> **The front portion of a bus, centrally located and slightly towards the right-hand side of the image, in front of pedestrians on the road.**

Tampered Content:  
<Types of objects or parts> **A vintage-style bus with distinctive features like a rounded body, large headlights, and a uniquely shaped grille and bumper.**

Visual Inconsistencies:  
<lighting> The lighting on the bus does not match the surrounding lighting; the tones and highlights differ slightly.  
<edges> The edges around the bus are too sharp and defined compared to the softer, more integrated edges of the surroundings.<resolution> There is a slight discrepancy in resolution and texture details between the bus and its environment.<shadows> The shadows cast by and on the bus do not align in direction and intensity with other objects in the scene.<boundary> The boundaries of the bus against the road and backdrop lack subtle blending, indicating insertion.  
Summary: The given image is {object tampered} based on the above analysis.

SIDA-13B



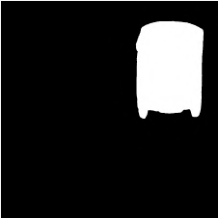
Type: {part tampered}

Areas:<tampered position> **The tampered area is located in the central part of the image, involving the bus and its occupants.**

Tampered Content:  
<Types of objects or parts> **The bus, its occupants, and the street scene**

Visual Inconsistencies:  
<lighting> The lighting on the bus and its occupants does not match the ambient lighting of the surrounding environment  
<edges> The edges around the bus and its occupants are unnaturally sharp and distinct  
<resolution> The resolution of the tampered area is slightly different from the rest of the image  
<shadows> Shadows and reflections on the bus do not align with the light sources in the scene  
<boundary> The boundaries of the bus and its occupants are overly defined and do not blend naturally with the surrounding environment  
Summary: The given image is {part tampered} based on the above analysis.

IFDL-VLM



Type: {object tampered}

Areas:  
<tampered position> **The bus is positioned on the road, appearing to move towards the viewer, located in the center-right portion of the image.**

Tampered Content:  
<Types of objects or parts> **A full bus, stationary or moving, depicted as larger and more rectangular than the surrounding environment.**

Visual Inconsistencies:  
<lighting> The bus is significantly brighter and more focused compared to the dimly lit and less defined surroundings.  
<edges> The edges of the bus are sharper and more distinct than those of the surrounding figures and objects.  
<resolution> The bus displays a higher resolution and clarity than the blurrier and less detailed background.  
<shadows> The shadows and reflections on the bus do not match the ambient lighting and environmental conditions.  
<boundary> The bus appears disproportionately large compared to the scale of the surrounding figures and urban environment.  
Summary: The given image is {object tampered} based on the above analysis.

Figure 8. Qualitative comparison example. We highlight the accurate descriptions from GT in red, while the areas where the model fails to identify the modified regions correctly are marked in blue.



GT



SIDA-13B



IFDL-VLM





Can you identify whether this image is real, full synthetic, or tampered? If the image is tampered, **output a mask for the tampered region and explain why.**

Type: {object tampered}

Areas:  
 <tampered position> **Centrally on the sandy surface, occupying the middle of the image.**

Tampered Content:  
 <Types of objects or parts> **An African Grey Parrot**

Visual Inconsistencies:  
 <lighting> The light source and shadowing on the parrot do not completely align with the environment.  
 <edges> The outline of the parrot, especially around the tail and feet, is unusually crisp compared to the sandy background.  
 <resolution> The resolution of the parrot is higher than the background, making the feathers and eyes appear more detailed.  
 <shadows> Minimal shadows under the parrot, which are inconsistent with the natural outdoor lighting.  
 <boundary> Presence of a green patch on the bird's wing, which is not characteristic of an African Grey Parrot, suggesting modification.  
 Summary: The given image is {object tampered} based on the above analysis.

Type: {part tampered}

Areas:  
 <tampered position> **On the bird's beak, specifically the upper part of the beak,** located in the central part of the image.

Tampered Content:  
 <Types of objects or parts> **The bird's beak**

Visual Inconsistencies:  
 <lighting> The color of the tampered area does not match the natural color of the bird's beak  
 <edges> The edges of the tampered area are unnaturally sharp and distinct<resolution> The tampered area appears slightly blurred compared to the rest of the image  
 <shadows> The lighting on the tampered area does not align with the lighting on the rest of the bird  
 <boundary> The boundaries of the tampered area are overly defined and do not blend naturally with the surrounding feathers  
 Summary: The given image is {part tampered} based on the above analysis.


Type: {object tampered}

Areas:  
 <tampered position> **The tampered area is the entire bird, centrally located on the ground against a sky background.**

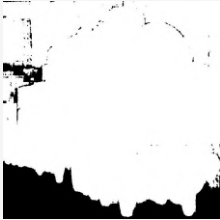
Tampered Content:  
 <Types of objects or parts> **A bird, specifically a species with a colorful body, predominantly green and gray, a prominent beak, and distinctive eye markings.**

Visual Inconsistencies:  
 <lighting> The bird's color saturation and shading do not match the muted tones of the surrounding environment.  
 <edges> The bird's outline is overly crisp compared to the softer focus of the background.  
 <resolution> The bird appears sharper than the blurred background, indicating a mismatch in resolution.  
 <shadows> The shadow under the bird does not align with the lighting direction seen in other parts of the image.  
 <boundary> The bird's presence on the ground against a sky background is contextually unusual and indicates tampering.  
 Summary: The given image is {object tampered} based on the above analysis.


Figure 9. Qualitative comparison example. We highlight the accurate descriptions from GT in red, while the areas where the model fails to identify the modified regions correctly are marked in blue.




GT




SIDA-13B



IFDL-VLM





Can you identify whether this image is real, full synthetic, or tampered? If the image is tampered, **output a mask for the tampered region and explain why.**

Type: **{object tampered}**

Areas: **<tampered position> Central and foreground portions of the image**, primarily involving the large structure which is the train.

Tampered Content: **<Types of objects or parts> The train locomotive including its main body, wheels, chimney, and front plow**

Visual Inconsistencies:

- <lighting> Uniform color and tone across the train and background suggest careful blending that might indicate tampering
- <edges> Potential mismatches in texture at the boundaries of the locomotive with the tracks and background
- <resolution> Differences in resolution between the train and other parts of the image
- <shadows> Possible inconsistencies in shadow directions and reflections relative to light sources
- <boundary> Inconsistencies in the texture and boundary integration with the environment

Summary: The given image is **{object tampered}** based on above analysis.

Type: **part tampered**

Areas:<tampered position> **On the front of the train, specifically on the left side of the locomotive, near the bottom center of the image.**

Tampered Content: **<Types of objects or parts> A large, metallic object resembling a part of the train's front**, possibly a mechanical or structural element.

Visual Inconsistencies:

- <lighting> The tampered object does not reflect light in the same manner as the rest of the train, showing a different glossiness or texture.
- <edges> The edges of the tampered object are sharper and less integrated with the surrounding textures.
- <resolution> The resolution of the tampered object is slightly different, showing more detail or less detail compared to the rest of the train.
- <shadows> Shadows and reflections on the tampered object do not align with the lighting conditions affecting other parts of the train.
- <boundary> The boundaries of the tampered object are not seamlessly integrated with the rest of the train, showing signs of artificial insertion.

Summary: The given image is part tampered based on the above analysis.

Type: **{object tampered}**

Areas: **<tampered position> The tampered area is the large locomotive centrally positioned in the image**, occupying a significant portion from the bottom to nearly the middle vertically.

Tampered Content: **<Types of objects or parts> A large, old-fashioned locomotive with detailed structure including a boiler, wheels, and various mechanical parts.**

Visual Inconsistencies:

- <lighting> The locomotive is darker and features more pronounced shadows compared to the surrounding environment.
- <edges> The edges of the locomotive are sharper and less integrated with the background.<resolution> The locomotive displays a higher level of detail and resolution compared to the softer and less defined background.
- <shadows> The shadows cast by the locomotive do not align with the lighting direction of the rest of the scene.
- <boundary> The integration of the locomotive with the surrounding scene lacks seamless continuity, indicating digital insertion.

Summary: The given image is **{object tampered}** based on the above analysis.

Figure 10. Qualitative comparison example. We highlight the accurate descriptions from GT in red, while the areas where the model fails to identify the modified regions correctly are marked in blue.



GT





Can you identify whether this image is real, full synthetic, or tampered? If the image is tampered, **output a mask for the tampered region and explain why.**

SIDA-13B



Type: {part tampered}

Areas: **<tampered position> On the laptop screen of the computer being used by the child, situated in the bottom right quarter of the image where the laptop screen appears center-right.**

Tampered Content:

<Types of objects or parts> **The laptop screen**

Visual Inconsistencies:

<lighting> The laptop screen displays an unnaturally uniform black color that does not match the lighting conditions seen in the surrounding environment

<edges> The edges of the manipulated area are distinct yet unnaturally sharp, lacking the natural transitions observed in real scenes

<resolution> The affected screen lacks the visual elements typical of a working computer screen, presenting a flat, unnatural appearance

<shadows> Not specifically mentioned, but the lack of reflections or light diffusion that would normally be present suggests improper shadow and light interaction

<boundary> The transition to the tampered area on the screen is sharply defined, indicating a clear alteration boundary

Summary: The given image is {part tampered} based on the above analysis.

IFDL-VLM



Type: {part tampered}

Areas: **<tampered position> On the laptop screen in front of the boy on the right, specifically located in the central part of the image, slightly towards the right side.**

Tampered Content:

<Types of objects or parts> **The laptop screen**

Visual Inconsistencies:

<lighting> The screen is unnaturally dark, lacking reflective properties, and does not match the ambient lighting of the room

<edges> The edges of the screen are too sharp and distinct, lacking the smooth integration seen in other parts of the laptop

<resolution> The resolution of the tampered screen area is noticeably lower, lacking the detail and clarity of the surrounding areas

<shadows> No reflective shadows or light interactions are visible on the screen, unlike other glossy surfaces in the image

<boundary> The boundaries of the tampered screen area are abrupt, without the gradual transition seen between different parts of the laptop

Summary: The given image is {part tampered} based on the above analysis.

Figure 11. Qualitative comparison example. We highlight the accurate descriptions from GT in red, while the areas where the model fails to identify the modified regions correctly are marked in blue.

13



GT



SIDA-13B



IFDL-VLM





Can you identify whether this image is real, full synthetic, or tampered? If the image is tampered, **output a mask for the tampered region and explain why.**

Type: **{object tampered}**

Areas:

<tampered position> In the forefront of the image, centrally located towards the bottom half, a few steps in front of a zebra under a tree.

Tampered Content:

<Types of objects or parts> **A horse, gray with a unique spotted pattern and unnaturally sharp, squared-off hooves.**

Visual Inconsistencies:

<lighting> The lighting on the horse does not match the rest of the scene, lacking harmony in color and intensity seen in the zebra and surrounding environment.

<edges> The edges around the horse are sharper and less naturally blended with the background.

<resolution> The resolution around the horse differs, with some areas less crisp, especially around the edges.

<shadows> The shadows cast by the horse do not align in direction or intensity with those of the zebra and trees.

<boundary> The horse features an unusual pattern that resembles an artificial creation rather than a natural coat, suggesting digital insertion.

Summary: The given image is **{object tampered}** based on the above analysis.

Type: **{part tampered}**

Areas:

<tampered position> **On the horse, specifically affecting its mane, located in the central part of the image, slightly towards the left side.**

Tampered Content:

<Types of objects or parts> **Mane of the horse**

Visual Inconsistencies:

<lighting> The lighting on the tampered mane does not match the natural lighting of the rest of the horse and the environment

<edges> The edges of the tampered mane are unnaturally sharp and distinct, lacking the natural gradient seen in the rest of the horse

<resolution> The resolution of the tampered mane is slightly different, showing a disparity in detail compared to the rest of the horse

<shadows> The shadows cast by the tampered mane do not align with the lighting direction and intensity of the rest of the image

<boundary> The boundary between the tampered mane and the rest of the horse shows an unnatural sharpness and lacks the natural blending seen in natural images

Summary: The given image is **{part tampered}** based on the above analysis.

Type: **{object tampered}**

Areas:

<tampered position> **Centrally located in the image, slightly overlapping with a zebra in the background, occupying the middle ground of the visual frame.**

Tampered Content:

<Types of objects or parts> **A hybrid animal combining features of a horse and a zebra**

Visual Inconsistencies:

<lighting> The lighting on the hybrid animal does not match the ambient lighting of the scene, with mismatched shadow directions and intensities

<edges> The edges of the hybrid animal are unnaturally sharp compared to the softer contours of the natural elements like the zebra and the tree

<resolution> The resolution of the hybrid animal appears crisper than that of the surrounding environment and other animals

<shadows> The shadows cast by the hybrid animal do not align with the light source affecting other objects in the image

<boundary> The boundaries of the hybrid animal show clear demarcations, indicating digital insertion

Summary: The given image is **{object tampered}** based on the above analysis.

Figure 12. Qualitative comparison example. We highlight the accurate descriptions from GT in red, while the areas where the model fails to identify the modified regions correctly are marked in blue.



GT





Can you identify whether this image is real, full synthetic, or tampered? If the image is tampered, **output a mask for the tampered region and explain why.**

SIDA-13B



Type: {part tampered}

Areas:

<tampered position> **Predominantly around the fireplace**, especially noticeable on the fireplace itself and portions of the walls adjacent to it, centrally located in the image, slightly to the right.

Tampered Content:

<Types of objects or parts> **The iron stove within the fireplace, parts of the fireplace (especially the arch and the hearth), and parts of the surrounding walls**

Visual Inconsistencies:

<lighting> The light from within the stove has different hue and intensity compared to the rest of the room's lighting

<edges> The boundaries around the fireplace and the iron stove are slightly blurred and do not align seamlessly with the texture of the surrounding walls

<resolution> The quality and sharpness around the tampered areas differ from other areas of the image

<shadows> Shadows cast by the fireplace elements do not match the lighting sources visible in other parts of the room

<boundary> Transition edges around the suspected tampered regions show possible overlays or alterations

Summary: The given image is {part tampered} based on the above analysis.

IFDL-VLM



Type: {part tampered}

Areas:

<tampered position> **On the wall behind the fireplace and on the door of a cabinet to the right of the fireplace**, predominantly in the central to right portion of the image.

Tampered Content:

<Types of objects or parts> **Wall behind fireplace and door of cabinet**

Visual Inconsistencies:

<lighting> Unnatural shadows and highlights on the tampered areas do not match the rest of the scene

<edges> Edges of the tampered areas are either too sharp or too soft compared to adjacent areas

<resolution> The tampered areas exhibit a different resolution, appearing either blurrier or sharper than surrounding elements

<shadows> Shadows and highlights on the tampered areas do not align with the light source direction in the rest of the image

<boundary> Boundaries of the tampered areas show unnatural transitions or abrupt texture changes

Summary: The given image is {part tampered} based on the above analysis.

Figure 13. Qualitative comparison example. We highlight the accurate descriptions from GT in **red**, while the areas where the model fails to identify the modified regions correctly are marked in **blue**.

15



GT





Can you identify whether this image is real, full synthetic, or tampered? If the image is tampered, **output a mask for the tampered region and explain why.**

SIDA-13B



Type: {part tampered}

Areas:  
<tampered position> **In the central part of the image, specifically on the child holding the umbrella, located in the middle of the scene.**

Tampered Content:  
<Types of objects or parts> **The child holding the umbrella, particularly the umbrella and the child's right hand.**

Visual Inconsistencies:  
<lighting> The lighting on the umbrella and the child's hand does not match the surrounding environment, suggesting different light sources or post-processing alterations.  
<edges> The edges around the umbrella and the child's hand are unnaturally sharp or blurred, indicating potential digital manipulation.  
<resolution> The resolution of the tampered area is inconsistent with the rest of the image, showing either higher or lower resolution.  
<shadows> Shadows and reflections in the tampered area do not align with the lighting conditions of the rest of the image.  
<boundary> The boundary around the tampered area shows signs of blending or feathering, indicating digital alteration.  
Summary: The given image is {part tampered} based on the above analysis.

IFDL-VLM



Type: {object tampered}

Areas:  
<tampered position> **In the foreground, centrally located towards the bottom half of the image, slightly towards the left side.**

Tampered Content:  
<Types of objects or parts> **A young child standing and holding an umbrella, facing towards the camera.**

Visual Inconsistencies:  
<lighting> The child is lit from the front, which does not match the overcast lighting of the background.  
<edges> The edges around the child are unnaturally smooth and do not blend well with the background.  
<resolution> The child appears sharper than the background, indicating a resolution mismatch.  
<shadows> No visible shadows or reflections that would normally be present due to the water spray.  
<boundary> The boundary around the child shows clear signs of tampering, with a faint outline distinguishing the inserted figure.  
Summary: The given image is {object tampered} based on the above analysis.

Figure 14. Qualitative comparison example. We highlight the accurate descriptions from GT in red, while the areas where the model fails to identify the modified regions correctly are marked in blue.



GT





Can you identify whether this image is real, full synthetic, or tampered? If the image is tampered, **output a mask for the tampered region and explain why.**

SIDA-13B



Type: {part tampered}

Areas:

<tampered position>**On the horizontal traffic signal bar, centrally located in the image, amidst the traffic light setup suspended over the road.**

Tampered Content:

<Types of objects or parts> **A figure of a person dressed in casual attire, sitting on the traffic signal bar facing away from the viewpoint.**

Visual Inconsistencies:

<lighting> The lighting and shadow direction on the person do not match those on the traffic signal poles.

<edges> The edges around the person are blurred and do not integrate well with the traffic signal bar.

<resolution> The resolution of the person differs subtly from the surrounding elements, suggesting a different origin.

<shadows> Shadows that should be cast by the person on the traffic light bar are missing or inconsistent.

<boundary> The proportions of the person relative to the traffic lights are not aligned, appearing disproportionately large.

Summary: The given image is {object tampered} based on the above analysis.

IFDL-VLM



Type: {object tampered}

Areas:

<tampered position>**The tampered area is located on a traffic light, specifically on the walk sign section where a human figure is seated.**

Tampered Content:

<Types of objects or parts> **A human figure seated in a relaxed position on the walk sign of a traffic light**

Visual Inconsistencies:

<lighting> The lighting on the figure does not match the ambient lighting of the scene, with discrepancies in shadow and highlight directions

<edges> The edges around the figure are slightly blurred and do not integrate well with the walk sign

<resolution> The resolution of the figure is slightly different from the surrounding elements, indicating a possible insertion from another source

<shadows> The shadows cast by the figure do not align with other shadows in the image, such as those from traffic light poles and streetlights

<boundary> The boundary around the figure shows signs of blending and feathering, typical of digital manipulations

Summary: The given image is {object tampered} based on the above analysis.

Figure 15. Qualitative comparison example. We highlight the accurate descriptions from GT in red, while the areas where the model fails to identify the modified regions correctly are marked in blue.

17



GT



SIDA-13B



IFDL-VLM





Can you identify whether this image is real, full synthetic, or tampered? If the image is tampered, **output a mask for the tampered region and explain why.**

Type: {object tampered}

Areas:<tampered position> **In the foreground of the image, involving a child directly interacting with a giraffe, towards the bottom center of the image.**

Tampered Content:  
 <Types of objects or parts> **The entire figure of a child reaching out or interacting with the giraffe's mouth, equipped with a backpack and wearing a cap, a short sleeve shirt, and dark shorts.**

Visual Inconsistencies:  
 <lighting> The lighting on the child does not match the surrounding environment, indicating inconsistent lighting conditions.  
 <edges> The edges around the child are too sharp or well-defined compared to the more naturally blending textures of the background.  
 <resolution> The clarity and detail level of the child's figure does not match the surrounding environment, indicating a disparity in image quality.  
 <shadows> Shadow alignment and reflections do not corroborate with the main light sources in the image, indicating potential tampering.  
 <boundary> Abrupt changes in texture, color, or noise patterns along the outline of the child, inconsistent with the rest of the scene.  
 Summary: The given image is {object tampered} based on the above analysis.

Type: {part tampered}

Areas:  
 <tampered position>**On the giraffe, specifically on the neck and head region, located in the central part of the image.**

Tampered Content:  
 <Types of objects or parts>**The giraffe's neck and head**

Visual Inconsistencies:  
 <lighting> The lighting on the tampered area does not match the natural lighting of the rest of the image  
 <edges> The edges around the tampered area are unnaturally sharp and distinct  
 <resolution> The resolution of the tampered area is slightly different from the rest of the image  
 <shadows> Shadows and reflections in the tampered area do not align with the light source  
 <boundary> The boundaries of the tampered area are overly defined and do not blend seamlessly with the surrounding area  
 Summary: The given image is {part tampered} based on the above analysis.

Type: {object tampered}

Areas:  
 <tampered position> **In the lower central part of the image, a child is standing next to a giraffe, interacting with it.**

Tampered Content:  
 <Types of objects or parts> **A child dressed in casual attire, reaching up towards the giraffe's neck**

Visual Inconsistencies:  
 <lighting> The lighting on the child does not match the sunlight's direction and intensity compared to the rest of the scene  
 <edges> The edges around the child are sharper and less blended with the background, indicating a possible cut-and-paste  
 <resolution> The resolution around the child appears slightly less detailed or smoother than the surrounding environment  
 <shadows> The shadows cast by the child do not align with the shadows of other objects in the scene  
 <boundary> The interaction between the child and the giraffe seems unnaturally close and intimate, lacking realistic boundaries  
 Summary: The given image is {object tampered} based on above analysis.

Figure 16. Qualitative comparison example. We highlight the accurate descriptions from GT in red, while the areas where the model fails to identify the modified regions correctly are marked in blue.



GT



SIDA-13B



IFDL-VLM





Can you identify whether this image is real, full synthetic, or tampered? If the image is tampered, **output a mask for the tampered region and explain why.**

Type: **{object tampered}**

Areas: **<tampered position> Centrally located upon a pile of colorful textiles in the middle of the image, spanning from the bottom to the lower-middle section.**

Tampered Content:

<Types of objects or parts> **A singular feline, specifically a cat seated with a calm and attentive expression.**

Visual Inconsistencies:

<lighting> The cat exhibits sharper, more direct lighting compared to the diffused natural light illuminating the surrounding textiles.

<edges> The edges around the cat are unnaturally sharp and well-defined, lacking natural shadowing or interaction with the textiles.

<resolution> The resolution and texture of the cat's fur are crisper and more detailed than the slightly blurrier and less detailed background.

<shadows> The shadows and interactions where the cat meets the textiles do not occur naturally, suggesting digital modification.

<boundary> The integration of the cat into the textiles is not seamless, contributing to an artificial appearance.

Summary: The given image is object tampered based on the above analysis.

Type: **{part tampered}**

Areas:<tampered position> **On the cat's face, specifically affecting its nose,** located in the central part of the image, slightly towards the left side.

Tampered Content:

<Types of objects or parts> **The cat's nose**

Visual Inconsistencies:

<lighting> The tampered area exhibits a different shade of orange, contrasting with the rest of the cat's nose

<edges> The edges around the tampered area are smoother and less detailed compared to the surrounding fur

<resolution> The tampered area shows a slight blur or pixelation, indicating a difference in resolution

<shadows> The tampered area lacks the expected shadowing and reflective properties of the rest of the cat's nose

<boundary> The boundaries of the tampered area are not seamlessly integrated with the surrounding fur, showing signs of manipulation

Summary: The given image is {part tampered} based on the above analysis.

Type: **{object tampered}**

Areas: **<tampered position> In the lower central portion of the image, specifically on a stack of fabric or material in a market stall.**

Tampered Content:

<Types of objects or parts> **An orange and white cat**

Visual Inconsistencies:

<lighting> The cat's color saturation and brightness slightly differ from the surrounding environment

<edges> The edges around the cat are slightly sharper than those of the surrounding objects

<resolution> The texture and clarity of the cat's fur are smoother and less detailed compared to the surrounding fabrics

<shadows> The shadows cast by the cat do not align with the lighting direction of the scene

<boundary> The boundaries around the cat show slight discrepancies in blending with the background

Summary: The given image is {object tampered} based on the above analysis.

Figure 17. Qualitative comparison example. We highlight the accurate descriptions from GT in red, while the areas where the model fails to identify the modified regions correctly are marked in blue.