Rethinking VLMs for Image Forgery Detection and Localization Supplementary Material

A. Appendix / Supplemental Material

A.1. Ablation on α for Region-aware Visual Feature Enhancement

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

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

where α controls the balance between the global semantics (from CLIP(x)) and region-specific forensic cues (from 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 α 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 α influences the model's overall performance.

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

α 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	**0.87**	0.67	**0.60**	**0.70**	**0.84**
0.7	0.80	**0.68**	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
IFDL-VLM (Ours)	**0.99**	**0.87**	**0.65**	**0.84**	**0.43**

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)	**0.580**	**0.434**	**0.290**	**0.030**

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.

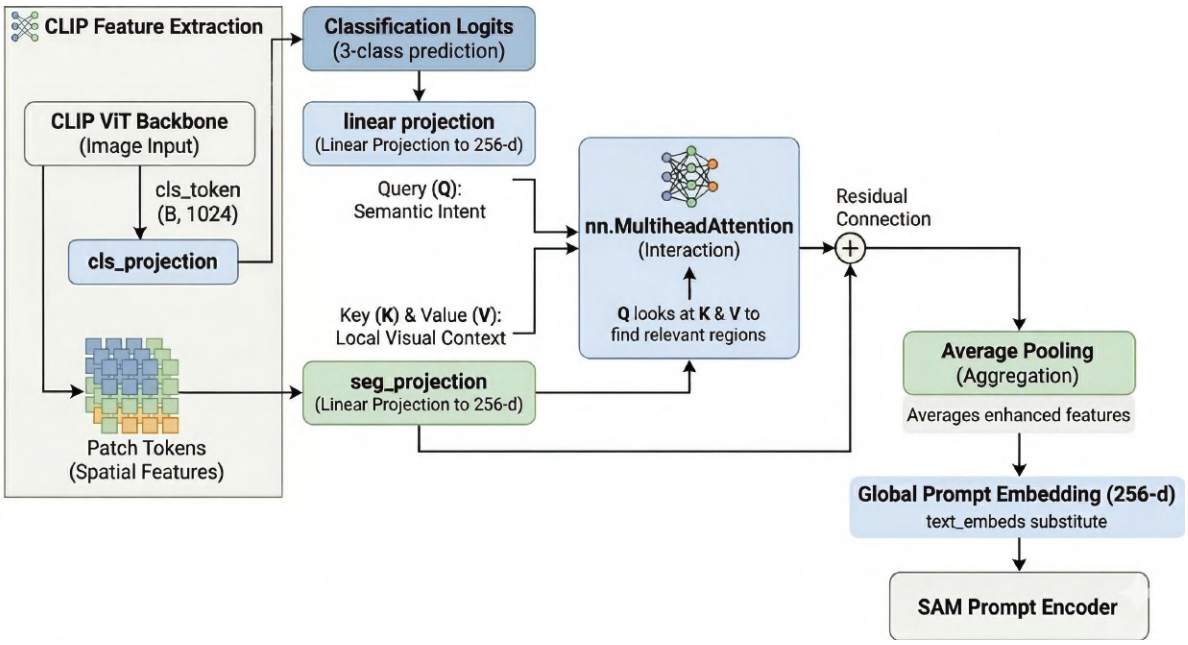


Figure 1. Attention details in the IFDL-VLM framework.

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:

$$\text{Overall} = 0.5 \cdot score_{Mask} + 0.5 \cdot Overall_{text}.$$

All results reported in Table 4 are averaged over all of SIDA validation images, with the mean ± 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)	**0.579**	**0.434**	0.290	**0.030**	0.842
PM (Boundary-Pert.)	0.574	0.431	**0.292**	0.020	0.842
GT	0.565	0.428	0.291	0.026	**0.853**

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.

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.

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.

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.

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.

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.

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.

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.

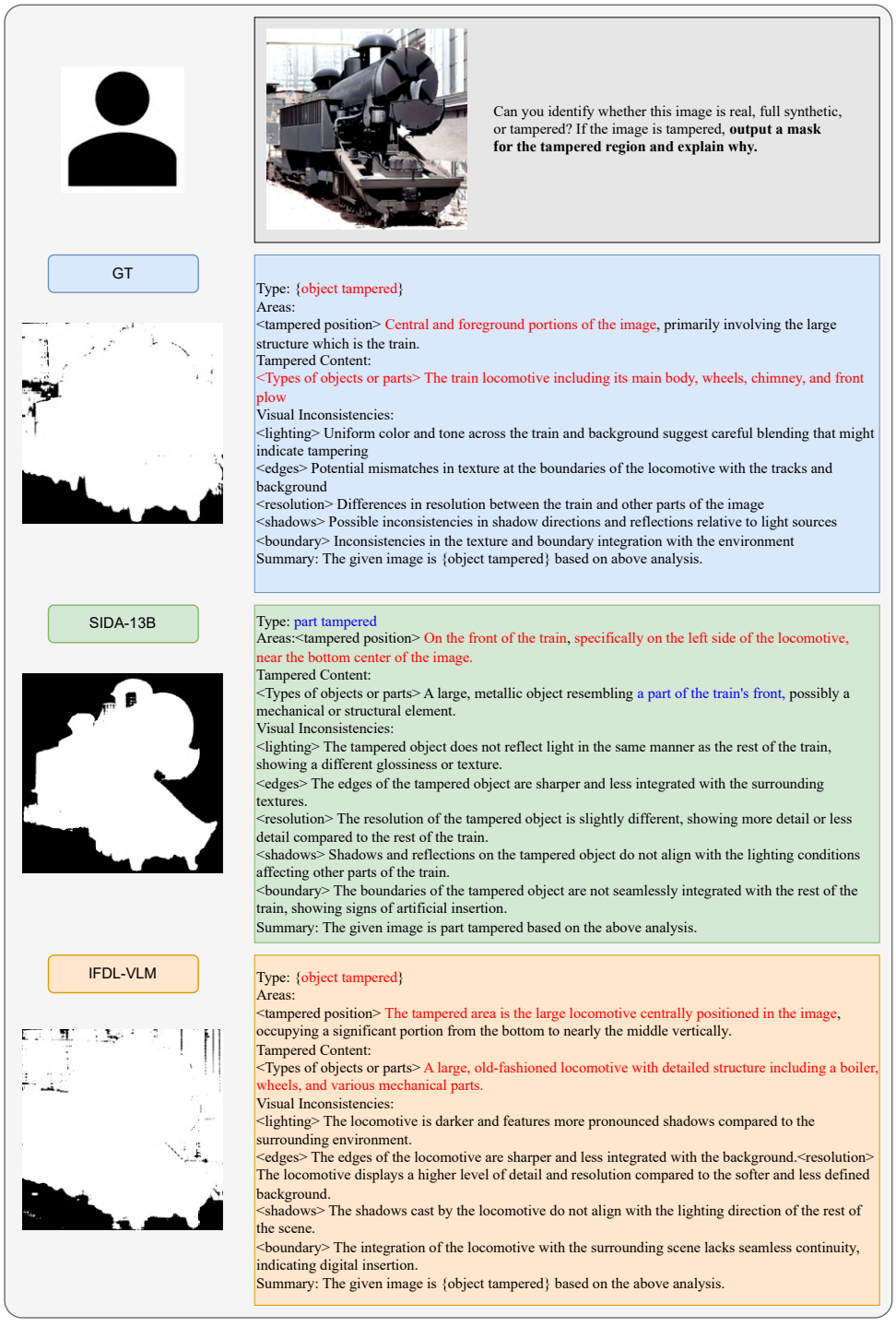


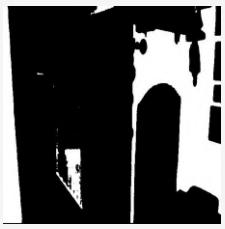
Figure 10. Qualitative comparison example. We highlight the accurate descriptions from GT in red, while the areas where the model fails to identify the modified regions correctly are marked in blue.

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.

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.

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

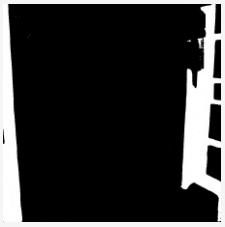


Type: {part tampered}
Areas:
<tampered position> Predominantly around the fireplace, especially noticeable on the fireplace itself and portions of the walls adjacent to it, centrally located in the image, slightly to the right.
Tampered Content:
<Types of objects or parts> The iron stove within the fireplace, 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.

SIDA-13B

Type: {part tampered}
Areas:
<tampered position> On the door frame, specifically on the upper left side of the door, extending slightly towards the center.
Tampered Content:
<Types of objects or parts> The entire upper left quadrant of the door frame, including the door frame and parts of the door frame.
Visual Inconsistencies:
<lighting> The lighting on the tampered area does not match the natural lighting of the rest of the door frame, showing a different hue or intensity.
<edges> The edges of the tampered area are unnaturally sharp or blurred, indicating digital manipulation.
<resolution> The resolution within the tampered area is either too high or too low compared to the rest of the door frame.
<shadows> Shadows or reflections within the tampered area do not align with the light source affecting other parts of the door frame.
<boundary> The boundary of the tampered area shows abrupt changes in texture or color, indicating a clear delineation from the original content.
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.

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.

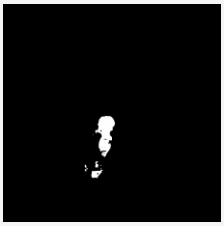
Figure 15. Qualitative comparison example. We highlight the accurate descriptions from GT in red, while the areas where the model fails to identify the modified regions correctly are marked in blue.

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

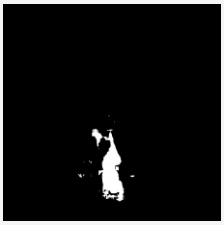
Type: {object tampered}
Areas:<tampered position> In the foreground of the image, involving a child directly interacting with a giraffe, towards the bottom center of the image.
Tampered Content:
<Types of objects or parts> The entire figure of a child reaching out or interacting with the giraffe's mouth, 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.

SIDA-13B



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.

IFDL-VLM



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.

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.