SIDA: Social Media Image Deepfake Detection, Localization and Explanation with Large Multimodal Model

Supplementary Material

Contents of the Appendices:

A. Experiment Settings

Detection Methods. We used AIGCDetecBenchmark[1] GitHub to test and re-train CnnSpot [4], FreDect [5], Fusing [8], Gram-Net [13], UnivFD [14], LGrad [16], and LNP [1]. For AntifakePrompt [2],we used the original training settings provided in the official GitHub repository[2]. During testing and training, we used only classification labels for these detection methods, as they cannot handle localization tasks.

We set noise (*e.g.* JPEG compression, blur, and resize) to None for testing each approach. For CNNSpot [4], FreDect [5], Fusing [8], and Gram-Net [13], we retrained them with the following hyperparameters: a blur probability of 0.1 with a sigma range of 0.0 to 3.0, a JPEG compression probability of 0.1, and JPEG quality ranging from 30 to 100. We used a batch size of 64, a crop size of 224, and Adam as the optimizer. We used different hyperparameters to achieve the best results for LGrad [16], LNP [1], and UnivFD [14], which require image pre-processing. Specifically, for LNP and LGrad, both the blur probability and JPEG compression probability were set to 0. For UnivFD, we used the same training settings as CNNSpot after pre-processing. For AntifakePrompt, we used the same hyperparameters and prompts as described in the original paper, recording and calculating performance across different classes in the results. All methods were trained for 10 epochs on a single NVIDIA A100 40GB GPU. Methods that did not require image pre-processing took approximately 36 hours to train, while LGrad, LNP, and UnivFD, which needed pre-processing, took around 48 hours.

Localization Methods. We used the pre-trained models for MVSS-Net [3] and HIFI-Net [6] to evaluate performance on SID-Set. For PSCC-Net [12], we used the same training settings as provided in the official GitHub repository[3].

[1]https://github.com/Ekko-zn/AIGCDetectBenchmark
[2]https://github.com/nctu-eva-lab/AntifakePrompt
[3]https://github.com/proteus1991/PSCC-Net/tree/main

Table 1. Comparison with existing related works. An (*) indicates methods that have created their own dataset.

Methods	Year	Has dataset*	Detection		Localization	Interpretation
			Binary	Muti-classification		
DIRE [17]	2023	✓	✓	✗	✗	✗
AntifakePrompt [2]	2024	✗	✓	✗	✗	✗
CnnSpott [4]	2021	✓	✓	✗	✗	✗
FreDect [5]	2020	✗	✓	✗	✗	✗
Fusing [8]	2022	✗	✓	✗	✗	✗
Gram-Net [13]	2020	✗	✓	✗	✗	✗
UnivFD [14]	2023	✓	✓	✗	✗	✗
LGrad [16]	2023	✗	✓	✗	✗	✗
LNP [1]	2023	✓	✓	✗	✗	✗
MVSS-Net [3]	2023	✗	✓	✗	✓	✗
HIFI-Net [6]	2023	✓	✓	✗	✓	✗
PSCC-Net [12]	2022	✓	✓	✗	✓	✗
FFAA [7]	2024	✓	✓	✗	✓	✓
FakeShield [18]	2024	✓	✓	✗	✓	✓
ForgeryGPT [10]	2024	✓	✓	✗	✓	✓
SIDA	2024	✓	✓	✓	✓	✓

For LISA [9], we used the LISA-7B-v1 version and fine-tuned it on SID-Set for comparison. Specifically, we set the learning rate to 0.0001, the batch size to 2, and the gradient accumulation steps to 10.

B. Detailed Comparison

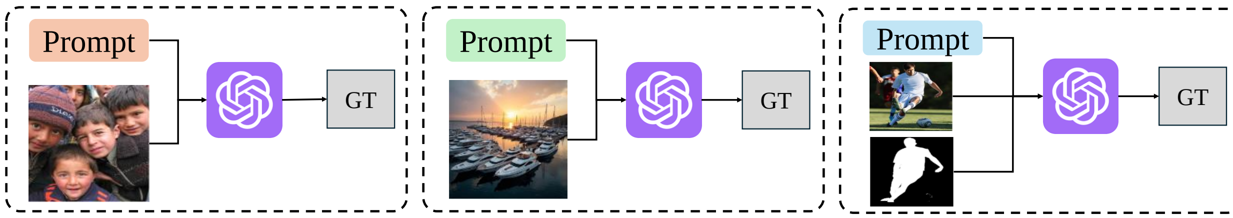
Due to page limitations, we selected only a few representative works for the main comparison. In this section, we present a more comprehensive comparison of SIDA with additional related works, as shown in Table 1.

Compared to detection methods [1, 2, 4, 5, 8, 13, 14, 16, 17], which often specialize in identifying specific generative techniques, SIDA is designed with a broader focus, capable of handling various manipulation types. This versatility allows SIDA to generalize better across different datasets and manipulations, making it more effective in real-world scenarios. Additionally, SIDA provides both detection and localization, offering a more comprehensive solution compared to detection-only models.

Compared to existing IFDL (Image Forgery Detection and Localization) methods [3, 6, 12], which primarily focus on detecting tampered versus real images, SIDA is capable of handling a broader range of scenarios, including fully synthetic, tampered, and real images. This allows SIDA to provide a more comprehensive detection capability. Furthermore, SIDA leverages LLMs to enhance the interpretability of its localization results, delivering not only segmentation masks but also detailed explanations. This combination improves precision and adds a valuable interpretative layer that existing methods lack, making it effective for understanding and addressing manipulations in complex scenarios.

Compared to other works that have explored the use of

CVPR #6596



a) Real image b) Fully synthetic image c) Tampered image

Figure 1. Generating ground truth descriptions for three different types of inputs.

LLMs in deepfake detection, our approach addresses multiple tasks, utilizes a larger dataset, and produces fine-grained outputs. For example, compared with FFAA [7], SID-Set is not limited to facial deepfake detection but is designed to tackle more complex scenarios commonly found on social media, such as object manipulation and partial tampering. Compared with FakeShield [18], our work not only includes more realistic images but also provides detailed fine-grained results, thereby enhancing detection accuracy and interpretability. Additionally, compared to ForgeryGPT [10], we have curated a large-scale, high-quality dataset that serves as a valuable resource to support and advance research in this domain.

Since some works have not released their code until paper submission [7, 10, 18], we chose PSCC-Net [12] and LISA [9] to demonstrate localization results due to their effective segmentation capabilities. We retrained both models on SID-Set for 10 epochs and obtained the output results. Compared to these methods, SIDA shows superior performance in detecting the borders of tampered areas, delivering more precise and clearer results, as illustrated in Figure 2.

C. Additional Visual Examples

In this section, we provide additional visual examples of SIDA. Figures 3 and 4 depict SIDA's outputs for tampered images, while Figure 5 highlights some failure cases.

The first row in Figure 5 illustrates instances where SIDA fails to detect tampered areas, with some cases resulting in no mask output at all. The second row demonstrates SIDA's inability to generate fine-grained masks for the tampered regions. We attribute these shortcomings to two primary factors. First, the current training data for tampered images may be insufficient. Although SID-Set provides 100k tampered images, this volume might still be inadequate for the LLM to effectively handle highly detailed and complex manipulations. Second, although SIDA surpasses other methods in detecting tampered regions, it may still lack the precision required for particularly challenging cases involving subtle or intricate tampering. These limitations indicate critical areas for future research. We aim to improve both the quality and quantity of training data, while also developing more sophisticated methodologies and enhancement strategies to better address the challenges posed by complex manipulation scenarios, ultimately enhancing detection accuracy and mask quality.

D. Detailed DataSet Creation Process

Prompts for Generating Descriptions. We designed prompts to generate different descriptions using GPT-4o. Separate prompts were crafted for real images, fully synthetic images, and tampered images. The prompts are illustrated in Figures 6, 7, and 8.

Examples of Generated Descriptions. we present examples of the output descriptions generated by SIDA. Cases of real images, fully synthetic images, and tampered images are shown in Figures 9, 10, and 11, respectively.

Details of Generative Process. We provide further details on the generation of fully synthetic and tampered images.

Fully Synthetic Images. We used FLUX[4] to generate fully synthetic images due to its high quality, utilizing original data from Flickr30k [15] and COCO [11]. The style prompt was set as "cinematic photo of prompt, 35mm photograph, film, professional, 4k, highly detailed," while the negative prompt included terms "deformed iris, deformed pupils, semi-realistic, cgi, 3d, render, sketch, cartoon, drawing, anime, text, cropped, out of frame, worst quality, low quality, jpeg artifacts, ugly, duplicate, morbid, mutilated, extra fingers, mutated hands, poorly drawn hands, poorly drawn face, mutation, deformed, blurry, dehydrated, bad anatomy, bad proportions, extra limbs, cloned face, disfigured, gross proportions, malformed limbs, missing arms, missing legs, extra arms, extra legs, fused fingers, too many fingers, long neck," to avoid unrealistic artifacts. All images were generated using 2 NVIDIA A100 GPUs with 40GB memory.

Tampered Images. We detail each step of the tampered image generation process. Two separate directories were set up: one for object replacement, where entire objects (e.g., animals, vehicles, household items) are swapped with similar classes to generate new scenarios, and another for attribute replacement, which modifies specific features or characteristics of objects (e.g., changing an animal's emotion or activity, such as making a "dog" appear "happy"or "running"). Figures 12 and 13 illustrate the detailed directo-

[4]https://github.com/black-forest-labs/flux

Figure 2. Visual comparison of SIDA with other localization methods. Both approaches were fine-tuned on the SID-Set for this evaluation.

ries for object and attribute replacements, respectively. Furthermore, we employed GPT-4o to generate ground truth descriptions for three distinct input types: (1) real images paired with prompts, (2) fully synthetic images accompanied by corresponding prompts, and (3) tampered images provided with both prompts and their associated tampered masks, as shown in Figure 1. Figure 3 elaborates on each step of the tampered image generation process using these replacement strategies. To further enrich the diversity of SID-Set, we integrated image segments from Magicbrush [19]. Additional information about Magicbrush is available on its project website[5].

E. Experts and Human Evaluation

We engaged five experts to undertake the following three tasks:

Model Selection: The experts examined approximately 1,000 images to assess the consistency and quality of the generated outputs. Based on their evaluations, we selected FLUX and latent-diffusion as our default generative models due to their superior performance.

Image Quality Assessment: Following image generation, the experts evaluated the realism of the outputs. To standardize this process, we introduced a five-point rating scale for image realism, ranging from 0 (lowest quality) to 5 (highest quality). Images scoring below 3 were flagged as unnatural or defective. These flagged images underwent a secondary review by the experts, after which all identified flawed images were excluded to maintain the dataset's overall quality.

Textual Description Evaluation: The experts systematically reviewed 3,000 textual descriptions produced by GPT-4 to verify their semantic accuracy and alignment with the corresponding images. This evaluation adhered to three key criteria: (1) Accuracy – ensuring the description accurately reflects the image's visual content; (2) Clarity – confirming the description is concise, unambiguous, and easily comprehensible; and (3) Consistency – verifying coherence with similar prompts or scenarios across the dataset.

These refinements have been incorporated into our revised approach. We appreciate the constructive feedback and valuable guidance provided, which have significantly strengthened this process.

[5] https://osu-nlp-group.github.io/MagicBrush/

CVPR
#6596

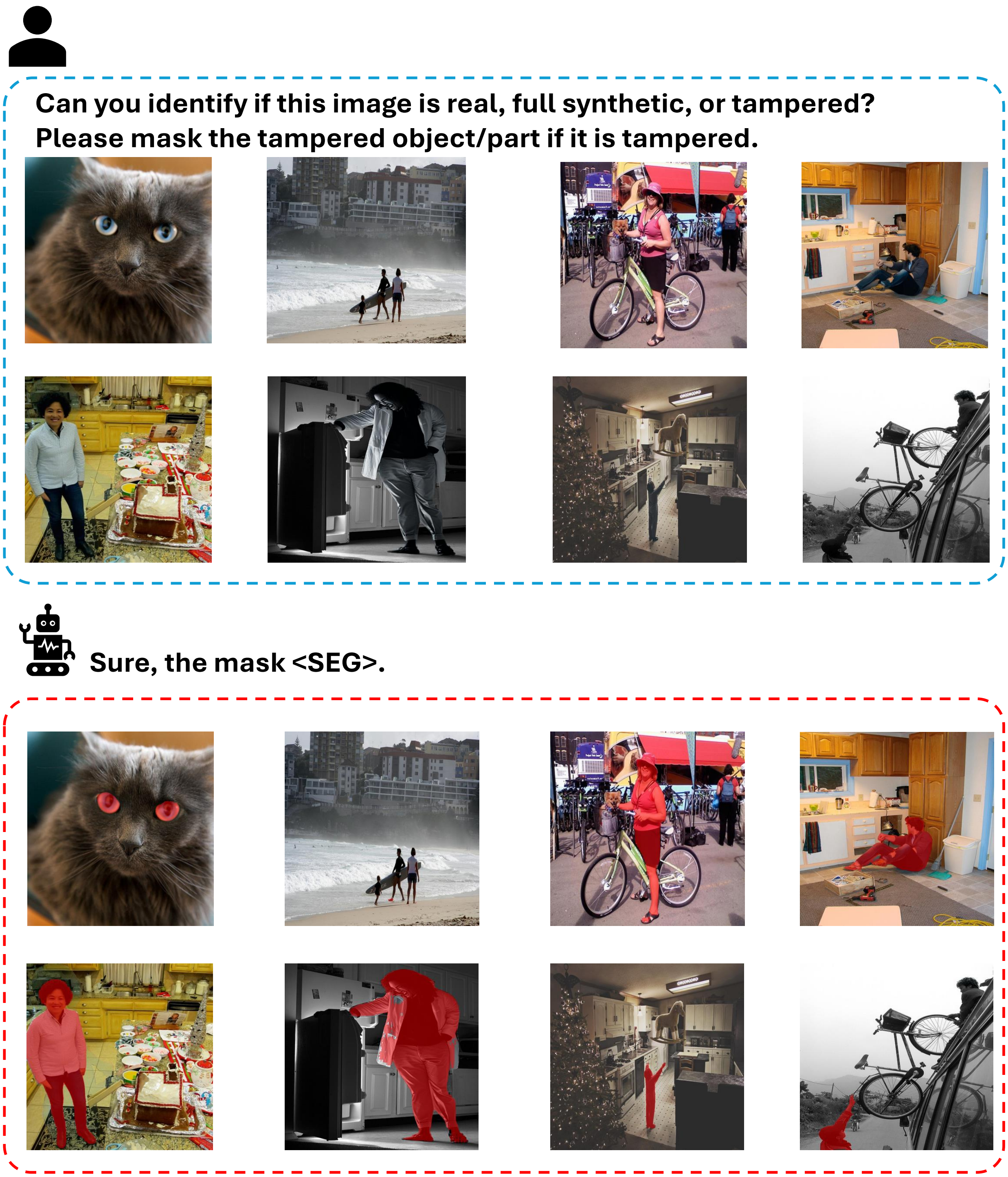


Figure 3. Visual examples of SIDA.

CVPR #6596

Figure 4. More Visual examples of SIDA.

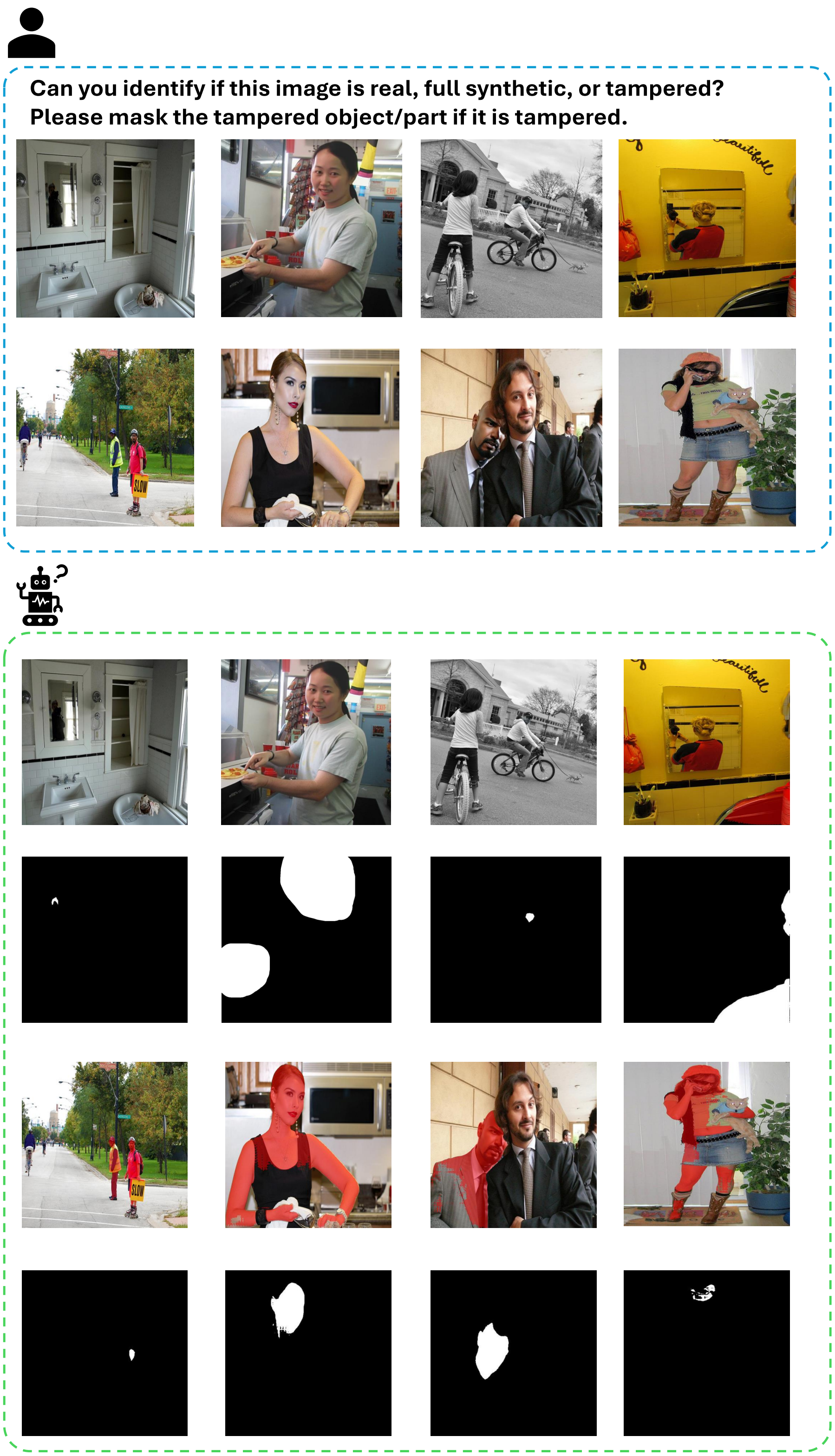


Figure 5. Failure cases of SIDA.

You are an AI visual authenticity expert. Your task is to analyze a genuine, untampered photograph and explain why it is authentic. Use a combination of technical and human-centered perspectives to provide a thorough analysis of the image.
1. Consistency
1.1 Lighting and Color
- Examine the overall style, color, and details for visual consistency.
- Ensure lighting, shadows, and colors are consistent across the entire image, especially among multiple objects or people.
- Identify subtle differences in light and color that contribute to the image's realism.

1.2 Edges and Pixels
- Examine the pixel distribution and edges for naturalness.
- Look for discontinuous or inconsistent edges that could indicate tampering.
- Check for obvious traces of clipping or compositing.

1.3 Resolution and Compression
- Assess the images resolution and look for compression artifacts.
- Ensure consistent quality across the image, avoiding unnatural pixel blurring, jaggedness, or excessive compression marks.

1.4 Reflections and Shadows
- Verify that reflections and shadows are consistent with the light sources.
- Ensure shadows fall naturally and match the objects casting them, considering their direction and softness.

1.5 Material Properties
- Analyze how different materials (fabric, skin, metal, etc.) interact with light.
- Ensure that materials have realistic properties, such as the proper reflection of skin or the sheen of metal, which supports authenticity.

1.6 Physical Laws and Perspective
- Confirm that objects adhere to physical laws, such as realistic movements and trajectories.
- Analyze perspective and scale relationships to ensure consistency, checking proportions and depth of field effects for realism.

2. Authenticity
2.1 Natural Imperfections
- Identify subtle flaws or imperfections that AI or editing tools might overlook.
- Describe how these natural inconsistencies-such as minor blemishes or environmental randomness-contribute to the image's authenticity.

2.2 Emotional Congruence
- Assess the emotional coherence of subjects expressions and body language.
- Explain how natural, congruent emotions help validate the authenticity of the image.

2.3 Environmental Interaction
- Analyze how subjects and objects interact naturally with their environment.
- Describe details that indicate genuine integration with surroundings, such as grass bending under a foot or reflections in nearby surfaces.

2.4 Temporal Consistency
- Look for elements that suggest a specific moment in time.
- Explain how coherent temporal details-like the movement of wind, ripples in water, or the positioning of objects-reinforce authenticity.

2.5 Cultural and Contextual Elements
- Identify culturally specific elements that appear naturally within the image.
- Explain how these details-such as attire, symbols, or social interactions-contribute to the credibility and realism of the photograph.

2.6 Unintended Elements
- Identify peripheral or background details that appear unplanned.
- Describe how these unintended inclusions support authenticity, as they are challenging to replicate intentionally.

2.7 Dynamic Range
- Assess the image's handling of highlights and shadows.
- Explain how the natural dynamic range-reflecting real-world lighting-contributes to the authenticity.

2.8 Micro-Expressions
- Look for subtle, fleeting expressions on subjects' faces.
- Describe how these micro-expressions-which are hard to fabricate-contribute to the authenticity.

3. Comparative Analysis
If you were offered a tampered images and were asked to judge whether this picture has been tampered with, describe from what angle you would analyze and judge it. Provide detailed reasons for your assessment, considering both consistency and authenticity aspects discussed above.
Final Assessment
Provide an overall evaluation of the image authenticity by synthesizing evidence from both technical consistency and human-centered authenticity perspectives. Clearly explain how these combined aspects lead to the conclusion that the image is genuine.
- Be clear and descriptive, providing specific examples from the image to support your analysis.
- Highlight both obvious and subtle indicators of authenticity, focusing on a holistic view of the image.

Figure 6. Prompts for real images.

You are an AI visual authenticity expert. I will provide you with a fully synthetic image that has been entirely generated or is a whole deepfake. Your task is to analyze this image and explain how you can tell it's fake. Use a combination of technical and human-centered perspectives to provide a thorough analysis of the image's artificiality.

1. Technical Analysis

1.1 Lighting and Color

- Analyze the overall lighting and color scheme for unnatural or unrealistic combinations.
- Identify inconsistencies in the application of lighting across multiple objects or people.
- Point out areas where shading and reflections are inconsistent or artificially smooth.

1.2 Edges and Pixels

- Describe any smoothness or uniformity in edges that appears overly consistent.
- Identify unnatural edge blending or lack of differentiation in sharpness.
- Highlight any uniform pixel artifacts or blurring patterns typically associated with GANs.

1.3 Resolution and Compression

- Assess the image for uniform resolution across all areas that seems artificial.
- Identify any compression artifacts like repeating patterns or unusual pixel formations.

1.4 Reflections and Shadows

- Point out reflections and shadows that don't correspond realistically with light sources.
- Describe any unrealistic shadow directions or exaggerated reflection intensities.

1.5 Material Properties

- Evaluate how different materials interact with light, noting if reflections appear uniformly glossy or unrealistic.
- Identify unrealistic reflections or specular highlights on surfaces.

1.6 Physical Laws and Perspective

- Highlight any objects, body parts, or facial features that don't adhere to real-world physical laws.
- Describe inconsistencies in depth perception, proportions, and alignment.

2. Human Perception Indicators

2.1 Natural Imperfections

- Describe overly smooth or uniform textures on skin, fabric, and other surfaces.
- Point out the absence of minor blemishes or natural wear that would be expected in a real image.

2.2 Emotional Coherence

- Assess the coherence of facial expressions and body language, noting any inconsistencies.
- Describe facial expressions that appear uniform or lack nuanced emotion.

2.3 Environmental Integration

- Analyze how subjects and objects interact with their surroundings, highlighting unnatural elements.
- Identify objects or subjects that appear unnaturally isolated from their environment.

2.4 Temporal Consistency

- Point out elements that lack a coherent sense of time or motion.
- Identify any discrepancies in the consistency of temporal details.

2.5 Cultural and Contextual Coherence

- Highlight cultural or contextual elements that appear overly uniform or lack diversity.
- Analyze the coherence of attire, symbols, or social interactions, noting any inconsistencies.

2.6 Background Naturalness

- Describe background elements that appear unnaturally consistent or lack variation.
- Identify instances where background elements lack natural integration with the foreground.

2.7 Dynamic Range Realism

- Assess the image's dynamic range, pointing out signs of artificial enhancement.
- Explain how the highlights or shadows appear unrealistic or overly enhanced.

2.8 Facial Feature Consistency

- Describe any unusual or exaggerated facial features.
- Identify subtle facial expressions or features that do not align naturally with the rest of the image.

3. Comparative Analysis

Describe the most prominent indicators that reveal this image as a fake. Provide detailed reasons for your assessment, considering both technical analysis and human perception indicators discussed above.

Final Assessment

Provide an overall evaluation of how you determined this image is fake by synthesizing evidence from both technical analysis and human perception indicators. Clearly explain which aspects were most crucial in identifying the image as artificially generated.

- Be clear and descriptive, providing specific examples from the image to support your analysis.
- Highlight both obvious and subtle indicators of artificial generation, focusing on a holistic view of the image.

Figure 7. Prompts for fully synthetic images.

You are an AI visual analysis expert specializing in detecting tampered images. You will receive two images: the first is the tampered image, and the second is the mask indicating potentially tampered areas. A value of 1 (white) in the mask represents tampered areas, while a value of 0 (black) represents untampered areas.
Your task:
Analyze the image for signs of manipulation, focusing on two types of tampering:
a) Object Tampering: An entire object artificially inserted or replaced.
b) Partial Tampering: Part of an object modified or altered.
Present your analysis as if examining only one image, without referencing any external tools or additional visual aids. Do not mention the "mask" or related terms in your analysis.
Your analysis should start by identifying the type of tampering in the image (object/partial). Then, provide a detailed description of the tampering, focusing on the following aspects:
Tampering Type and Location:
Identify whether this is an object tampering (entire object inserted) or a partial tampering (part of an object modified).
Describe the location of the tampered area in natural language.
Provide both relative and absolute positions: a) Relative position: Describe the location in relation to other elements in the image (e.g., "above the crowd," "on the wall," "in the sky"). b) Absolute position: Describe the location within the entire image (e.g., "left side," "bottom right corner," "upper half").
Be specific and avoid ambiguous descriptions.
If multiple tampered areas are present, identify all locations in detail.
Tampered Content:
Describe in detail the content of the tampered area.
For object tampering, describe the entire inserted object, including characteristics like size, color, texture, and orientation.
For partial tampering, focus on the specific part of the object that has been modified, describing attributes like shape, size, and changes compared to other similar parts.
Include information about: a) Types of objects or parts. b) Number of objects (if applicable). c) Actions or poses of objects. d) Properties and characteristics of the tampered elements (e.g., material, texture, shading).
Mention whether the tampered content appears natural or whether any oddities make it stand out compared to surrounding objects or regions.
Visual Inconsistencies:
Analyze and describe any visible details that indicate tampering, such as: a) Lighting and Color inconsistencies: Look for unnatural lighting, shadows, or discrepancies in color that make the tampered area appear out of place. b) Edges and Texture anomalies: Evaluate if edges of the tampered region appear overly smooth, too sharp, or inconsistent with the rest of the image. Note any texture mismatches. c) Perspective and Proportions misalignments: Determine if the tampered object or region does not match the perspective or proportions of surrounding elements. d) Resolution and Quality differences: Identify any differences in resolution or image quality, such as pixelation or blurriness, in the tampered areas compared to the rest of the image. e) Contextual inconsistencies: Look for elements that seem out of context (e.g., an object that doesn't belong in the given scene or an improbable configuration). f) Shadows and Reflections irregularities: Check if the tampered elements cast shadows or reflections consistent with the rest of the scene. Lack of proper shadows or incorrect reflection angles can indicate manipulation. g) Boundary analysis (for partial tampering): Evaluate whether the boundary of the modified area is consistent with natural transitions in the image.
Additional Considerations:
When analyzing object tampering, compare the characteristics of the inserted object to those of surrounding objects to determine if it blends naturally.
When analyzing partial tampering, assess whether the modified parts match the object's original material, texture, and appearance.
Mention any subtle indicators of tampering, such as slight color variations, mismatched noise, or irregularities in object consistency.
Assess if the manipulation could be intended to mislead or convey false information. Consider the impact of the tampered area on the overall context of the image.
Important Instructions:
Provide a thorough and detailed analysis of the image, focusing on areas that appear manipulated or out of place.
Be specific in your descriptions, highlighting both obvious and subtle indicators of tampering.
Do not mention or refer to any additional images, masks, or analysis tools in your response.
Base your analysis solely on what you observe in the provided image.
Treat all information provided to you as part of the single image being analyzed.

Figure 8. Prompts for tampered images.

Can you identify if this image is real, full synthetic, or tampered? Please mask the tampered object/part if it is tampered.

Real Image

1. Consistency
1.1 Lighting and Color
Lighting is consistent, with natural light shading the group uniformly. Color tones are realistic, as expected in outdoor photos.
1.2 Edges and Pixels
Edges are smooth, especially around challenging areas like hair and clothing. No sharp or unnatural transitions suggest tampering.
1.3 Resolution and Compression
Resolution is uniform with no conflicting detail levels. Compression artifacts are consistent throughout.
1.4 Reflections and Shadows
Shadows and reflections align correctly with the light source, indicating a cohesive environment.
1.5 Material Properties
Textures (e.g., hats, clothing, skin) interact naturally with light, with believable reflections on noses and cheeks.
1.6 Physical Laws and Perspective
Proportions, placement, and spacing among subjects follow realistic perspective rules, enhancing authenticity.
2. Authenticity
2.1 Natural Imperfections
Minor imperfections (e.g., reddened cheeks, skin texture) add to the image's genuineness, often removed in manipulations.
2.2 Emotional Congruence
Joy and camaraderie are naturally expressed, creating an emotionally coherent scene.
2.3 Environmental Interaction
Subjects' interactions (e.g., overlapping arms) are convincingly real, difficult to reproduce through manipulation.
2.4 Temporal Consistency
The motion-blurred hand and genuine expressions suggest spontaneity, which is challenging to fabricate convincingly.
2.5 Cultural and Contextual Elements
Attire and accessories fit the context, suggesting a natural outdoor scene among youths.
2.6 Unintended Elements
Background details like untidy hair and partially visible faces support a candid, unstaged environment.
2.7 Dynamic Range
The dynamic range handles lighting well, with no unnatural "hot spots" or shadows.
2.8 Micro-Expressions
Subtle details, such as slight nose scrunching and blended smiles, add layers of authenticity, often missed in edited images.
Final Assessment
Based on technical consistency and human-centered authenticity, this image appears genuine. Consistent lighting, pixel integrity, and emotional authenticity suggest the image is untampered and a true representation of a captured moment among children.

Figure 9. Example output from SIDA when analyzing a real image.

CVPR
#6596

Can you identify if this image is real, full synthetic, or tampered? Please mask the tampered object/part if it is tampered.

Fully Synthetic Image

Technical Analysis
1.1 Lighting and Color
Lighting Consistency: The lighting is inconsistent; boats are uniformly lit despite a setting sun, indicating an artificial light source.
Color Saturation: Colors are overly vibrant, lacking the subtle variation typical of a natural sunset.
1.2 Edges and Pixels
Edges: Boat and mast edges are overly sharp, missing depth-related blurring.
Artifacts: Smudging around small details, like ropes, suggests manipulation.
1.3 Resolution and Compression
Resolution: Uniform across the scene, missing depth-based variations.
1.4 Reflections and Shadows
Shadows: Shadow directions are inconsistent with natural lighting, indicating manipulation.
1.5 Material Properties
Reflections: Highlights on water are too even; natural water would reflect with more variation.
1.6 Perspective
Perspective Accuracy: Boat sizes are inconsistent relative to distance, suggesting unrealistic spatial arrangement.
Human Perception Indicators
2.1 Natural Imperfections
Texture: Boat and sea textures are too smooth, lacking real-world granularity.
2.2 Environmental Integration
Object Interaction: Boats appear overlaid on the water without realistic interaction, such as ripples.
2.3 Temporal Consistency
Lighting: Sunlight consistency does not match the portrayed time of day.
2.4 Background Naturalness
Atmosphere: Distant hills are overly defined, lacking the atmospheric perspective seen in natural scenes.
2.5 Dynamic Range
Realism: The dynamic range is too high, with exaggerated detail in both glare and shadowed areas.
Comparative Analysis
Unnatural lighting, shadow inconsistencies, overly sharp edges, and hyper-real textures suggest a digitally generated or altered image rather than a natural photo.
Final Assessment
The image appears artificially generated or heavily edited. Inconsistent lighting, abnormal shadows, uniform sharpness, and unrealistic environmental interactions indicate it was likely constructed digitally.

Figure 10. Example output from SIDA when analyzing a fully synthetic image.

Can you identify if this image is real, full synthetic, or tampered? Please mask the tampered object/part if it is tampered.

Object Tampered Image

Analyzing the provided images, the type of tampering detected in this image is an "object tampered," which involves an entire object artificially inserted or replaced within the image.

1. Tampering Type and Location:

- **Relative Position:** The entire inserted fake object is the main soccer player who appears central in the frame, positioned between additional players in the background.
- **Absolute Position:** The tampered area covers the central portion of the image, extending through the middle horizontally and from the bottom to near the top vertically.

2. Tampered Content:

- **Type of Object: The** object is a young male soccer player.
- **Number of Objects:** There is one object identified as tampered.
- **Actions or Poses:** The soccer player is engaged in an active soccer play, dribbling the ball downwards with his foot slightly raised about to make contact with the ball.
- **Properties and Characteristics:** The player is wearing a standard white jersey paired with navy blue shorts, white socks, and dark-colored soccer cleats. His pose and outfit are entirely consistent with playing soccer.

3. Visual Inconsistencies:

- **Lighting and Color Inconsistencies:** The lighting on the soccer player closely matches that of the background, which lowers suspicion initially. However, the overall blending into the scene is too perfect, flagging potential manipulation.
- **Edges and Textures Anomalies:** The outline of the player is suspiciously clean and slightly too distinct relative to the background elements, suggesting digital insertion.
- **Perspective and Proportions Misalignments:** The proportions of the player align with the image, but a detailed examination of placement suggests possible discrepancies in scale and interaction with the grass and ball.
- **Resolution and Quality Differences:** There is a slight variance in resolution between the tampered player and the background, indicating possible manipulation.
- **Contextual Inconsistencies:** The player's integration within the field context looks fitting at a glance but upon deeper inspection, the contact point with the ball lacks tangible interaction effects (e.g., grass disturbance).
- **Shadows and Reflections Irregularities:** Shadows under the player do conform to lighting but might be overly pronounced or not perfectly aligned, raising further suspicions.
- **Boundary Analysis:** Keeping focus on the player's edges reveals a stark transition from the player to the background, hinting at retouching or insertion postures.

The analysis indicates that the soccer player in this image has been digitally inserted into this scene, demonstrating typical characteristics of an "object tampered."

Figure 11. Example output from SIDA when analyzing a tampered image.

Object	Replacements
Person	*Person, Baby, Boy, Girl, Man, Woman*
Dog	*Cat, Rabbit, Squirrel, Horse, Cow, Elephant, Giraffe, Monkey*
Cat	*Dog, Rabbit, Squirrel, Fox, Owl, Lion, Tiger, Cheetah*
Bird	*Owl, Duck, Chicken, Goose, Parrot, Eagle, Sparrow, Penguin*
Horse	*Sheep, Goat, Ox, Bull, Deer*
Elephant	*Giraffe, Zebra, Rhinoceros, Hippopotamus*

a)Animals Replacement

Object	Replacements
Bicycle	*Motorcycle, Scooter, Mountain Bike, Racing Bike*
Car	*Sports Car, Taxi, Police Car, Pickup Truck, Motorcycle, Bus*
Motorcycle	*Bicycle, Scooter, Moped, Dirt Bike*
Airplane	*Jet, Helicopter, Glider, Drone*
Bus	*Subway Train, Freight Train, Monorail*
Train	*Giraffe, Zebra, Rhinoceros, Hippopotamus*
Boat	*Sailboat, Yacht, Canoe, Kayak, Jet Ski*

b) Vehicles Replacement

Object	Replacements
Chair	*Armchair, Bench, Stool, Office Chair*
Couch	*Sofa, Loveseat, Recliner*
Bed	*Bunk Bed, Sofa Bed, Daybed*
Dining Table	*Coffee Table, Picnic Table, Desk*
TV	*Computer Monitor, Projector*
Laptop	*Tablet, Notebook, Desktop Computer*
Microwave	*Oven, Toaster, Stove*
Refrigerator	*Freezer, Mini Fridge*

c) Household Items Replacement

Object	Replacements
Pizza	*Calzone, Sandwich, Flatbread*
Sandwich	*Burger, Hot Dog, Wrap*
Cake	*Pie, Brownie, Cupcake*
Apple	*Pear, Banana, Grape*
Orange	*Lemon, Grapefruit, Tangerine*
Carrot	*Potato, Radish, Celery*
Donut	*Cookie, Cake, Muffin*

d) Food Items Replacement

Object	Replacements
Backpack	*School Bag, Hiking Bag, Briefcase*
Handbag	*Purse, Tote Bag, Clutch*
Hat	*Cap, Helmet, Beanie*
Traffic Light	*Stop Sign, Streetlight, Yield Sign*
Fire Hydrant	*Water Pump, Fire Hose*
Stop Sign	*Yield Sign, Speed Limit Sign*
Clock	*Wall Clock, Watch, Alarm Clock*
Teddy Bear	*Stuffed Toy, Doll, Plush Animal*

e) General Items Replacement

Figure 12. Object replacement directories for SID-Set.

Object	Attributes
Bathroom	*Clean, Messy, With towels, With toiletries, With a mirror*
Toilet	*Clean, Closed lid, Open lid, With a mat, With a paper holder*
Sink	*Full of water, With soap, Shiny, Dirty, Leaking*
Shower	*Running water, With curtain, With soap bottles, Glass door*
Bathtub	*Filled with water, Empty, With bubbles, With toys*
Mirror	*Foggy, Clean, With a frame, Reflecting light*
Vanity	*With makeup, With brushes, Organized, Messy*

a) Bathroom-Related Objects

Object	Attributes
Car	*Red, Blue, Dirty, Shiny, Parked, Moving*
Bicycle	*With basket, Without basket, Rusty, New, Red, Blue*
Bus	*Yellow, Red, With passengers, Empty, Old, Shiny*
Truck	*Pickup, Blue, Dirty, Carrying load, Empty*
Motorcycle	*Black, Red, With helmet, Without helmet, Parked*
Boat	*White, Small, Large, In water, Docked*
Train	*Moving, Stopped, With graffiti, Clean*

b) Vehicles

Object	Attributes
Dog	*Happy, Angry, Sleeping, Running, Barking*
Cat	*Curious, Sleepy, Angry, Playing, Sitting*
Horse	*Galloping, Standing, Brown, White, With saddle*
Bird	*Flying, Perched, Singing, Looking up, With spread wings*

c) Animals

Object	Attributes
Man	*Smiling, Wearing glasses, Sitting, Standing, Running*
Woman	*Happy, Wearing a hat, Holding a bag, Walking, Sitting*
Person	*Smiling, Frowning, Wearing sunglasses, Holding a book*
People	*Talking, Walking, Sitting, Gathered, Dancing*
Family	*Smiling, Posing, Sitting, Standing, Hugging*
Gentleman	*Wearing a suit, Smiling, Holding a cane, Walking*
Baby	*Sleeping, Crawling, Smiling, Wrapped in a blanket*
Father	*Holding a child, Smiling, Sitting, Walking*
Lady	*Wearing a dress, Sitting, Standing, Holding a purse*
Kids	*Playing, Running, Laughing, Sitting*
Girl	*Playing, Smiling, Holding a doll, Running*

d) People

Object	Attributes
Kitchen	*Clean, Messy, With utensils, With food on the counter*
Table	*Wooden, With a tablecloth, With plates, Empty, With food*
Oven	*Hot, Cool, Open, Closed, Dirty*
Microwave	*With food, Clean, Open, Closed*
Refrigerator	*Open, Closed, White, With magnets*
Counter	*Clean, With food, With utensils, Cluttered*
Stove	*With pots, On, Off, Dirty, Clean*
Dishwasher	*Open, Closed, With dishes, Empty*
Chair	*Wooden, Metal, With cushion, Without cushion, Broken*
Bench	*Wooden, Metal, Painted, With a person, Empty*
Bed	*With pillows, Unmade, Made, With a blanket, Without a blanket*
Couch	*With cushions, Without cushions, Red, Blue, Occupied*

e) Home-Related Items and Furniture

Object	Attributes
Cake	*Chocolate, With frosting, Sliced, Whole, With candles*
Bowl	*With fruit, Empty, With soup, With cereal*
Food	*Delicious, Half-eaten, On a plate, Served, Hot*
Fruit	*Fresh, Cut, In a bowl, Ripe, Mixed*
Pizza	*With extra cheese, With pepperoni, Half-eaten, Vegetarian*
Vase	*With flowers, Empty, Blue, Red, Ceramic*
Clock	*Wall-mounted, Digital, Analog, Showing noon, Showing midnight*
Skateboard	*Black, With stickers, In motion, New, Old*
Snowboard	*Blue, With designs, On snow, Leaning against a wall*
Kite	*Flying, On the ground, Colorful, Broken*
Frisbee	*Red, Blue, In the air, On the ground*
Toys	*Scattered, Organized, Plastic, Wooden*
Decorations	*Colorful, On the wall, Hanging, Festive*
Photographs	*Framed, Black and white, Color, Hanging*
Areas	*Green, With flowers, Muddy, Dry*

f) Food and Miscellaneous Objects

Figure 13. Attribute modification directories for SID-Set.

References

[1] Xiuli Bi, Bo Liu, Fan Yang, Bin Xiao, Weisheng Li, Gao Huang, and Pamela C. Cosman. Detecting generated images by real images only. *Arxiv*, 2023. 1

[2] You-Ming Chang, Chen Yeh, Wei-Chen Chiu, and Ning Yu. Antifakeprompt: Prompt-tuned vision-language models are fake image detectors. *Arxiv*, 2023. 1

[3] Chengbo Dong, Xinru Chen, Ruohan Hu, Juan Cao, and Xirong Li. Mvss-net: Multi-view multi-scale supervised networks for image manipulation detection. *IEEE TPAMI*, 2023. 1

[4] Joel Frank and Thorsten Holz. Cnn-generated images are surprisingly easy to spot...for now. *Arxiv*, 2021. 1

[5] Joel Frank, Thorsten Eisenhofer, Lea Schönherr, Asja Fischer, Dorothea Kolossa, and Thorsten Holz. Leveraging frequency analysis for deep fake image recognition. In *Proceedings of the 37th International Conference on Machine Learning (ICML)*, 2020. 1

[6] Xiao Guo, Xiaohong Liu, Zhiyuan Ren, Steven Grosz, Iacopo Masi, and Xiaoming Liu. Hierarchical fine-grained image forgery detection and localization. In *CVPR*, 2023. 1

[7] Zhengchao Huang, Bin Xia, Zicheng Lin, Zhun Mou, and Wenming Yang. FFAA: multimodal large language model based explainable open-world face forgery analysis assistant. *Arxiv*, abs/2408.10072, 2024. 1, 2

[8] Yan Ju, Shan Jia, Lipeng Ke, Hongfei Xue, Koki Nagano, and Siwei Lyu. Fusing global and local features for generalized ai-synthesized image detection. In *ICIP*, 2022. 1

[9] Xin Lai, Zhuotao Tian, Yukang Chen, Yanwei Li, Yuhui Yuan, Shu Liu, and Jiaya Jia. LISA: reasoning segmentation via large language model. In *CVPR*, 2024. 1, 2

[10] Jiawei Li, Fanrui Zhang, Jiaying Zhu, Esther Sun, Qiang Zhang, and Zheng-Jun Zha. Forgerygpt: Multimodal large language model for explainable image forgery detection and localization. *Arxiv*, 2024. 1, 2

[11] Tsung-Yi Lin, Michael Maire, Serge J. Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft COCO: common objects in context. In *ECCV*, 2014. 2

[12] Xiaohong Liu, Yaojie Liu, Jun Chen, and Xiaoming Liu. Pscc-net: Progressive spatio-channel correlation network for image manipulation detection and localization. *IEEE TCSVT*, 2022. 1, 2

[13] Zhengzhe Liu, Xiaojuan Qi, and Philip H. S. Torr. Global texture enhancement for fake face detection in the wild. In *CVPR*, 2020. 1

[14] Utkarsh Ojha, Yuheng Li, and Yong Jae Lee. Towards universal fake image detectors that generalize across generative models. In *CVPR*, 2023. 1

[15] Bryan A. Plummer, Liwei Wang, Chris M. Cervantes, Juan C. Caicedo, Julia Hockenmaier, and Svetlana Lazebnik. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. *IJCV*, 2017. 2

[16] Chuangchuang Tan, Yao Zhao, Shikui Wei, Guanghua Gu, and Yunchao Wei. Learning on gradients: Generalized artifacts representation for gan-generated images detection. In *CVPR*, 2023. 1

[17] Zhendong Wang, Jianmin Bao, Wengang Zhou, Weilun Wang, Hezhen Hu, Hong Chen, and Houqiang Li. DIRE for diffusion-generated image detection. In *ICCV*, 2023. 1

[18] Zhipei Xu, Xuanyu Zhang, Runyi Li, Zecheng Tang, Qing Huang, and Jian Zhang. Fakeshield: Explainable image forgery detection and localization via multi-modal large language models. *Arxiv*, 2024. 1, 2

[19] Kai Zhang, Lingbo Mo, Wenhu Chen, Huan Sun, and Yu Su. Magicbrush: A manually annotated dataset for instruction-guided image editing. In *NeurIPS*, 2023. 3