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

Zhenglin Huang[1] Jinwei Hu[1] Xiangtai Li[2 †] Yiwei He[1]
Xingyu Zhao[3] Bei Peng[1] Baoyuan Wu[4] Xiaowei Huang[1] Guangliang Cheng[1 †]

[1]University of Liverpool, UK [2]Nanyang Technological University, SG
[3] WMG, University of Warwick [4]The Chinese University of Hong Kong, Shenzhen, Guangdong, China
Project Page: https://hzlsaber.github.io/projects/SIDA/
[†] Corresponding author. E-mail: guangliang.cheng@liverpool.ac.uk xiangtai94@gmail.com

Abstract

*The rapid advancement of generative models in creating highly realistic images poses substantial risks for misinformation dissemination. For instance, a synthetic image, when shared on social media, can mislead extensive audiences and erode trust in digital content, resulting in severe repercussions. Despite some progress, academia has not yet created a large and diversified deepfake detection dataset for social media, nor has it devised an effective solution to address this issue. In this paper, we introduce the **S**ocial media **I**mage **D**etection data**Set** (SID-Set), which offers three key advantages: (1) **extensive volume**, featuring 300K AI-generated/tampered and authentic images with comprehensive annotations, (2) **broad diversity**, encompassing fully synthetic and tampered images across various classes, and (3) **elevated realism**, with images that are predominantly indistinguishable from genuine ones through mere visual inspection. Furthermore, leveraging the exceptional capabilities of large multimodal models, we propose a new image deepfake detection, localization, and explanation framework, named SIDA (**S**ocial media **I**mage **D**etection, localization, and explanation **A**ssistant). SIDA not only discerns the authenticity of images, but also delineates tampered regions through mask prediction and provides textual explanations of the model's judgment criteria. Compared with state-of-the-art deepfake detection models on SID-Set and other benchmarks, extensive experiments demonstrate that SIDA achieves superior performance among diversified settings. The code, model, and dataset will be released.*

1. Introduction

Recent advances in generative AI [9, 57, 75] have significantly improved the ability to generate highly realistic images, making it easier to create content that closely resembles real-world events. However, these advancements also bring new risks of malicious misuse, particularly in creating deceptive content aimed at misleading public opinion or distorting historical records. Such concerns have motivated the computer vision community to develop more sophisticated deepfake detection techniques. Contemporary methods [32, 70] primarily focus on assessing the authenticity of facial images (i.e., *real* or *fake*), while an emerging subset aims to detect and localize facial manipulations [21, 22]. These methods are typically trained on datasets containing real and fake images, aiming to detect images as real or fake, or to localize the tampered regions. Consequently, the quality and diversity of the datasets used for training and evaluation play a crucial role in achieving high accuracy in deepfake detection and localization. A well-curated dataset can enable models to learn nuanced features, improving robustness and generalization in real-world scenarios.

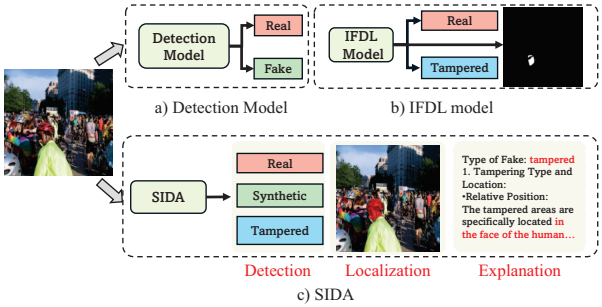

Figure 1. The framework comparisons. Existing deepfake methods (a-b) are limited to detection, localization, or both. In contrast, SIDA (c) offers a more comprehensive solution, capable of handling detection, localization, and explanation tasks.

However, existing deepfake detection and localization datasets face two main challenges: **1) Insufficient Diversity.** The majority of existing datasets for deepfake detection focus mainly on facial imagery [7, 24]. However, given the growing capabilities of generative AI, the issue of non-

facial image falsification on social media cannot be overlooked. While researchers [66, 86] have developed relatively large datasets based on ImageNet for image deepfake detection, these datasets typically consist of images from simple scenarios that do not specifically focus on social media. Additionally, they often utilize somewhat outdated image-generation techniques, which can result in less convincing forgeries that are easier for both humans and models to detect. Currently, there is a substantial lack of large-scale image deepfake datasets specifically designed for social media that leverage the latest generative methods. **2) Limited Comprehensiveness.** Existing datasets are typically suited either for deepfake detection or for tampered region localization [21–23, 73], focusing on specific types of generative methods or image manipulations. However, an ideal deepfake dataset should encompass a wide range of scenarios to reflect the complexity of real social media content, where fake images may be fully generated or manipulated through image editing strategies [75]. Furthermore, most existing datasets primarily focus on binary real/fake classification or tampered region localization, with limited emphasis on explaining the cues that models use to make these decisions.

To address these challenges, we introduce the **S**ocial media **I**mage **D**etection data**Set** (**SID-Set**), which consists of 300K images (*i.e.*, 100K real, 100K synthetic, and 100K tampered images), providing a comprehensive resource for the deepfake detection community. Additionally, we include textual descriptions explaining the model's judgment basis. As shown in Figures 2 and 4, synthetic and tampered images are indistinguishable to humans. In particular, **challenges** for the SID-Set include: 1) subtle alterations of just dozens of pixels; 2) natural-looking local manipulations; 3) complex scenes in datasets. To our knowledge, SID-Set is the first dataset of its scale with extensive annotations, making it the largest and most comprehensive dataset for social media deepfake detection to date. Compared to existing datasets in Table 1, SID-Set addresses the challenges of limited diversity and outdated generative techniques by providing a more comprehensive set of high-quality and diverse images. Accordingly, we propose a new VLMs-based deepfake detection framework, named the **S**ocial media **I**mage **D**etection, localization, and explanation **A**ssistant (**SIDA**), which achieves the state-of-the-art (SOTA) performance on SID-Set and generalizes effectively across other benchmarks. SIDA can serve as a baseline model on SID-Set, offering a new framework for tackling social media image deepfake detection and localization.

The main contributions of this paper are as follows:

- We establish SID-Set, a comprehensive benchmark for detecting, localizing, and explaining deepfakes in social media images, featuring multiple image types and extensive annotations. SID-Set holds the potential for advancing the field of deepfake detection and ensuring robust model performance in complex real-world scenarios.
- We propose SIDA, a new image deepfake detection, localization, and explanation framework that not only detects images with high accuracy but also localizes and explains potential manipulations, enhancing the transparency and utility of deepfake detection technologies.
- Extensive experiments demonstrate that SIDA effectively identifies and delineates tampered areas within images, supporting the development of more robust and interpretable deepfake detection systems. Notably, SIDA demonstrates superior or equivalent performance on the SID-Set and other benchmarks.

Table 1. Comparison with existing image deepfake datasets.

Dataset	Content	Data Source	Generator Year	Multiclasses	Masks	Explanation
OHImg [39]	Overhead	Google Map	2023	✗	✗	✗
FakeSpotter [63]	Face	CelebA, FFHQ	2020	✗	✓	✗
ForgeryNet [24]	Face	CREMA-D	2021	✗	✓	✗
DCFace [27]	Face	FFHQ	2023	✗	✗	✗
DFF [58]	Face	IMDB-WIKI	2023	✗	✗	✗
RealFaces [48]	Face	Prompts	2023	✗	✗	✗
M3Dsynth [87]	Biology	DDPM, CycleGAN	2023	✗	✓	✗
CNNSpot [64]	Object	SDXL	2020	✓	✗	✗
CiFAKE [3]	Object	CIFAR	2023	✓	✗	✗
CASIA 2.0 [12]	General	Corel	2022	✗	✓	✗
ArtiFact [53]	General	COCO, FFHQ, LSUN	2023	✗	✗	✗
IMD2020 [45]	General	Places2	2020	✗	✓	✗
AIGCD [84]	General	LSUN, COCO, FFHQ	2023	✗	✗	✗
GenImage [86]	General	ImageNet	2023	✗	✗	✗
SID-Set	General	COCO, Flickr30k, MagicBrush	2024	✓	✓	✓

2. Related Work

2.1. Image Deepfake Datasets

In the realm of deepfake detection, the primary focus has historically centered on the identification of facial deepfakes. Renowned datasets such as ForgeryNet [24], DeepFakeFace [58], and DFFD [7] have been pivotal in this area. As the field evolves, there is a growing shift among researchers towards exploring non-facial deepfake data. Advanced methodologies involving text-to-image or image-to-image generation techniques [76], utilizing GANs or the stable diffusion series, have facilitated the creation of expansive deepfake datasets like GenImage [86], HiFi-IFDL [21], and DiffForensics [66]. These datasets are characterized by their enlarged data volumes, diversified generation methodologies, and enriched annotation details. Furthermore, beyond the conventional real/fake annotations, certain datasets [21, 22, 73] now include more granular annotations. Table 1 delineates a detailed comparison among various deepfake datasets, highlighting key differences in generation scenarios and annotation types supported. It shows that SID-Set is particularly tailored towards social media data, incorporating the latest SOTA generation models, emphasizing high-quality production, and providing much more comprehensive and diverse annotations.

2.2. Image Deepfake Detection and Localization

Deepfake detection methods [4, 6, 38, 49, 62] are typically approached as classification tasks within the data-driven

Figure 2. SID-Set examples. The 1st row is the synthetic images, while the 2nd row shows tampered images. (Zoom in to view)

paradigm. These strategies primarily leverage diverse architectures [38, 49], including Convolutional Neural Networks (CNNs) and Transformers, to detect distinctive artifacts. Some scholars have attempted to achieve relatively high precision and generality by employing strategies such as employing techniques such as data augmentation [62], adversarial training [6], reconstruction [4], etc. On the other hand, some researchers [25, 60] have explored extracting features from the frequency domain for deepfake identification. Efforts [13, 65] have also been made to fuse features from both spatial and frequency domains to obtain a more comprehensive set of discriminative features for deepfake detection. Although these methods have shown some progress, they still struggle with issues of generalization. Furthermore, some scholars [20, 40, 44, 61, 80, 83] have gone beyond the basic classification between real and fake by gradually constructing datasets annotated with masks of locally tampered areas, thereby addressing both image deepfake detection and localization tasks. However, these datasets are concentrated mainly on facial data, with fewer datasets available for non-facial deepfake detection and localization, and even fewer large and public datasets for realistic social media data. As a result, our work aims to address these critical gaps by providing a new, comprehensive dataset that includes diverse manipulations beyond facial data, particularly focusing on social media images.

2.3. Large Multimodal Models

The progress in large language models (LLMs) [14–16, 41] and vision-language models (VLMs) [14–16, 29, 71, 78] has notably improved multimodal comprehension, seamlessly integrating visual and textual data. The LLaMA series [14–16] optimize language understanding with a compact, high-performance design using fewer parameters than prior models. LLaVA series [34, 35] enhance visual question answering by synchronizing visual features with textual data. Additionally, LISA series [29, 71] employ LLM for accurate image segmentation, merging visual perception with linguistic insights to precisely segment the targeted areas. Several grounding large multimodal models [50, 54, 55, 67, 68, 72, 74, 77, 81] have been proposed to localize the contents based on linguistic information.

Advancements [5, 10, 19, 69, 82] in integrating multimodal data such as visual and linguistic information have also significantly improved the performance of models in detecting and pinpointing deepfakes. For instance, AntifakePrompt [5] approaches deepfake detection by formulating it as a visual question answering problem, which adjusts soft prompts for InstructBLIP [10], enabling it to determine whether a query image is real or fake. ForgeryGPT [31] enhances image forgery detection and localization by integrating advanced forensic knowledge with a mask-aware forgery extractor that targets pixel-level fraud detection. As a concurrent work, FakeShield [69] leverages the capabilities of LLaVA to identify and localize altered regions while providing interpretable insights into the findings. Our method diverges significantly from existing approaches by creating the most extensive deepfake dataset tailored specifically for social media images with comprehensive annotations. Additionally, we also set a new standard for social media image deepfake detection by utilizing the visual interpretation strengths of VLMs to boost detection accuracy, pinpoint forgeries, and provide clearer explanations within a comprehensive framework.

3. Benchmark

3.1. Motivation

In the realm of deepfake detection and localization, research has predominantly focused on facial deepfakes due to their substantial societal impact. However, with the advancements in generative technology, the scope of deepfakes has extended beyond facial content to include non-facial manipulations. Historically, non-facial deepfakes were less prevalent, largely limited by technological constraints that produced low-quality, easily detectable forgeries. Although some datasets, such as GenImage [86] and AIGCD [84], have been constructed, they suffer from several limitations: **1)** They often utilize relatively outdated generative technologies, resulting in lower quality data easily distinguished by humans. **2)** They primarily focus on text-to-image or image-to-image generation, neglecting the need for data involving manipulations of specific regions, objects, or parts. Such tampered manipulations can be especially insidious as they introduce subtle misinformation, making existing SOTA deepfake detection methods less effective. **3)** They lack well-defined criteria for content authenticity, limiting their effectiveness in providing interpretative insights and educating the public on distinguishing synthetic content.

Addressing these limitations is crucial for improving the transparency and utility of deepfake detection systems. Furthermore, most existing datasets emphasize either fully synthetic or tampered images, as shown in Table 1, but in real applications, we don't know the image deepfake type in advance. Effective deepfake detection and localization methods should be capable of addressing both scenarios, as

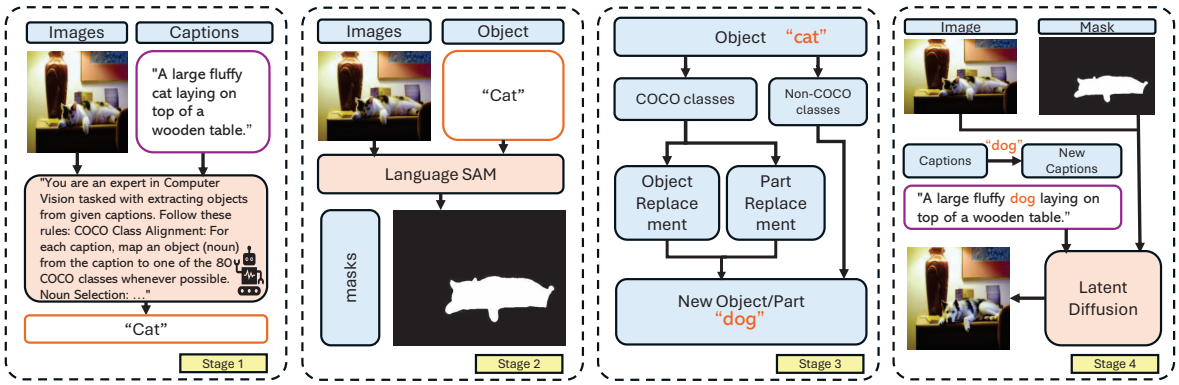

Figure 3. Tampered image generation pipeline: It consists of four stages—extracting objects from captions using GPT-4o, obtaining object masks with Language-SAM, setting up replacement dictionaries for generating tampered images, and generating new images using Latent Diffusion. This figure illustrates an example of object replacement (e.g., "cat" to "dog") and attribute modification.

social media images often involve complex combinations of synthetic and tampered content. Therefore, developing a comprehensive benchmark for detecting and localizing deepfakes in social media images is essential. We propose SID-Set, which encompasses comprehensive, high-quality annotations for detection and localization, along with detailed textual explanations of the judgment criteria.

3.2. Benchmark Construction

Data Details. To develop an effective benchmark for detecting and localizing images on social media, we created SID-Set, a dataset with real, synthetic, and tampered images reflecting diverse real-world scenarios. Our benchmark assesses whether models can differentiate among real, synthetic, and tampered images, as well as accurately identify altered regions in tampered images.

Real Images: 100K images from OpenImages V7[1], with a wide range of scenarios reflecting real-world diversity.

Synthetic Images: 100K images generated through FLUX [43], specifically designed to challenge identification due to their high-quality, highly realistic appearance.

Tampered Images: 100K tampered images, with specific objects or regions replaced or altered; the detailed generation process is shown in Figure 4.

Data Generation. To generate highly realistic synthetic images, we experimented with several open-source SOTA generative models, such as FLUX [43], Kandinsky 3.0 [1], SDXL [52], AbsoluteReality [42], and others. Following a review by five human experts of 1,000 images from each generative model, FLUX emerged as the top performer, producing highly convincing images that were indistinguishable from real ones to human experts. Consequently, we employed FLUX to create 100K synthetic images based on

[1] https://storage.googleapis.com/openimages/web/index.html

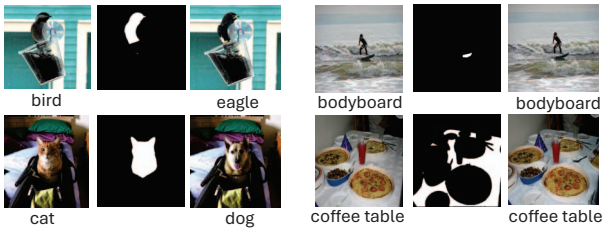

Figure 4. Examples of tampered images. (Zoom in to view)

the Flickr30k [51] and COCO [33]. The image tampering process depicted in Figure 4 follows four distinct stages, utilizing the COCO image as an example.

Stage 1: We extract objects from an image's caption using GPT-4o [47]. For instance, from the caption "A large fluffy cat laying on top of a wooden table", GPT-4o identifies relevant COCO class objects or retains nouns if no match exists. This extraction is documented in an "Image-Caption-Object" JSON file.

Stage 2: Employing Language-SAM [30], we generate masks for identified objects as training ground truth.

Stage 3: We establish dictionaries for full and partial image tampering using COCO classes for object replacement and attribute modifications, respectively. For example, replacing "dog" with animals like "cat" or adding attributes such as "happy" or "angry" to the "dog" class. For more details, please refer to the Appendix.

Stage 4: Utilizing Latent Diffusion [56], we modify captions and regenerate images, either replacing or retaining original objects based on availability. An example modification is altering "cat" to "dog" in the image caption.

In total, we generated 80,000 object-tampered images and 20,000 partially tampered images. To demonstrate the explainability of SIDA, we used GPT-4o to generate textual descriptions for the judgment basis of 3,000 images from

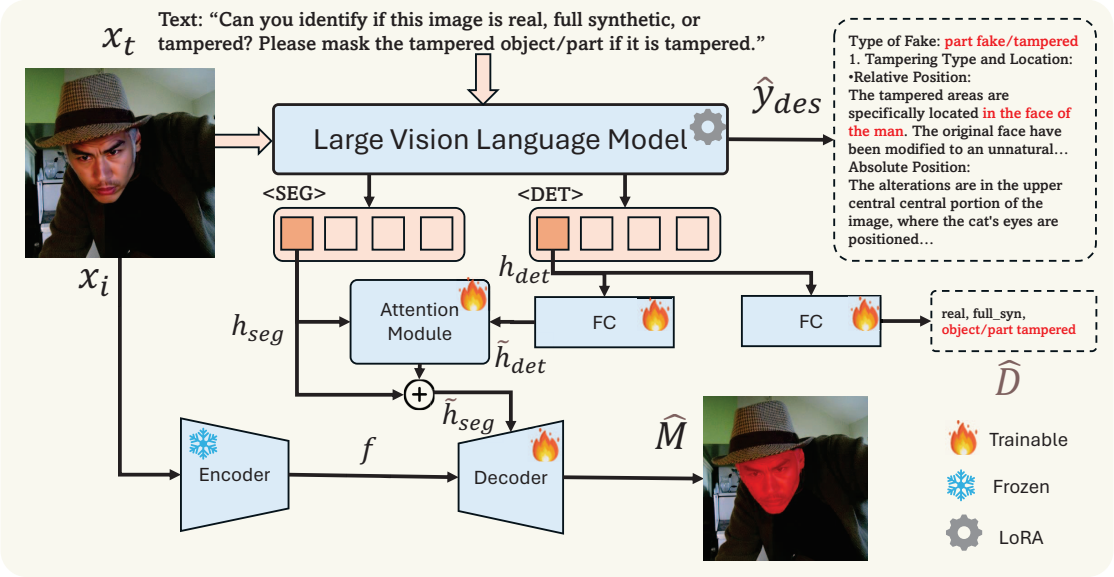

Figure 5. The pipeline of SIDA: Given an image x_i and the corresponding text input x_t, the last hidden layer for the <DET> token provides the detection result. If the detection result indicates a tampered image, SIDA extracts the <SEG> token to generate masks for the tampered regions. This figure shows an example where the man's face has been manipulated.

the SID-Set, divided equally among three types. Additionally, to ensure the realism of synthetic images, tampered images, and their textual descriptions, we engaged 5 annotation experts for quality control and adjustments. Detailed prompts and generation pipelines used to create descriptions for each type of image are provided in the Appendix.

4. Method

In this section, we first present the model architecture of SIDA in Section 4.1, followed by an introduction to the training process of our method in Section 4.2.

4.1. Architecture

Large vision-language models have demonstrated remarkable capabilities in understanding the alignment between textual and visual information. For instance, LLaVA [34] leverages language alone to achieve a comprehensive understanding of both visual and linguistic content. Building on LLaVA, LISA [29] extends this capability by providing fine-grained segmentation masks along with corresponding textual descriptions. However, to effectively detect and localize synthetic images, VLMs must not only be capable of multimodal understanding but also possess the ability to identify and segment manipulated regions, providing detailed explanations for both synthetic and tampered images.

To this end, we propose SIDA to tackle the task of synthetic image detection and tampered region localization. The pipeline of our method is illustrated in Figure 5. Inspired by previous approaches [28, 29, 85], we expand the original vocabulary of VLMs by adding two new tokens, <DET> and <SEG>, to enable the model to extract detection and segmentation information. Given an image x_i and a text prompt x_t, such as "Can you identify if this image is real, fully synthetic, or tampered? Please mask the tampered object/part if it is tampered." We feed them into the VLM. The VLM then outputs a text description $\hat{y}_{\text{des}}$, while the last hidden layer h_{hid} contain the <DET> and <SEG> tokens. This process can be formulated as follows:

$$\hat{y}_{\text{des}} = \text{VLM}(x_i, x_t). \tag{1}$$

Next, we extract the <DET> token from the last hidden layer h_{hid} to obtain h_{det}. The representation h_{det} is then passed through a detection head $F_{\texttt{det}}$ to determine whether the image is real, fully synthetic, or tampered. We denote the final detection result by $\hat{\texttt{D}}$:

$$\hat{\texttt{D}} = F_{\texttt{det}}(h_{\text{det}}), \tag{2}$$

$F_{\texttt{det}}$ is the detection head that processes the extracted <DET> representation h_{det} to produce the detection output.

If the detection result indicates that the image has been tampered with, SIDA will then predict a mask for the tampered regions. The h_{seg} feature is extracted from the hidden layer h_{hid}, similar to the extraction process for the <DET> token. Given that the <DET> token encapsulates crucial information that can aid in generating fine-grained segmentation masks, the representation h_{det} is transformed using a fully connected layer F to align with the dimensions of h_{seg}. To further capture the relationship between the h_{det} and h_{seg}

features, we apply a single-layer Multihead Attention, facilitating effective feature interaction and enhancing mask quality. The process can be formulated as follows:

$$\begin{aligned} \tilde{h}_{\text{det}} &= F(h_{\text{det}}), \\ \tilde{h}_{\text{seg}} &= \texttt{MSA}(\tilde{h}_{\text{det}}, h_{\text{seg}}), \\ \tilde{h}_{\text{seg}} &= \tilde{h}_{\text{seg}} + h_{\text{seg}}. \end{aligned} \tag{3}$$

In this formulation, we treat the detection features as the query, and the segmentation features as the key and value. A residual connection is employed to combine both the original and transformed information, thereby enhancing the representation for precise segmentation.

Finally, we employ a frozen image encoder $F_{\texttt{enc}}$ to extract visual features from the input image x_i, resulting in visual features f. The segmentation embedding $\tilde{h}_{\text{seg}}$ is then combined with the visual features f and fed into a decoder to produce the final mask $\hat{M}$. This can be formulated as:

$$\begin{aligned} f &= F_{\texttt{enc}}(x_i), \\ \hat{\texttt{M}} &= F_{\texttt{dec}}(\tilde{h}_{\text{seg}}, f). \end{aligned} \tag{4}$$

4.2. Training

Training Objectives. The training loss, $\mathcal{L}$, for the SIDA consists of three components: the detection loss $\mathcal{L}_{det}$, the text generation loss $\mathcal{L}_{txt}$, and the segmentation mask loss $\mathcal{L}_{mask}$. Initially, SIDA is trained in an end-to-end manner by employing the detection loss and the segmentation loss. For detection, we use CrossEntropy loss, while for the segmentation task, we use a weighted combination of binary cross-entropy (BCE) and DICE loss, with respective loss weights λbce and λ_{dice}. This can be formulated as:

$$\begin{aligned} \mathcal{L} &= \lambda_{det}\mathcal{L}_{det} + \lambda_{mask}\mathcal{L}_{mask}, \\ \mathcal{L}_{det} &= \mathcal{L}_{\text{CE}}(\hat{\texttt{D}}, \texttt{D}), \\ \mathcal{L}_{mask} &= \lambda_{bce}\mathcal{L}_{BCE}(\hat{\texttt{M}}, \texttt{M}) + \lambda_{dice}\mathcal{L}_{DICE}(\hat{\texttt{M}}, \texttt{M}). \end{aligned} \tag{5}$$

After completing the training phase, we proceed to fine-tune the SIDA model by utilizing detailed textural descriptions from 3,000 images as the ground truth, represented by y_{des}. This phase focuses on optimizing the text generation component, $\mathcal{L}_{txt}$, to improve its ability in textural interpretability. The final loss function is as follows:

$$\begin{aligned} \mathcal{L}_{txt} &= \mathcal{L}_{\text{CE}}(y_{\hat{des}}, y_{des}), \\ \mathcal{L}_{total} &= \lambda_{det}\mathcal{L}_{det} + \lambda_{mask}\mathcal{L}_{mask} + \lambda_{txt}\mathcal{L}_{txt}, \end{aligned} \tag{6}$$

where λ_{det}, λ_{mask}, and λ_{txt} are the weighting factors that balance the contributions of the detection, segmentation, and text generation losses, respectively.

Training Data. We utilize the SID-Set, consisting of 300k images, to train SIDA. To further enhance diversity, we incorporate the MagicBrush dataset [79] after filtering out low-quality images. The combined dataset supports robust training for both the detection and localization of synthetic content. Additionally, we generate descriptions for 3,000 randomly selected images using LLMs.

5. Experiments

Implementation Details. We choose LISA as the base large vision language model due to its strong capability for reasoning-based localization. We fine-tuned both LISA-7B-v1 and LISA-13B-v1 on the SID-Set using LoRA, setting α at 16 and the dropout rate at 0.05. The input images are resized to 1024×1024. The loss weights in Eq. (6) for the detection (λ_{det}), the text generation (λ_{txt}), and the localization (λ_{mask}) are set to 1.0, respectively. The localization loss weights in Eq. (5) for the λ_{bce} and λ_{dice} are set to 2.0 and 0.5, respectively. To determine the optimal weight configuration, we perform ablation studies as detailed in Section 5.5. During the detection and localization training stage, the image encoder is frozen, and all other modules are trainable. For the text generation stage, only vision-language models are fine-tuned using the LoRA strategy. The initial learning rate is set to 1×10^{-4}, with a batch size of 2 per device and a gradient accumulation step of 10. We use two NVIDIA A100 GPUs (40GB each). Training for SIDA-7B and SIDA-13B took 48 hours and 72 hours, respectively.

Evaluation Metrics. We evaluate detection using image-level accuracy and F1 scores. For forgery localization, our metrics include Area Under the Curve (AUC), F1 scores, and Intersection over Union (IoU).

5.1. Detection Evaluation

We compare SIDA against other SOTA deepfake detection methods on SID-Set, including CnnSpot [17], Antifake-Prompt [5], FreDect [18], Fusing [26], Gram-Net [37], UnivFD [46], LGrad [59], and LNP [2]. To ensure a fair comparison, we first evaluate these models on our dataset using their original pre-trained weights, then retrain them with the SID-Set to assess performance improvements. Table 2 demonstrates that SIDA achieves better or comparable results among all the evaluated methods. Notably, LGrad [59] achieves the highest accuracy and F1 score on tampered images after retraining, but this comes at the expense of lower performance in other metrics. Our analysis indicates that LGrad's high recall and false positive rates stem from its propensity to misclassify other types as tampered. The training details are provided in the Appendix.

5.2. Localization Results

Table 3 presents the forgery localization performance on the SID-Set. We selected PSCC-Net [36], MVSS-Net [11], and HIFI-Net [21] as representative IFDL methods. Additionally, we chose LISA [29] as a representative LLM due to

Table 2. Comparison of SIDA with other deepfake detection methods. Values outside parentheses are evaluated with open-source models directly on the SID-Set; values in parentheses indicate performance changes after training with the SID-Set. The best results are in bold.

Methods	Year	Real		Fully synthetic		Tampered		Overall	
		Acc	F1	Acc	F1	Acc	F1	Acc	F1
AntifakePrompt [5]	2024	64.8(↑24.1)	78.6(↑10.5)	93.8(↑3.7)	96.8(↑1.1)	30.8(↑60.1)	47.2(↑33.2)	63.1(↑29.1)	69.3(↑23.4)
CnnSpott [17]	2021	79.8(↑9.2)	88.7(↑2.1)	39.5(↑51.2)	56.6(↑31.5)	6.9(↑61.2)	12.9(↑51.1)	42.1(↑39.3)	69.6(↑20.7)
FreDect [18]	2020	83.7(↓37.7)	91.1(↓43.5)	16.8(↑44.1)	28.8(↑37.2)	11.9(↑25.2)	21.3(↑31.7)	37.4(↑33.6)	23.4(↑47.2)
Fusing [26]	2022	85.1(↑4.1)	**92.0**(↑0.7)	34.0(↑54.1)	50.7(↑38.4)	2.7(↑24.3)	5.3(↑26.1)	40.1(↑33.1)	29.1(↑40.3)
Gram-Net [37]	2020	70.1(↑19.1)	82.4(↑9.3)	93.5(↑4.4)	96.6(↑2.0)	0.8(↑89.1)	1.6(↑85.3)	55.0(↑37.1)	58.0(↑35.1)
UnivFD [46]	2023	68.0(↑0.3)	67.4(↑1.1)	62.1(↑24.3)	87.5(↑10.5)	64.0(↑28.5)	85.3(↑4.7)	64.0(↑21.7)	85.3(↑4.5)
LGrad [59]	2023	64.8(↓2.8)	78.6(↓2.5)	83.5(↓25.5)	91.0(↓23.7)	**6.8**(↑**92.3**)	**12.7**(↑**86.1**)	51.8(↑20.2)	55.5(↑23.9)
LNP [2]	2023	71.2(↓56.8)	83.2(↓60.2)	91.8(↓55.6)	95.7(↓60.1)	2.9(↑90.4)	5.7(↑88.9)	55.2(↓7.6)	58.2(↑4.1)
SIDA-7B	2024	89.1	91.0	**98.7**	98.6	91.2	91.0	93.5	**93.5**
SIDA-13B	2024	**89.6**	91.1	98.5	**98.7**	92.9	91.2	**93.6**	**93.5**

its segmentation reasoning capabilities. We used LISA-7B-v1 and fine-tuned it on SID-Set. The results indicate that SIDA achieves the best performance. We suppose that while LISA possesses strong general segmentation capabilities, it lacks the specific specialized features required to detect subtle manipulations, ultimately limiting the effectiveness of fine-tuning for precise forgery localization.

Table 3. Comparison between SIDA and other IFDL approaches. * indicates the use of the pre-trained model from the original paper due to unavailable training code.

Methods	Years	Tampered		
		AUC	F1	IOU
MVSS-Net* [11]	2023	48.9	31.6	23.7
HIFI-Net* [21]	2023	64.0	45.9	21.1
PSCC-Net [36]	2022	82.1	71.3	35.7
LISA-7B-v1 [29]	2024	78.4	69.1	32.5
SIDA-7B	2024	**87.3**	**73.9**	**43.8**

5.3. Robustness Study

We further evaluate the robustness of SIDA against common image perturbations found in social media, such as JPEG compression, resizing, and Gaussian noise. Table 4 shows our model's performance on the SID-Set under six degradation scenarios: JPEG compression (with quality levels of 70 and 80), resizing (with scaling factors of 0.5 and 0.75), and Gaussian noise (with variances of 5 and 10). Despite not being explicitly trained on degraded data, SIDA demonstrates resilience to these low-level distortions. The model's stable performance against common social media perturbations highlights its robustness and practical applicability.

5.4. Test on Other Benchmark

In this stage, we evaluate SIDA on DMimage [8] dataset to assess its generalization capabilities. We compare SIDA with CNNSpot [17], Fusing [26], Gram-Net [37], LNP [2], UnivFD [46], and AntifakePrompt [5]. For these methods, we use the original hyperparameter settings and pre-trained weights provided by the authors. The results in Table 5 show that SIDA achieves superior performance, demonstrating its strong adaptability.

Table 4. Performance of SIDA under different perturbations.

	Detection		Localization		
	ACC	F1	AUC	F1	IOU
JPEG 70	89.4	90.1	86.2	71.8	42.3
JPEG 80	88.7	89.5	85.8	71.1	41.7
Resize 0.5	89.3	91.1	86.8	72.5	43.2
Resize 0.75	89.9	91.6	87.1	73.0	43.5
Gaussian 10	86.9	89.3	84.1	70.2	41.0
Gaussian 5	88.4	89.9	85.3	71.0	41.5
SIDA-7B	93.5	93.5	87.3	73.9	43.8

Table 5. Comparison with other deepfake detection methods on DMimage [8]. We used original weights for each method.

Methods	Real		Fake		Overall	
	Acc	F1	Acc	F1	Acc	F1
CNNSpot [17]	87.8	88.4	28.4	44.2	40.6	43.3
Gram-Net [37]	62.8	54.1	78.8	88.1	67.4	79.4
Fusing [26]	87.7	86.1	15.5	27.2	40.4	36.5
LNP [2]	63.1	67.4	56.9	72.5	58.2	68.3
UnivFD [46]	89.4	88.3	44.9	61.2	53.9	60.7
AntifakePrompt [5]	91.3	92.5	89.3	**91.2**	90.6	91.2
SIDA-7B	**92.9**	**93.1**	**90.7**	91.0	**91.8**	**92.4**

5.5. Ablation Study

Attention Module. We conducted ablation experiments to assess the importance of the attention module. Variants included removing the attention module entirely and replacing it with fully connected (FC) layers. Results in Table 6 show that removing the attention module or replacing

Table 6. Ablation study results for attention module in SIDA.

	Detection		Localization		
	ACC	F1	AUC	F1	IOU
FC	91.1	90.3	84.3	71.6	38.9
w/o Attention	90.3	89.9	84.1	71.3	38.8
SIDA	93.5	93.5	87.3	73.9	43.8

it with FC significantly reduces performance, underscoring the critical role of attention in enhancing feature interaction and improving detection and localization accuracy.

Training Weights. SIDA training utilizes weighted losses to balance task contributions. In the detection and localization stages, detection loss is adjusted by a weight λ_{det}, and localization loss by binary cross-entropy (BCE) and DICE losses, with weights λ_{bce} and λ_{dice} respectively. For our experiments, we set λ_{det} to 1, λ_{bce} to 2.0, and λ_{dice} to 0.5 to maintain balance between detection and localization, enhancing model stability and performance. We summarize the outcomes of various weight configurations in Table 7.

Table 7. Different weight configurations in SIDA training.

λ_{det}	λ_{bce}	λ_{dice}	Acc	F1 Score
1.0	2.0	0.5	93.56	91.01
1.0	4.0	1.0	93.49	90.86

5.6. Qualitative Results

In this section, we present examples of SIDA's output for tampered images, showcasing its detection, localization, and explanation capabilities. SIDA accurately identifies tampered regions and provides explanations for its decisions. We also include some challenging failure cases where SIDA was unable to accurately detect the tampered regions, highlighting areas for future improvement. Additional visualization results are available in the Appendix.

Figure 6. Visual results of SIDA on tampered images.

6. Conclusion and Discussions

In this work, we present SID-Set for social image deepfake detection, localization, and explanation tasks, consisting of 100k real images, 100k fully synthetic images, and 100k tampered images. Furthermore, we propose a new VLMs-based deepfake detection framework, SIDA, to address these tasks. SIDA demonstrates its ability to detect fake types, localize tampered regions, and provide explanations for its decisions. We believe that the integration of VLMs into deepfake detection tasks, as demonstrated by SIDA, offers promising new avenues for future research.

Although the development of the SID-Set and the introduction of the SIDA framework have yielded favorable outcomes in deepfake detection tasks, we recognize some potential limitations and will optimize them in future research.

Dataset Size. While SID-Set includes 100k fully synthetic and 100k tampered images, the complexity of real social media environments demands a larger dataset. Therefore, expanding the dataset with additional images is a crucial objective for future research.

Data Domain. Other methods often generate social media images that lack authenticity or are easily identifiable. Consequently, we used only FLUX to generate fake images due to its superior quality. However, depending exclusively on one method may lead to issues of data skew, although this was not significantly evident in our experiments. Such skew could potentially impair performance on diverse datasets. Moving forward, we plan to explore additional generative methods and integrate various other generation techniques to produce a more varied and higher-quality set of images.

Localization Results. Although SIDA demonstrates relatively strong performance on our dataset, there is still room for improvement. Certain tampered regions are not reliably detected, underscoring the need for further advancements.

Acknowledgments

This work is supported by The Alan Turing Institute (UK) through the project 'Turing-DSO Labs Singapore Collaboration' (SDCfP2\100009).

References

[1] Vladimir Arkhipkin, Andrei Filatov, Viacheslav Vasilev, Anastasia Maltseva, Said Azizov, Igor Pavlov, Julia Agafonova, Andrey Kuznetsov, and Denis Dimitrov. Kandinsky 3.0 technical report. *Arxiv*, 2023. 4

[2] 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. 6, 7

[3] Jordan J. Bird and Ahmad Lotfi. CIFAKE: image classification and explainable identification of ai-generated synthetic images. *IEEE Access*, 2024. 2

[4] Junyi Cao, Chao Ma, Taiping Yao, Shen Chen, Shouhong Ding, and Xiaokang Yang. End-to-end reconstruction-classification learning for face forgery detection. In *CVPR*, 2022. 2, 3

[5] You-Ming Chang, Chen Yeh, Wei-Chen Chiu, and Ning Yu. Antifakeprompt: Prompt-tuned vision-language models are fake image detectors. *Arxiv*, 2023. 3, 6, 7

[6] Liang Chen, Yong Zhang, Yibing Song, Lingqiao Liu, and Jue Wang. Self-supervised learning of adversarial example: Towards good generalizations for deepfake detection. In *CVPR*, 2022. 2, 3

[7] Harry Cheng, Yangyang Guo, Tianyi Wang, Liqiang Nie, and Mohan S. Kankanhalli. Diffusion facial forgery detection. In *Proceedings of the 32nd ACM International Conference on Multimedia*, 2024. 1, 2

[8] Riccardo Corvi, Davide Cozzolino, Giada Zingarini, Giovanni Poggi, Koki Nagano, and Luisa Verdoliva. On the detection of synthetic images generated by diffusion models. In *ICASSP*, 2023. 7

[9] Florinel-Alin Croitoru, Vlad Hondru, Radu Tudor Ionescu, and Mubarak Shah. Diffusion models in vision: A survey. *IEEE TPAMI*, 2023. 1

[10] Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale Fung, and Steven C. H. Hoi. Instructblip: Towards general-purpose vision-language models with instruction tuning. In *NeurIPS*, 2023. 3

[11] 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. 6, 7

[12] Jing Dong, Wei Wang, and Tieniu Tan. CASIA image tampering detection evaluation database. In *ISIP*, 2013. 2

[13] Junxian Duan, Yuang Ai, Jipeng Liu, Shenyuan Huang, Huaibo Huang, Jie Cao, and Ran He. Test-time forgery detection with spatial-frequency prompt learning. *IJCV*, 2024. 3

[14] Abhimanyu Dubey et al. The llama 3 herd of models. *Arxiv*, 2024. 3

[15] Hugo Touvron et al. Llama: Open and efficient foundation language models. *Arxiv*, 2023.

[16] Hugo Touvron et al. Llama 2: Open foundation and fine-tuned chat models. *Arxiv*, 2023. 3

[17] Joel Frank and Thorsten Holz. Cnn-generated images are surprisingly easy to spot...for now. *Arxiv*, 2021. 6, 7

[18] 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. 6, 7

[19] Zhaopeng Gu, Bingke Zhu, Guibo Zhu, Yingying Chen, Ming Tang, and Jinqiao Wang. Anomalygpt: Detecting industrial anomalies using large vision-language models. In *AAAI*, 2024. 3

[20] Fabrizio Guillaro, Davide Cozzolino, Avneesh Sud, Nicholas Dufour, and Luisa Verdoliva. Trufor: Leveraging all-round clues for trustworthy image forgery detection and localization. In *CVPR*, 2023. 3

[21] 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, 2, 6, 7

[22] Xiao Guo, Xiaohong Liu, Iacopo Masi, and Xiaoming Liu. Language-guided hierarchical fine-grained image forgery detection and localization. *Arxiv*, 2024. 1, 2

[23] Ruidong Han, Xiaofeng Wang, Ningning Bai, Yihang Wang, Jianpeng Hou, and Jianru Xue. Hdf-net: Capturing homogeny difference features to localize the tampered image. *IEEE TPAMI*, 2024. 2

[24] Yinan He, Bei Gan, Siyu Chen, Yichun Zhou, Guojun Yin, Luchuan Song, Lu Sheng, Jing Shao, and Ziwei Liu. Forgerynet: A versatile benchmark for comprehensive forgery analysis. In *CVPR*, 2021. 1, 2

[25] Yonghyun Jeong, Doyeon Kim, Youngmin Ro, and Jongwon Choi. Frepgan: Robust deepfake detection using frequency-level perturbations. In *AAAI*, 2022. 3

[26] 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. 6, 7

[27] Minchul Kim, Feng Liu, Anil K. Jain, and Xiaoming Liu. Dcface: Synthetic face generation with dual condition diffusion model. In *CVPR*, 2023. 2

[28] Wonjae Kim, Bokyung Son, and Ildoo Kim. Vilt: Vision-and-language transformer without convolution or region supervision. In *Proceedings of the 38th International Conference on Machine Learning (ICML)*, 2021. 5

[29] 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. 3, 5, 6, 7

[30] lang-sam team. Lang-segment-anything. https://github.com/luca-medeiros/lang-segment-anything, 2024. Accessed: 2024-11-15. 4

[31] 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. 3

[32] Li Lin, Neeraj Gupta, Yue Zhang, Hainan Ren, Chun-Hao Liu, Feng Ding, Xin Wang, Xin Li, Luisa Verdoliva, and Shu Hu. Detecting multimedia generated by large AI models: A survey. *Arxiv*, 2024. 1

[33] 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. 4

[34] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. In *NeurIPS*, 2023. 3, 5

[35] Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. In *CVPR*, 2024. 3

[36] 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. 6, 7

[37] Zhengzhe Liu, Xiaojuan Qi, and Philip H. S. Torr. Global texture enhancement for fake face detection in the wild. In *CVPR*, 2020. 6, 7

[38] Momina Masood, Marriam Nawaz, Khalid Mahmood Malik, Ali Javed, Aun Irtaza, and Hafiz Malik. Deepfakes generation and detection: state-of-the-art, open challenges, countermeasures, and way forward. *Appl. Intell.*, 2023. 2, 3

[39] Brandon B. May, Kirill Trapeznikov, Shengbang Fang, and Matthew C. Stamm. Comprehensive dataset of synthetic and manipulated overhead imagery for development and evaluation of forensic tools. In *IH&MMSec*, 2023. 2

[40] Changtao Miao, Qi Chu, Zhentao Tan, Zhenchao Jin, Wanyi Zhuang, Yue Wu, Bin Liu, Honggang Hu, and Nenghai Yu. Multi-spectral class center network for face manipulation detection and localization. *Arxiv*, 2023. 3

[41] Shervin Minaee, Tomás Mikolov, Narjes Nikzad, Meysam Chenaghlu, Richard Socher, Xavier Amatriain, and Jianfeng Gao. Large language models: A survey. *Arxiv*, 2024. 3

[42] Absolutereality model team. Absolutereality model repository. https://huggingface.co/jochemstoel/absolutereality-model-repository, 2024. Accessed: 2024-11-15. 4

[43] Flux model team. Flux model. https://huggingface.co/black-forest-labs/FLUX.1-dev, 2024. Accessed: 2024-11-15. 4

[44] Dat Nguyen, Nesryne Mejri, Inder Pal Singh, Polina Kuleshova, Marcella Astrid, Anis Kacem, Enjie Ghorbel, and Djamila Aouada. Laa-net: Localized artifact attention network for quality-agnostic and generalizable deepfake detection. In *CVPR*, 2024. 3

[45] Adam Novozámský, Babak Mahdian, and Stanislav Saic. IMD2020: A large-scale annotated dataset tailored for detecting manipulated images. In *WACVW*, 2020. 2

[46] Utkarsh Ojha, Yuheng Li, and Yong Jae Lee. Towards universal fake image detectors that generalize across generative models. In *CVPR*, 2023. 6, 7

[47] OpenAI. GPT-4 technical report. *Arxiv*, 2023. 4

[48] Lorenzo Papa, Lorenzo Faiella, Luca Corvitto, Luca Maiano, and Irene Amerini. On the use of stable diffusion for creating realistic faces: from generation to detection. In *IWBF*, 2023. 2

[49] Gan Pei, Jiangning Zhang, Menghan Hu, Guangtao Zhai, Chengjie Wang, Zhenyu Zhang, Jian Yang, Chunhua Shen, and Dacheng Tao. Deepfake generation and detection: A benchmark and survey. *Arxiv*, 2024. 2, 3

[50] Zhiliang Peng, Wenhui Wang, Li Dong, Yaru Hao, Shaohan Huang, Shuming Ma, and Furu Wei. Kosmos-2: Grounding multimodal large language models to the world. *Arxiv*, 2023. 3

[51] 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. 4

[52] Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. SDXL: improving latent diffusion models for high-resolution image synthesis. In *ICLR*, 2024. 4

[53] Md Awsafur Rahman, Bishmoy Paul, Najibul Haque Sarker, Zaber Ibn Abdul Hakim, and Shaikh Anowarul Fattah. Artifact: A large-scale dataset with artificial and factual images for generalizable and robust synthetic image detection. In *ICIP*, 2023. 2

[54] Hanoona Abdul Rasheed, Muhammad Maaz, Sahal Shaji Mullappilly, Abdelrahman M. Shaker, Salman H. Khan, Hisham Cholakkal, Rao Muhammad Anwer, Eric P. Xing, Ming-Hsuan Yang, and Fahad Shahbaz Khan. Glamm: Pixel grounding large multimodal model. In *CVPR*, 2024. 3

[55] Zhongwei Ren, Zhicheng Huang, Yunchao Wei, Yao Zhao, Dongmei Fu, Jiashi Feng, and Xiaojie Jin. Pixellm: Pixel reasoning with large multimodal model. In *CVPR*, 2024. 3

[56] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *CVPR*, 2022. 4

[57] Xincheng Shuai, Henghui Ding, Xingjun Ma, Rongcheng Tu, Yu-Gang Jiang, and Dacheng Tao. A survey of multimodal-guided image editing with text-to-image diffusion models. *Arxiv*, 2024. 1

[58] Haixu Song, Shiyu Huang, Yinpeng Dong, and Wei-Wei Tu. Robustness and generalizability of deepfake detection: A study with diffusion models. *Arxiv*, 2023. 2

[59] 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. 6, 7

[60] Chuangchuang Tan, Yao Zhao, Shikui Wei, Guanghua Gu, Ping Liu, and Yunchao Wei. Frequency-aware deepfake detection: Improving generalizability through frequency space domain learning. In *AAAI*, 2024. 3

[61] Dragos-Constantin Tântaru, Elisabeta Oneata, and Dan Oneata. Weakly-supervised deepfake localization in diffusion-generated images. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, 2024. 3

[62] Chengrui Wang and Weihong Deng. Representative forgery mining for fake face detection. In *CVPR*, 2021. 2, 3

[63] Run Wang, Felix Juefei-Xu, Lei Ma, Xiaofei Xie, Yihao Huang, Jian Wang, and Yang Liu. Fakespotter: A simple yet robust baseline for spotting ai-synthesized fake faces. In *IJCAI*, 2020. 2

[64] Sheng-Yu Wang, Oliver Wang, Richard Zhang, Andrew Owens, and Alexei A Efros. Cnn-generated images are surprisingly easy to spot...for now. In *CVPR*, 2020. 2

[65] Yuan Wang, Kun Yu, Chen Chen, Xiyuan Hu, and Silong Peng. Dynamic graph learning with content-guided spatial-frequency relation reasoning for deepfake detection. In *CVPR*, 2023. 3

[66] Zhendong Wang, Jianmin Bao, Wengang Zhou, Weilun Wang, Hezhen Hu, Hong Chen, and Houqiang Li. DIRE for diffusion-generated image detection. In *ICCV*, 2023. 2

[67] Cong Wei, Haoxian Tan, Yujie Zhong, Yujiu Yang, and Lin Ma. Lasagna: Language-based segmentation assistant for complex queries. *Arxiv*, 2024. 3

[68] Zhuofan Xia, Dongchen Han, Yizeng Han, Xuran Pan, Shiji Song, and Gao Huang. GSVA: generalized segmentation via multimodal large language models. In *CVPR*, 2024. 3

[69] 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. 3

[70] Zhiyuan Yan, Yong Zhang, Xinhang Yuan, Siwei Lyu, and Baoyuan Wu. Deepfakebench: A comprehensive benchmark of deepfake detection. In *NeurIPS*, 2023. 1

[71] Senqiao Yang, Tianyuan Qu, Xin Lai, Zhuotao Tian, Bohao Peng, Shu Liu, and Jiaya Jia. An improved baseline for reasoning segmentation with large language model. *Arxiv*, 2023. 3

[72] Haoxuan You, Haotian Zhang, Zhe Gan, Xianzhi Du, Bowen Zhang, Zirui Wang, Liangliang Cao, Shih-Fu Chang, and Yinfei Yang. Ferret: Refer and ground anything anywhere at any granularity. In *ICLR*, 2024. 3

[73] Zeqin Yu, Jiangqun Ni, Yuzhen Lin, Haoyi Deng, and Bin Li. Diffforensics: Leveraging diffusion prior to image forgery detection and localization. In *CVPR*, 2024. 2

[74] Yuhang Zang, Wei Li, Jun Han, Kaiyang Zhou, and Chen Change Loy. Contextual object detection with multimodal large language models. *Arxiv*, 2023. 3

[75] Fangneng Zhan, Yingchen Yu, Rongliang Wu, Jiahui Zhang, Shijian Lu, Lingjie Liu, Adam Kortylewski, Christian Theobalt, and Eric P. Xing. Multimodal image synthesis and editing: The generative AI era. *IEEE TPAMI*, 2023. 1, 2

[76] Chenshuang Zhang, Chaoning Zhang, Mengchun Zhang, and In So Kweon. Text-to-image diffusion models in generative AI: A survey. *Arxiv*, 2023. 2

[77] Hao Zhang, Hongyang Li, Feng Li, Tianhe Ren, Xueyan Zou, Shilong Liu, Shijia Huang, Jianfeng Gao, Leizhang, Chunyuan Li, and Jainwei Yang. Llava-grounding: Grounded visual chat with large multimodal models. In *ECCV*, 2024. 3

[78] Jingyi Zhang, Jiaxing Huang, Sheng Jin, and Shijian Lu. Vision-language models for vision tasks: A survey. *IEEE TPAMI*, 2024. 3

[79] Kai Zhang, Lingbo Mo, Wenhu Chen, Huan Sun, and Yu Su. Magicbrush: A manually annotated dataset for instruction-guided image editing. In *NeurIPS*, 2023. 6

[80] Lingzhi Zhang, Zhengjie Xu, Connelly Barnes, Yuqian Zhou, Qing Liu, He Zhang, Sohrab Amirghodsi, Zhe Lin, Eli Shechtman, and Jianbo Shi. Perceptual artifacts localization for image synthesis tasks. In *ICCV*, 2023. 3

[81] Tao Zhang, Xiangtai Li, Hao Fei, Haobo Yuan, Shengqiong Wu, Shunping Ji, Change Loy Chen, and Shuicheng Yan. Omg-llava: Bridging image-level, object-level, pixel-level reasoning and understanding. In *NeurIPS*, 2024. 3

[82] Yue Zhang, Ben Colman, Ali Shahriyari, and Gaurav Bharaj. Common sense reasoning for deep fake detection. *Arxiv*, 2024. 3

[83] Yi Zhang, Changtao Miao, Man Luo, Jianshu Li, Wenzhong Deng, Weibin Yao, Zhe Li, Bingyu Hu, Weiwei Feng, Tao Gong, and Qi Chu. MFMS: learning modality-fused and modality-specific features for deepfake detection and localization tasks. In *Proceedings of the 32nd ACM International Conference on Multimedia*, 2024. 3

[84] Nan Zhong, Yiran Xu, Zhenxing Qian, and Xinpeng Zhang. Rich and poor texture contrast: A simple yet effective approach for ai-generated image detection. *Arxiv*, 2023. 2, 3

[85] Luowei Zhou, Hamid Palangi, Lei Zhang, Houdong Hu, Jason J. Corso, and Jianfeng Gao. Unified vision-language pre-training for image captioning and VQA. In *AAAI*, 2020. 5

[86] Mingjian Zhu, Hanting Chen, Qiangyu Yan, Xudong Huang, Guanyu Lin, Wei Li, Zhijun Tu, Hailin Hu, Jie Hu, and Yunhe Wang. Genimage: A million-scale benchmark for detecting ai-generated image. In *NeurIPS*, 2023. 2, 3

[87] Giada Zingarini, Davide Cozzolino, Riccardo Corvi, Giovanni Poggi, and Luisa Verdoliva. M3DSYNTH: A dataset of medical 3d images with ai-generated local manipulations. In *ICASSP*, 2024. 2