technologies

MDPI

Article

RADAR: Reasoning AI-Generated Image Detection for Semantic Fakes

Haochen Wang [1,†], Xuhui Liu [2,†], Ziqian Lu [3,*], Cilin Yan [2], Xiaolong Jiang [4], Runqi Wang [5] and Efstratios Gavves [1]

[1] Informatics Institute, University van Amsterdam, 1098 XH Amsterdam, The Netherlands; h.wang3@uva.nl (H.W.); egavves@uva.nl (E.G.)
[2] School of Automation Science and Electrical Engineering, Beihang University, Beijing 100191, China; xuhui_cc@126.com (X.L.); clyan@buaa.edu.cn (C.Y.)
[3] School of Computer Science and Technology (School of Artificial Intelligence), Zhejiang Sci-Tech University, Hangzhou 310018, China
[4] Xiaohongshu Inc., Beijing 100029, China; laieg@xiaohongshu.com
[5] School of Computer Science & Technology, Beijing Jiaotong University, Beijing 100091, China; rqwang@bjtu.edu.cn
* Correspondence: ziqianlu@zju.edu.cn; Tel.:+86-1530-658-1718
† These authors contributed equally to this work.

Abstract

As modern generative models advance rapidly, AI-generated images exhibit higher resolution and lifelike details. However, the generated images may not adhere to world knowledge and common sense, as there is no such awareness and supervision in the generative models. For instance, the generated images could feature a penguin walking in the desert or a man with three arms, scenarios that are highly unlikely to occur in real life. Current AI-generated image detection methods mainly focus on low-level features, such as detailed texture patterns and frequency domain inconsistency, which are specific to certain generative models, making it challenging to identify the above-mentioned general semantic fakes. In this work, (1) we propose a new task, reasoning AI-generated image detection, which focuses on identifying semantic fakes in generative images that violate world knowledge and common sense. (2) To benchmark the new task, we collect a new dataset Spot the Semantic Fake (STSF). STSF contains 358 images with clear semantic fakes generated by three different modern diffusion models and provides bounding boxes as well as text annotations to locate the fakes. (3) We propose RADAR, a reasoning AI-generated image detection assistor, to locate semantic fakes in the generative images and output corresponding text explanations. Specifically, RADAR contains a specialized multimodal LLM to process given images and detect semantic fakes. To improve the generalization ability, we further incorporate ChatGPT as an assistor to detect unrealistic components in grounded text descriptions. The experiments on the STSF dataset show that RADAR effectively detects semantic fakes in modern generative images.

Keywords: AI-generated detection; multimodal large language models; semantic-level fake detection

check for updates

Academic Editor: Loris Nanni

Received: 25 April 2025
Revised: 20 June 2025
Accepted: 27 June 2025
Published: 2 July 2025

Citation: Wang, H.; Liu, X.; Lu, Z.; Yan, C.; Jiang, X.; Wang, R.; Gavves, E. RADAR: Reasoning AI-Generated Image Detection for Semantic Fakes. *Technologies* **2025**, *13*, 280. https://doi.org/10.3390/technologies13070280

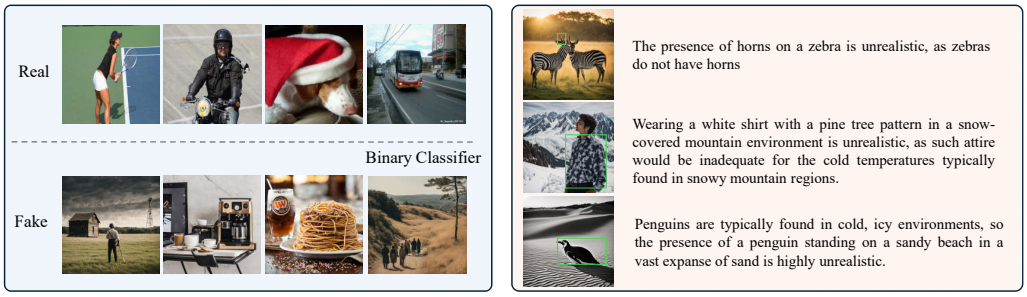
Copyright: © 2025 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

1. Introduction

The development of generative models opens up new possibilities in entertainment, media, and content creation and helps with our daily lives. However, the advancement of generative models has led to remarkable progress in creating photorealistic images, also significantly lowering the expertise and effort needed to produce fake images. This

capability extends beyond static images to include videos, audio, and even text. The ease of producing such realistic fakes also raises significant ethical and security concerns, as it becomes increasingly difficult to detect manipulated content. This evolution necessitates advancements in detection technologies and regulatory measures to mitigate potential misuse while harnessing the creative potential of these powerful tools.

Diffusion-based generation models usually focus on low-level details of the generation image. Therefore, even though the resolution, texture, and detail of AI-generated images continuously improve, there is no semantic-level supervision or guidance in modern generative models, leading to a common issue that the generative images violate world knowledge and common sense. As shown in Figure 1b, the AI-generated images contain penguins walking in the desert, zebras with horns, and persons with shirts on snow mountains. The current AI-generated image detection methods lack the ability to capture such semantic-level issues. Specifically, artifact-feature-based methods [1,2] train deep learning networks as binary classifiers on both real and AI-generated images. Spectrum-feature-based methods [3] transform images into a spectrum domain using Fourier transform before performing recognition. Those methods are more adept at identifying detailed flaws and inconsistencies in the frequency domain while having no reasoning ability to understand high-structured information, such as the relation between objects, events, and causality for semantic error detection.



(a) General AI-Generated Image Detection (b) Reasoning AI-Generated Image Detection

Figure 1. Comparison between (**a**) general AI-generated image detection and (**b**) reasoning AI-generated image detection. (**a**) General AI-generated image detection aims to recognize AI-generated images from real images with binary classification. (**b**) Reasoning AI-generated image detection aims to detect regions with semantic fakes violating world knowledge, and give text descriptions.

In this work, we propose reasoning AI-generated image detection, a new task focusing on identifying semantic fakes in generative images that violate world knowledge and common sense. This task pioneers a new perspective on detecting AI-generated images, complementing traditional methods by uncovering semantic-level fakes in generative images. To accomplish this task, the detection models are supposed to locate the semantic-fake regions and generate corresponding explanations in natural language, thereby potentially helping to improve the quality of generative images with the text feedback.

To tackle this new task, we introduce RADAR: a reasoning AI-generated image detection assistor, to detect semantic fakes in AI-generated images. Specifically, we conduct instruction tuning of a multimodal LLM to generate boxes of semantic fakes and corresponding text descriptions. To improve the generalization ability over novel semantic fakes, we incorporate ChatGPT-3.5 as a general AI assistor, which detects unrealistic components in the grounded descriptions given by the multimodal LLM. Finally, the results are obtained by combining the box–text responses from both the multimodal LLM and ChatGPT assistor.

To benchmark reasoning AI-generated image detection, we assemble a dataset STSF, Spot the Semantic Fake, which contains 358 images with clear semantic-level issues. Specifically, we use three modern generative models and tools, SDXL [4], SDXL-Turbo [5], and

Stable-Diffusion-3 [6] to generate more than 20,000 initial images. We then manually select 358 images that violate world knowledge and common sense, annotate the fake regions with bounding boxes, and provide corresponding text descriptions, as shown in Figure 1b. Compared with general AI-generated image detection datasets [7–9], the STSF dataset focuses on semantic-level fakes in the generative images. We also provide detailed regional and textual information instead of solo binary signals, enhancing the interpretability of the recognition and regional detection, and promoting the development of generative models as a consequence. Experimentally, RADAR shows a promising ability to detect semantic errors in our collected STSF dataset.

To sum up, our contributions are as follows:

- We propose a new task, reasoning AI-generated image detection, which focuses on identifying semantic fakes in generative images that contradict world knowledge and common sense.
- We collect a new dataset STSF that contains AI-generated images with clear semantic fakes to be a proper testbed for reasoning AI-generated image detection.
- We introduce RADAR: a reasoning AI-generated image detection assistor, a specialized multimodal LLM to detect semantic errors in AI-generated images with ChatGPT assistor. Experiments on STSF show that RADAR could detect semantic fakes in generative images, which is an extensive complement to traditional AI-generated image detection methods focusing on low-level details.

2. Related Work

Diffusion Models

Recent advancements in generative models, especially diffusion models, have revolutionized the field of image generation, facilitating the production of realistic and superior-quality images and videos. Diffusion models (DMs) [10,11], generating samples by gradually denoising images corrupted by Gaussian noise, have demonstrated their superior performance in various image synthesis, such as text-to-image (T2I) generation in Stable Diffusion [12], GLIDE [13], DALLE-2 [14], and Imagen [15], speech synthesis [16,17], and super-resolution in SR3 [18] and IDM [19]. Moreover, DMs have been applied to text-to-3D synthesis in DreamFusion [20] and subsequent research [21–25], video synthesis [26–30], text-to-motion generation [31], etc. On the other hand, DiT [32] provides better potential for diffusion models based on vision transformers. Pixart [33], Stable-Diffusion-3 [6], Hunyuan-DiT [34], and Lumina-T2X [35] further enhance the stability and scalability of DiT, expanding its applicability across various complex visual tasks.

Multimodal Large Language Models

Recent research has witnessed a surge of interest in multimodal large language models (MLLMs). Benefiting from the strong reasoning capabilities of large language models (LLMs), MLLMs exhibit promising performance in multimodal generation tasks, such as image captioning and Visual Question Answering (VQA). Collaborative models, such as Visual ChatGPT [36], MM-REACT [37], and HuggingGPT [38], often use pretrained tools to translate visual information into text inputs of LLMs. End-to-end trained models including BLIP series [39,40], LLaVA series [41], and MiniGPT-4 [42] enable the LLMs to comprehend the visual inputs by training projection layers to map the image contents to the text space of LLMs.

AI-Generated Image Detection

Early works primarily centered around face forgery detection, which usually focuses on capturing the low-level artifacts [43–45] and frequency domain inconsistencies [46–48].

Similar strategies have also been applied to general fake image detection [1,3,49–51]. More recently, DIRE [52] utilizes the reconstruction error between the input and its reconstruction counterpart by a diffusion model to train the detector. RIGID [53] proposes a training-free method by comparing the representation similarity of the real and fake images. In general, these methods concentrate on capturing the artifacts of image features (e.g., image stitching and cropping), ignoring specific semantic-level fakes (e.g., logical or physical errors). In contrast, our RADAR leverages the inference ability of an LLM to not only locate semantic forgery in generated images but also provide semantic level explanations.

To benchmark AI-generated image detection, many works use advanced generative models to collect synthetic datasets. Early datasets [54,55] are primarily constructed using manipulated facial images. For general datasets, CNNSpot [1] introduces ProGAN [56] to obtain synthetic data. DeepArt [57], DE-FAKE [58], and CiFAKE [59] employ advanced diffusion models to produce more realistic images. Nonetheless, these datasets only have simple classification labels, indicating whether the given images are real or fake. In contrast, our collected STSF is specially collected with distinct semantic-level issues, complete with annotations in the form of bounding boxes and textual explanations.

3. Method

3.1. Task Setup

Given a set $\{I_i\}_{i=1}^{N}$ containing AI-generated and real images, traditional AI-generated image detection aims to distinguish AI-generated images from real images. In other words, a binary classifier $y = f_\theta(I)$ is required to recognize real images $(y = 0)$ and AI-generated images $(y = 1)$. Given the binary labels, the classifiers are usually trained end-to-end with neural networks [43–45,48]. Therefore, the models tend to learn shortcuts that recognize specific artifacts from certain generative models that may not even be visible to the naked eye.

We claim that a common issue of current generative models is the lack of reasoning ability. In other words, the generative models have no awareness of world knowledge and common sense. As a result, the generative images may possess semantic-level fakes. Figure 2 shows cases, a man features three hands and a person is floating in the air, which are unlikely to happen in the real world. In this work, we propose reasoning AI-generated image detection, which aims to locate those semantic unrealistic regions in given images, and also provide corresponding text explanations:

$$\{\mathbf{s}, \mathbf{b}, \mathbf{t}\} = f_\theta(I), \tag{1}$$

where $\mathbf{b} \in \mathbb{R}^{L\times 4}$ is the bounding boxes of the target regions, and $\mathbf{t}$ is the corresponding text explanations of each box. $\mathbf{s}$ is a binary score to indicate the confidence of the box–text response, which could be aggregated and used for binary classification as in traditional methods.

To sum up, traditional AI-generated image detection aims to recognize generative images with binary classifiers, without the requirements of interpretability, while reasoning AI-generated image detection focuses on detecting regions with semantic fakes in generative images, which could benefit the generative models to produce more realistic images given the feedback of the semantic fakes.

We design three sub-tasks for reasoning AI-generated image detection. (a) Binary Classification aims to identify generative images from real images, the same as the traditional AI-generated image detection. We utilize AUC and AP as the evaluation metrics, following existing work [53]. (b) Regional Semantic Fake Detection aims to locate the regions with semantic issues in given images. We adopt average precision (AP) and average recall (AR)

as the evaluation metrics. (c) Text Generation aims to output reasonable text descriptions for semantic fakes. As for the evaluation, we adopt LAVE (LLM-Assisted VQA Evaluation) [60] to score the similarity of generated text and ground truth text with the help of large language models.

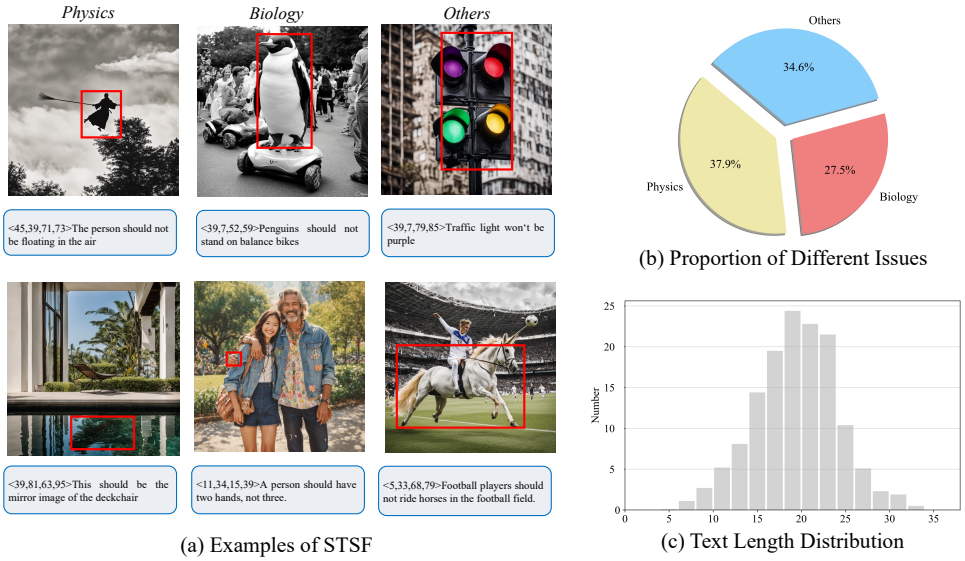


Figure 2. Statistics of STSF dataset. (**a**) Annotation cases in STSF dataset. (**b**) Proportions of three different groups of semantic issues, (**c**) Distribution of annotated text length for each bounding box.

3.2. STSF Dataset

3.2.1. Collection and Annotation

To benchmark reasoning AI-generated image detection, we establish a dataset named STSF (Spot the Semantic Fake), which contains real images and generative images with semantic errors. Specifically, we start with collecting 2000 text prompts from widely used image caption datasets, COCOCaption [61] and LLaVA-Instruct-158K [41], as the inputs of generative models. Then we obtain corresponding generative images given text inputs with three modern diffusion-based models SDXL [4], SDXL-Turbo [5], and Stable-Diffusion-3 [6]. Some of the generative images have no obvious semantic fakes, which is not the objective of this work. Therefore, we manually select generative images with obvious semantic fakes that violate world knowledge and common sense, and remove the rest as a consequence, forming a collection of 358 generative images with obvious semantic fakes. The resolution is set to 1024×1024. To provide detailed information for the semantic fakes, we thoroughly annotate the bounding boxes of the fake regions as well as the text descriptions, which increases the interpretability compared with traditional AI-generated image detection datasets. We further randomly sample 968 real images from MS-COCO [62] to serve as negative samples in binary classification task.

3.2.2. Dataset Statistics

Specifically, the length of the text descriptions varies from 4 to 34 words. We group all data into several commonly seen issues, as shown in Figure 2. (i) Physics: events, interactions, and states violate laws of physics, such as a person floating in the air, and a mirror showing different objects; (ii) Biology: organisms have unreal structures or are in the wrong habitats, such as a person with three hands, and a penguin on a balance bike. (iii) Others: images violate all other common senses, such as a traffic light having 4 different colors, and a man riding a horse in the football field.

Compared to traditional AI-generated image detection benchmarks, such as LSUN-BEDROOM [7], Genimage [8], and Wildfake [9], we provide multimodal annotations rather

than binary labels to intuitively explain unrealistic parts, which could benefit the generative models as a consequence with the feedback from the semantic fakes. For training and evaluation purposes, STSF is split into a train set and a validation set, which contain 843 and 1157 images, respectively.

3.3. RADAR

After constructing the STSF dataset, we here present RADAR, a reasoning AI-generated image detection assistor, to detect the semantic fakes in generative images, as shown in Figure 3. Specifically, we conduct instruction tuning of a pretrained multimodal LLM (MLLM) on the STSF dataset. Then the finetuned MLLM is enabled to generate bounding boxes and text descriptions of semantic fakes in given images. To improve the generalization ability on novel semantic fakes, we propose to incorporate ChatGPT as an AI assistor to detect general unrealistic components in grounded image descriptions given by MLLM.

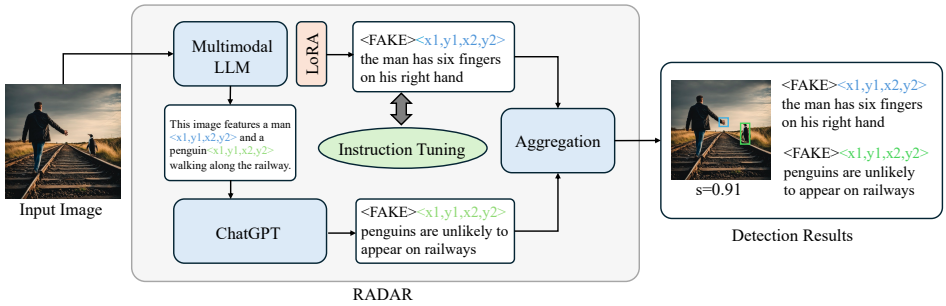


Figure 3. Overview of RADAR. Given an input image, a finetuned multimodal LLM is adopted to predict the bounding boxes and text explanation of unrealistic regions, which is accomplished by instruction tuning on the STSF training set. To improve the generalization ability of RADAR on various types of semantic fakes, we utilize ChatGPT as an assistor with more adequate knowledge backups and reasoning ability, to discover semantic fakes in the grounded descriptions given by the multimodal LLM. RADAR then aggregates and refines results from both the multimodal LLM and ChatGPT to obtain the final results. The image classification score s is obtained by averaging the scoring of each box–text response.

3.3.1. Instruction Tuning of MLLM

To detect semantic fakes in AI-generated images, knowledge awareness and reasoning ability are required. Inspired by the achievements of recent large language models, we adopt an instruction tuning scheme over a pretrained MLLM to accomplish reasoning AI-generated image detection. We start with building a visual-to-text (V2T) mapping to align the visual inputs with text inputs. Specifically, an EVA [63] model is adopted as the visual backbone to extract visual features from the input images. Then a trainable linear projection layer is used to project visual features into text domains. The projected features are then concatenated with the text prompts as the inputs to a pretrained LLaMA2-chat(7B) [64]. We follow the LLaMA-2 conversation template design and adapt it for the reasoning AI-generated image detection template. The template is shown as follows:

```
USER: <Img> <ImgFeature> </Img> Please locate the regions that look unreal
(violate world knowledge and common sense), and explain why.
RADAR: <flag ><x1, y1, x2, y2 ><reason>.
```

In this template, `USER` refers to the human role, and `RADAR` refers to assistor role. `<flag>` identifies if the image is AI-generated. For AI-generated images, `<flag>` is set to `<FAKE>`, while for real images, `<flag>` is set to `<REAL>`. `<FAKE>` and `<REAL>` are two special tokens;

we conduct softmax over logits of those two tokens to obtain the classification score s of this box–text response:

$$s = \text{softmax}(logits_{[real]}), logits_{[fake]}) \quad (2)$$

`<x1, y1, x2, y2>` represents the coordinate top-left corner and bottom-right corner of the semantic-fake regions. The coordinates are encoded by integers normalized in range [0, 100]. The `<reason>` is the text description of the semantic fake. As for real images or images without semantic level issues, we set `<x1, y1, x2, y2>` and `<reason>` to be empty.

Based on the designed template, we conduct the instruction tuning on the STSF training set. Specifically, we use the pretrained MiniGPT-v2 [65] after stage 2 training as the starting point of instruction tuning. During the instruction tuning, we combine STSF and datasets used in the MiniGPT-v2 stage 3 for joint training to maintain the general ability of the MLLM, such as grounding and conversation.

3.3.2. ChatGPT Assistor

Reasoning AI-generated image detection is a completely new task for the MLLM, and there are not enough training samples to cover all kinds of semantic fakes during the instruction tuning process. Moreover, the original reasoning ability of the LLM is harmed during the alignment with the visual domain in the multimodal LLM, which makes it challenging to generalize to unseen semantic issues requiring complex knowledge reasoning ability. To improve the generalization ability of RADAR on various semantic fakes, we introduce ChatGPT as a zero-shot assistor to help locate complex semantic issues. The process is shown in Figure 4. As a general LLM, such as ChatGPT, cannot take images as input, we first ask the multimodal LLM to provide detailed and grounded descriptions in text. After obtaining the grounded description of the images, the LLM assistor could understand the details of objects and events in the given images and find illogical components with only text as inputs.

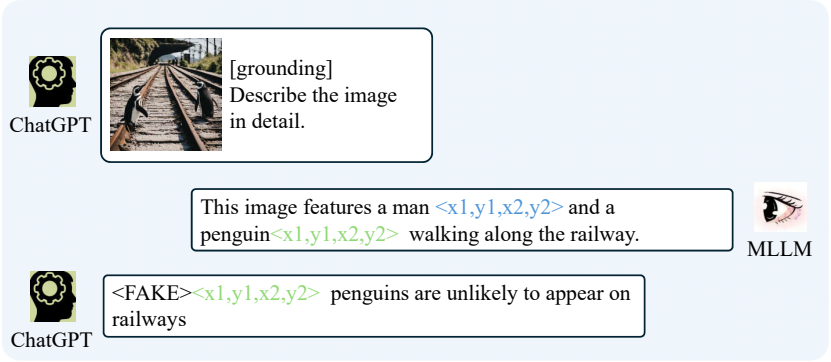


Figure 4. Demonstration of ChatGPT assistor.

Result Aggregation Finally, the outputs of the ChatGPT assistor are aggregated with the original outputs from the MLLM. Specifically, NMS (Non-Maximum Suppression) is applied to remove duplicated boxes. Then we utilize ChatGPT to refine the text generated by the MLLM. The final image classification score is obtained by averaging the classification score s of all valid box–text responses. By combining the results from the MLLM and ChatGPT assistor, RADAR could locate more complex semantic fakes requiring comprehensive knowledge reasoning, performing robust generalization ability over unseen semantic fakes.

4. Experiments

4.1. Implementation Details

Following MiniGPT-v2 [65], RADAR adopts LLaMA2-7B [64] as the LLM and a ViT vision encoder to extract visual tokens. We initialize RADAR with the parameters of pretrained MiniGPT-v2 after stage two. During the entire training process, the visual backbone remains frozen. We finetune the LLM with LoRA [66] on our collected STSF dataset, in which the rank r is set to 64. The images are resized to 448×448 during the training. We use AdamW [67] optimizer with a cosine learning rate scheduler to train RADAR. The learning rate is set to 1×10^{-5}. We train on $8 \times$ A100 GPUs for 50,000 iterations with a global batch size of 48 for the STSF dataset. As for the results aggregation, we generate five responses from the MLLM and utilize all results from ChatGPT. Then NMS is conducted to remove duplicated boxes. The NMS threshold is set to 0.5. During training, the datasets used in MiniGPT-v2, such as LLaVA-Instruction [41], Flicker 30k [68], and RefCOCO [69], are combined with the STSF dataset for joint training to maintain the general ability of the MLLM, such as text generation and grounded image caption. As for regional semantic fake detection evaluation, the threshold of IoU between predicted boxes and ground truth boxes is fixed to a small value of 0.3, as we focus on rough semantic fake regions.

4.2. Results on STSF Datasets

We conduct extensive experiments on the STSF dataset, see Table 1. In terms of binary classification, two existing AI-generated image detection methods, DIRE [52] and RIGID [53], achieve 71.2/73.1 and 78.4/77.6 on the STSF validation set, while RADAR achieves 81.4/82.0, outperforming GIRE and RIGID by 10.2/8.9 and 3.0/4.4. Moreover, the existing AI-generated image detection methods can only generate a binary classification signal, without the ability to detect the semantic fake regions and generate corresponding text descriptions. RADAR outperforms traditional object detection methods Faster-RCNN [70] and DETR [71] by over 20 AP/AR in terms of detection performance. The reason is that object detection methods lack general knowledge reasoning ability, and thus suffer from poor generalization ability when novel issues occur. Some visualizations of RADAR are shown in Figure 5. For instance, in the last image of the first row, RADAR outputs "The turtle looks unreal". The bounding box precisely locates the turtle's head, revealed to be a lion's head, which is the specific attribute inconsistency that RADAR identified. This localization helps to pinpoint the visual evidence for the detected implausibility.

Table 1. The performance comparison on the STSF dataset. Methods above the horizontal lines are traditional detection-based methods finetuned on the STSF training set, while the methods below the horizontal lines are based on LLMs. IT means instruction tuning.

Methods	Backbone	Classification		Detection		Text
		AUC	AP	AP	AR	LAVE
DIRE	ADM	71.2	73.1	-	-	-
RIGID	DINOv2	78.4	77.6	-	-	-
Faster-RCNN	Resnet-50	57.6	56.9	9.4	10.2	-
DETR	Resnet-50	62.3	63.5	12.2	13.7	-
MLLM	LLaMA-7B	53.4	51.2	8.4	7.5	4.8
MLLM(IT)	LLaMA-7B	78.5	77.6	27.3	25.4	28.0
RADAR	LLaMA-7B	81.4	82.0	35.3	38.5	42.1

To further demonstrate the rationality of the proposed RADAR, we conduct a small-scale survey, inviting 50 participants to independently rate the reasonableness of the outputs

from RADAR and ChatGPT on a scale of 0 to 5 (where 0 indicates completely unreasonable and 5 indicates completely reasonable). As shown in the following Figure 6, compared to vision GPT, RADAR achieves superior results and its performance is close to the gold standard of human judgment. This indicates that the proposed RADAR can accurately reflect the semantic-level fakes of AI-generated images.

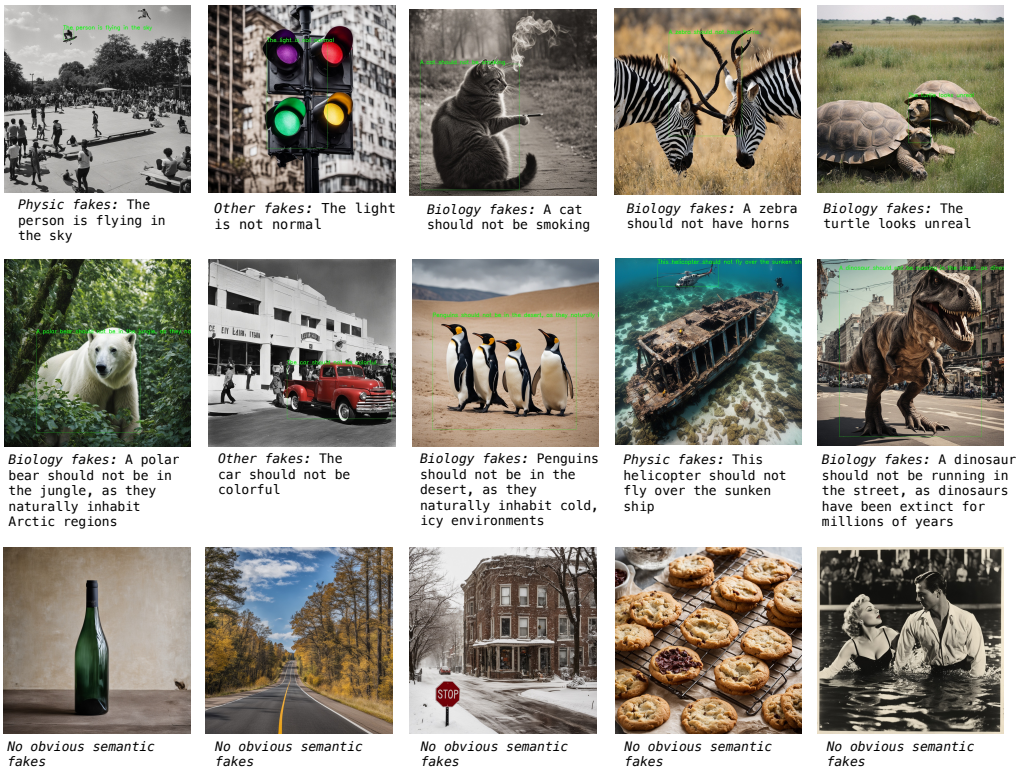


Figure 5. Visualizations of RADAR's outputs on the STSF dataset. The third row shows some generative images without obvious semantic fakes. Italicized text indicates the semantic issues.



Figure 6. Manual scoring of model output rationality on STSF dataset.

4.3. Ablation Studies

w/o ChatGPT Assistor As shown in Table 1, with the ChatGPT assistor, RADAR achieves more than a 10 percent improvement in terms of detection AP/AR and text similarity. We

claim that the pretrained multimodal large language models, such as LLaVA [41], focus on visual–language alignment and have poor text generation ability. The raw text generation and reasoning ability of large language models are harmed during the visual–language joint training. As shown in the first row of Figure 7, without the ChatGPT assistor, RADAR tends to generate simple text descriptions with the same pattern in the training set, while the detected boxes tend to be commonly seen issues such as extra hands and legs. With the help of the ChatGPT assistor, RADAR is enabled to generate reasonable text descriptions and locate novel mistakes without seeing any similar fakes during the training. However, the ChatGPT assistor fails to detect fake regions that will not appear in the grounded captions. For example, as for the cases in Figure 7, the MLLM will generate grounded captions like: "A black man wearing a yellow T-shirt is playing tennis", while the extra legs/arms are not included. Therefore, the ChatGPT assistor has no ability to locate such issues. To sum up, the combination of MLLM and ChatGPT will help to locate both detailed and high-level semantic fakes.

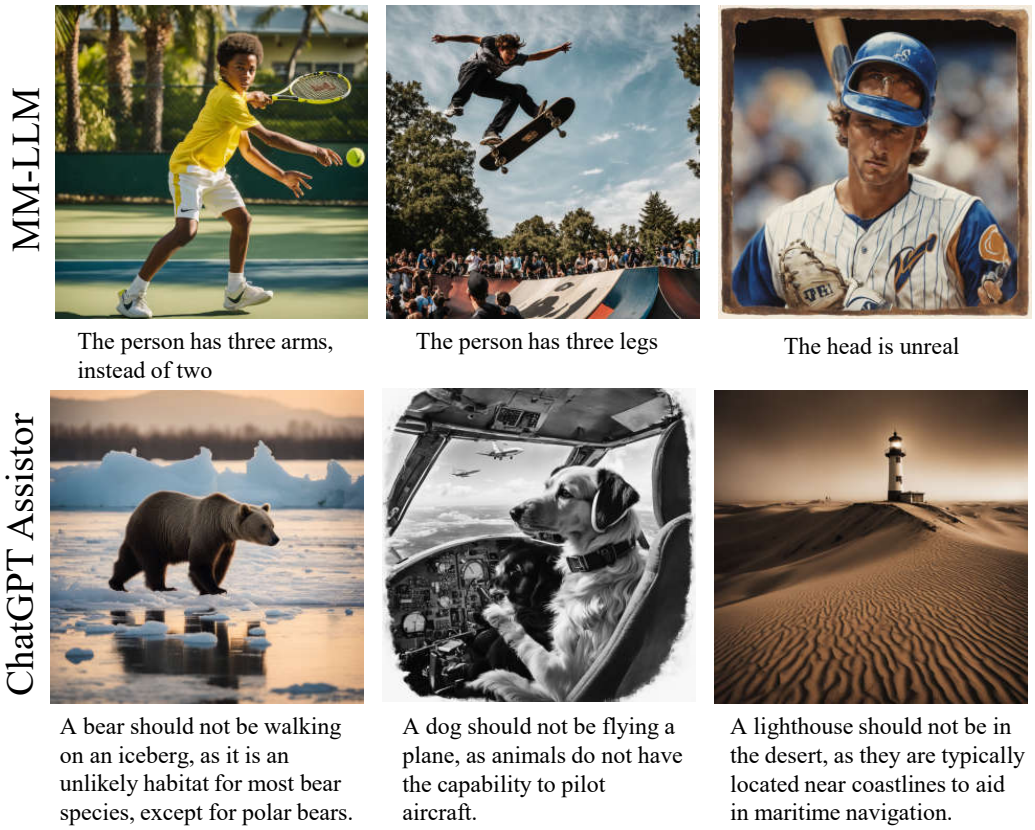


Figure 7. Generated text comparison between MLLM and ChatGPT Assistor. The texts generated by MLLM tend to be brief and short, while texts generated by ChatGPT tend to be detailed and more reasonable.

Cross-validation We show the experiment results of RADAR and MLLM (without the ChatGPT assistor) with different training and evaluation groups in Table 2. As shown, without the ChatGPT Assistor, the MLLM trained on certain groups of training sets has serious overfitting, and cannot generalize to the rest of the groups. By utilizing the general knowledge reasoning ability of the ChatGPT assistor, RADAR is enabled to detect general semantic fakes given grounded text descriptions. As a result, RADAR achieves relatively stable performance in different validation groups.

Table 2. The detection results comparison of different training and validation groups.

Methods	Training Group	Physics		Biology		Others	
		AP	AR	AP	AR	AP	AR
MLLM(IT)	Physics	30.4	28.5	10.2	10.1	6.8	7.3
	Biology	11.6	12.1	24.7	28.5	7.4	7.9
	Others	12.2	11.6	11.4	11.0	15.6	17.2
RADAR	Physics	34.5	33.1	23.2	22.7	19.2	19.8
	Biology	26.3	26.5	30.2	31.4	18.3	19.5
	Others	27.5	28.8	25.3	25.0	26.8	28.7

4.4. Limitation

The main goal of RADAR is to locate and explain the semantic fakes in AI-generated images, thus helping generate more realistic images that conform to world knowledge and common sense. As shown in the third row of Figure 5, the AI-generated images may not have semantic-level fakes. In this case, RADAR has no ability to distinguish AI-generated images from real images. As a result, for traditional deepfake detection, RADAR could only serve as a complement to existing binary classification methods to detect semantic-level issues.

5. Conclusions

In this work, we propose a new task, reasoning AI-generated image detection, which focuses on identifying semantic fakes in generative images that violate world knowledge and common sense. To benchmark reasoning AI-generated image detection, we collect a dataset, STSF, short for Spot the Semantic Fake. STSF contains 358 images with clear semantic errors generated by three different modern diffusion models and provides bounding boxes as well as text annotations to locate the fake regions. We propose RADAR, a reasoning AI-generated image detection assistor, to locate semantic fakes in the generated images and offer corresponding text explanations. Extensive experiment results on STSF show that RADAR has promising abilities to locate and explain semantic fakes in AI-generated images, which could help to improve the quality of the generative images.

6. Discussion

While RADAR demonstrates promising capabilities in detecting semantic fakes, several considerations and avenues for future research are important for advancing the field. A critical aspect for practical deployment is robustness against adversarial fakes, i.e., subtle manipulations designed to deceive detection methods. Future work should investigate the resilience of VLM-based approaches like RADAR to such attacks and develop mitigation strategies to ensure reliability. Furthermore, enhancing how semantic errors are represented in datasets warrants exploration. While the STSF dataset utilizes bounding boxes, adopting more flexible annotation methods such as segmentation masks could provide more precise localization of errors, and graph-based representations could explicitly model complex incorrect relationships between multiple entities. These richer annotation schemes could lead to models with a deeper understanding of intricate semantic fakes and improve the explainability of their results.

Future Work: Building upon the foundation of RADAR and the STSF dataset, further advancements in semantic fake detection can be achieved by enhancing dataset scale, diversity, and the granularity of semantic error representation. One such direction involves exploring novel generation and annotation paradigms. Specifically, leveraging dynamically generated sequences where visual content exhibits progressively increasing semantic implausibility holds considerable potential. This method could streamline annotation

processes, as annotators could more readily identify the onset and nature of semantic errors by comparing distorted elements against a consistent baseline within the sequence. Such an approach might not only facilitate the efficient collection of diverse semantic fakes but also enable the creation of datasets with fine-grained labels capturing the severity or specific type of semantic deviation.

Ethical Impacts: While RADAR aims to address the challenge of detecting semantic fakes in AI-generated images, it is crucial to acknowledge its inherent limitations and potential ethical implications. The method, like any AI tool, cannot entirely avoid misidentification, leading to false positives or false negatives, which could discredit creators or allow misleading content to proliferate. This risk is compounded by the subjective nature of "common sense" and the continuous evolution of generative models. Furthermore, there is a potential for misuse of the technology, such as selective censorship or weaponizing of fake labels, if not deployed responsibly. A core component of RADAR involves large language models (LLMs), which are known to be prone to hallucinations and biases. Such hallucinations could result in inaccurate textual explanations, while inherent biases in LLM training data might skew RADAR's performance across different cultural contexts or image types. Consequently, the explanations provided by RADAR should be critically evaluated, and the tool itself is best utilized as an assistor requiring human oversight in critical applications, rather than a definitive arbiter of semantic truth. The generalizability of RADAR is also dependent on the diversity of its training data, and its performance may vary on novel types of semantic fakes or generative models not represented in the current STSF dataset.

Author Contributions: Conceptualization and writing, H.W. and Z.L.; methodology and data curation, X.L.; software, C.Y.; validation, X.J. and C.Y.; review and editing, R.W.; project administration and funding acquisition, E.G. All authors have read and agreed to the published version of the manuscript.

Funding: This work is funded by the European Union (ERC, EVA, 950086).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: We declare that this study does not involve humans.

Data Availability Statement: Datasets can be found at https://drive.google.com/drive/folders/1i2PnPeSBITTmHoZmoCUMZmXTmJH3l8yn?usp=sharing (accessed on 26 June 2025).

Conflicts of Interest: Xiaolong Jiang was employed by Xiaohongshu. The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

References

1. Wang, S.Y.; Wang, O.; Zhang, R.; Owens, A.; Efros, A.A. CNN-generated images are surprisingly easy to spot... for now. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Virtual, 14–19 June 2020; pp. 8695–8704.
2. Gragnaniello, D.; Cozzolino, D.; Marra, F.; Poggi, G.; Verdoliva, L. Are GAN generated images easy to detect? A critical analysis of the state-of-the-art. In Proceedings of the 2021 IEEE International Conference on Multimedia and Expo (ICME), Shenzhen, China, 5–9 July 2021; IEEE: Piscataway, NJ, USA, 2021; pp. 1–6.
3. Sinitsa, S.; Fried, O. Deep image fingerprint: Towards low budget synthetic image detection and model lineage analysis. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, Waikoloa, HI, USA, 3–7 January 2024; pp. 4067–4076.
4. Podell, D.; English, Z.; Lacey, K.; Blattmann, A.; Dockhorn, T.; Müller, J.; Penna, J.; Rombach, R. Sdxl: Improving latent diffusion models for high-resolution image synthesis. *arXiv* **2023**, arXiv:2307.01952.
5. Sauer, A.; Lorenz, D.; Blattmann, A.; Rombach, R. Adversarial diffusion distillation. *arXiv* **2023**, arXiv:2311.17042.

6. Esser, P.; Kulal, S.; Blattmann, A.; Entezari, R.; Müller, J.; Saini, H.; Levi, Y.; Lorenz, D.; Sauer, A.; Boesel, F.; et al. Scaling rectified flow transformers for high-resolution image synthesis. In Proceedings of the Forty-first International Conference on Machine Learning, Vienna, Austria, 21–27 July 2024.
7. Yu, F.; Seff, A.; Zhang, Y.; Song, S.; Funkhouser, T.; Xiao, J. Lsun: Construction of a large-scale image dataset using deep learning with humans in the loop. *arXiv* **2015**, arXiv:1506.03365.
8. Zhu, M.; Chen, H.; Yan, Q.; Huang, X.; Lin, G.; Li, W.; Tu, Z.; Hu, H.; Hu, J.; Wang, Y. Genimage: A million-scale benchmark for detecting ai-generated image. In Proceedings of the Advances in Neural Information Processing Systems, Vancouver, BC, Canada, 9–15 December 2024; Volume 36.
9. Hong, Y.; Zhang, J. Wildfake: A large-scale challenging dataset for ai-generated images detection. *arXiv* **2024**, arXiv:2402.11843.
10. Ho, J.; Jain, A.; Abbeel, P. Denoising diffusion probabilistic models. In Proceedings of the NeurIPS, Virtual, 6–12 December 2020.
11. Vahdat, A.; Kreis, K.; Kautz, J. Score-based generative modeling in latent space. In Proceedings of the NeurIPS, Virtual, 6–14 December 2021.
12. Rombach, R.; Blattmann, A.; Lorenz, D.; Esser, P.; Ommer, B. High-resolution image synthesis with latent diffusion models. In Proceedings of the CVPR, New Orleans, LA, USA, 19–24 June 2022.
13. Nichol, A.; Dhariwal, P.; Ramesh, A.; Shyam, P.; Mishkin, P.; McGrew, B.; Sutskever, I.; Chen, M. Glide: Towards photorealistic image generation and editing with text-guided diffusion models. *arXiv* **2021**, arXiv:2112.10741.
14. Ramesh, A.; Dhariwal, P.; Nichol, A.; Chu, C.; Chen, M. Hierarchical text-conditional image generation with clip latents. *arXiv* **2022**, arXiv:2204.06125.
15. Saharia, C.; Chan, W.; Saxena, S.; Li, L.; Whang, J.; Denton, E.L.; Ghasemipour, K.; Gontijo Lopes, R.; Karagol Ayan, B.; Salimans, T.; et al. Photorealistic text-to-image diffusion models with deep language understanding. In Proceedings of the NeurIPS, New Orleans, LA, USA, 28 November–9 December 2022.
16. Kong, Z.; Ping, W.; Huang, J.; Zhao, K.; Catanzaro, B. Diffwave: A versatile diffusion model for audio synthesis. *arXiv* **2020**, arXiv:2009.09761.
17. Liu, J.; Li, C.; Ren, Y.; Chen, F.; Zhao, Z. Diffsinger: Singing voice synthesis via shallow diffusion mechanism. In Proceedings of the AAAI, Vancouver, BC, Canada, 22 February–1 March 2022.
18. Saharia, C.; Ho, J.; Chan, W.; Salimans, T.; Fleet, D.J.; Norouzi, M. Image super-resolution via iterative refinement. *IEEE Trans. Pattern Anal. Mach. Intell.* **2022**, *45*, 4713–4726. [CrossRef]
19. Gao, S.; Liu, X.; Zeng, B.; Xu, S.; Li, Y.; Luo, X.; Liu, J.; Zhen, X.; Zhang, B. Implicit diffusion models for continuous super-resolution. In Proceedings of the CVPR, Vancouver, BC, Canada, 18–22 June 2023.
20. Poole, B.; Jain, A.; Barron, J.T.; Mildenhall, B. Dreamfusion: Text-to-3d using 2d diffusion. *arXiv* **2022**, arXiv:2209.14988.
21. Lin, C.H.; Gao, J.; Tang, L.; Takikawa, T.; Zeng, X.; Huang, X.; Kreis, K.; Fidler, S.; Liu, M.Y.; Lin, T.Y. Magic3D: High-Resolution Text-to-3D Content Creation. *arXiv* **2022**, arXiv:2211.10440.
22. Wang, H.; Du, X.; Li, J.; Yeh, R.A.; Shakhnarovich, G. Score jacobian chaining: Lifting pretrained 2d diffusion models for 3d generation. In Proceedings of the ICCV, Paris, France, 2–6 October 2023.
23. Tang, J.; Wang, T.; Zhang, B.; Zhang, T.; Yi, R.; Ma, L.; Chen, D. Make-it-3d: High-fidelity 3d creation from a single image with diffusion prior. *arXiv* **2023**, arXiv:2303.14184.
24. Wang, Z.; Lu, C.; Wang, Y.; Bao, F.; Li, C.; Su, H.; Zhu, J. ProlificDreamer: High-Fidelity and Diverse Text-to-3D Generation with Variational Score Distillation. *arXiv* **2023**, arXiv:2305.16213.
25. Zhu, J.; Zhuang, P. HiFA: High-fidelity Text-to-3D with Advanced Diffusion Guidance. *arXiv* **2023**, arXiv:2305.18766.
26. Ho, J.; Salimans, T.; Gritsenko, A.; Chan, W.; Norouzi, M.; Fleet, D.J. Video diffusion models. In Proceedings of the NeurIPS, New Orleans, LA, USA, 28 November–9 December 2022.
27. Ho, J.; Chan, W.; Saharia, C.; Whang, J.; Gao, R.; Gritsenko, A.; Kingma, D.P.; Poole, B.; Norouzi, M.; Fleet, D.J.; et al. Imagen video: High definition video generation with diffusion models. *arXiv* **2022**, arXiv:2210.02303.
28. Yu, S.; Sohn, K.; Kim, S.; Shin, J. Video probabilistic diffusion models in projected latent space. In Proceedings of the CVPR, Vancouver, BC, Canada, 18–22 June 2023.
29. Blattmann, A.; Rombach, R.; Ling, H.; Dockhorn, T.; Kim, S.W.; Fidler, S.; Kreis, K. Align your latents: High-resolution video synthesis with latent diffusion models. In Proceedings of the CVPR, Vancouver, BC, Canada, 18–22 June 2023.
30. Wu, J.Z.; Ge, Y.; Wang, X.; Lei, S.W.; Gu, Y.; Shi, Y.; Hsu, W.; Shan, Y.; Qie, X.; Shou, M.Z. Tune-a-video: One-shot tuning of image diffusion models for text-to-video generation. In Proceedings of the ICCV, Paris, France, 2–6 October 2023.
31. Tevet, G.; Raab, S.; Gordon, B.; Shafir, Y.; Cohen-Or, D.; Bermano, A.H. Human motion diffusion model. In Proceedings of the ICLR, Kigali, Rwanda, 1–5 May 2023.
32. Peebles, W.; Xie, S. Scalable Diffusion Models with Transformers. *arXiv* **2022**, arXiv:2212.09748.
33. Chen, J.; Yu, J.; Ge, C.; Yao, L.; Xie, E.; Wu, Y.; Wang, Z.; Kwok, J.; Luo, P.; Lu, H.; et al. Pixart-α: Fast training of diffusion transformer for photorealistic text-to-image synthesis. *arXiv* **2023**, arXiv:2310.00426.

34. Li, Z.; Zhang, J.; Lin, Q.; Xiong, J.; Long, Y.; Deng, X.; Zhang, Y.; Liu, X.; Huang, M.; Xiao, Z.; et al. Hunyuan-DiT: A Powerful Multi-Resolution Diffusion Transformer with Fine-Grained Chinese Understanding. *arXiv* **2024**, arXiv:2405.08748.
35. Gao, P.; Zhuo, L.; Lin, Z.; Liu, C.; Chen, J.; Du, R.; Xie, E.; Luo, X.; Qiu, L.; Zhang, Y.; et al. Lumina-T2X: Transforming Text into Any Modality, Resolution, and Duration via Flow-based Large Diffusion Transformers. *arXiv* **2024**, arXiv:2405.05945.
36. Wu, C.; Yin, S.; Qi, W.; Wang, X.; Tang, Z.; Duan, N. Visual chatgpt: Talking, drawing and editing with visual foundation models. *arXiv* **2023**, arXiv:2303.04671.
37. Yang, Z.; Li, L.; Wang, J.; Lin, K.; Azarnasab, E.; Ahmed, F.; Liu, Z.; Liu, C.; Zeng, M.; Wang, L. Mm-react: Prompting chatgpt for multimodal reasoning and action. *arXiv* **2023**, arXiv:2303.11381.
38. Shen, Y.; Song, K.; Tan, X.; Li, D.; Lu, W.; Zhuang, Y. Hugginggpt: Solving ai tasks with chatgpt and its friends in hugging face. In Proceedings of the Advances in Neural Information Processing Systems, Vancouver, BC, Canada, 9–15 December 2024; Volume 36.
39. Li, J.; Li, D.; Xiong, C.; Hoi, S. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In Proceedings of the International Conference on Machine Learning, PMLR, Baltimore, MD, USA, 17–23 July 2022; pp. 12888–12900.
40. Li, J.; Li, D.; Savarese, S.; Hoi, S. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In Proceedings of the International Conference on Machine Learning, PMLR, Honolulu, HI, USA, 23–29 July 2023; pp. 19730–19742.
41. Liu, H.; Li, C.; Wu, Q.; Lee, Y.J. Visual instruction tuning. In Proceedings of the Advances in Neural Information Processing Systems, Vancouver, BC, Canada, 9–15 December 2024; Volume 36.
42. Zhu, D.; Chen, J.; Shen, X.; Li, X.; Elhoseiny, M. Minigpt-4: Enhancing vision-language understanding with advanced large language models. *arXiv* **2023**, arXiv:2304.10592.
43. Bayar, B.; Stamm, M.C. A deep learning approach to universal image manipulation detection using a new convolutional layer. In Proceedings of the 4th ACM Workshop on Information Hiding and Multimedia Security, Vigo, Spain, 25–27 June 2016; pp. 5–10.
44. Afchar, D.; Nozick, V.; Yamagishi, J.; Echizen, I. Mesonet: A compact facial video forgery detection network. In Proceedings of the 2018 IEEE International Workshop on Information Forensics and Security (WIFS), Hong Kong, China, 11–13 December 2018; IEEE: Piscataway, NJ, USA, 2018; pp. 1–7.
45. Liu, H.; Li, X.; Zhou, W.; Chen, Y.; He, Y.; Xue, H.; Zhang, W.; Yu, N. Spatial-phase shallow learning: Rethinking face forgery detection in frequency domain. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 20–25 June 2021; pp. 772–781.
46. Durall, R.; Keuper, M.; Keuper, J. Watch your up-convolution: Cnn based generative deep neural networks are failing to reproduce spectral distributions. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020; pp. 7890–7899.
47. Qian, Y.; Yin, G.; Sheng, L.; Chen, Z.; Shao, J. Thinking in frequency: Face forgery detection by mining frequency-aware clues. In Proceedings of the European Conference on Computer Vision, Glasgow, UK, 23–28 August 2020; Springer: Berlin/Heidelberg, Germany, 2020; pp. 86–103.
48. Li, J.; Xie, H.; Li, J.; Wang, Z.; Zhang, Y. Frequency-aware Discriminative Feature Learning Supervised by Single-Center Loss for Face Forgery Detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 20–25 June 2021; pp. 6458–6467.
49. Zhang, X.; Karaman, S.; Chang, S.F. Detecting and simulating artifacts in gan fake images. In Proceedings of the 2019 IEEE international workshop on information forensics and security (WIFS), Delft, The Netherlands, 9–12 December 2019; IEEE: Piscataway, NJ, USA, 2019; pp. 1–6.
50. Dzanic, T.; Shah, K.; Witherden, F. Fourier spectrum discrepancies in deep network generated images. *Adv. Neural Inf. Process. Syst.* **2020**, *33*, 3022–3032.
51. Chandrasegaran, K.; Tran, N.T.; Cheung, N.M. A closer look at fourier spectrum discrepancies for cnn-generated images detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 20–25 June 2021; pp. 7200–7209.
52. Wang, Z.; Bao, J.; Zhou, W.; Wang, W.; Hu, H.; Chen, H.; Li, H. Dire for diffusion-generated image detection. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Paris, France, 2–6 October 2023; pp. 22445–22455.
53. He, Z.; Chen, P.Y.; Ho, T.Y. RIGID: A Training-free and Model-Agnostic Framework for Robust AI-Generated Image Detection. *arXiv* **2024**, arXiv:2405.20112.
54. Korshunov, P.; Marcel, S. Deepfakes: A new threat to face recognition? assessment and detection. *arXiv* **2018**, arXiv:1812.08685.
55. Li, Y.; Yang, X.; Sun, P.; Qi, H.; Lyu, S. Celeb-df: A large-scale challenging dataset for deepfake forensics. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020; pp. 3207–3216.
56. Karras, T.; Aila, T.; Laine, S.; Lehtinen, J. Progressive growing of gans for improved quality, stability, and variation. *arXiv* **2017**, arXiv:1710.10196.

57. Wang, Y.; Huang, Z.; Hong, X. Benchmarking deepart detection. *arXiv* **2023**, arXiv:2302.14475.
58. Sha, Z.; Li, Z.; Yu, N.; Zhang, Y. De-fake: Detection and attribution of fake images generated by text-to-image generation models. In Proceedings of the 2023 ACM SIGSAC Conference on Computer and Communications Security, Copenhagen, Denmark, 26–30 November 2023; pp. 3418–3432.
59. Bird, J.J.; Lotfi, A. Cifake: Image classification and explainable identification of ai-generated synthetic images. *IEEE Access* **2024**, 12, 15642–15650. [CrossRef]
60. Mañas, O.; Krojer, B.; Agrawal, A. Improving automatic vqa evaluation using large language models. In Proceedings of the AAAI Conference on Artificial Intelligence, Vancouver, BC, Canada, 20–27 February 2024; Volume 38, pp. 4171–4179.
61. Chen, X.; Fang, H.; Lin, T.Y.; Vedantam, R.; Gupta, S.; Dollár, P.; Zitnick, C.L. Microsoft coco captions: Data collection and evaluation server. *arXiv* **2015**, arXiv:1504.00325.
62. Lin, T.Y.; Maire, M.; Belongie, S.; Hays, J.; Perona, P.; Ramanan, D.; Dollár, P.; Zitnick, C.L. Microsoft coco: Common objects in context. In Proceedings of the Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, 6–12 September 2014; Proceedings, Part V 13; Springer: Berlin/Heidelberg, Germany, 2014; pp. 740–755.
63. Fang, Y.; Wang, W.; Xie, B.; Sun, Q.; Wu, L.; Wang, X.; Huang, T.; Wang, X.; Cao, Y. Eva: Exploring the limits of masked visual representation learning at scale. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Vancouver, BC, Canada, 18–22 June 2023; pp. 19358–19369.
64. Touvron, H.; Martin, L.; Stone, K.; Albert, P.; Almahairi, A.; Babaei, Y.; Bashlykov, N.; Batra, S.; Bhargava, P.; Bhosale, S.; et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv* **2023**, arXiv:2307.09288.
65. Chen, J.; Zhu, D.; Shen, X.; Li, X.; Liu, Z.; Zhang, P.; Krishnamoorthi, R.; Chandra, V.; Xiong, Y.; Elhoseiny, M. Minigpt-v2: Large language model as a unified interface for vision-language multi-task learning. *arXiv* **2023**, arXiv:2310.09478.
66. Hu, E.J.; Shen, Y.; Wallis, P.; Allen-Zhu, Z.; Li, Y.; Wang, S.; Wang, L.; Chen, W. Lora: Low-rank adaptation of large language models. *arXiv* **2021**, arXiv:2106.09685.
67. Loshchilov, I.; Hutter, F. Fixing weight decay regularization in adam. *arXiv* **2018**, arXiv:1711.05101.
68. Young, P.; Lai, A.; Hodosh, M.; Hockenmaier, J. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *Trans. Assoc. Comput. Linguist.* **2014**, *2*, 67–78. [CrossRef]
69. Yu, L.; Poirson, P.; Yang, S.; Berg, A.C.; Berg, T.L. Modeling context in referring expressions. In Proceedings of the Computer Vision—ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, 11–14 October 2016; Proceedings, Part II 14; Springer: Berlin/Heidelberg, Germany, 2016; pp. 69–85.
70. Girshick, R. Fast r-cnn. In Proceedings of the IEEE International Conference on Computer Vision, Santiago, Chile, 7–13 December 2015; pp. 1440–1448.
71. Carion, N.; Massa, F.; Synnaeve, G.; Usunier, N.; Kirillov, A.; Zagoruyko, S. End-to-end object detection with transformers. In Proceedings of the European Conference on Computer Vision, Glasgow, UK, 23–28 August 2020; Springer: Berlin/Heidelberg, Germany, 2020; pp. 213–229.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.