This CVPR paper is the Open Access version, provided by the Computer Vision Foundation.
Except for this watermark, it is identical to the accepted version;
the final published version of the proceedings is available on IEEE Xplore.

OpenSDI: Spotting Diffusion-Generated Images in the Open World

Yabin Wang
Xi'an Jiaotong University
iamwangyabin@stu.xjtu.edu.cn

Zhiwu Huang*
University of Southampton
Zhiwu.Huang@soton.ac.uk

Xiaopeng Hong
Harbin Institute of Technology
hongxiaopeng@ieee.org

Abstract

This paper identifies OpenSDI, a challenge for spotting diffusion-generated images in open-world settings. In response to this challenge, we define a new benchmark, the OpenSDI dataset (OpenSDID), which stands out from existing datasets due to its diverse use of large vision-language models that simulate open-world diffusion-based manipulations. Another outstanding feature of OpenSDID is its inclusion of both detection and localization tasks for images manipulated globally and locally by diffusion models. To address the OpenSDI challenge, we propose a Synergizing Pretrained Models (SPM) scheme to build up a mixture of foundation models. This approach exploits a collaboration mechanism with multiple pretrained foundation models to enhance generalization in the OpenSDI context, moving beyond traditional training by synergizing multiple pretrained models through prompting and attending strategies. Building on this scheme, we introduce MaskCLIP, an SPM-based model that aligns Contrastive Language-Image Pre-Training (CLIP) with Masked Autoencoder (MAE). Extensive evaluations on OpenSDID show that MaskCLIP significantly outperforms current state-of-the-art methods for the OpenSDI challenge, achieving remarkable relative improvements of 14.23% in IoU (14.11% in F1) and 2.05% in accuracy (2.38% in F1) compared to the second-best model in localization and detection tasks, respectively. Our dataset and code are available at https://github.com/iamwangyabin/OpenSDI.

1. Introduction

The rise of advanced text-to-image (T2I) generation models, such as Stable Diffusion [9, 41, 45] and other diffusion-based frameworks [27, 47, 61], has profoundly transformed digital content creation, enabling users to produce remarkably lifelike images from simple text prompts. This capability, while expanding creative possibilities, has blurred the boundary between real and AI-generated content, making it increasingly difficult to ensure content authenticity. Addressing this challenge requires detectors that can handle three essential open-world settings (Figure 1): (1) the wide range of user preferences (*user diversity*), covering diverse styles, subjects, and creative intentions; (2) the rapid evolution of diffusion models (*model innovation*), introducing significant variability in generated images; and (3) the extensive T2I generation (*manipulation scope*), from a broad spectrum of global image synthesis to precise local changes, object insertions, and background manipulations [46, 62, 63]. We identify this challenge as *Open-world Spotting of Diffusion Images (OpenSDI)*.



Figure 1. Three open-world settings of the OpenSDI challenge: (1) user diversity by simulating a range of user preferences, (2) model innovation through the use of multiple advanced diffusion models, and (3) a full manipulation scope that enables both global and local generation.

In response to the OpenSDI challenge, we introduce OpenSDID, a large-scale dataset specifically designed to address the scarcity of suitable public datasets for benchmarking OpenSDI (Table 1). Our approach incorporates the three essential dimensions of OpenSDI to create a comprehensive and realistic dataset. To reflect *user diversity*, we use multiple large vision-language models (VLMs) [1,

*Corresponding author.

30, 33, 53] to generate a broad range of natural language prompts, simulating how users iteratively refine their creative ideas through diverse, open-world textual descriptions. These prompts are designed to mirror real-world user behavior and preferences, capturing the variety seen in different creative intentions and styles. To address *model innovation*, we incorporate state-of-the-art (SOTA) T2I diffusion models [9, 27, 41, 45], each with distinct architectures and parameter configurations. This variety in diffusion models allows our dataset to encompass the unique visual characteristics and nuances associated with advanced diffusion models, enhancing the dataset's relevance for evaluating detection and localization across a spectrum of T2I diffusion models. Finally, for *manipulation scope*, we leverage the semantic understanding of VLMs in combination with Segment Anything Model (SAM) [24] and Florence 2 [60] to translate textual instructions into precise region masks. These masks enable a wide range of manipulations, from local region edits to global image synthesis, simulating the types of modifications that make detection challenging in open-world settings. By combining global and localized manipulations, OpenSDID aims at embodying the complexity required for benchmarking the OpenSDI challenge.

To develop effective approaches for addressing OpenSDI, we face two main barriers: (1) unifying detection and localization frameworks, and (2) generalizing to open-world settings. The unification is challenging due to the high heterogeneity of detection and location. Detection relies on image-level semantics, whereas localization requires pixel-level precision. To tackle these issues, we suggest an approach of building up a mixture of pretrained foundation models. While employing multiple expert models holds significant potential, achieving synergy among them remains a complex challenge. To address this, we introduce a novel framework, Synergizing Pretrained Models (SPM). It is designed to optimize the collaboration between pretrained foundation models for enhanced performance and maintaining generalization capability across both detection and localization tasks in open-world settings. In particular, this approach leverages the strengths of multiple pretrained models through prompt tuning and attention mechanisms, creating a mixture of pretrained foundation models that enhances generalization across diverse real-world scenarios. Building on this SPM approach, we introduce MaskCLIP, a mixture of pretrained models that aligns the capabilities of Contrastive Language-Image Pre-Training (CLIP) with a Masked Autoencoder (MAE) model, outperforming SOTA techniques with considerable relative improvements in both detecting and localizing diffusion-generated image content on the suggested OpenSDI dataset.

In summary, our main contributions are as follows:

- We define the OpenSDI challenge, and introduce a large-scale dataset (OpenSDID) designed to address the lack of suitable public datasets for OpenSDI.
- We introduce an SPM framework and its derived MaskCLIP to effectively address the OpenSDI challenge.
- Through extensive experiments, MaskCLIP demonstrates markedly superior performance compared to SOTA methods for OpenSDI.

Dataset	# Real	# Fake	User	Model	Full
DiffusionDB [57]	–	14M	✓	✗	✗
GenImage [64]	1.3M	1.4M	✗	✓	✗
AutoSplice [21]	2.3K	3.6K	✗	✗	✓
CocoGlide [12]	–	512	✗	✗	✗
HiFi-IFDL [13]	–	1M*	✓	✗	✓
GIM [5]	1.1M	1.1M	✗	✓	✗
TGIF [38]	3.1K	75K	✗	✓	✗
OpenSDID	300K	450K	✓	✓	✓

Table 1. Overview of existing diffusion image datasets and the proposed OpenSDID. User: Simulates diverse preferences. Model: Leverages multiple advanced diffusion models published after 2023. Full: Includes both global and local generation. Note: The asterisk (*) indicates HiFi-IFDL is built from multiple prior datasets. A more comprehensive version is in the suppl. material.

2. Related Works

Diffusion Image Benchmark Datasets. Existing datasets fall short in addressing the OpenSDI challenge due to their significant limitations. Many focus on traditional manipulation techniques such as splicing and copy-move operations. Widely used datasets like Columbia [18], NIST MFC [11], and CASIA [8] concentrate on these conventional methods. HiFi-IFDL [13] created a million-scale dataset that combines various generative models and datasets. However, the manipulation types in HiFi-IFDL are mainly centered around traditional methods and facial editing, and do not involve recent T2I diffusion generators.

To the best of our knowledge, only a few datasets incorporate text-guided editing using diffusion models such as DiffusionDB [57], GenImage [64], CocoGlide [12], AutoSplice [21], TGIF [38] and GIM [5]. DiffusionDB uses a single diffusion model, while CocoGlide and AutoSplice are relatively small in scale and restrict their scope to specific models (Glide [39] and DALL-E2 [43]). While TGIF and GIM offer a broader range of diffusion models, they lack diversity in user preferences as their editing instructions come from a single source. Furthermore, these datasets primarily focus on partial edits rather than comprehensive image generation, limiting their scope. As shown in Table 1, none of these datasets fully addresses all three essential aspects of the OpenSDI challenge.

Detection and Localization Approaches. Early detection methods [3, 54] simply use CNN-based architecture, such

as ResNet [15] and Xception [7], as binary classifiers for distinguishing between real and synthetic images. Several works [10, 35, 49–51] exploit distinctive frequency patterns that differentiate synthetic from real images to improve model generalization performance. However, these methods often struggle with generalization across different types of generators, as they tend to overfit to specific artifacts present in their training data. Recent methods [22, 32, 40, 55, 56] employ large pretrained models, particularly CLIP [42], to construct binary classifiers for deepfake detection. While they evince promising detection performance, they are limited to binary predictions and lack the ability to precisely localize tampered regions.

Another family of approaches (e.g., [19, 58, 59]) provides both detection results and accurate localization of manipulated regions. Early methods like [4, 12, 26, 34] design complex feature extraction and fusion modules across spatial, frequency, and noise domains to enhance the precision of forged image localization. Recently, vision transformer based methods [14, 36, 52] clearly improved the accuracy and efficiency of detection and localization. While these approaches primarily rely on ImageNet pretrained models, newer innovations such as [28, 29] are pushing boundaries by leveraging large-scale vision-language models like CLIP for more robust image forgery detection and localization.

However, most existing methods focus on traditional image manipulations, overlooking those generated globally and locally by advanced T2I diffusion models, leaving a substantial gap in addressing OpenSDI. Furthermore, they have limited generalization when faced with new types of AI-generated images, underscoring a clear limitation in adapting to rapidly evolving image generation technologies.

3. Dataset Creation

We introduce OpenSDID, a large-scale dataset specifically curated for the OpenSDI challenge. Our dataset design addresses the three core requirements essential for open-world spotting of AI-generated content: user diversity, model innovation, and manipulation scope.

User diversity. To simulate the variety of user intentions and styles in real-world settings, we employ an automated pipeline driven by multiple state-of-the-art VLMs, including LLaMA3 Vision [1], LLaVA [33], InternVL2 [6] and Qwen2 VL [53]. These VLMs simulate human-like editing behavior, generating diverse text prompts that mirror the wide range of real-world image manipulation intentions. Based on an authentic set of images from Megalith-10M [2], which is a resource of wholesome, unedited, copyright-free licensed images[1], our pipeline utilizes VLMs to produce varied manipulation instructions. These prompts emulate common user-driven edits, such as adding or removing objects, altering backgrounds, or changing key image elements. This approach ensures that OpenSDID accurately captures the breadth of user-like diversity in T2I generation, producing a dataset with rich and varied samples that reflect real-world creative diversity.

Model Innovation. As diffusion models evolve rapidly, it is crucial to account for the diverse visual styles and architectural innovations of different T2I diffusion models. To achieve this, we leverage a suite of state-of-the-art generative models—including multiple versions of Stable Diffusion (SD1.5 [45], SD2.1 [45], SDXL [41], SD3 [9]) and Flux.1 [27]. These models span various architectures, parameter configurations, and generation characteristics, providing our dataset with a wide range of model-specific image outputs. By using different models to execute the manipulations specified in the VLM-generated prompts, we ensure that OpenSDID covers a comprehensive spectrum of T2I outputs, reflecting the progression and variety within diffusion-based image synthesis. Our dataset includes distinct manipulation styles and visual attributes unique to each model version, supporting more rigorous testing of detection systems across evolving T2I diffusion models.

Manipulation Scope. As illustrated in Figure 1, we leverages T2I diffusion models for both global image synthesis and precise local edits. To simulate this full range of manipulations, our pipeline integrates SAM [24] and Florence 2 [60] to translate textual prompts into detailed region masks, enabling accurate localized edits. For each image, SAM and Florence generate segmentation masks that delineate specific areas for modifications. During the editing phase, we use the AutoPipelineForInpainting[2], a standard practice in image edition, to ensure that only the masked regions are modified. To further enhance dataset diversity, we produce multiple variations of each manipulation by using different random seeds. To enhance dataset quality and ensure that generated images align precisely with intended instructions, we employ the CLIP model for optimal image selection. We further diversify our dataset by incorporating traditional editing techniques, such as copy-paste and splicing operations [38], applied at random intervals. Additionally, we include fully synthetic images to ensure comprehensive coverage of diffusion image types. This fusion of global and local generation ensures that OpenSDID captures the complexity of real-world T2I forgeries, covering a comprehensive manipulation spectrum.

The dataset statistics are presented in Table 2, with some samples provided in Figure 2. OpenSDID comprises 300,000 images, evenly distributed between real and fake samples, with 200,000 images (100,000 real from Megalith-10M and 100,000 fake generated by SD1.5) used for train-

[1] The images are licensed under: United States Government Work, Public Domain Dedication (CC0), No known copyright restrictions (Flickr commons), or Public Domain Mark.

[2] https://huggingface.co/docs/diffusers/using-diffusers/inpaint



Figure 2. OpenSDID examples of authentic and manipulated images with their corresponding ground-truth masks, and entire generated images.

Model	Training Set		Test Set		Total
	Real	Fake	Real	Fake	Images
SD1.5	100K	100K	10K	10K	220K
SD2.1	-	-	10K	10K	20K
SDXL	-	-	10K	10K	20K
SD3	-	-	10K	10K	20K
Flux.1	-	-	10K	10K	20K
Total	100K	100K	50K	50K	300K

Table 2. Dataset Statistics.

ing, and 100,000 images (20,000 each from SD1.5, SD2.1, SDXL, SD3, and Flux.1) reserved for testing, ensuring a balanced representation across multiple generative models. This design follows common practices in AI-generated image detection [12, 54], evaluating generalization to unseen generators. Additional analysis and details about this dataset, including considerations of ethical and bias issues, are provided in the suppl. material.

4. Proposed Approach

SPM Framework. To tackle the OpenSDI challenge, we focus on two primary issues: (1) unifying detection and localization frameworks, and (2) generalizing to open-world environments. Detection and localization are inherently heterogeneous—the former relies on image-level semantics to classify an image, while the latter demands pixel-level accuracy to identify specific manipulated regions. To address them, we propose a scheme of Synergizing Pretrained Models (SPM) to optimize collaboration among pretrained models, each excelling in learning image- or pixel-level semantics while showing strong generalization capabilities.

The suggested SPM scheme achieves this synergy through two key mechanisms: *prompting* and *attending*. The prompting strategy introduces an additional parametrization to the network input, enabling the model to encode new knowledge without significantly altering the pretrained network components. This approach typically involves minimal fine-tuning or even freezing the original foundation model, ensuring that the pretrained knowledge is largely preserved. The attending strategy, on the other hand, facilitates collaboration between multiple pretrained models via a cross-attention mechanism, aligning feature representations from different models to improve performance. We demonstrate the effectiveness of the SPM scheme with MaskCLIP, which is based on a mixture of CLIP for image-level semantic understanding and MAE for fine-grained spatial precision.

Preliminaries. The CLIP [42] model is a dual-encoder architecture consisting of a text encoder E_t and an vision encoder E_v. The pretrained CLIP has demonstrated impressive zero-shot classification performances by computing cosine similarity between the visual features extracted from images and text features generated from class descriptions.

Given an input image $X \in \mathbb{R}^{H\times W\times 3}$ and text description T, the vision encoder splits the input image X into N image patches $\{x_i\}_{i=1}^N$. Each patch x_i is flattened and projected to obtain visual tokens $v_i \in \mathbb{R}^D$. This set of input visual tokens can be defined as $V_{patch} = \{v_i\}_{i=1}^N$. A learnable [CLS] token $v_{cls} \in \mathbb{R}^D$ is prepended to the sequence of visual tokens to form the complete input sequence, to facilitate a global image representation. Thus the input sequence for layer l is formed by concatenating the CLS token and visual tokens as $V^l = [v^l_{cls}; V^l_{patch}] \in \mathbb{R}^{(N+1)\times D}$.

Through the self-attention of the transformer, the v_{cls} token iteratively aggregates information from all visual tokens, resulting in a global representation of the entire image by the final layer. The [CLS] token v^L_{cls} from the final layer L is linearly projected to get the global image representation $\mathbf{g}$. Similarly, the text encoder embeds the text description T through the transformer to produce text embeddings $\mathbf{t}$. The image and text feature $\mathbf{g}$ and $\mathbf{t}$ are then normalized and used to compute the cosine similarity between image-text pairs.

MAEs [17] are self-supervised vision transformers that learn rich visual representations through a reconstruction objective of masked image regions. Unlike CLIP, the MAE encoder E_m processes the input image X without a [CLS] token, generating only visual patch tokens $V_m = \{v_{m,i}\}_{i=1}^N$ where $v_{m,i} \in \mathbb{R}^{D_m}$. The decoder attempts to reconstruct the original image patches corresponding to the masked tokens, encouraging the model to learn meaningful features from the spatial context available in the visible tokens.

MaskCLIP Overview. As shown in Figure 3, MaskCLIP integrates the CLIP visual encoder, the CLIP text encoder, and the MAE encoder. Following the MAE study [17], we replace the vanilla MAE decoder with a Feature Pyramid Network (FPN) [31] based decoder, as done in [36], for more precise pixel-level predictions. Based on the SPM scheme, we align the CLIP and MAE components through *prompting* and *attending* to address the OpenSDI challenge. Rather than full fine-tuning, we employ a prompt-tuning strategy [20, 23, 55] with the textual prompting block (Fig-

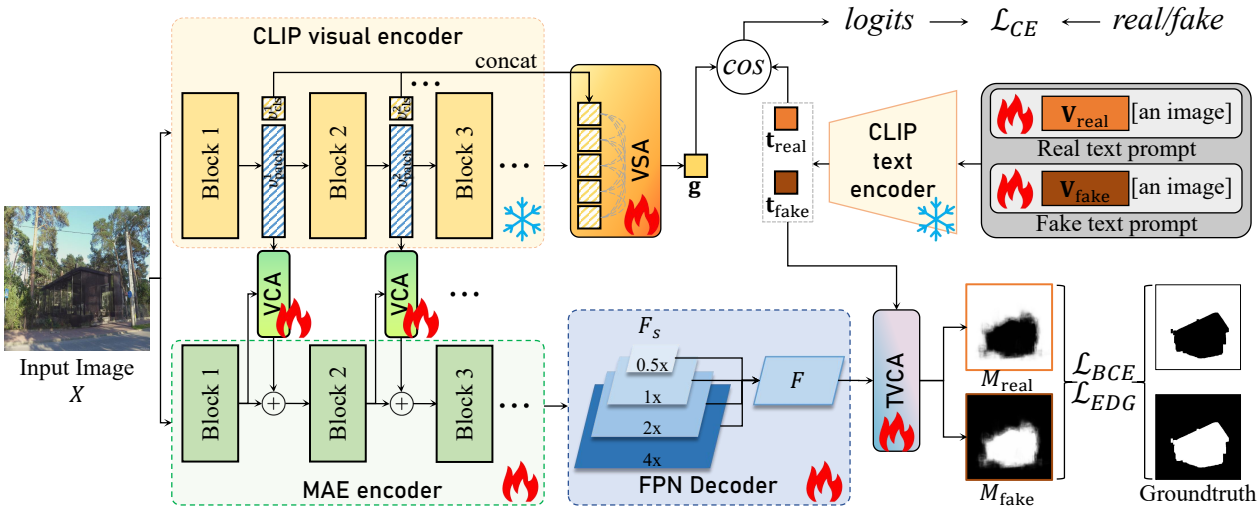


Figure 3. MaskCLIP overview. MaskCLIP inherits core components from pretrained CLIP and MAE: the CLIP vision and text encoders, the MAE encoder, and uses an FPN-style decoder for precise pixel-level predictions. To synergize CLIP and MAE, MaskCLIP introduces one prompting block (top right) and three attention blocks: VCA (visual cross-attention), TVCA (textual-visual cross-attention), and VSA (visual self-attention). Frozen components are marked with snowflakes, while tunable components (marked with flames) are optimized using a balanced objective with $\mathcal{L}_{\text{CE}}$ (cross-entropy), $\mathcal{L}_{\text{BCE}}$ (binary cross-entropy), and $\mathcal{L}_{\text{EDG}}$ (edge-weighted loss).

ure 3), learning a set of continuous prompt vectors for real/fake concepts while preserving CLIP's original parameters to retain its extensive visual-linguistic knowledge. To effectively integrate the prompted CLIP with MAE, the *attending* process is critical. For this purpose, we introduce two types of attending blocks (Figure 3): (1) visual cross-attention (VCA), which aligns the CLIP visual encoding features with the MAE encoding features, and (2) textual-visual cross-attention (TVCA), which coordinates the CLIP textual encoding features with the enhanced MAE decoding features. Additionally, we add a visual self-attention (VSA) block to concatenate global features from each CLIP encoding layer for more robust feature extraction. During training, all the prompting and attending blocks (VCA, TVCA, VSA) are jointly optimized.

Prompting. MaskCLIP learns a pair of continuous prompt vectors $\mathbf{V}_c \in \mathbb{R}^{M\times D}$ for $c \in \{\text{real}, \text{fake}\}$, where M denotes the prompt length and D represents the embedding dimension. These learned prompts function analogously to hand-crafted text prompts like "`a photo of a [real/fake]`", but employ learnable continuous vectors to capture the semantic context more effectively. These learnable prompts $\mathbf{V}_c$ are then processed by the text encoder to generate class-specific embeddings $\mathbf{t}_{\text{real}}$ and $\mathbf{t}_{\text{fake}} \in \mathbb{R}^D$, representing the real and fake class concepts, respectively.

Attending. To align CLIP with MAE, MaskCLIP employs the VCA block that combines CLIP's semantic understanding with MAE's spatial reconstruction capabilities. VCA is deployed across multiple network layers to facilitate hierarchical feature fusion between CLIP and MAE. At each fusion layer l, VCA processes spatial features V_m^l from the MAE encoder alongside the corresponding CLIP visual token features V_{patch}^l. To align these features, we first apply bilinear interpolation to resize V_{patch}^l to the spatial size of MAE features. Subsequently, a 1×1 convolution is employed to project the feature dimension D to D_m. The aligned CLIP features serve as query vectors $\mathbf{Q}$, while the MAE features V_m^l are projected to form keys and values $\mathbf{K}, \mathbf{V} \in \mathbb{R}^{N\times D_m}$, respectively. We use the multi-head attention mechanism within VCA is computed as:

$$\mathbf{G}^l = \text{Softmax}\left(\frac{\mathbf{Q}\mathbf{K}^\top}{\sqrt{d}}\right)\mathbf{V}, \tag{1}$$

where d is the feature dimension for the attention head. To enhance the fusion features, a residual connection is introduced, and the input features for layer $l+1$ of the MAE is $V_m^{l+1} = V_m^l + \mathbf{G}^l$.

In order to further align the CLIP textual features with the decoding features F, we suggest incorporating a TVCA block to integrate textual information into the decoding process. Given a textual embedding $\mathbf{t}_{\text{real}}, \mathbf{t}_{\text{fake}}$, the decoder performs cross-attention between $\mathbf{t}$ and the decoder's output feature F. First, F is linearly projected to match the dimension of $\mathbf{t}$. The cross-attention operation uses $\mathbf{t}$ as queries and F as keys and values to get the attention feature map F'. Finally we use a 1×1 convolutional layer to generate the segmentation logits M_{fake} and $M_{\text{real}} \in \mathbb{R}^{H\times W}$.

In addition, the conventional approaches [40, 55] of utilizing only the final layer's [CLS] token v_{cls}^L from the CLIP vision encoder fails to fully leverage the rich hierarchical features learned throughout the network. This limita-

tion is particularly critical for AI-generated image detection, where both low-level artifacts and high-level semantic inconsistencies play crucial roles in discrimination. Thus, rather than relying solely on the last layer output [CLS] token v_{cls}^{L} of the CLIP vision encoder, we introduce the VSA block to attend [CLS]s from multiple the CLIP encoder layers. VSA leverages a multi-head self-attention mechanism that collects [CLS] token embeddings $v_{cls}^{l} \in \mathbb{R}^{D}$ from multiple layers $l \in L$, forming a sequence $V_{cls} = \{v_{cls}^{l}\}_{l \in L}$. The output features are subsequently aggregated via global average pooling and transformed through a learned linear projection to derive the final image representation $\mathbf{g}$.

Essential Add-ons. Inspired by prior works [13, 17, 36], we employ an FPN-based decoder [16, 31] that processes the MAE encoder's output features V_{m}^{L} at multiple scales (0.5×, 2×, 4×), enabling the capture of both fine and coarse details. At each resolution, a 3×3 convolutional layer is applied to extract information. These multi-scale features F_s are then resized to the original spatial shape (H, W), and are concatenated along the channel dimension to form the feature map F.

To train MaskCLIP for localization, we adopt a combination of binary cross-entropy loss $\mathcal{L}_{\text{BCE}}$ and edge-weighted loss $\mathcal{L}_{\text{EDG}}$ applied to M_{fake}, following previous works [12, 36]. The total loss function incorporates the detection loss (i.e., cross-entropy loss term, $\mathcal{L}_{\text{CE}}$) to optimize the prompts, VSA, VCA, FPN Decoder, TVCA and the MAE encoder jointly. The weights for all loss terms are set to 1.

During inference, the binary detection is performed by the computing of cosine similarities between the image representation $\mathbf{g}$ and class-specific text embeddings $\mathbf{t}_{\text{real}}, \mathbf{t}_{\text{fake}}$. In parallel, the TVCA block generates the forgery localization map M_{fake} to predict potentially manipulated regions.

5. Experiments

Implementation Details. The training of MaskCLIP is conducted with batch size 64 and learning rate $1e-4$ for 20 epoch using the Adam optimizer. For data augmentation, we adopt a set of standard settings [12, 34, 36, 37], including Gaussian blurring, JPEG compression, random scaling, horizontal and vertical flips, and random cropping to 512×512. However, for the CLIP encoder, images are resized to 224×224 pixels to maintain compatibility with its pretrained configuration. In the main experiments, we choose the OpenAI ViT-L/14[3] in line with the previous works [25, 40, 48]. For the MAE encoder, we use the ImageNet pretrained weights ViT-B/32 as in previous works [36]. We set the prompt length M to be 10 as previous works [44, 55]. All training is conducted on 4 NVIDIA H100 GPUs.

[3]We also explore CLIP variants and alternative MAE models. To ensure reliability, we conduct experiments with varying hyperparameters and multiple runs. More details are provided in the suppl. material.

State-of-the-art Methods. These methods include classical image manipulation localization methods, including CAT-Net [26], MVSS-Net [4], PSCC-Net [34] and TruFor [12]. For transformer-based methods, we also include ObjectFormer [52], IML-ViT [36] and DeCLIP [48]. In addition to these pixel-level based methods, we also include SOTA image-level based (detection) methods, include CNNDet [54], UniFD [40], NPR [49], GramNet [35], FreqNet [51], and RINE [25]. All the methods were implemented using the IMDL-BenCo framework [37] and refer to their official codes if possible. We trained these methods on the OpenSDID's training data (Table 2), carefully tuning their hyperparameters to ensure optimal performance that matches their reported benchmarks. A brief description of these methods is provided in the supplementary.

Evaluation Metrics. Following standard practices in [12, 37], we evaluate methods using both pixel-level and image-level metrics on the OpenSDID's test data (Table 2). For pixel-level evaluation, we employ the F1-score (F1) and Intersection over Union (IoU). At the image level, we assess performance using F1 and accuracy (Acc).

Localization Results. In Table 3, we show the pixel-level localization performance in terms of F1 and IoU metrics. For the in-domain localization, specifically on the SD1.5 subset, MaskCLIP achieves an IoU of 0.6712 and an F1 score of 0.7563, surpassing all other methods. Notably, our method outperforms IML-ViT, the second best, by **0.92%** in IoU and **2.73%** in F1 score. In terms of the cross-domain generalization, MaskCLIP shows superior performance across various unseen generators. Our method outperforms previous best methods by significant margins: SD2.1 (**1.52%** IoU, **0.91%** F1), SDXL (**16.69%** IoU, **16.17%** F1), SD3 (**23.07%** IoU, **21.73%** F1), and most notably on Flux.1 (**44.69%** IoU, **42.34%** F1). The average performance across all datasets further emphasizes the superiority of MaskCLIP. It achieves an average IoU of 0.4271 and an average F1 score of 0.4941, surpassing the second best by **14.23%** in IoU and **14.11%** in F1. The markedly superior in-domain and cross-domain performance suggests that MaskCLIP does not overfit to specific content or manipulation patterns but learns generalizable features crucial for detecting and localizing forgeries across diverse open-world scenarios.

Detection Results. Table 4 displays image-level detection performance (F1-score and Accuracy) across various generative models, separating methods focused solely on image-level detection from those addressing both image and pixel-level tasks. The results illustrate that MaskCLIP achieves the best average performance among all evaluated methods, demonstrating both strong in-domain detection and robust cross-domain generalization capabilities. For the in-domain detection MaskCLIP shows competitive performance with the others. However, in terms of the cross-domain general-

Method	SD1.5		SD2.1		SDXL		SD3		Flux.1		AVG	
	IoU	F1	IoU	F1	IoU	F1	IoU	F1	IoU	F1	IoU	F1
MVSS-Net [4]	0.5785	0.6533	0.4490	0.5176	0.1467	0.1851	0.2692	0.3271	0.0479	0.0636	0.2983	0.3493
CAT-Net [26]	0.6636	0.7480	0.5458	0.6232	0.2550	0.3074	0.3555	0.4207	0.0497	0.0658	0.3739	0.4330
PSCC-Net [34]	0.5470	0.6422	0.3667	0.4479	0.1973	0.2605	0.2926	0.3728	0.0816	0.1156	0.2970	0.3678
ObjectFormer [52]	0.5119	0.6568	0.4739	0.4144	0.0741	0.0984	0.0941	0.1258	0.0529	0.0731	0.2414	0.2737
TruFor [12]	0.6342	0.7100	0.5467	0.6188	0.2655	0.3185	0.3229	0.3852	0.0760	0.0970	0.3691	0.4259
DeCLIP [48]	0.3718	0.4344	0.3569	0.4187	0.1459	0.1822	0.2734	0.3344	0.1121	0.1429	0.2520	0.3025
IML-ViT [36]	0.6651	0.7362	0.4479	0.5063	0.2149	0.2597	0.2363	0.2835	0.0611	0.0791	0.3251	0.3730
MaskCLIP	**0.6712**	**0.7563**	**0.5550**	**0.6289**	**0.3098**	**0.3700**	**0.4375**	**0.5121**	**0.1622**	**0.2034**	**0.4271**	**0.4941**

Table 3. Pixel-level (localization) performance on OpenSDID. Blue Block: in-domain test, Red Block: cross-domain test, **Bold**: best results, Underline: second best results. More results are in the supplementary material.

Method	SD1.5		SD2.1		SDXL		SD3		Flux.1		AVG	
	F1	Acc	F1	Acc	F1	Acc	F1	Acc	F1	Acc	F1	Acc
CNNDet [54]	0.8460	0.8504	0.7156	0.7594	0.5970	0.6872	0.5627	0.6708	0.3572	0.5757	0.6157	0.7087
GramNet [35]	0.8051	0.8035	0.7401	0.7666	0.6528	0.7076	0.6435	0.7029	0.5200	0.6337	0.6723	0.7229
FreqNet [51]	0.7588	0.7770	0.6097	0.6837	0.5315	0.6402	0.5350	0.6437	0.3847	0.5708	0.5639	0.6631
NPR [49]	0.7941	0.7928	0.8167	0.8184	0.7212	0.7428	**0.7343**	0.7547	**0.6762**	**0.7136**	0.7485	0.7645
UniFD [40]	0.7745	0.7760	0.8062	0.8192	0.7074	0.7483	0.7109	0.7517	0.6110	0.6906	0.7220	0.7572
RINE [25]	**0.9108**	**0.9098**	**0.8747**	**0.8812**	**0.7343**	**0.7876**	0.7205	**0.7678**	0.5586	0.6702	**0.7598**	**0.8033**
MVSS-Net [4]	0.9347	0.9365	0.7927	0.8233	0.5985	0.7042	0.6280	0.7213	0.2759	0.5678	0.6460	0.7506
CAT-Net [26]	**0.9615**	0.9615	0.7932	0.8246	0.6476	0.7334	0.6526	0.7361	0.2266	0.5526	0.6563	0.7616
PSCC-Net [34]	0.9607	0.9614	0.7685	0.8094	0.5570	0.6881	0.5978	0.7089	0.5177	0.6704	0.6803	0.7676
ObjectFormer [52]	0.7172	0.7522	0.6679	0.7255	0.4919	0.6292	0.4832	0.6254	0.3792	0.5805	0.5479	0.6626
TruFor [12]	0.9012	**0.9773**	0.3593	0.5562	0.5804	0.6641	0.5973	0.6751	0.4912	0.6162	0.5859	0.6978
DeCLIP [48]	0.8068	0.7831	0.8402	0.8277	0.7069	0.7055	0.6993	0.6840	0.5177	0.6561	0.7142	0.7313
IML-ViT [36]	0.9447	0.7573	0.6970	0.6119	0.4098	0.4995	0.4469	0.5125	0.1820	0.4362	0.5361	0.5635
MaskCLIP	0.9264	0.9272	**0.8871**	**0.8945**	**0.7802**	**0.8122**	**0.7307**	**0.7801**	**0.5649**	**0.6850**	**0.7779**	**0.8198**

Table 4. Image-level (detection) performance on OpenSDID. Blue Block: in-domain test, Red Block: cross-domain test, Top Section: image-level detection methods, Bottom Section: pixel-level detection and localization methods. **Bold**: best results, Underline: second best results. More results are in the supplementary material.

ization, MaskCLIP exhibits clearly better performance than all the other methods. On the SD2.1 and SDXL subsets, MaskCLIP even outperforms the SOTA image-level deepfake detectors, RINE, by **1.24%** and **4.59%**, respectively. For the Flux.1 test data, while image-level detector NPR shows better performance compared to our multi-task detectors, MaskCLIP achieves the best results among all localization methods. For the average result, MaskCLIP significantly outperforms all other methods by **2.38%** on F1 and by **2.05%** on Accuracy.

Qualitative Comparisons. We show some output localization masks of MaskCLIP and other SOTA approaches in Figure 4. As shown across different generators, our method produces masks with well-defined object boundaries and shapes. For instance, the ground truth comparison demonstrates that our approach successfully captures vehicle shapes (SD2.1), clothing textures (SD3), and object shape (SDXL). In contrast, competing methods often exhibit limitations such as over-segmentation of objects or missed detections, highlighting the robustness of our framework. More results are available in the suppl. materials.

Robustness Analysis. We conducted extensive experiments to evaluate detection robustness under different image degradation scenarios. Following standard practices [12, 37], we employed two common degradation methods: Gaussian blur and JPEG compression, applying them at various intensity levels. The results are presented in Figure 5, and we report the pixel-level F1 scores. For in-domain evaluation (SD1.5), our MaskCLIP (shown in red line) demonstrates superior robustness across both degradation types. The proposed MaskCLIP exhibits strong robustness in cross-domain scenarios (SD3), though with a slightly reduced margin compared to in-domain results. This is particularly evident in the JPEG compression test, where our method maintains steady performance even at lower-quality settings. These comprehensive evaluations demonstrate our method's superior robustness in both in-domain and cross-domain scenarios, particularly under challenging image quality conditions.

Ablation Study. To thoroughly evaluate the effectiveness of each component in our proposed method, we conduct comprehensive ablation experiments. Starting with a baseline architecture that combines UniFD's [40] classifier with DeCLIP's [48] decoder, we progressively incorporate our key components. Table 5 presents the quantitative results on the OpenSDID SD1.5 test set. The baseline achieves a

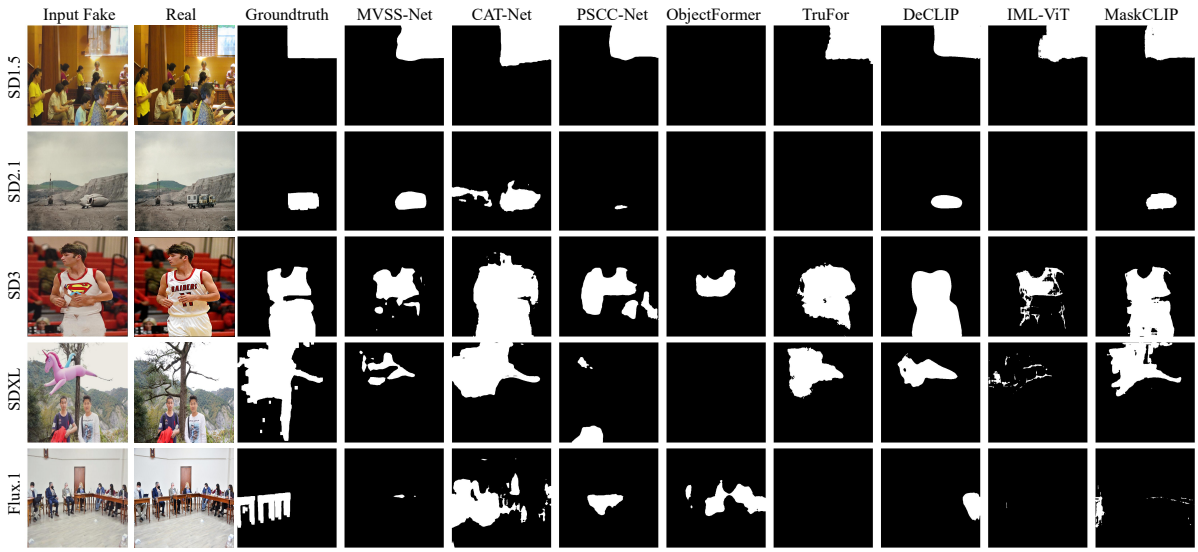


Figure 4. Qualitative results on OpenSDID.

Model Variation	Pixel F1	Image F1
Baseline	0.4290	0.7613
+ Prompt Tuning	0.4485	0.8576
+ VSA	0.4259	0.9083
+ VCA	0.6532	0.9343
+ TVCA (Full MaskCLIP)	0.7563	0.9264

Table 5. Ablation study of the main components in MaskCLIP.

pixel-level F1 of 0.4290 and an image-level F1 of 0.7613. Prompt tuning improves image F1 to 0.8576, showing effective adaptation of CLIP representations. VSA further boosts image-level F1 to 0.9083. The VCA effectively combines CLIP's semantic and MAE's spatial features, which significantly enhances pixel-level F1. Finally, TVCA completes our MaskCLIP model, achieving the highest pixel F1 of 0.7563 while maintaining strong image F1 at 0.9264. These results demonstrate that each component, particularly VCA and TVCA, contributes substantially to the model's overall performance. In the suppl. materials, we replace MAE with SAM [24], achieving a remarkable 0.7873 Pixel F1 on SD1.5, which further demonstrates the scalability of the proposed SPM scheme.

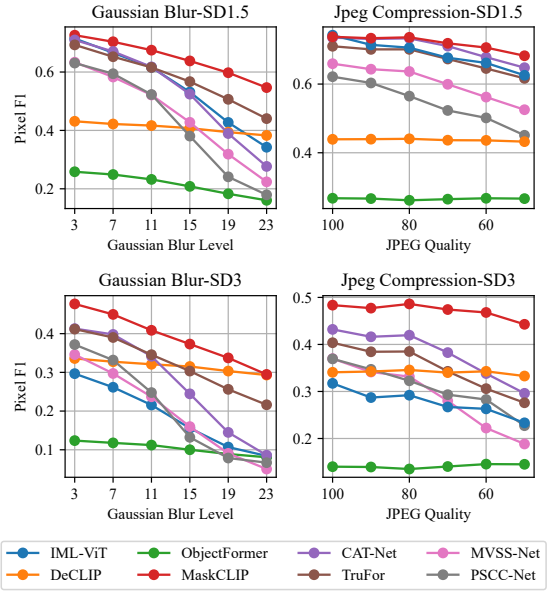


Figure 5. Robustness evaluation of different SOTA methods under image degradation. It compares performance across varying levels of Gaussian Blur (left) and JPEG Compression (right) on both in-domain (SD1.5) and cross-domain (SD3) test sets. Results on the rest test data are provided in the suppl. material.

6. Conclusions and Outlook

This paper addresses the OpenSDI challenge of detecting and localizing diffusion-generated images in open-world settings by presenting the OpenSDID benchmark dataset. OpenSDID stands out from existing datasets due to its use of diverse vision-language models and advanced diffusion models to replicate real-world local and global manipulations. To address this challenge, we proposed the approach of Synergizing Pretrained Models (SPM) and introduced MaskCLIP, an SPM-based model that synergizes CLIP and MAE. This enhances generalization at both image and pixel levels, achieving state-of-the-art detection and localization on OpenSDID. We envision OpenSDID as an open-ended dataset, capable of incorporating images generated by increasingly advanced generative AI models and serving as a standard benchmark for this evolving domain.

7. Acknowledgment

This research was supported by the ECS Fund (504079140) at the University of Southampton, National Natural Science Foundation of China (62376070, 62076195), and the Fundamental Research Funds for the Central Universities (AUGA-5710011522). Yabin Wang was supported by the China Scholarship Council (CSC) Grant No. 202306280130. This work was primarily conducted at the University of Southampton. The authors acknowledge the use of the IRIDIS X High Performance Computing Facility and the Southampton-Wolfson AI Research Machine (SWARM) GPU cluster, funded by the Wolfson Foundation, together with the associated support services at the University of Southampton in the completion of this work.

References

[1] Meta AI. Llama 3.2-vision: Multimodal large language model. Hugging Face Model Hub, 2024. Model release date: September 25, 2024. 1, 3

[2] Ollin Boer Bohan. Megalith-10m: A dataset of public domain photographs. https://huggingface.co/datasets/madebyollin/megalith-10m, 2024. Accessed: 2024-10-07. 3

[3] Lucy Chai, David Bau, Ser-Nam Lim, and Phillip Isola. What makes fake images detectable? understanding properties that generalize. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXVI 16*, pages 103–120. Springer, 2020. 2

[4] Xinru Chen, Chengbo Dong, Jiaqi Ji, Juan Cao, and Xirong Li. Image manipulation detection by multi-view multi-scale supervision. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 14185–14193, 2021. 3, 6, 7

[5] Yirui Chen, Xudong Huang, Quan Zhang, Wei Li, Mingjian Zhu, Qiangyu Yan, Simiao Li, Hanting Chen, Hailin Hu, Jie Yang, et al. Gim: A million-scale benchmark for generative image manipulation detection and localization. *arXiv preprint arXiv:2406.16531*, 2024. 2

[6] Zhe Chen, Weiyun Wang, Hao Tian, Shenglong Ye, Zhangwei Gao, Erfei Cui, Wenwen Tong, Kongzhi Hu, Jiapeng Luo, Zheng Ma, et al. How far are we to gpt-4v? closing the gap to commercial multimodal models with open-source suites. *Science China Information Sciences*, 67(12):220101, 2024. 3

[7] François Chollet. Xception: Deep learning with depthwise separable convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1251–1258, 2017. 3

[8] Jing Dong, Wei Wang, and Tieniu Tan. Casia image tampering detection evaluation database. In *2013 IEEE China summit and international conference on signal and information processing*, pages 422–426. IEEE, 2013. 2

[9] Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, et al. Scaling rectified flow transformers for high-resolution image synthesis. In *Forty-first International Conference on Machine Learning*, 2024. 1, 2, 3

[10] Joel Frank, Thorsten Eisenhofer, Lea Schönherr, Asja Fischer, Dorothea Kolossa, and Thorsten Holz. Leveraging frequency analysis for deep fake image recognition. In *International conference on machine learning*, pages 3247–3258. PMLR, 2020. 3

[11] Haiying Guan, Mark Kozak, Eric Robertson, Yooyoung Lee, Amy N Yates, Andrew Delgado, Daniel Zhou, Timothee Kheyrkhah, Jeff Smith, and Jonathan Fiscus. Mfc datasets: Large-scale benchmark datasets for media forensic challenge evaluation. In *2019 IEEE Winter Applications of Computer Vision Workshops (WACVW)*, pages 63–72. IEEE, 2019. 2

[12] Fabrizio Guillaro, Davide Cozzolino, Avneesh Sud, Nicholas Dufour, and Luisa Verdoliva. Trufor: Leveraging all-round clues for trustworthy image forgery detection and localization. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 20606–20615, 2023. 2, 3, 4, 6, 7

[13] Xiao Guo, Xiaohong Liu, Zhiyuan Ren, Steven Grosz, Iacopo Masi, and Xiaoming Liu. Hierarchical fine-grained image forgery detection and localization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3155–3165, 2023. 2, 6

[14] 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 Transactions on Pattern Analysis and Machine Intelligence*, 2024. 3

[15] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2016. 3

[16] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 2961–2969, 2017. 6

[17] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16000–16009, 2022. 4, 6

[18] Yu-Feng Hsu and Shih-Fu Chang. Detecting image splicing using geometry invariants and camera characteristics consistency. In *2006 IEEE International Conference on Multimedia and Expo*, pages 549–552. IEEE, 2006. 2

[19] Xuefeng Hu, Zhihan Zhang, Zhenye Jiang, Syomantak Chaudhuri, Zhenheng Yang, and Ram Nevatia. Span: Spatial pyramid attention network for image manipulation localization. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXI 16*, pages 312–328. Springer, 2020. 3

[20] Menglin Jia, Luming Tang, Bor-Chun Chen, Claire Cardie, Serge Belongie, Bharath Hariharan, and Ser-Nam Lim. Visual prompt tuning. In *European Conference on Computer Vision*, pages 709–727. Springer, 2022. 4

[21] Shan Jia, Mingzhen Huang, Zhou Zhou, Yan Ju, Jialing Cai, and Siwei Lyu. Autosplice: A text-prompt manipulated image dataset for media forensics. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 893–903, 2023. 2

[22] Sohail Ahmed Khan and Duc-Tien Dang-Nguyen. Clipping the deception: Adapting vision-language models for universal deepfake detection. *arXiv preprint arXiv:2402.12927*, 2024. 3

[23] Muhammad Uzair Khattak, Hanoona Rasheed, Muhammad Maaz, Salman Khan, and Fahad Shahbaz Khan. Maple: Multi-modal prompt learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19113–19122, 2023. 4

[24] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4015–4026, 2023. 2, 3, 8

[25] Christos Koutlis and Symeon Papadopoulos. Leveraging representations from intermediate encoder-blocks for synthetic image detection. In *Computer Vision – ECCV 2024*, pages 394–411, Cham, 2025. Springer Nature Switzerland. 6, 7

[26] Myung-Joon Kwon, Seung-Hun Nam, In-Jae Yu, Heung-Kyu Lee, and Changick Kim. Learning jpeg compression artifacts for image manipulation detection and localization. *International Journal of Computer Vision*, 130(8):1875–1895, 2022. 3, 6, 7

[27] Black Forest Labs. Flux.1: A 12 billion parameter rectified flow transformer for text-to-image generation. https://github.com/black-forest-labs/flux, 2024. A 12 billion parameter rectified flow transformer capable of generating images from text descriptions. 1, 2, 3

[28] Yingxin Lai, Zhiming Luo, and Zitong Yu. Detect any deepfakes: Segment anything meets face forgery detection and localization. In *Chinese Conference on Biometric Recognition*, pages 180–190. Springer, 2023. 3

[29] Dong Li, Jiaying Zhu, Xueyang Fu, Xun Guo, Yidi Liu, Gang Yang, Jiawei Liu, and Zheng-Jun Zha. Noise-assisted prompt learning forimage forgery detection and localization. In *Computer Vision – ECCV 2024*, pages 18–36, Cham, 2025. Springer Nature Switzerland. 3

[30] Yanghao Li, Haoqi Fan, Ronghang Hu, Christoph Feichtenhofer, and Kaiming He. Scaling language-image pre-training via masking. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 23390–23400, 2023. 2

[31] Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2117–2125, 2017. 4, 6

[32] Bo Liu, Fan Yang, Xiuli Bi, Bin Xiao, Weisheng Li, and Xinbo Gao. Detecting generated images by real images. In *European Conference on Computer Vision*, pages 95–110. Springer, 2022. 3

[33] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. In *NeurIPS*, 2023. 2, 3

[34] Xiaohong Liu, Yaojie Liu, Jun Chen, and Xiaoming Liu. Pscc-net: Progressive spatio-channel correlation network for image manipulation detection and localization. *IEEE Transactions on Circuits and Systems for Video Technology*, 32 (11):7505–7517, 2022. 3, 6, 7

[35] Zhengzhe Liu, Xiaojuan Qi, and Philip HS Torr. Global texture enhancement for fake face detection in the wild. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8060–8069, 2020. 3, 6, 7

[36] Xiaochen Ma, Bo Du, Xianggen Liu, Ahmed Y Al Hammadi, and Jizhe Zhou. Iml-vit: Image manipulation localization by vision transformer. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2024. 3, 4, 6, 7

[37] Xiaochen Ma, Xuekang Zhu, Lei Su, Bo Du, Zhuohang Jiang, Bingkui Tong, Zeyu Lei, Xinyu Yang, Chi-Man Pun, Jiancheng Lv, et al. Imdl-benco: A comprehensive benchmark and codebase for image manipulation detection & localization. *arXiv preprint arXiv:2406.10580*, 2024. 6, 7

[38] Hannes Mareen, Dimitrios Karageorgiou, Glenn Van Wallendael, Peter Lambert, and Symeon Papadopoulos. Tgif: Text-guided inpainting forgery dataset. In *2024 IEEE International Workshop on Information Forensics and Security (WIFS)*, pages 1–6. IEEE, 2024. 2, 3

[39] Alexander Quinn Nichol, Prafulla Dhariwal, Aditya Ramesh, Pranav Shyam, Pamela Mishkin, Bob Mcgrew, Ilya Sutskever, and Mark Chen. Glide: Towards photorealistic image generation and editing with text-guided diffusion models. In *International Conference on Machine Learning*, pages 16784–16804. PMLR, 2022. 2

[40] Utkarsh Ojha, Yuheng Li, and Yong Jae Lee. Towards universal fake image detectors that generalize across generative models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 24480–24489, 2023. 3, 5, 6, 7

[41] 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 *The Twelfth International Conference on Learning Representations*, 2024. 1, 2, 3

[42] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021. 3, 4

[43] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 1 (2):3, 2022. 2

[44] Shuhuai Ren, Aston Zhang, Yi Zhu, Shuai Zhang, Shuai Zheng, Mu Li, Alexander J Smola, and Xu Sun. Prompt pretraining with twenty-thousand classes for open-vocabulary visual recognition. *Advances in Neural Information Processing Systems*, 36, 2024. 6

[45] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022. 1, 2, 3

[46] Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Yael Pritch, Michael Rubinstein, and Kfir Aberman. Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 22500–22510, 2023. 1

[47] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, et al. Photorealistic text-to-image diffusion models with deep language understanding. *Advances in neural information processing systems*, 35:36479–36494, 2022. 1

[48] Stefan Smeu, Elisabeta Oneata, and Dan Oneata. Declip: Decoding clip representations for deepfake localization. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, 2025. 6, 7

[49] Chuangchuang Tan, Huan Liu, Yao Zhao, Shikui Wei, Guanghua Gu, Ping Liu, and Yunchao Wei. Rethinking the up-sampling operations in cnn-based generative network for generalizable deepfake detection, 2023. 3, 6, 7

[50] Chuangchuang Tan, Yao Zhao, Shikui Wei, Guanghua Gu, and Yunchao Wei. Learning on gradients: Generalized artifacts representation for gan-generated images detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12105–12114, 2023.

[51] Chuangchuang Tan, Yao Zhao, Shikui Wei, Guanghua Gu, Ping Liu, and Yunchao Wei. Frequency-aware deepfake detection: Improving generalizability through frequency space learning, 2024. 3, 6, 7

[52] Junke Wang, Zuxuan Wu, Jingjing Chen, Xintong Han, Abhinav Shrivastava, Ser-Nam Lim, and Yu-Gang Jiang. Objectformer for image manipulation detection and localization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2364–2373, 2022. 3, 6, 7

[53] Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, et al. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*, 2024. 2, 3

[54] Sheng-Yu Wang, Oliver Wang, Richard Zhang, Andrew Owens, and Alexei A Efros. Cnn-generated images are surprisingly easy to spot... for now. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8695–8704, 2020. 2, 4, 6, 7

[55] Yabin Wang, Zhiwu Huang, and Xiaopeng Hong. S-prompts learning with pre-trained transformers: An occam's razor for domain incremental learning. *Advances in Neural Information Processing Systems*, 35:5682–5695, 2022. 3, 4, 5, 6

[56] Zhendong Wang, Jianmin Bao, Wengang Zhou, Weilun Wang, Hezhen Hu, Hong Chen, and Houqiang Li. Dire for diffusion-generated image detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 22445–22455, 2023. 3

[57] Zijie J Wang, Evan Montoya, David Munechika, Haoyang Yang, Benjamin Hoover, and Duen Horng Chau. Diffusiondb: A large-scale prompt gallery dataset for text-to-image generative models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 893–911, 2023. 2

[58] Haiwei Wu and Jiantao Zhou. Iid-net: Image inpainting detection network via neural architecture search and attention. *IEEE Transactions on Circuits and Systems for Video Technology*, 32(3):1172–1185, 2021. 3

[59] Yue Wu, Wael AbdAlmageed, and Premkumar Natarajan. Mantra-net: Manipulation tracing network for detection and localization of image forgeries with anomalous features. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9543–9552, 2019. 3

[60] Bin Xiao, Haiping Wu, Weijian Xu, Xiyang Dai, Houdong Hu, Yumao Lu, Michael Zeng, Ce Liu, and Lu Yuan. Florence-2: Advancing a unified representation for a variety of vision tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4818–4829, 2024. 2, 3

[61] Zeyue Xue, Guanglu Song, Qiushan Guo, Boxiao Liu, Zhuofan Zong, Yu Liu, and Ping Luo. Raphael: Text-to-image generation via large mixture of diffusion paths. *Advances in Neural Information Processing Systems*, 36, 2024. 1

[62] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3836–3847, 2023. 1

[63] Shihao Zhao, Dongdong Chen, Yen-Chun Chen, Jianmin Bao, Shaozhe Hao, Lu Yuan, and Kwan-Yee K Wong. Uni-controlnet: All-in-one control to text-to-image diffusion models. *Advances in Neural Information Processing Systems*, 36, 2024. 1

[64] 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. *Advances in Neural Information Processing Systems*, 36, 2024. 2