OpenSDI: Spotting Diffusion-Generated Images in the Open World

Yabin Wang [1] Zhiwu Huang [2] Xiaopeng Hong [3]

[1]Xi'an Jiaotong University [2]University of Southampton [3]Harbin Institute of Technology
iamwangyabin@stu.xjtu.edu.cn Zhiwu.Huang@soton.ac.uk hongxiaopeng@ieee.org

Problem & Challenge

The rapid advancement of Text-to-Image (T2I) diffusion models blurs the line between real and AI-generated content, posing a significant challenge to content authenticity. We identify this as the **Open-world Spotting of Diffusion Images (OpenSDI)** challenge, characterized by three key open-world settings:

- **User Diversity:** Wide range of user preferences in styles and intentions.
- **Model Innovation:** Rapid evolution of diffusion models with diverse architectures.
- **Manipulation Scope:** Broad spectrum from global synthesis to precise local edits.

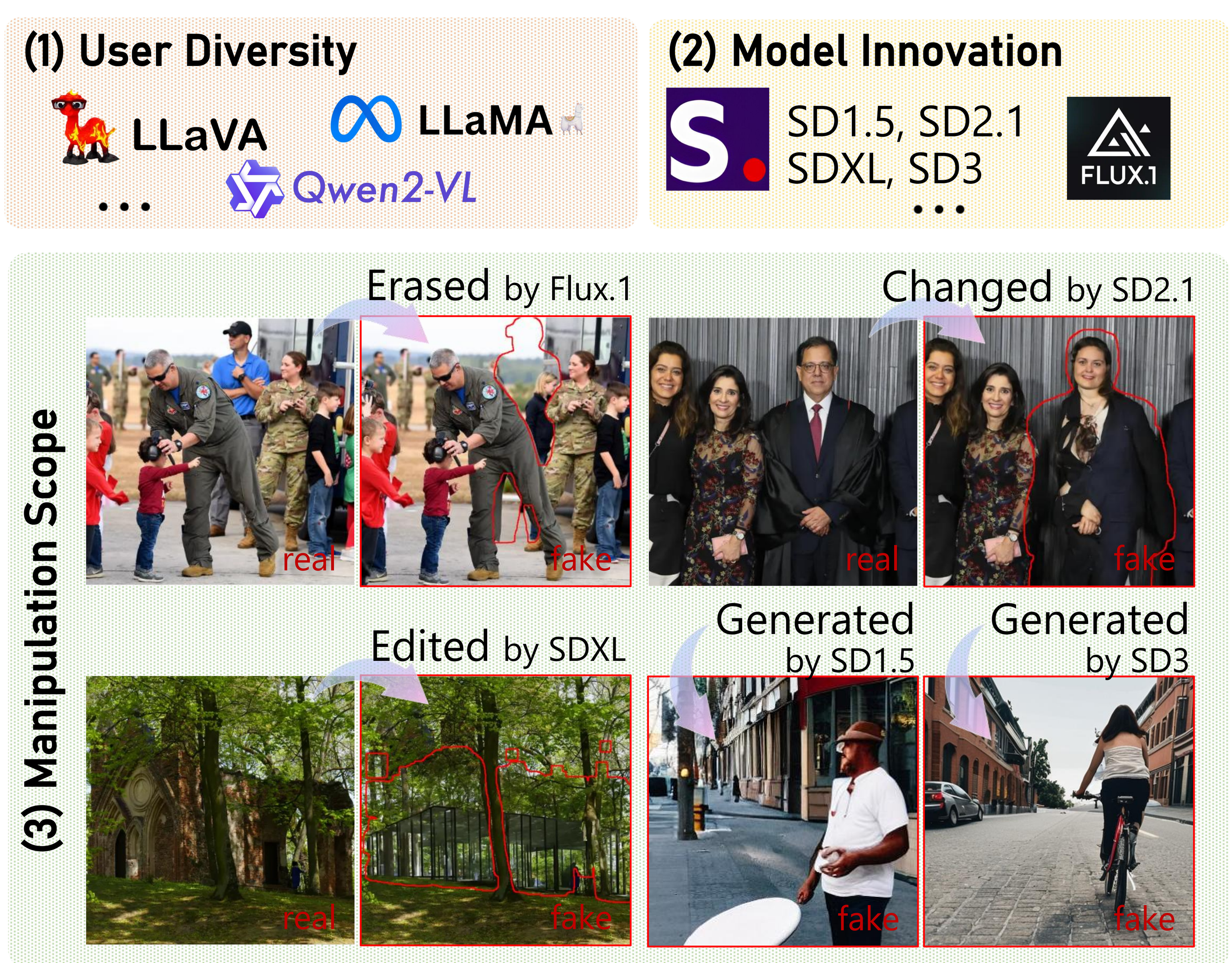

Figure 1. OpenSDI Challenge Settings: (1) User Diversity, (2) Model Innovation, and (3) Full Manipulation Scope.

How Do We Create Dataset?

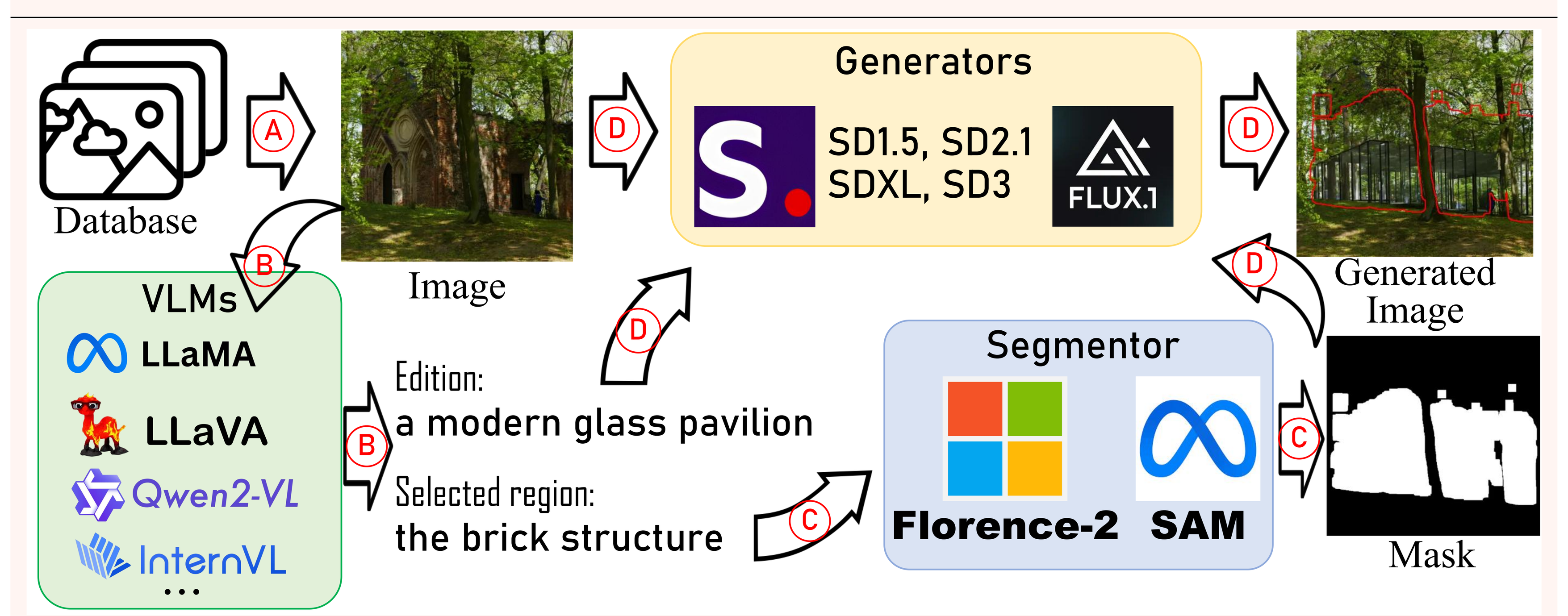

Figure 2. OpenSDID Pipeline: Local Editing via (A) Sampling, (B) Instruction, (C) Masking, (D) Generation; Global Generation with (B) and (D), no real images or masks.

OpenSDID: A Benchmark for Open-World Diffusion Image Detection

Key Features of OpenSDID:

- **User Diversity:** Simulating real-world user intentions via diverse VLM prompts.
- **Model Innovation:** Incorporating SOTA diffusion models (SD1.5, SD2.1, SDXL, SD3, Flux.1).
- **Manipulation Scope:** Covering global synthesis and local edits.

Dataset	# Real	# Fake	User	Model	Full
DiffusionDB	–	14M	✓	✗	✗
GenImage	1.3M	1.4M	✗	✓	✗
AutoSplice	2.3K	3.6K	✗	✗	✓
CocoGlide	–	512	✗	✗	✗
HiFi-IFDL	–	1M*	✓	✗	✓
GIM	1.1M	1.1M	✗	✓	✗
TGIF	3.1K	75K	✗	✓	✗
OpenSDID	300K	450K	✓	✓	✓

Table 1. Overview of existing diffusion image datasets and the proposed OpenSDID.

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 on OpenSDID.

Synergizing Pretrained Models (SPM): MaskCLIP

Tackling the OpenSDI challenge requires **detection** and **localization**, while generalizing to diverse, open-world scenarios. We introduce Synergizing Pretrained Models (SPM), a novel framework that achieves this by:

- **Prompting**: Efficiently adapting pretrained models to the OpenSDI task using learned prompts, preserving their existing knowledge.
- **Attending**: Creating synergy between multiple models through cross-attention, enhancing overall performance.

Leveraging SPM, we developed **MaskCLIP**[a], a model that strategically fuses the strengths of **CLIP** and **MAE**

- **Visual Cross-Attention (VCA)**: Aligns CLIP and MAE visual features.
- **Textual-Visual Cross-Attention (TVCA)**: Integrates text semantics into localization.
- **Visual Self-Attention (VSA)**: Enhances CLIP feature extraction.

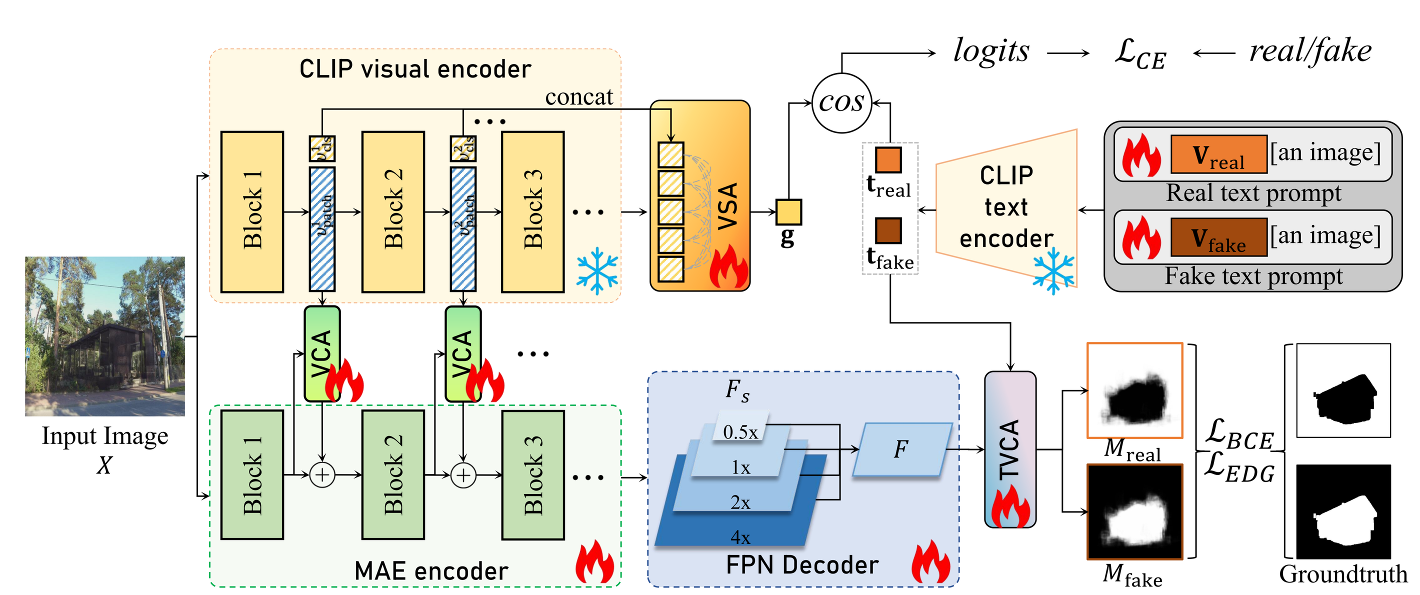

Figure 3. MaskCLIP Architecture: Synergizing CLIP and MAE with Prompting and Attention Mechanisms.

[a]We also developed SAMCLIP by replacing MAE with SAM, achieving a 0.7873 Pixel F1 on SD1.5, demonstrating the scalability of our SPM scheme.

Experiments

Key Performance Highlights:

- **DRAMATICALLY SUPERIOR LOCALIZATION:** Achieves a remarkable **+14.23% IoU** and **+14.11% F1** relative improvement in average localization accuracy, outperforming the second-best method by a significant margin. This demonstrates MaskCLIP's unparalleled precision in pinpointing manipulated regions.
- **NEW SOTA DETECTION:** Demonstrates clear improvement over existing state-of-the-art image-level image detection methods.
- **ROBUST OPEN-WORLD GENERALIZATION:** Exhibits outstanding generalization, consistently surpassing state-of-the-art methods across all *unseen* diffusion models (SD2.1, SDXL, SD3, Flux.1) at both image and pixel levels. This highlights MaskCLIP's ability to adapt and perform reliably in truly open-world scenarios.

Quantitative Supremacy:

The tables below showcase a detailed performance breakdown.

Table 3. Pixel-level Localization Performance Comparison (IoU & F1)

Method	SD1.5		SD2.1		SDXL		SD3		Flux.1		AVG	
	IoU	F1	IoU	F1	IoU	F1	IoU	F1	IoU	F1	IoU	F1
2nd Best	0.665	0.736	0.448	0.506	0.215	0.260	0.236	0.284	0.0611	0.079	0.325	0.373
MaskCLIP	**0.671**	**0.756**	**0.555**	**0.629**	**0.310**	**0.370**	**0.438**	**0.512**	**0.162**	**0.203**	**0.427**	**0.494**

Table 4. Image-level Detection Performance Comparison (F1 & Accuracy)

Method	SD1.5		SD2.1		SDXL		SD3		Flux.1		AVG	
	F1	Acc	F1	Acc	F1	Acc	F1	Acc	F1	Acc	F1	Acc
2nd Best	0.911	0.910	0.875	0.881	0.734	0.788	0.721	0.768	0.559	0.670	0.760	0.803
MaskCLIP	**0.926**	**0.927**	**0.887**	**0.895**	**0.780**	**0.812**	**0.731**	**0.780**	**0.565**	**0.685**	**0.778**	**0.820**

Visual Confirmation:

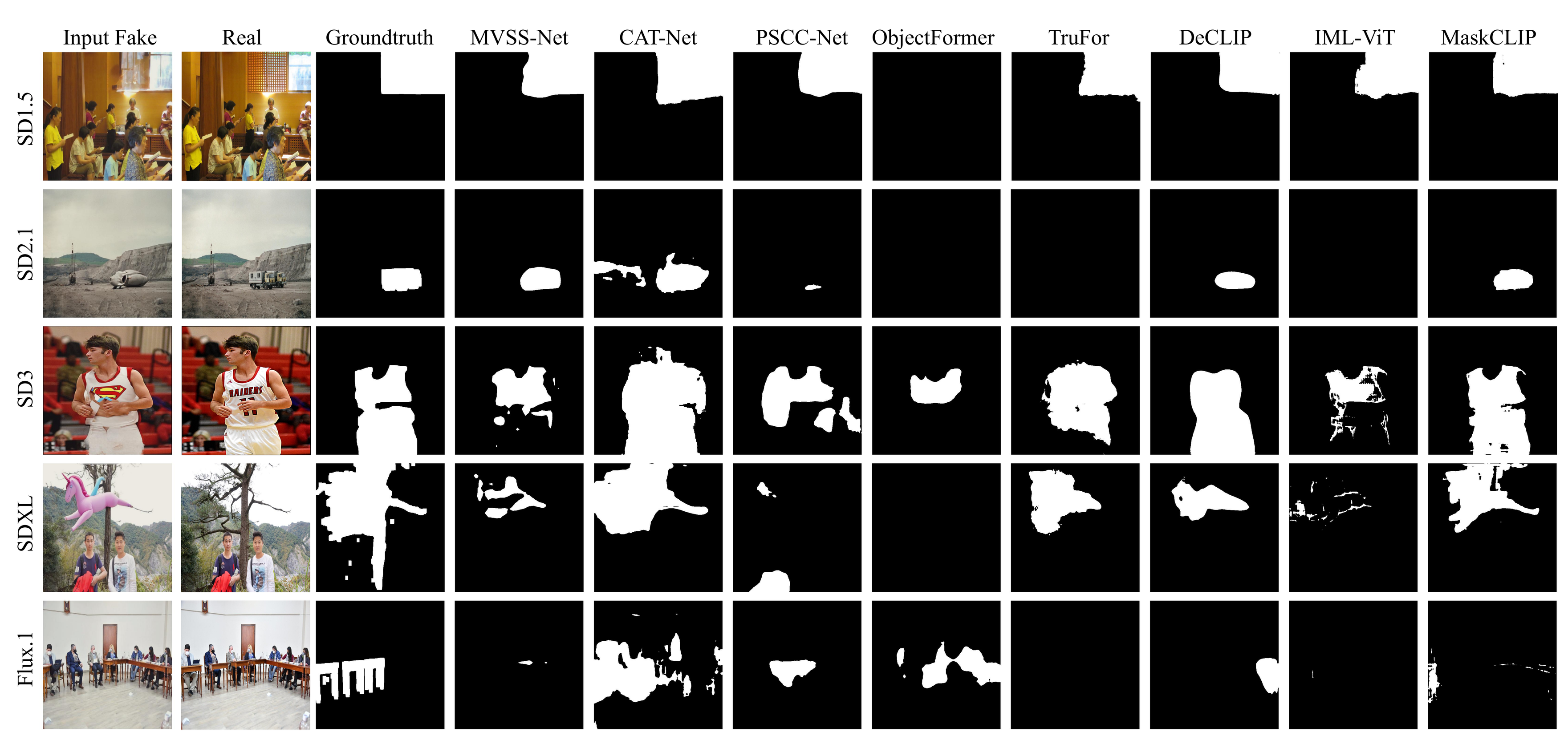

Figure 4. Qualitative Localization Results on OpenSDID.