Training-Free Image Manipulation Localization Using Diffusion Models

Zhenfei Zhang, Ming-Ching Chang, Xin Li

Department of Computer Science, University at Albany, State University of New York, New York, USA, 12222
zzhang45@albany.edu, mchang2@albany.edu, xli48@albany.edu

Abstract

Image manipulation localization (IML) is a critical technique in media forensics, focusing on identifying tampered regions within manipulated images. Most existing IML methods require extensive training on labeled datasets with both image-level and pixel-level annotations. These methods often struggle with new manipulation types and exhibit low generalizability. In this work, we propose a training-free IML approach using diffusion models. Our method adaptively selects an appropriate number of diffusion timesteps for each input image in the forward process and performs both *conditional* and *unconditional* reconstructions in the backward process without relying on external conditions. By comparing these reconstructions, we generate a localization map highlighting regions of manipulation based on inconsistencies. Extensive experiments were conducted using sixteen state-of-the-art (SoTA) methods across six IML datasets. The results demonstrate that our training-free method outperforms SoTA unsupervised and weakly-supervised techniques. Furthermore, our method competes effectively against fully-supervised methods on novel (unseen) manipulation types.

Introduction

Image manipulation localization (IML) aims to locate tampered regions within an image. This technology has become increasingly important due to the advancements in media editing and generation methods, such as Photoshop and Generative AI techniques (Qiao et al. 2019; Xu et al. 2018; Zhang and Chang 2023; Dhariwal and Nichol 2021a), to ensure media authentication. Traditional image manipulation types fall into three categories: *removal*, where media content is removed and synthesized; *splicing*, which involves inserting content from a different source into an image; and *copy-move*, which involves relocating content within the same image.

Even though fully-supervised IML methods have achieved satisfactory localization performance on some common IML datasets, they still have several drawbacks. First, they require extensive training with datasets including image and pixel-level annotations, which are costly. Second, these methods perform poorly when localizing manipulation types different from those in the training datasets, resulting in low generalizability and unsatisfactory performance in real-world scenarios. Given the numerous and ever-growing types of tampering, it is impractical to create datasets that fully encompasses all tampering types for model training.

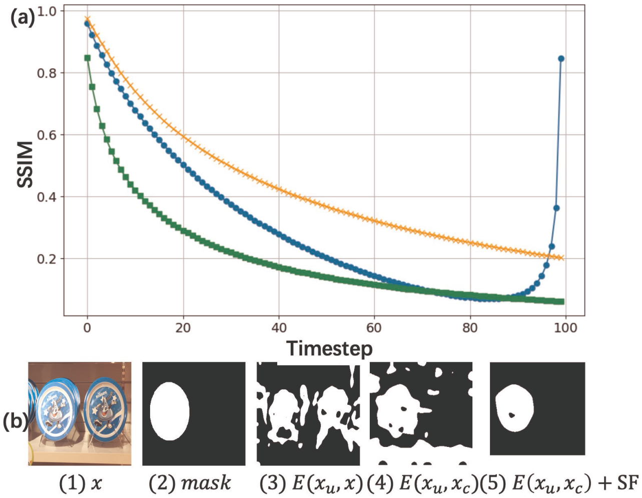

Figure 1: **(a)** SSIM scores at various timesteps are shown for forward and backward diffusion processes. For the forward process, results with a high-pass filter are indicated by a green line, and without a high-pass filter by an orange line. The backward diffusion process is depicted with a blue line. These scores are averaged across CASIAv1 (Dong, Wang, and Tan 2010), Coverage (Wen et al. 2016), and Columbia (Hsu and Chang 2006) datasets. **(b)** From left to right: the tampered image, ground-truth mask, and three error masks (unconditional reconstruction *vs.* input, unconditional *vs.* conditional reconstruction, and unconditional *vs.* conditional reconstruction with self-attention guidance).

To address the aforementioned issues as well as improve the generalizability of IML methods for real-world scenarios, this work explores the possibility of a training-free method for IML that does not require any training datasets for learning. Our initial experiment is inspired by **diffusion purification** methods (Nie et al. 2022; Wang et al. 2022), which have demonstrated that *diffusion models (DM)*, having learned the clean data distribution, can effectively re-

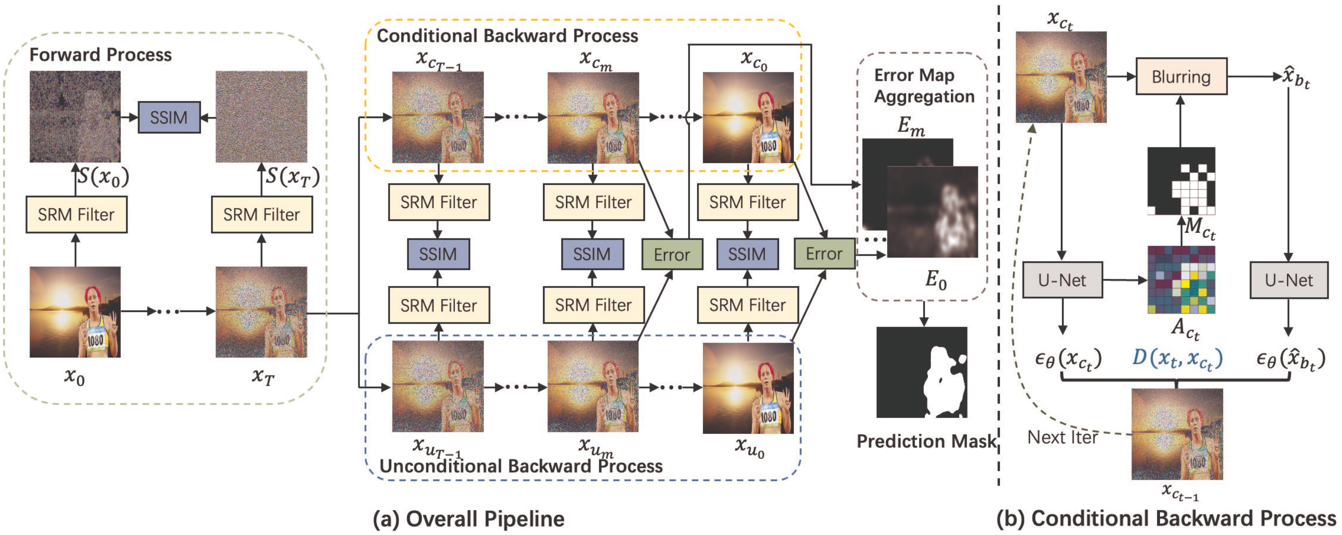

Figure 2: **(a)** Overview of our IML method. In the forward process, $S(x_0)$ is compared with each $S(x_t)$ using SSIM scores. These scores help choose the appropriate T to remove manipulation traces while preserving the input image's structure. Two backward processes then aggregate the error maps starting from the backward timestep m, where SSIM is lowest. **(b)** The conditional denoising is guided by both self-attention and similarity.

move adversarial attacks. As an extension of image purification, the work of (Tailanián et al. 2024) shows that DM can hide manipulation traces, resulting in decreased performance of IML methods. Based on this idea and its successful results, we propose the following hypothesis: *Since DM learns the clean data distribution using authentic images, and forensic traces can be hidden after the diffusion reverse process, can we use this property to locate possible manipulations through the reconstruction inconsistencies?*

To verify this hypothesis, we start with feeding tampered images into an unconditional DM, akin to the diffusion purification, to obtain the reconstructed images. A key issue before this is choosing the appropriate number of *diffusion timesteps* T. If T is too large, the reconstructed image may deviate significantly from the input, introducing unwanted artifacts. Conversely, if T is too small, the method might not effectively remove the tampered traces. Previous purification methods often use a fixed T for all images, which is clearly suboptimal. Inspired by the high-pass (HP) filters (Fridrich and Kodovsky 2012; Bayar and Stamm 2018) commonly used in IML to enhance performance by filtering out image content, we use HP filters to assess whether tampering traces have been effectively removed. The green and orange lines in Fig. 1(a) illustrate the Structural Similarity Index Metric (SSIM) (Wang et al. 2004) at different timesteps in the diffusion forward process, with and without the application of HP filters. The SSIM scores are calculated between each time-step-noised sample and the original input. Both trends show a consistent decrease, indicating that more noise leads to greater deviation from the original input. The SSIM with HP filters (green curve) drop more rapidly, which helps in selecting T that effectively removes tampered traces while preserving the structure of the input image.

We obtained the reconstruction error map by comparing the original input against the reconstructed image. Unfortunately, the results did not align with our expectations and assumptions, as shown in Fig. 1(b3), where the error map covers the entire foreground region. This observation shows that using DM directly cannot differentiate between the tampered and authentic image regions. The underlying issue is that while the DM can reconstruct the tampered image to align with a clean distribution, leading to inconsistencies in tampered regions, it fails to accurately reconstruct the authentic pixel values, resulting in unexpected inconsistencies in the authentic regions as well. To address this issue, we modify the diffusion reverse process to start from the same noised image x_T and perform both **conditional** and **unconditional** reconstructions. The conditional reconstruction is guided by the forged image, using similarity scores SSIM (Wang et al. 2004) to reconstruct the tampered traces, while the unconditional reconstruction generates a clean image devoid of manipulation traces. We seek for a diffusion reconstruction that minimizes inconsistencies in authentic pixels, while ensuring that the error is concentrated solely on the tampered regions, such that IML can be achieved. We also ensure that both reverse diffusion processes use the same random noise in the sampling step to minimize the impact of noise randomness of the results.

The error mask between two backward processes focuses more on the tampered region rather than the entire foreground, as shown in the example in Fig. 1(b4). However, there is one more challenge to overcome: Due to the global guidance of the conditional generation by SSIM, the result still contains many false alarms. To address these false positives, inspired by the *self-attention (SF) guidance diffusion model* (Hong et al. 2023), which demonstrates that self-attention masks from DM overlap with high-frequency regions. We incorporate the guidance from both SF and

SSIM into the conditional branch to direct the reconstruction more precisely towards the tampered regions. The final error mask, shown in Fig. 1(b5), contains much less false positives, achieving the best IML effects. Unlike traditional guided diffusion methods, our conditional backward process does not require any external conditions (such as class labels or text), thereby demonstrating strong generalizability.

We also observed that the SSIM values between unconditional and conditional samplings along backward timestamps, shown by the blue curve in Fig. 1(a), initially decrease and then increase. This pattern is similar to what was observed in (Che et al. 2024) with external conditions (image-level labels). Based on this observation, and following the approach in (Che et al. 2024), we obtain the final error mask by aggregating the error maps starting from the backward timestep when SSIM reaches its minimum. This approach produces the best performance in our experiments. Fig. 2 overviews our method.

We evaluated our IML method on six public datasets: five standard datasets with common tampering types and one novel dataset with unseen and more complex manipulation types. The results show that our training-free method outperforms State-of-The-Art (SoTA) unsupervised and weakly-supervised approaches. Additionally, our method competes effectively with fully-supervised methods on unseen, novel manipulation types, demonstrating stronger generalizability.

The contributions can be summarized as follows:

- We present a novel image manipulation localization approach that does not require any training or training data.
- The conditional backward process in our method operates without relying on external conditions, making the approach more generalizable.
- We conducted comprehensive evaluations of sixteen SoTA methods using six IML datasets, encompassing unsupervised, weakly-supervised, and fully-supervised approaches. The results demonstrate superior performance on both standard and novel tampered datasets compared to existing SoTA methods.

Related Work

Denoising Diffusion Probabilistic Model

The Denoising Diffusion Probabilistic Model (DDPM) (Ho, Jain, and Abbeel 2020) has become popular because of its superior generative capabilities compared to earlier generative models such as GANs (Goodfellow et al. 2020) and VAEs (Kingma and Welling 2013). DDPM involves two main processes: the forward process adds noise to the image, while the backward process removes the noise to produce a clean image. DDPM has been widely used in media generation and editing (Dhariwal and Nichol 2021b; Kawar et al. 2023; Zhang et al. 2024b), image segmentation (Wolleb et al. 2022; Amit et al. 2021), and image classification (Yang et al. 2023). In 2024, the study by (Yu et al. 2024) employed DDPM as the decoder and SegFormer (Xie et al. 2021) as the encoder for IML task, necessitating extensive training with labeled datasets. In contrast, our method is training-free and relies solely on DDPM, without requiring any external conditions.

Image Manipulation Localization (IML)

IML methods can be organized into three types: unsupervised, weakly-supervised, and fully-supervised approaches. Most unsupervised methods apply hand-crafted features such as noise inconsistency (Mahdian and Saic 2009; Lyu, Pan, and Zhang 2014; Wagner 2015), color filter array (Ferrara et al. 2012; Dirik and Memon 2009; Choi, Choi, and Lee 2011), local mosaic consistency (Bammey, Gioi, and Morel 2020), JPEG compression (Li, Yuan, and Yu 2009) and camera fingerprint (Cozzolino and Verdoliva 2019). The work by (Zhang et al. 2025) utilized implicit neural representation (Sitzmann et al. 2020), for both unsupervised and weakly supervised methods. Additionally, the study in (Zhai et al. 2023) introduced self-consistency learning, another approach to weakly supervised learning. Fully-supervised IML methods (Chen et al. 2021; Yang et al. 2020; Wu, AbdAlmageed, and Natarajan 2019; Liu et al. 2022; Guo et al. 2023) require large datasets with both image and pixel-level annotations. Most of these methods learn manipulation traces by detecting anomalous features. (Kwon et al. 2022; Zhang, Li, and Chang 2024) proposed a two-branch network that can effectively detect both image editing and double JPEG compression artifacts.

Most SoTA methods require extensive training. Although some unsupervised methods do not require training, they often perform poorly even on standard manipulation types. In this paper, we introduce a novel, simple, yet effective method that does not require any training or datasets. Our method demonstrates strong generalizability and enhanced performance in real-life scenarios.

Preliminaries

The Denoising Diffusion Probabilistic Model (DDPM): The DDPM (Ho, Jain, and Abbeel 2020) consists of two main processes: the forward process, which adds noise, and the backward process, which removes noise. In the forward process, Gaussian noise is gradually added to the image x_0 to obtain the noised image x_t. The formula for the DDPM forward process is:

$$x_t = \sqrt{\bar{\alpha}_t}x_0 + \sqrt{1-\bar{\alpha}_t}\epsilon, \epsilon \sim \mathcal{N}(0, I), \tag{1}$$

where $\alpha_t = 1 - \beta_t$ and $\bar{\alpha}_t = \prod_{i=1}^{t} \alpha_i$. ϵ is the random noise from normal distribution, and β_t is the predefined variance schedule at a timestep t.

In the backward process, the model removes the noise from x_t to obtain x_{t-1}. The formula for the DDPM backward process is represented as follows:

$$x_{t-1} = \frac{1}{\sqrt{\bar{\alpha}_t}}\left(x_t - \frac{\beta_t}{\sqrt{1-\bar{\alpha}_t}}\epsilon_\theta(x_t, t)\right) + \sigma_t z, \tag{2}$$

where $z \sim \mathcal{N}(0, I)$ and $\sigma_t^2 = \beta_t$. $\epsilon_\theta(x_t, t)$ represents the predicted noise of x_t using trained U-Net (Ronneberger, Fischer, and Brox 2015).

The training objective is to minimize the difference between the predicted and ground-truth noise, as shown by the following equation:

$$\mathcal{L}(\theta) = \mathbb{E}_{t, x_{0 \sim data, \epsilon \sim (0, I)}}[\|\epsilon - \epsilon_\theta(x_t, t)\|^2]. \tag{3}$$

Classifier and classifier-free guidance: Traditional DDPM often produce random outputs that may not meet specific real-world needs. To address this, conditional DDPM are introduced using either classifier guidance (Dhariwal and Nichol 2021b) and classifier-free guidance (Ho and Salimans 2022). For classifier guidance, a separate classifier $p(c|x_t)$ is trained to predict a condition c from x_t. Let s_c denote the classifier guiding scale and $\widetilde{\epsilon}(\mathbf{x}_t, c, t)$ denote the conditional output based on condition c on timestep t. The classifier guidance is given by:

$$\widetilde{\epsilon}(\mathbf{x}_t, c, t) = \epsilon_\theta(\mathbf{x}_t, t) - s_c \cdot \sigma_t \nabla_{\mathbf{x}_t} \log p(c|\mathbf{x}_t). \quad (4)$$

The main drawback of classifier guidance is the need to train a standalone classifier. To address this, a classifier-free method is introduced in (Ho and Salimans 2022). Let s_f denote the classifier-free guiding scale. The classifier-free guidance is:

$$\tilde{\epsilon}(\mathbf{x}_t, c, t) = \epsilon_\theta(\mathbf{x}_t, t) + s_f \cdot (\epsilon_\theta(\mathbf{x}_t, c, t) - \epsilon_\theta(\mathbf{x}_t, t)). \quad (5)$$

Refer to the original papers (Ho, Jain, and Abbeel 2020; Dhariwal and Nichol 2021b; Ho and Salimans 2022) for detailed derivation.

Method

Fig. 2(a) shows the overall pipeline of our method, which is both training-free and condition-free. Our method includes a single *forward process* that adds noise to the image and two *backward processes* that reconstruct the image with and without manipulation traces. Let t denote a timestep where $t \in [0, T]$, with T being the final timestep. During the forward process, samples are denoted as x_t. In the *conditional* and *unconditional* backward processes, samples are denoted as x_{c_t} and x_{u_t}, respectively. Let A_{c_t} denote the self-attention map, and M_{c_t} denote the corresponding attention mask.

In the forward process, let $S(\cdot)$ denote the steganalysis rich model (SRM) filters (Fridrich and Kodovsky 2012). The SSIM values between $S(x_0)$ and each $S(x_t)$ are used to adaptively select an appropriate T to add noise, with the goal of removing tampered traces while preserving the overall structure of the input image.

Starting from the same noised image x_T, the two backward processes diverge: *unconditional denoising* produces a clean image without tampered traces, as the pre-trained diffusion model has learned a clean data distribution from untampered images. In contrast, *conditional denoising* reconstructs the image with tampered traces, guided by the forged input and self-attention. SSIM scores are then calculated between the two denoised samples, $S(x_{c_t})$ and $S(x_{u_t})$, to determine the appropriate reverse timestep m for starting the aggregation of error maps, following the approach of (Che et al. 2024). The error map is computed using squared error.

Adaptive Number of Diffusion Timesteps Selection

The number of diffusion timestep T for adding noise plays a crucial role in our method. If T is too large, it would cause significant deviation from the input image, resulting in unexpected artifacts. Conversely, if T is too small, it cannot effectively remove the tampered areas, leaving manipulation traces in the unconditional reconstruction. Previous purification methods (Wang et al. 2022; Nie et al. 2022; Tailanián et al. 2024) select a fixed T for all images, which is clearly inappropriate. Inspired by the previous IML methods (Bayar and Stamm 2018; Chen et al. 2021; Zhou et al. 2018) that use high-pass filters to suppress image content, based on the idea that manipulation traces are more likely to be detected in the filtered results rather than in the image content. As discussed in the introduction and Fig. 1(a), high-frequency information is removed more quickly than image content in the diffusion forward process. Therefore, we use SRM filters (Fridrich and Kodovsky 2012) to process each x_t as $S(x_t)$, and the SSIM scores between each $S(x_t)$ and $S(x_0)$ are used to adaptively select an appropriate T. The basic idea is that when the SSIM using high-pass filters approaches 0, the SSIM without high-pass filters remains higher. This ensures that the forward process effectively removes manipulation traces while preserving the overall structure of the image. As illustrated in the examples in Fig. 2(a), the adaptively selected T causes the filtered image $S(x_T)$ to resemble random noise, while the image sample x_T still retains the overall structure. The reason for using SSIM is that it provides a clear cut-off value to indicate when two images are dissimilar (when the score is 0), whereas other metrics, such as mean square error (MSE), do not offer this property. We select 0.2 as the SSIM threshold to determine the appropriate T, meaning that when SSIM falls below 0.2, that timestep is selected as T.

Conditional Backward Process

Due to unexpected inconsistencies in authentic pixels when calculating the error directly from the unconditional reconstruction and the input, we modified the diffusion backward process into two branches, with the error now calculated between these two backward processes. By applying both conditional and unconditional backward processes to the same noised image x_T and ensuring that both use the same random noise at each timestep, we aimed to resolve these inconsistencies and improve localization performance.

For unconditional denoising, the process simply starts from x_T and gradually removes noise without any guidance, as described in Eq. (2). Our primary contribution lies in introducing conditional denoising without apply any external conditions, aiming to reconstruct an image that retains manipulation traces. This allows the inconsistency between the conditional and unconditional branches to effectively highlight the tampered regions.

Similarity guidance is employed to direct the reconstruction using the forged input, with the aim of guiding the model to reconstruct the image as closely as possible to the forged one, thereby preserving the tampered traces. Similar to (Wang et al. 2022; Tailanián et al. 2024), we define the similarity metric as $D(\cdot)$ and the similarity guidance is given by:

$$\widetilde{\epsilon}(x_{c_t}, d, t) = \epsilon_\theta(x_{c_t}, t) - s_{d_t} \cdot \sigma_t \nabla_{x_{c_t}} D(x_t, x_{c_t}), \quad (6)$$

which is similar to classifier guidance shown in Eq. (4). The key difference is that the separate classifier is replaced by the similarity metric. Here, $\tilde{\epsilon}(x_{c_t}, d, t)$ represents the conditional output guided by the similarity d. x_t and x_{c_t} are samples in forward and conditional backward process, respectively. s_{d_t} is the guidance scale that is proportional to added noise and it can be expressed as $s_{d_t} = s_d \cdot \sqrt{1-\bar{\alpha}_t}/\sqrt{\bar{\alpha}_t}$, Where s_d is a pre-defined initial guidance scale.

Self-attention Guidance: Using solely Eq. (6) for conditional reconstruction results in unsatisfactory localization outcomes because the similarity guidance is applied globally to the entire image, leading to false alarms in the untampered regions. To address this issue and focus the reconstruction error more on the tampered region, we draw inspiration from (Hong et al. 2023), which demonstrates that the self-attention map from the diffusion U-Net overlaps with high-frequency details in the image. Since manipulation traces are most likely found in high-frequency regions, such as edge inconsistencies, we incorporate self-attention guidance into the conditional reconstruction. The self-attention in U-Net is implemented as multi-head self-attention (Vaswani et al. 2017), with the number of attention heads denoted by N. Let Q_t^h denote the query, K_t^h denote the key and V_t^h denote the value. The attention on the hth head at timestep t is:

$$A(Q_t^h, K_t^h, V_t^h) = \mathrm{softmax}(Q_t^h (K_t^h)^T / \sqrt{d}) \cdot V_t^h. \quad (7)$$

The stacked self-attention maps across all attention heads at timestep t is $A_{s_t} \in \mathbb{R}^{N \times (HW) \times (HW)}$, where H and W denote the height and width, respectively. Then, A_{s_t} is processed by global average pooling (GAP), reshaping Reshape$(\cdot)$ and upsampling Upsample$(\cdot)$ to match the dimensions of image sample x_{c_t}. The final aggregated attention A_{c_t} from all attention heads at timestep t is:

$$A_{c_t} = \mathrm{Upsample}(\mathrm{Reshape}(\mathrm{GAP}(A_{s_t}))). \quad (8)$$

As shown in Fig. 2(b), once we have the attention map, we can use the activated information to guide the generation, thus the reconstruction can focus more on these regions. The basic idea is to apply Gaussian blur only to the activated regions and then use the residual information between the blurred and unblurred image samples to guide the generation in a classifier-free manner. Let $M_{c_t}^i$ denote the binary mask value, and $A_{c_t}^i$ denote the self-attention map value at the ith pixel. Given an attention mask threshold τ, we first threshold A_{c_t} to a binary mask M_{c_t} using:

$$M_{c_t} = \begin{cases} M_{c_t}^i = 1, & \text{if } A_{c_t}^i > \tau, \\ M_{c_t}^i = 0, & \text{otherwise.} \end{cases} \quad (9)$$

For the Gaussian blur process, we follow the method outlined in (Hong et al. 2023) to generate blurred samples x_{b_t} from x_{c_t}. This approach helps mitigate the side effects of reducing Gaussian noise when applying Gaussian blur, as discussed in (Hong et al. 2023). Finally, M_{c_t} is used to obtain the masked blurred samples $\hat{x}_{b_t}$, where only the regions with high activation in self-attention are blurred. The residual information is then used to guide the generation. Let $\odot$ denote element-wise multiplication, and $\tilde{\epsilon}(x_{c_t}, a, t)$ denote the guided output using self-attention guidance a, and s_f be the self-attention guidance scale. The masking and final self-attention guiding process is:

$$\hat{x}_{b_t} = (1 - M_{c_t}) \odot x_{c_t} + M_{c_t} \odot x_{b_t}, \quad (10)$$

$$\tilde{\epsilon}(x_{c_t}, a, t) = \epsilon_\theta(x_{c_t}, t) + s_f \cdot (\epsilon_\theta(x_{c_t}, t) - \epsilon_\theta(\hat{x}_{b_t}, t)). \quad (11)$$

This allows using the masked residual information to guide the generation, making the denoising process concentrate more on the masked region.

The complete conditional denoising process incorporates both similarity and self-attention guidance, applying guidance from both global and local perspectives. The final conditional generation output guided by both a and d is:

$$\begin{aligned}\tilde{\epsilon}(x_{c_t}, a, d, t) = \; & \epsilon_\theta(x_{c_t}, t) + s_f \cdot (\epsilon_\theta(x_{c_t}, t) \\ & - \epsilon_\theta(\hat{x}_{b_t}, t)) - s_{d_t} \cdot \sigma_t \nabla_{x_{c_t}} D(x_t, x_{c_t}).\end{aligned} \quad (12)$$

Error Map Aggregation

Unlike the forward process, where SSIM consistently decreases, in the backward process, SSIM first decreases and then increases. This behavior, also noted in (Che et al. 2024) using external conditions, occurs because the unconditional branch initially reconstructs tampered information into a clean distribution, while the conditional branch works to reverse manipulation traces, leading to a decrease in SSIM. Once SSIM reaches its minimum, both branches have reconstructed the tampered regions and start to reconstruct the original information, causing SSIM to rise. Following (Che et al. 2024), we aggregate error maps starting from the reverse timestep m (where SSIM is lowest). We calculate the error map using the squared error formula: $(x_{c_t} - x_{u_t})^2$. The final aggregated error map is the average of all error maps from reverse timestep m to 0. The final localization map $E(x_u, x_c)$ is obtained by:

$$E(x_u, x_c) = \frac{\sum_{t=0}^{m} (x_{c_t} - x_{u_t})^2}{m+1}. \quad (13)$$

Experimental Results

We first present the experimental setup, including implementation details, datasets, and evaluation metrics. We then compare the IML performance of our method against State-of-The-Art approaches. Finally, we provide ablation studies.

Experimental Setup

Datasets: We use six IML datasets for evaluation: CASIAv1 (Dong, Wang, and Tan 2013), Colombia (Hsu and Chang 2006), Coverage (Wen et al. 2016), NIST16 (Guan et al. 2019), CIMD (Zhang, Li, and Chang 2024) and MagicBrush (Zhang et al. 2024a). The first five datasets contain only standard manipulation types, which are splicing, copy-move, and removal. MagicBrush, however, is a novel instruction-guided manipulation dataset that features previously unseen and more complex tampered types, such as color changes, action changes, and object alterations. This dataset is closer to real-world manipulations and is particularly valuable for evaluating a model's generalizability. For the CIMD dataset, we applied the uncompressed subset,

Method	CASIAv1		Columbia		Coverage		NIST16		CIMD		Average	
	AUC	AP	AUC	AP	AUC	AP	AUC	AP	AUC	AP	AUC	AP
NOI1	0.586	0.140	0.539	0.387	0.580	0.168	0.511	0.115	0.680	0.060	0.579	0.174
CFA1	0.498	0.100	0.641	0.445	0.533	0.133	0.503	0.101	0.427	0.016	0.520	0.159
MCA	0.542	0.117	0.513	0.270	0.536	0.124	0.520	0.083	0.521	0.020	0.526	0.123
NoisePrint	0.514	0.091	0.563	0.359	0.515	0.123	0.450	0.114	0.543	0.018	0.517	0.141
IVC	0.531	0.109	0.511	0.291	0.532	0.140	0.532	0.092	0.561	0.021	0.533	0.131
CFA2	0.531	0.104	0.530	0.411	0.524	0.143	0.480	0.095	0.510	0.017	0.515	0.154
NOI2	0.574	0.135	0.559	0.353	0.598	0.161	0.519	0.089	0.506	0.018	0.551	0.151
NOI4	0.535	0.106	0.536	0.313	0.537	0.130	0.494	0.085	0.634	0.043	0.547	0.135
BLK	0.541	0.112	0.624	0.416	0.598	0.154	**0.583**	0.136	0.494	0.027	0.568	0.169
Ours	**0.587**	**0.162**	**0.682**	**0.461**	**0.622**	**0.208**	0.556	**0.160**	**0.690**	**0.068**	**0.627**	**0.212**

Table 1: Evaluation results of unsupervised methods for the **Standard Manipulation task**. Average scores are calculated across five datasets, with the best and second-best performances highlighted in bold and underlined.

Method	Training Data Size	MagicBrush	
		AUC	AP
Mantra-Net	64K	0.426	0.156
PSCC-Net	100K	0.375	0.140
CAT-Net	858K	0.392	0.155
Hifi-Net	1,710K	0.480	0.169
CR-CNN	12.5K	0.515	0.193
MVSS-Net	12.5K + NMA	**0.578**	**0.270**
WSCL	12.5K	0.516	0.170
Ours	**None**	0.543	0.206

Table 2: Evaluation results for the **Novel Manipulation task** for both fully supervised and weakly supervised methods. NMA refers to Naive Manipulation Augmentation, which includes techniques such as cropping and pasting squared areas, and utilizing OpenCV inpainting functions (Telea 2004; Bertalmio, Bertozzi, and Sapiro 2001). The best and second-best performances are highlighted in bold and underline, respectively.

which is intended for evaluating image editing IML methods.

Evaluation metrics: We use two thresholding-agnostic metrics for evaluation: Area Under the Receiver Operating Characteristic curve (AUC) and Average Precision (AP). These two evaluation metrics do not require predefined thresholds, making the evaluation more fair.

Implementation details: Our method does not require training or external conditions. We used the pre-trained diffusion model from (Dhariwal and Nichol 2021b), which was trained on ImageNet (Deng et al. 2009). The method is implemented using Pytorch (Paszke et al. 2019) on an A40 GPU. For the diffusion model itself, we did not modify any of the diffusion settings except for the diffusion timestep T. For our proposed components, we set the initial similarity scale to $s_d = 10^4$, and the threshold for selecting the appropriate T is set to 0.2. In self-attention guidance, the guidance scale s_f is set to 1.3, the attention threshold τ is 1.3, and the blur sigma is 3.

Comparison with SoTA Methods

We conducted a comprehensive comparison with sixteen state-of-the-art (SoTA) methods, spanning unsupervised, weakly-supervised, and fully-supervised approaches. Crucially, all selected methods have open-source code, ensuring a fair evaluation. The unsupervised methods include (Mahdian and Saic 2009; Lyu, Pan, and Zhang 2014; Wagner 2015; Ferrara et al. 2012; Dirik and Memon 2009; Li, Yuan, and Yu 2009; Bammey, Gioi, and Morel 2020; Choi, Choi, and Lee 2011; Cozzolino and Verdoliva 2019), with the first six being implemented by MKLab (Zampoglou, Papadopoulos, and Kompatsiaris 2017). For weakly-supervised methods, we evaluate (Zhai et al. 2023). The fully-supervised methods include (Bayar and Stamm 2018; Kwon et al. 2022; Liu et al. 2022; Wu, AbdAlmageed, and Natarajan 2019; Guo et al. 2023; Chen et al. 2021). Using their open-source code, we generated localization maps and applied the same evaluation code to obtain quantitative results, maintaining consistency for a fair comparison. Abbreviations for each method follow those used in prior work.

Comparison using standard manipulation datasets: Table 1 provides evaluation results on five standard IML datasets for unsupervised methods. In most cases, our training-free method achieves the best localization performance across almost all datasets, except for the AUC score on the NIST16 dataset. Regarding the AUC performance on NIST16, our method does not outperform BLK, as all images in NIST16 are JPEG compressed, and BLK is specifically designed for JPEG format. Additionally, our method achieves significantly higher average performance than other approaches, demonstrating much stronger localization ability.

Comparison on the MagicBrush dataset: Table 2 shows the evaluation results comparing fully-supervised and weakly-supervised IML methods on MagicBrush (Zhang et al. 2024a), a recent IML dataset containing new tampered types. For methods trained on a dataset size of 12.5K, CASIAv2 (Dong, Wang, and Tan 2013) was used as the training set, while other methods used their own synthetic datasets. Table 2 shows results demonstrating that even fully-supervised methods trained on large datasets often

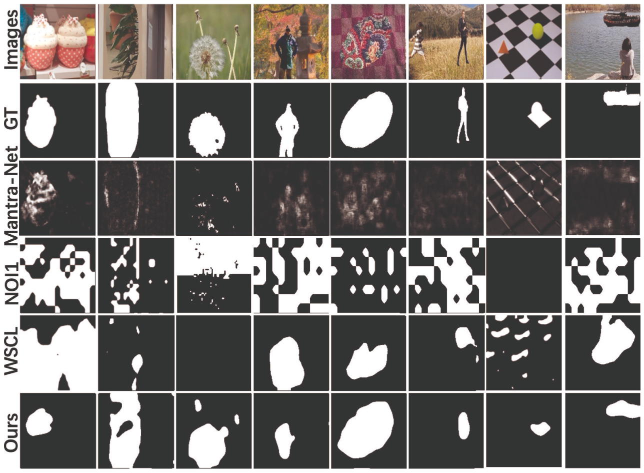

Figure 3: Visualization results are shown from top to bottom: the tampered images, ground-truth masks, results of the fully-supervised method Mantra-Net (Wu, AbdAlmageed, and Natarajan 2019), the unsupervised method NOI1 (Mahdian and Saic 2009), the weakly-supervised method WSCL (Zhai et al. 2023), and our training-free method.

struggle to adapt to new manipulation types, exhibiting low generalizability. In contrast, our training-free method delivers competitive results without any training images. Although MVSS-Net outperforms our method, it employs naive manipulation augmentation (NMA), such as cropping and pasting squared areas, and utilizing OpenCV inpainting functions (Telea 2004; Bertalmio, Bertozzi, and Sapiro 2001), thereby increasing its training data diversity beyond the 12.5K samples.

Visualization: Fig. 3 presents the visualization of the IML results. Compared to other methods, our approach offers improved coverage of the tampered regions, despite not requiring any training. However, because our method is training-free and does not rely on datasets or pixel-level masks for supervision, it struggles to define the edges of the tampered regions precisely. We plan to address this limitation in our future work.

Ablation Study

We conduct ablation studies using CASIAv1 (Dong, Wang, and Tan 2013) to demonstrate the effectiveness of the proposed components.

Unconditional	Similarity	Self-attention	AUC	AP
✓			0.544	0.102
✓	✓		0.514	0.109
✓		✓	0.573	0.127
✓	✓	✓	**0.587**	**0.162**

Table 3: Ablation study on different conditions.

T	$T = 10$	$T = 50$	$T = 100$	$T = 200$	Adap
AUC	0.532	0.577	0.558	0.467	**0.587**
AP	0.119	0.155	0.143	0.107	**0.162**

Table 4: IML performance comparisons using various fixed timesteps T and our adaptive T.

Effectiveness of conditional guidance: We assess the impact of conditional guidance, focusing on similarity and self-attention guidance. As shown in Table 3, using only similarity guidance does not produce satisfactory results. This is because similarity guidance directs reconstruction globally, leading to unintended false alarms, as explained in the introduction and method sections. On the other hand, using only self-attention guidance significantly improves performance. The best results are achieved when both similarity and self-attention guidance are combined, underscoring the importance of both. In summary, similarity guidance increases the inconsistency between the two branches in the tampered area, while self-attention guidance focuses more on the tampered area and reduces false positives. Both are essential for optimal performance.

Adaptive diffusion timestep selection: Our method adaptively selects an appropriate diffusion timesteps T to add noise to the input image, thereby avoiding the issue of T being too low or too high. As shown in Table 4, increasing T initially improves performance but eventually causes a decline. Although there might be an optimal fixed T, finding it would require extensive experimentation. In contrast, our adaptive approach achieves the best performance without the need for such experiments.

Conclusion

In this work, we introduce a novel training-free method for Image Manipulation Localization (IML) using diffusion models. Our method adaptively selects an appropriate number of diffusion timesteps for each input image, adding noise in the forward process. In the reverse process, starting from the same noised sample, we perform both conditional and unconditional reconstructions without relying on external conditions. The localization maps are generated from inconsistencies between the two reverse processes. We conducted comprehensive evaluations against 16 state-of-the-art (SoTA) methods across six IML datasets. Our method not only demonstrated superior performance on standard image manipulation types but also showed remarkable generalizability to unseen manipulation types, all without the need for model training or reliance on external datasets.

Limitations of this work include the inaccurate localization boundaries due to the absence of pixel-level annotation for supervision. **Future work** could focus on developing an effective method to address the issue of inaccurate localization edges in a training-free manner.

Acknowledgements

This work is supported by the DARPA Semantic Forensics (SemaFor) Program under contract HR001120C0123 and NSF CCSS-2348046.

References

Amit, T.; Shaharbany, T.; Nachmani, E.; and Wolf, L. 2021. Segdiff: Image segmentation with diffusion probabilistic models. *arXiv preprint arXiv:2112.00390*.

Bammey, Q.; Gioi, R. G. v.; and Morel, J.-M. 2020. An adaptive neural network for unsupervised mosaic consistency analysis in image forensics. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 14194–14204.

Bayar, B.; and Stamm, M. C. 2018. Constrained convolutional neural networks: A new approach towards general purpose image manipulation detection. *IEEE Transactions on Information Forensics and Security*, 13(11): 2691–2706.

Bertalmio, M.; Bertozzi, A. L.; and Sapiro, G. 2001. Navier-stokes, fluid dynamics, and image and video inpainting. In *Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. CVPR 2001*, volume 1, I–I. IEEE.

Che, Y.; Rafsani, F.; Shah, J.; Siddiquee, M. M. R.; and Wu, T. 2024. AnoFPDM: Anomaly Segmentation with Forward Process of Diffusion Models for Brain MRI. *arXiv preprint arXiv:2404.15683*.

Chen, X.; Dong, C.; Ji, J.; Cao, J.; and Li, X. 2021. Image manipulation detection by multi-view multiscale supervision. In *IEEE/CVF International Conference on Computer Vision*, 14185–14193.

Choi, C.-H.; Choi, J.-H.; and Lee, H.-K. 2011. CFA pattern identification of digital cameras using intermediate value counting. In *Proceedings of the thirteenth ACM multimedia workshop on Multimedia and security*, 21–26.

Cozzolino, D.; and Verdoliva, L. 2019. Noiseprint: A CNN Based Camera Model Fingerprint. *IEEE Transactions on Information Forensics and Security*, 15: 144–159.

Deng, J.; Dong, W.; Socher, R.; Li, L.-J.; Li, K.; and Fei-Fei, L. 2009. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, 248–255. Ieee.

Dhariwal, P.; and Nichol, A. 2021a. Diffusion models beat gans on image synthesis. *Advances in neural information processing systems*, 34: 8780–8794.

Dhariwal, P.; and Nichol, A. 2021b. Diffusion models beat gans on image synthesis. *Advances in neural information processing systems*, 34: 8780–8794.

Dirik, A. E.; and Memon, N. 2009. Image tamper detection based on demosaicing artifacts. In *2009 16th IEEE International Conference on Image Processing (ICIP)*, 1497–1500. IEEE.

Dong, J.; Wang, W.; and Tan, T. 2010. CASIA Image Tampering Detection Evaluation Database. http://forensics.idealtest.org.

Dong, J.; Wang, W.; and Tan, T. 2013. Casia image tampering detection evaluation database. In *2013 IEEE China summit and international conference on signal and information processing*, 422–426. IEEE.

Ferrara, P.; Bianchi, T.; De Rosa, A.; and Piva, A. 2012. Image forgery localization via fine-grained analysis of CFA artifacts. *IEEE Transactions on Information Forensics and Security*, 7(5): 1566–1577.

Fridrich, J.; and Kodovsky, J. 2012. Rich models for steganalysis of digital images. *IEEE Transactions on information Forensics and Security*, 7(3): 868–882.

Goodfellow, I.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; and Bengio, Y. 2020. Generative adversarial networks. *Communications of the ACM*, 63(11): 139–144.

Guan, H.; Kozak, M.; Robertson, E.; Lee, Y.; Yates, A. N.; Delgado, A.; Zhou, D.; Kheyrkhah, T.; Smith, J.; and Fiscus, J. 2019. MFC datasets: Large-scale benchmark datasets for media forensic challenge evaluation. In *IEEE/CVF Winter Conference on Applications of Computer Vision Workshops*, 63–72. IEEE.

Guo, X.; Liu, X.; Ren, Z.; Grosz, S.; Masi, I.; and Liu, X. 2023. Hierarchical fine-grained image forgery detection and localization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 3155–3165.

Ho, J.; Jain, A.; and Abbeel, P. 2020. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33: 6840–6851.

Ho, J.; and Salimans, T. 2022. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*.

Hong, S.; Lee, G.; Jang, W.; and Kim, S. 2023. Improving sample quality of diffusion models using self-attention guidance. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 7462–7471.

Hsu, Y.-F.; and Chang, S.-F. 2006. Detecting image splicing using geometry invariants and camera characteristics consistency. In *2006 IEEE International Conference on Multimedia and Expo*, 549–552. IEEE.

Kawar, B.; Zada, S.; Lang, O.; Tov, O.; Chang, H.; Dekel, T.; Mosseri, I.; and Irani, M. 2023. Imagic: Text-based real image editing with diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 6007–6017.

Kingma, D. P.; and Welling, M. 2013. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*.

Kwon, M.-J.; Nam, S.-H.; Yu, I.-J.; Lee, H.-K.; and Kim, C. 2022. Learning jpeg compression artifacts for image manipulation detection and localization. *International Journal of Computer Vision*, 1875–1895.

Li, W.; Yuan, Y.; and Yu, N. 2009. Passive detection of doctored JPEG image via block artifact grid extraction. *Signal Processing*, 89(9): 1821–1829.

Liu, X.; Liu, Y.; Chen, J.; and Liu, X. 2022. 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.

Lyu, S.; Pan, X.; and Zhang, X. 2014. Exposing region splicing forgeries with blind local noise estimation. *International journal of computer vision*, 110: 202–221.

Mahdian, B.; and Saic, S. 2009. Using noise inconsistencies for blind image forensics. *Image and vision computing*, 27(10): 1497–1503.

Nie, W.; Guo, B.; Huang, Y.; Xiao, C.; Vahdat, A.; and Anandkumar, A. 2022. Diffusion models for adversarial purification. *arXiv preprint arXiv:2205.07460*.

Paszke, A.; Gross, S.; Massa, F.; Lerer, A.; Bradbury, J.; Chanan, G.; Killeen, T.; Lin, Z.; Gimelshein, N.; Antiga, L.; et al. 2019. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32.

Qiao, T.; Zhang, J.; Xu, D.; and Tao, D. 2019. Mirrorgan: Learning text-to-image generation by redescription. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 1505–1514.

Ronneberger, O.; Fischer, P.; and Brox, T. 2015. U-net: Convolutional networks for biomedical image segmentation. In *Medical image computing and computer-assisted intervention–MICCAI 2015: 18th international conference, Munich, Germany, October 5-9, 2015, proceedings, part III 18*, 234–241. Springer.

Sitzmann, V.; Martel, J.; Bergman, A.; Lindell, D.; and Wetzstein, G. 2020. Implicit neural representations with periodic activation functions. *Advances in neural information processing systems*, 33: 7462–7473.

Tailanián, M.; Gardella, M.; Pardo, A.; and Musé, P. 2024. Diffusion models meet image counter-forensics. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 3925–3935.

Telea, A. 2004. An image inpainting technique based on the fast marching method. *Journal of graphics tools*, 9(1): 23–34.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, Ł.; and Polosukhin, I. 2017. Attention is all you need. *Advances in neural information processing systems*.

Wagner, J. 2015. Noise analysis for image forensics. Online.

Wang, J.; Lyu, Z.; Lin, D.; Dai, B.; and Fu, H. 2022. Guided diffusion model for adversarial purification. *arXiv preprint arXiv:2205.14969*.

Wang, Z.; Bovik, A. C.; Sheikh, H. R.; and Simoncelli, E. P. 2004. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4): 600–612.

Wen, B.; Zhu, Y.; Subramanian, R.; Ng, T.-T.; Shen, X.; and Winkler, S. 2016. COVERAGE – A NOVEL DATABASE FOR COPY-MOVE FORGERY DETECTION. In *IEEE International Conference on Image processing (ICIP)*.

Wolleb, J.; Sandkühler, R.; Bieder, F.; Valmaggia, P.; and Cattin, P. C. 2022. Diffusion models for implicit image segmentation ensembles. In *International Conference on Medical Imaging with Deep Learning*, 1336–1348. PMLR.

Wu, Y.; AbdAlmageed, W.; and Natarajan, P. 2019. ManTra-Net: Manipulation Tracing Network for Detection and Localization of Image Forgeries With Anomalous Features. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 9543–9552.

Xie, E.; Wang, W.; Yu, Z.; Anandkumar, A.; Alvarez, J. M.; and Luo, P. 2021. SegFormer: Simple and efficient design for semantic segmentation with transformers. *Advances in neural information processing systems*, 34: 12077–12090.

Xu, T.; Zhang, P.; Huang, Q.; Zhang, H.; Gan, Z.; Huang, X.; and He, X. 2018. Attngan: Fine-grained text to image generation with attentional generative adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 1316–1324.

Yang, C.; Li, H.; Lin, F.; Jiang, B.; and Zhao, H. 2020. Constrained R-CNN: A General Image Manipulation Detection Model. In *IEEE International Conference on Multimedia and Expo (ICME)*, 1–6. IEEE.

Yang, Y.; Fu, H.; Aviles-Rivero, A. I.; Schönlieb, C.-B.; and Zhu, L. 2023. Diffmic: Dual-guidance diffusion network for medical image classification. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 95–105. Springer.

Yu, Z.; Ni, J.; Lin, Y.; Deng, H.; and Li, B. 2024. DiffForensics: Leveraging Diffusion Prior to Image Forgery Detection and Localization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 12765–12774.

Zampoglou, M.; Papadopoulos, S.; and Kompatsiaris, Y. 2017. Large-scale evaluation of splicing localization algorithms for web images. *Multimedia Tools and Applications*, 76(4): 4801–4834.

Zhai, Y.; Luan, T.; Doermann, D.; and Yuan, J. 2023. Towards Generic Image Manipulation Detection with Weakly-Supervised Self-Consistency Learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 22390–22400.

Zhang, K.; Mo, L.; Chen, W.; Sun, H.; and Su, Y. 2024a. Magicbrush: A manually annotated dataset for instruction-guided image editing. *Advances in Neural Information Processing Systems*, 36.

Zhang, Z.; and Chang, M.-C. 2023. Two-stage dual augmentation with clip for improved text-to-sketch synthesis. In *2023 IEEE 6th International Conference on Multimedia Information Processing and Retrieval (MIPR)*, 1–6. IEEE.

Zhang, Z.; Huang, T.-W.; Su, G.-M.; Chang, M.-C.; and Li, X. 2024b. Text-Driven Synchronized Diffusion Video and Audio Talking Head Generation. In *2024 IEEE 7th International Conference on Multimedia Information Processing and Retrieval (MIPR)*, 61–67. IEEE.

Zhang, Z.; Li, M.; and Chang, M.-C. 2024. A New Benchmark and Model for Challenging Image Manipulation Detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 7405–7413.

Zhang, Z.; Li, M.; Li, X.; Chang, M.-C.; and Hsieh, J.-W. 2025. Image Manipulation Detection with Implicit Neural Representation and Limited Supervision. In *European Conference on Computer Vision*, 255–273. Springer.

Zhou, P.; Han, X.; Morariu, V. I.; and Davis, L. S. 2018. Learning rich features for image manipulation detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 1053–1061.