SYNTHETIC IMAGE DETECTION VIA CURVATURE OF DIFFUSION PROBABILITY FLOWS

Anonymous authors
Paper under double-blind review

ABSTRACT

Synthetic image detection (SID) faces two major challenges: high computational cost from reconstruction-based methods and insufficient generalization. To address these issues, we propose a novel SID paradigm that leverages the ODE formulation of diffusion models. Rather than reconstructing images, our method analyzes the probability flow trajectories from data distributions to a Gaussian prior. We show that the discrete-step distances on the Wasserstein manifold inherently encode reconstruction error, and that real and synthetic images diverge most significantly in the early half of the diffusion inversion. Real images exhibit higher curvature variance with extreme deviations, whereas synthetic ones follow smoother, more consistent trajectories. Building on this insight, we introduce curvature features of probability flow trajectories as a new discriminative signal. To the best of our knowledge, this is the first work to exploit probability flow curvature for SID. Extensive experiments demonstrate that our method generalizes robustly to unseen models, achieves SOTA results across multiple benchmarks, and does so with less than half the computational cost of full diffusion inversion.

1 INTRODUCTION

Continuous-time dynamic models (CTDMs) have emerged as a dominant class of generative models, evolving from energy-based and score-matching models (Hyvärinen & Dayan, 2005), through diffusion probabilistic models (Ho et al., 2020; Nichol & Dhariwal, 2021; Song et al., 2020a), Score-Based SDEs and Probability Flow ODEs (Song et al., 2020b; 2021; De Bortoli et al., 2021), to recent Flow Matching methods (Lipman et al., 2022; Liu et al., 2022; Albergo et al., 2023), achieving increasingly realistic image synthesis. However, the rapid proliferation of generative architectures has heightened concerns about the malicious use of synthetic media, motivating the need for detection frameworks that generalize across diverse and unseen generators.

Prior forensic methods (Corvi et al., 2023b; Ojha et al., 2023; Tan et al., 2024b; Sha et al., 2023; Liu et al., 2024) show strong performance on GAN-generated images but often fail to generalize to diffusion-based or newer generative models. Consequently, this work focuses on CTDMs. Among existing approaches, methods such as Wang et al. (2023); Cazenavette et al. (2024); Ricker et al. (2024); Chu et al. (2025); Guillaro et al. (2025) introduced reconstruction error as a discriminative feature, yet most rely on replaying the full diffusion trajectory, without questioning whether the reconstruction error is already implicitly encoded in the probability flow.

In this paper, we adopt a unified ODE perspective and propose a novel detection paradigm that leverages the curvature of the probability flow velocity field as the primary discriminative feature (Fig. 1), complemented by diagonal high-frequency components extracted via wavelet transforms. Following Song et al. (2020a), a CTDM can be understood at the sample level as evolving a single data point along a probability flow ODE, which defines a continuous velocity field and maps noise distributions to data distributions. At the macroscopic level, this velocity field satisfies a continuity equation, describing how the probability density evolves smoothly over time. Integrating the reverse-time ODE then yields the instantaneous velocity at any intermediate time starting from a clean image.

We define the integration of the probability flow ODE over a data distribution as an *ODE pipeline*, which maps data distributions to a Gaussian prior. When analyzing this pipeline in the Wasserstein manifold, the upper bound of the discrete-step distances inherently encodes the reconstruction error,

which can be quantified as the cumulative sum of non-optimality terms across each discrete step. Our analysis reveals that synthetic images reside in regions of the model manifold that are easier to represent. Consequently, during the latter half of diffusion inversion—from the Gaussian back to the data distribution—the difference in accumulated non-optimality terms between real and synthetic images is negligible. This indicates that the main contribution to reconstruction error arises in the first half of the diffusion inversion. Building on this insight, we focus on extracting discriminative information from this stage by computing curvature features of the probability flow trajectories. Empirically, real images exhibit larger variance in curvature values and more extreme deviations, whereas synthetic images follow smoother and more consistent trajectories.

To the best of our knowledge, this is the first work to exploit curvature features of probability flow trajectories for synthetic image detection. Our method demonstrates strong robustness across SOTA generative models while being trained on a single dataset. It generalizes to a wide range of unseen models and achieves this with less than half the computational cost of full diffusion inversion. Compared to prior SOTA methods, it improves ACC by +10.6% and AUCROC by +8.2% across multiple benchmarks.

To summarize our contributions:

- We propose a new paradigm for synthetic image detection based on curvature features of ODE-defined velocity fields.
- We characterize the optimal transport properties of the ODE pipeline and show that synthetic images exhibit lower total kinetic energy, with probability flow trajectories closer to the optimal transport path than those of real images.
- We demonstrate that the model inherently represents lower-energy distributions during reconstruction, embedding reconstruction error within the velocity-field tensors.
- We introduce a pseudo-Gaussian curvature to compress the temporal dimension of curvature features, enhancing their discriminative effectiveness.
- We complement curvature features with diagonal high-frequency wavelet components, capturing fine-grained artifacts and further improving robustness and generalization.

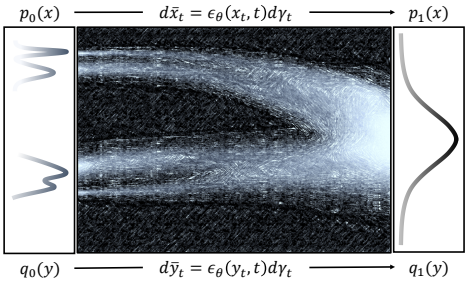

Figure 1: The figure shows the visualization of the velocity field, where brighter regions correspond to higher probability density. On the left are the initial distributions p_0, q_0 at time $t = 0$, which can be transformed into a Gaussian distribution by integrating the backward ODE.

2 Related Work

Artifact Detectors. Prior works have explored diverse strategies for synthetic image detection. CNN-based methods like Wang et al. (2020) train a ResNet-50 on ProGAN outputs with JPEG and blur augmentations. Frequency-based approaches, such as FrePGAN (Jeong et al., 2022) and FreqNet (Tan et al., 2024a), exploit high-frequency artifacts via model-specific analysis or FFT. UniFD (Ojha et al., 2023) decouples feature extraction and classification using a frozen CLIP encoder with a linear classifier. NPR (Tan et al., 2024b) targets autocorrelations induced by upsampling, while FatFormer (Liu et al., 2024) combines semantic contrastive learning with wavelet-based artifact extraction.

Reconstruction Error. DIRE: Wang et al. (2023) propose Diffusion Reconstruction Error (DIRE), which differentiates real from DM-generated images by measuring reconstruction error. AEROBLADE: Ricker et al. (2024) utilize autoencoder reconstruction error from latent DMs for a simple, training-free approach. FakeInversion: Cazenavette et al. (2024) detect images generated by unseen text-to-image DMs using text-conditioned inversion. Luo et al. (2024) propose LaRE2, leveraging Latent Reconstruction Error (LaRE) with an Error-Guided Feature Refinement module for more distinct error feature extraction. B-Free (Guillaro et al., 2025) constructs an unbiased dataset and employs a Vision Transformer to extract discriminative features.

In this paper, we focus on and extend the second line of related work discussed above. While these approaches are all implemented as variants of reconstruction error, we question whether performing the entire reconstruction pipeline is truly necessary. The inversion process requires a full forward pass of the U-Net at every denoising or noising step, resulting in substantial computational and time overhead. Motivated by this limitation, we investigate whether certain inconsistency features can be directly captured within a single noising pass, thereby providing effective discriminative signals while avoiding the cost of full reconstruction.

3 BACKGROUND

3.1 ODE-BASED PROBABILITY FLOW AND VELOCITY FIELD

Let $p_0(x)$ denote the data distribution. In DDPM (Ho et al., 2020), the forward process is formulated as a stochastic differential equation (SDE) that gradually transforms $p_0(x)$ into a standard Gaussian distribution $p_T(x)$:

$$dx_t = f(x_t, t)dt + g(t)dw_t, \quad x_0 \sim p_0(x) \tag{1}$$

where $f : \mathbb{R}^D \to \mathbb{R}^D$ is the drift coefficient, $g(t) \in \mathbb{R}$ is the diffusion coefficient, and $w_t \in \mathbb{R}$ is a standard Wiener process. By sampling $x_T \sim p_T(x)$ and solving the reverse-time SDE, one can recover samples $x_0 \sim p_0(x)$:

$$dx_t = \big[f(x_t, t) - g(t)^2 \nabla_x \log p_t(x_t)\big]dt + g(t)d\bar{w}_t \tag{2}$$

Here, $\bar{w}_t$ denotes a standard Wiener process evolving backward from T to 0. By setting the diffusion coefficient in the reverse process to zero, the stochastic trajectory becomes deterministic while preserving the same marginal distributions as in Eq. 2. This yields the Probability Flow ODE (PF-ODE) (Song et al., 2020b):

$$dx_t = \big[f(x_t, t) - \frac{1}{2} g(t)^2 \nabla_x \log p_t(x_t)\big]dt \tag{3}$$

Recent works (Lipman et al., 2022; Liu et al., 2022; Albergo et al., 2023) move beyond marginal distribution matching and instead directly learn a velocity field $\frac{dx_t}{dt} = v_\theta(x_t, t)$,which models the instantaneous velocity at each timestep to construct a continuous flow between distributions (illustrated in Figure 1). Under this view, the PF-ODE can be equivalently interpreted as modeling the velocity field using a score function:

$$v_\theta(x_t, t) = f(x_t, t) - \frac{1}{2} g(t)^2 \nabla_x \log p_t(x_t) \tag{4}$$

From this perspective, the velocity field v_θ satisfies the continuity equation:

$$\frac{\partial p_t(x)}{\partial t} + \nabla \cdot \big(v_\theta(x_t, t)\, p_t(x)\big) = 0 \tag{5}$$

3.2 OPTIMAL TRANSPORT AND W-DISTANCE

According to optimal transport theory, the cost of transporting one distribution $p(x)$ to another $q(x)$ is defined as:

$$C[p, q] = \inf_{\gamma \in \Pi(p,q)} \mathbb{E}_{(x,y)\sim\gamma}\big[c(x, y)\big] \tag{6}$$

where $\Pi(p, q)$ denotes the set of all couplings of $p(x)$ and $q(x)$, and $c(\cdot, \cdot)$ is the transport cost function. The Wasserstein-ρ distance is then given by

$$W_\rho(p, q) = \left(\inf_{\gamma \in \Pi(p,q)} \mathbb{E}_{(x,y)\sim\gamma}\big[d(x, y)^\rho\big]\right)^{1/\rho} \tag{7}$$

where $d(x, y)$is typically the Euclidean distance, and the cost is defined as its ρ-th power.

The Wasserstein distance provides a metric over probability distributions that preserves the underlying geometry of the sample space, making it particularly suitable for comparing distributions with partially non-overlapping supports. Furthermore, under certain conditions, the probability space endowed with the Wasserstein distance (i.e., the Wasserstein space) can be regarded as a Riemannian

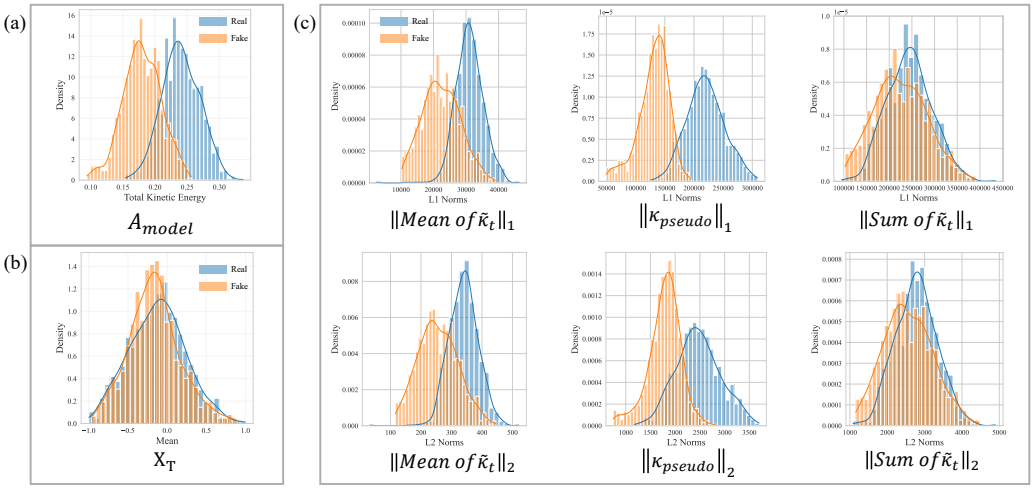

Figure 2: Both real and synthetic images (1,000 each) are randomly sampled from all datasets described in Section 6. The resulting histograms are: **(a)** total kinetic energy computed according to Theorem 2; **(b)** image means at time T, closely following a standard Gaussian distribution; and **(c)** curvature signal $\tilde{\kappa}_t$ under three different temporal compression strategies, highlighting differences in discriminative effectiveness.

manifold. This interpretation makes it a natural choice for characterizing the continuous temporal evolution of probability densities.

From an optimal transport viewpoint, diffusion training corresponds to gradient descent on the KL divergence functional $\mathcal{F}[q] = KL(q\|p_{data})$ in Wasserstein space. The neural network learns a velocity field $v_\theta(x_t, t)$ that approximates the steepest descent direction, thereby constructing minimal-energy trajectories with smooth velocity fields. Consequently, generated samples can be seen as natural endpoints of such geodesic-like flows.

These theoretical foundations highlight the close connection between PF-ODE formulations and optimal transport geometry. Building on this connection, we later analyze how the energy characteristics of probability flow trajectories reveal discriminative differences between real and synthetic images.

4 Trajectory Analysis in Wasserstein Space

4.1 Velocity Field as W_2 Distance Estimation

We consider deterministic sampling trajectories induced by the PF-ODE corresponding to the marginal distributions of the forward VP-SDE. In particular, for the ADM model (Dhariwal & Nichol, 2021), the PF-ODE takes the form:

$$dx_t = \Big[-\tfrac{1}{2}\beta(t)x_t - \tfrac{1}{2}\beta(t)\nabla_x \log p_t(x_t)\Big]dt \tag{8}$$

where the score function $\nabla_x \log p_t(x_t)$ is estimated via a noise prediction network $\epsilon_\theta(x_t, t)$:

$$\nabla_x \log p_t(x_t) \approx -\tfrac{1}{\sigma(t)}\epsilon_\theta(x_t, t) \tag{9}$$

with $\beta(t)$ are time-dependent constants, $\beta(t) = -\frac{d}{dt}log\bar{\alpha}_t$ and $\sigma(t)^2 = 1 - \bar{\alpha}_t$.

Solving Eq. 8 from 0 to T as an ODE pipeline, defines what we refer to as an ODE pipeline, which transports an initial distribution into an approximate Gaussian. As shown in Fig. 2(b), at terminal time T, both real and synthetic images are mapped close to a standard Gaussian. In the following, we use the term ODE pipeline with the default assumption that all trajectories are derived from the same CTDMs.

Theorem 1. *For an ODE pipeline applied to $x_0 \sim p_0$, the Wasserstein distance between two intermediate marginals is bounded by the mean kinetic energy at time t:*

$$W_2[p_t, p_{t+\Delta t}] \leq \Delta t \sqrt{\mathbb{E}_{x\sim p_t} \|v_\theta(x_t, t)\|^2}$$

Theorem 2. *Over the full pipeline from 0 to T, the cumulative one-step W_2 distances are bounded by the total kinetic energy A_{model}:*

$$\sum_t W_2^2[p_t, p_{t+\Delta t}] \leq \delta t \int_0^{T\delta t} \mathbb{E}_{x\sim p_s} \|v_\theta(x_s, s)\|^2 ds$$

$$A_{model} := \int_0^{T\delta t} \mathbb{E}_{x\sim p_s} \|v_\theta(x_s, s)\|^2 ds$$

(Proofs are provided in Appendix A.1)

Theorems 1 and 2 formally characterize the transformation of marginals on the Wasserstein manifold under an ODE pipeline: Theorem 1 establishes a one-step bound, while Theorem 2 provides a cumulative bound that implicitly captures diffusion reconstruction error.

Previous work (Wang et al., 2023; Ricker et al., 2024; Cazenavette et al., 2024; Luo et al., 2024; Chu et al., 2025) has shown that synthetic images yield lower reconstruction error. By Theorem 2, this corresponds to velocity fields with smaller kinetic energy. Thus, synthetic trajectories tend to be straighter than those of real images. To verify, we sample 1,000 real and 1,000 synthetic images (datasets in Sec. 6) and compute their total kinetic energy. As shown in Fig. 2(a), real images exhibit significantly larger energy, though the distributions overlap considerably, which motivates us to seek more effective statistical measures.

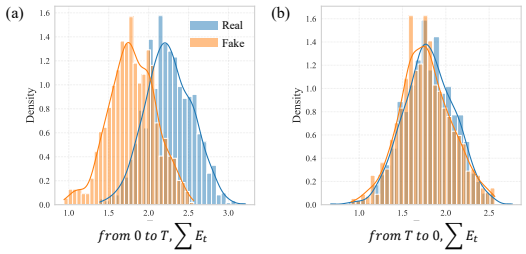

Figure 3: The data sampling is the same as in Fig. 2. The left panel corresponds to the first half of the diffusion reconstruction, and the right panel corresponds to the second half.

To address error accumulation, we define the non-optimality term E_t at each step as (more details in Appendix A.2):

$$E_t = \Delta t \sqrt{\mathbb{E}_{x\sim p_t} \|v_\theta(x_t, t)\|^2} - W_2[p_t, p_{t+\Delta t}] \tag{10}$$

namely, the difference between the step-wise upper bound and the true optimal transport.

Data sampling follows Fig. 2. Fig. 3 reports the cumulative non-optimality across forward and reverse diffusion. Interestingly, although the ODE pipeline is theoretically bijective, real-image trajectories tend to collapse into lower-energy regions upon reconstruction, effectively yielding mappings $p_0 \to \mathcal{N}(0, I) \to q_0$ for real images, and $q_0 \to \mathcal{N}(0, I) \to q_0$ for synthetic ones. As a result, synthetic images consistently achieve smaller reconstruction errors, suggesting that they lie on manifold regions more easily represented by the model. Crucially, this also resolves our initial concern: the latter half of the trajectory, $\mathcal{N}(0, I) \to q_0$, contributes little discriminative information and can be discarded, offering substantial savings in U-Net computation.

4.2 ODE Taylor Expansion and Curvature Surrogate

Following the time reparameterization of Dockhorn et al. (2022), Eq. 8 becomes:

$$d\bar{x}_t = \epsilon_\theta(x_t, t) d\gamma_t \tag{11}$$

where $\gamma_t = \sqrt{\frac{1-\bar{\alpha}_t^2}{\bar{\alpha}_t}}$, $\bar{x}_t = x_t \sqrt{1+\gamma_t^2}$ and $\epsilon_\theta(x_t, t) = -\sigma(t) \nabla_x \log p_t(x_t)$

From the previous analysis, synthetic images tend to exhibit smoother velocity fields. This motivates curvature-based descriptors of trajectory flatness. The geometric curvature is defined as:

$$\kappa(\gamma) = \frac{\|\bar{x}''_\gamma - (\bar{x}''_\gamma \cdot \hat{\epsilon_\theta})\hat{\epsilon_\theta}\|}{\|\epsilon_\theta\|^2} \tag{12}$$

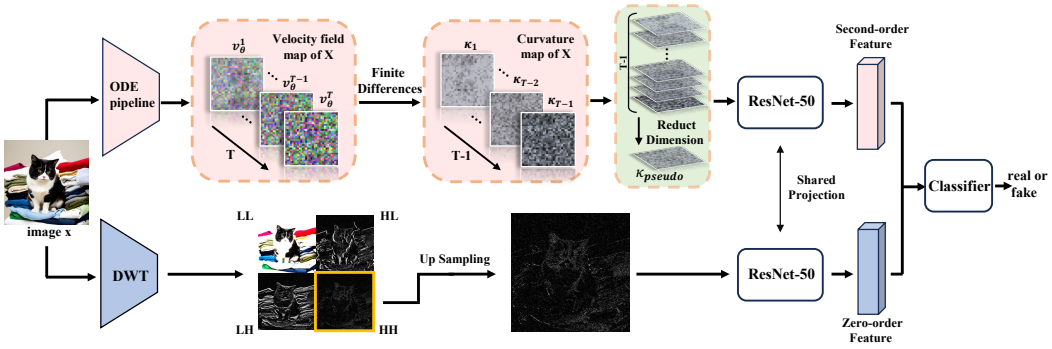

Figure 4: **Proposed method.** The model consists of two pipelines: one extracts second-order curvature features, and the other extracts zeroth-order diagonal high-frequency features. For the second-order features, we perform temporal dimensionality reduction using the proposed pseudo-Gaussian curvature. The two ResNet-50 backbones share the final projection layer to align features from both pipelines.

We approximate curvature via a second-order truncated Taylor method (TTM):

$$\bar{x}_{t_{n+1}} = \bar{x}_{t_n} + h_n \epsilon_\theta(x_{t_n}, t_n) + \frac{1}{2} h_n^2 \left. \frac{d\epsilon_\theta}{d\gamma_t} \right|_{(x_{t_n}, t_n)} \tag{13}$$

with step size $h_n = \gamma_{n+1} - \gamma_n$, The second-order term

$$\begin{aligned} \frac{d\epsilon_\theta(x_t,t)}{d\gamma_t} &= \frac{\partial \epsilon_\theta(x_t,t)}{\partial x_t} \frac{dx_t}{d\gamma_t} + \frac{\partial \epsilon_\theta(x_t,t)}{\partial t} \frac{dt}{d\gamma_t} \\ &= \frac{1}{\sqrt{\gamma_t^2+1}} \frac{\partial \epsilon_\theta(x_t,t)}{\partial x_t} \epsilon_\theta(x_t,t) - \frac{\gamma_t}{1+\gamma_t^2} \frac{\partial \epsilon_\theta(x_t,t)}{\partial x_t} x_t + \frac{\partial \epsilon_\theta(x_t,t)}{\partial t} \frac{dt}{d\gamma_t} \end{aligned} \tag{14}$$

involves two Jacobian-vector products (JVPs) $\frac{\partial \epsilon_\theta(x_t,t)}{\partial x_t} \epsilon_\theta(x_t,t)$, $\frac{\partial \epsilon_\theta(x_t,t)}{\partial x_t} x_t$, requiring additional backward passes and thus heavy computation.

To balance accuracy and efficiency, we adopt the correction term $\frac{1}{2} h_t^2 \frac{d\epsilon_\theta}{d\gamma_t}$ as a surrogate curvature signal, and approximate it via finite differences:

$$\frac{1}{2} h_t^2 \frac{d\epsilon_\theta}{d\gamma_t} \approx -\frac{\Delta \gamma_t^2 \bar{\alpha} t^2 \gamma_t \Delta \epsilon \theta_t}{(1+\bar{\alpha}_t^2) \Delta \bar{\alpha}_t} := \tilde{\kappa}_t \tag{15}$$

This surrogate captures the dominant acceleration while avoiding costly JVPs, and empirically preserves trajectory characteristics (details in Appendix A.1.3).

Figure 2(c) shows histograms of $\tilde{\kappa}_t$. To enhance discriminability, we compare several temporal compression strategies, including sum, mean, and pseudo-Gaussian aggregation, each evaluated under both L1 and L2 norms. The results demonstrate that the pseudo-Gaussian curvature with the L1 norm exhibits the least distributional overlap between real and synthetic images, thereby providing the strongest discriminative power (more details in Appendix A.3). This motivates us to adopt κ_{pseudo} as the primary curvature descriptor in our subsequent analysis. Analogous to Gaussian curvature as the product of principal curvatures, we define a pseudo-Gaussian curvature κ_{pseudo} as the product of the maximal and minimal trajectory curvatures observed along time:

$$\kappa_{pseudo} := \left(\max_{t \in [0,T]} (\tilde{\kappa}_t) \right) \cdot \left(\min_{t \in [0,T]} (\tilde{\kappa}_t) \right) \tag{16}$$

5 METHOD

As shown in Figure 4, our full framework consists of two pipelines. Pipeline 1 is employed to extract curvature features. Since curvature is a second-order quantity and highly sensitive to variations in the input, we additionally use Pipeline 2 to extract zero-order information as an auxiliary signal.

5.1 PIPELINE 1: CURVATURE FEATURE EXTRACTION.

Given a clean image x_0, we apply the ODE pipeline with a first-order Euler approximation to obtain intermediate states x_t. The corresponding first-order TTM from Eq. 13 reduces to:

$$\bar{x}_{t_{n+1}} = \bar{x}_{t_n} + h_n \epsilon_\theta(x_{t_n}, t_n) \tag{17}$$

At each time step, curvature features are computed using Eq. 15 and subsequently compressed along the temporal axis via the pseudo-Gaussian formulation in Eq. 16, yielding the pseudo-Gaussian curvature κ_{pseudo}. Although the L1 norm of κ_{pseudo} already provides sufficient discriminative power for classification, to capture more comprehensive and fine-grained characteristics, we parameterize these features using a convolutional neural network. Therefore, κ_{pseudo} is fed into a ResNet-50 backbone, which encodes the tensor into a compact curvature feature vector.

5.2 PIPELINE 2: IMAGE-BASED REPRESENTATION EXTRACTION.

Previous work (Corvi et al., 2023a; Tan et al., 2024a; Chu et al., 2025; Guillaro et al., 2025) has shown that synthetic images often contain frequency-domain artifacts. Based on these observations, we select the diagonal high-frequency components from the wavelet transform as zero-order information to complement the curvature-based features. Specifically, we employ a one-level discrete wavelet transform (DWT) using the Biorthogonal 1.3 ("bior1.3") basis with symmetric boundary extension. The DWT decomposes $x \in \mathbb{R}^{B\times C\times H\times W}$ into:

$$Y_\ell \in \mathbb{R}^{B\times C\times \frac{H}{2}\times \frac{W}{2}}, \tag{18}$$

$$\{\, Y_h^{(1)},\, Y_h^{(2)},\, Y_h^{(3)}\} \subset \mathbb{R}^{B\times C\times \frac{H}{2}\times \frac{W}{2}}. \tag{19}$$

where Y_ℓ is the low-frequency approximation and $\{\, Y_h^{(1)},\, Y_h^{(2)},\, Y_h^{(3)}\}$ are the horizontal (LH), vertical (HL), and diagonal (HH) detail subbands, respectively.

We select the diagonal detail coefficients $Y_h^{(3)}$, which capture edge and texture variations along oblique orientations. To align this feature map with the original spatial resolution, we upsample via bilinear interpolation:

$$\tilde{Y}_h^{(3)} = \text{Interp}\left(Y_h^{(3)};\ [H, W]\right) \tag{20}$$

The resulting output $\tilde{Y}_h^{(3)}$, now spatially aligned with the original image, are then encoded by a ResNet-50 to produce hidden representations.

5.3 PROJECTION AND CLASSIFICATION.

In both pipelines, the final fully connected layers of the ResNet-50 networks project their outputs into a shared 512-dimensional subspace, aligning curvature-based and frequency-domain features. These two 512-d vectors are concatenated and passed to the final classification layer. Notably, the diagonal high-frequency band (HH subband) is particularly sensitive to oblique variations and irregular details, capturing fine-grained, directionally structured textures that standard convolutions often overlook. By integrating these complementary modalities, our framework assesses how frequency-domain artifacts correlate with less-flat ODE trajectories, enabling cross-validation of cues and substantially improving robustness and generalization.

6 EXPERIMENTS

Datasets. We follow the same evaluation setup as FakeInversion (Cazenavette et al., 2024) for evaluation data. Some synthetic images, such as those from Imagen (Saharia et al., 2022), Midjourney (mid, 2022), and DALL·E 3, are obtained via Hugging Face using KPI. Additionally, we generate thousands of high-fidelity images using open-source text-to-image models based on COCO (Lin et al., 2014) prompts. All settings are kept consistent with those in FakeInversion. The evaluation datasets includes fake images from Kandinsky 2 (Arseniy Shakhmatov & Dimitrov, 2023), Kandinsky 3 (Vladimir Arkhipkin & Dimitrov, 2023), PixArt (Chen et al., 2023), SDXL-DPO (Wallace et al., 2024), SDXL (Podell et al., 2023), SegMoE (Harish Prabhala Yatharth Gupta, 2024),

Eval Set	CNNDet	DMDet	UFD	FakeInv.	B-Free	ours	TPR@5%FPR
DALL·E 2	0.624 / 0.680	0.618 / 0.672	0.700 / 0.776	0.678 / 0.747	**0.906 / 0.969**	0.851 / 0.953	0.762
DALL·E 3	0.659 / 0.716	0.461 / 0.415	0.473 / 0.480	0.698 / 0.759	**0.912 / 0.972**	0.860 / 0.961	0.759
Midjourney v5/6	0.595 / 0.630	0.485 / 0.484	0.558 / 0.592	0.606 / 0.664	0.946 / 0.988	**0.965 / 0.993**	0.966
Imagen	0.674 / 0.714	0.521 / 0.573	0.538 / 0.575	0.720 / 0.807	0.908 / 0.970	**0.960 / 0.991**	0.983
Kandinsky 2	0.574 / 0.600	0.483 / 0.478	0.541/ 0.562	0.652 / 0.699	0.778 / 0.860	**0.950 / 0.995**	0.979
Kandinsky 3	0.609 / 0.659	0.588 / 0.614	0.600 / 0.637	0.684 / 0.743	0.801 / 0.884	**0.948 / 0.991**	0.980
PixArt-α	0.591 / 0.627	0.523 / 0.580	0.606 / 0.647	0.669 / 0.730	0.830 / 0.911	**0.974 / 0.997**	0.982
Playground 2.5	0.553 / 0.582	0.502 / 0.517	0.562 / 0.587	0.591 / 0.625	0.796 / 0.879	**0.863 / 0.899**	0.810
SDXL-DPO	0.761 / 0.843	0.515 / 0.563	0.647 / 0.702	0.801 / 0.881	0.647 / 0.759	**0.843 / 0.957**	0.776
SDXL	0.735 / 0.814	0.549 / 0.568	0.620 / 0.663	0.737 / 0.807	0.651 / 0.776	**0.867 / 0.962**	0.798
Seg-MOE	0.625 / 0.663	0.480 / 0.476	0.586 / 0.620	0.664 / 0.713	0.705 / 0.777	**0.963 / 0.995**	0.978
SSD-1B	0.665 / 0.726	0.583 / 0.556	0.585 / 0.628	0.724 / 0.794	0.833 / 0.919	**0.967 / 0.996**	0.984
Stable-Cascade	0.652 / 0.705	0.539 / 0.565	0.633 / 0.682	0.694 / 0.749	0.824 / 0.906	**0.963 / 0.996**	0.981
Segmind Vega	0.676 / 0.742	0.564 / 0.540	0.587 / 0.623	0.733 / 0.811	0.819 / 0.901	**0.937 / 0.983**	0.965
Würstchen 2	0.580 / 0.610	0.640 / 0.675	0.640 / 0.697	0.658 / 0.705	0.807 / 0.890	**0.871 / 0.916**	0.877
ADM	0.582 / 0.740	0.697 / 0.746	0.682 / 0.779	0.676 / 0.700	0.725 / 0.843	**0.998 / 1.000**	0.999
Glide	0.580 / 0.732	0.784 / 0.857	0.640 / 0.685	0.749 / 0.786	0.700 / 0.739	**0.982 / 0.998**	0.990
VQDM	0.552 / 0.671	0.528 / 0.520	0.840 / 0.876	0.648 / 0.681	0.885 / 0.928	**0.945 / 0.995**	0.976
FLUX	0.498 / 0.540	0.512 / 0.603	0.599 / 0.637	0.651 / 0.656	0.862 / 0.900	**0.963 / 0.996**	0.983
Stable Diffusion 1.4	0.502 / 0.558	0.599 / 0.702	0.651 / 0.663	0.597 / 0.612	0.997 / 1.000	**0.999 / 1.000**	0.999
Stable Diffusion 1.5	0.512 / 0.603	0.585 / 0.653	0.647 / 0.684	0.639 / 0.675	0.995 / 0.997	**0.999 / 1.000**	0.999
Stable Diffusion 3	0.506 / 0.570	0.590 / 0.644	0.613 / 0.638	0.600 / 0.646	0.996 / 0.997	**0.998 / 0.999**	0.999
Average	0.605 / 0.669	0.561 / 0.591	0.616 / 0.656	0.676 / 0.727	0.833 / 0.899	**0.939 / 0.981**	0.933

Table 1: **ACC / AUCROC** comparisons with SOTA methods and **TPR@5%FPR** of our detection. All detectors trained on SD+LAION, except for B-Free, which is trained on its own debiased dataset.

SSD-1B (Gupta et al., 2024), Stable-Cascade (Pablo Pernias & Aubreville, 2023), Segmind-Vega (Gupta et al., 2024), Würstchen 2 (Pablo Pernias & Aubreville, 2023), ADM (Dhariwal & Nichol, 2021), GLIDE (Nichol et al., 2021), VQDM (Gu et al., 2022), FLUX (Labs., 2024), Stable Diffusion 1.4, 1.5 and 3 (Rombach et al., 2022).

Metrics. We report detection ACC, AUC-ROC as the primary metrics, and additionally provide TPR at 5% FPR as a supplementary measure.

NFEs	trained on			Average
	SD + LAION	ProGAN + LSUN	ADM + LAION	
5	0.876 / 0.912	0.885 / 0.928	0.890 / 0.933	0.883 / 0.924
10	0.939 / 0.981	0.894 / 0.946	0.927 / 0.971	0.920 / 0.966
15	0.934 / 0.979	0.892 / 0.945	0.922 / 0.969	0.916 / 0.964
20	0.919 / 0.956	0.874 / 0.919	0.897 / 0.959	0.897 / 0.945
50	0.910 / 0.948	0.875 / 0.923	0.898 / 0.959	0.894 / 0.943

Table 2: Ablation results ACC / AUCROC on different diffusion steps and different training dataset.

Baselines. We use recent methods with publicly available code and pretrained weights as our baselines. DMDet (Corvi et al., 2023b) is a state-of-the-art RGB-only method. UniFD (Ojha et al., 2023) decoupling feature learning from classification. Their framework employs a frozen CLIP encoder (Radford et al., 2021) to extract domain-agnostic embeddings. FakeInversion (Cazenavette et al., 2024) uses text embeddings encoded by CLIP to guide the diffusion-based reconstruction. It builds detection features by combining the reconstruction error, noise map, and the original image. While it achieves strong performance, it incurs high computational cost due to the use of CLIP, BLIP, and U-Net components. B-Free (Guillaro et al., 2025) constructs an unbiased dataset and employs a Vision Transformer to extract discriminative features. For further implementation details of our model, please refer to Appendix A.4.

6.1 Main results

Table 1 presents a comprehensive comparison across a broad range of recent generative models. Our method consistently achieves superior performance on nearly all benchmark datasets, with an average ACC/AUCROC of 0.939 / 0.981, substantially outperforming SOTA baselines (B-Free: 0.833 / 0.899). For instance, on high-fidelity diffusion models such as ADM, Glide, and Stable Diffusion 1.4/1.5/3, our framework reaches near-perfect detection (ACC 0.998 to 0.999, AUCROC 0.998 to 1.000), while baseline methods show noticeable gaps (e.g., B-Free ACC 0.700 to 0.997). This demonstrates the efficiency of our approach in capturing the distinctive generative characteristics of a wide spectrum of diffusion models, including text-to-image systems such as Midjourney v5/6 and Imagen. Unlike conventional methods that rely on model-specific features or complex reconstruction pipelines, our framework generalizes well without requiring text prompts or additional large

models. By leveraging curvature-based features derived from the velocity field of the underlying differential equations, together with high-frequency components extracted via wavelet transforms, our approach effectively detects subtle inconsistencies introduced during generation.

Additionally, Table 1 reports the true positive rate (TPR) at a fixed false positive rate (FPR), which is particularly informative for practical deployment (set to 5% in this study). Our TPR values are consistently high across all datasets, with an average of 0.933. Although some models show slight variations (e.g., Playground 2.5: TPR 0.810), the overall performance remains balanced, highlighting the robustness of the proposed framework.

Ablations. We studied the effects of dataset selection and diffusion steps on model performance. As shown in Table 2, SD+LAION achieves the best training results. Although ProGAN+LSUN lags behind in comparison, it still outperforms the baseline methods. While GAN-based models can be interpreted as diffusion processes along the temporal dimension from a gradient flow perspective (Yi et al., 2023), their stronger model-specific characteristics may compromise generalization. Regarding the number of function evaluations (NFEs), 10 and 15 steps yield the best results, showing nearly identical performance. With 5 steps, performance drops noticeably, likely because the selected ODE pipeline does not support one-step generation, resulting in significant error when NFEs fall below 10. Beyond 20 steps, performance fluctuations are minor, and the additional computational cost from more diffusion steps is not justified.

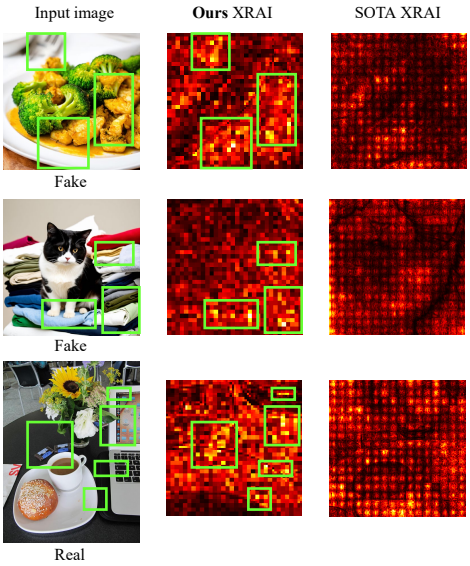

Figure 5: **Saliency Analysis.** The green boxes highlight the most salient regions identified by our model. We visualize these regions using the post-hoc explainability method XRAI (Kapishnikov et al., 2019). Areas in synthetic images with incorrect lighting or perspective exhibit the highest saliency in our model.

Interpretability. Figure 5 visualizes the most salient regions identified by our model. In synthetic images, areas with inconsistent lighting, incorrect perspective, or structural anomalies receive the highest saliency, whereas real images display abundant evidence supporting authenticity. This dual ability underpins the model's balanced detection and indicates potential for localized anomaly detection and artifact correction. In contrast, the baseline B-Free emphasizes global weighted aggregation rather than specific generative artifacts, and its performance depends heavily on large-scale training (360k images: 51k real, 309k fake) with ViT backbones. Our model, by comparison, achieves strong results using only 80k images (40k real, 40k fake).

7 Conclusions

We present a novel framework for synthetic image detection that moves beyond traditional reconstruction-based approaches by leveraging curvature features derived from the velocity field of diffusion models. By formalizing reconstruction error in the optimal transport framework, we identify that synthetic images follow lower-energy trajectories closer to the optimal transport path, and embed this insight into pseudo-Gaussian curvature features. These curvature features, combined with high-frequency components extracted via discrete wavelet transforms as a zeroth-order complement, enable the model to capture subtle generative artifacts with high discriminative power. Our approach not only improves generalization across diverse diffusion models but also reduces computational overhead compared to full reconstruction pipelines. We hope that this work will inspire further research in multimedia forensics and foster progress in the community.

REFERENCES

Midjourney, https://www.midjourney.com/home/. 2022.

Michael S Albergo, Nicholas M Boffi, and Eric Vanden-Eijnden. Stochastic interpolants: A unifying framework for flows and diffusions. *arXiv preprint arXiv:2303.08797*, 2023.

Aleksandr Nikolich Vladimir Arkhipkin Igor Pavlov Andrey Kuznetsov Arseniy Shakhmatov, Anton Razzhigaev and Denis Dimitrov. kandinsky 2.2. 2023.

George Cazenavette, Avneesh Sud, Thomas Leung, and Ben Usman. Fakeinversion: Learning to detect images from unseen text-to-image models by inverting stable diffusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10759–10769, 2024.

Junsong Chen, Jincheng Yu, Chongjian Ge, Lewei Yao, Enze Xie, Yue Wu, Zhongdao Wang, James Kwok, Ping Luo, Huchuan Lu, et al. Pixart-α: Fast training of diffusion transformer for photorealistic text-to-image synthesis. *arXiv preprint arXiv:2310.00426*, 2023.

Beilin Chu, Xuan Xu, Xin Wang, Yufei Zhang, Weike You, and Linna Zhou. Fire: Robust detection of diffusion-generated images via frequency-guided reconstruction error. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 12830–12839, 2025.

Riccardo Corvi, Davide Cozzolino, Giovanni Poggi, Koki Nagano, and Luisa Verdoliva. Intriguing properties of synthetic images: from generative adversarial networks to diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 973–982, 2023a.

Riccardo Corvi, Davide Cozzolino, Giada Zingarini, Giovanni Poggi, Koki Nagano, and Luisa Verdoliva. On the detection of synthetic images generated by diffusion models. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1–5. IEEE, 2023b.

Marco Cuturi. Sinkhorn distances: Lightspeed computation of optimal transport. *Advances in neural information processing systems*, 26, 2013.

Valentin De Bortoli, James Thornton, Jeremy Heng, and Arnaud Doucet. Diffusion schrödinger bridge with applications to score-based generative modeling. *Advances in neural information processing systems*, 34:17695–17709, 2021.

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

Tim Dockhorn, Arash Vahdat, and Karsten Kreis. Genie: Higher-order denoising diffusion solvers. *Advances in Neural Information Processing Systems*, 35:30150–30166, 2022.

Shuyang Gu, Dong Chen, Jianmin Bao, Fang Wen, Bo Zhang, Dongdong Chen, Lu Yuan, and Baining Guo. Vector quantized diffusion model for text-to-image synthesis. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 10696–10706, 2022.

Fabrizio Guillaro, Giada Zingarini, Ben Usman, Avneesh Sud, Davide Cozzolino, and Luisa Verdoliva. A bias-free training paradigm for more general ai-generated image detection. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 18685–18694, 2025.

Yatharth Gupta, Vishnu V Jaddipal, Harish Prabhala, Sayak Paul, and Patrick Von Platen. Progressive knowledge distillation of stable diffusion xl using layer level loss. *arXiv preprint arXiv:2401.02677*, 2024.

Vishnu V Jaddipal Harish Prabhala Yatharth Gupta. Segmoe: Segmind mixture of diffusion experts, https : / / github . com / segmind / segmoe. 2024.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020.

Aapo Hyvärinen and Peter Dayan. Estimation of non-normalized statistical models by score matching. *Journal of Machine Learning Research*, 6(4), 2005.

Yonghyun Jeong, Doyeon Kim, Youngmin Ro, and Jongwon Choi. Frepgan: robust deepfake detection using frequency-level perturbations. In *Proceedings of the AAAI conference on artificial intelligence*, volume 36, pp. 1060–1068, 2022.

Andrei Kapishnikov, Tolga Bolukbasi, Fernanda Viégas, and Michael Terry. Xrai: Better attributions through regions. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 4948–4957, 2019.

Black Forest Labs. Flux. In *https://huggingface.co/black-forestlabs/FLUX.1- schnell*, 2024.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, pp. 740–755. Springer, 2014.

Yaron Lipman, Ricky TQ Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow matching for generative modeling. *arXiv preprint arXiv:2210.02747*, 2022.

Huan Liu, Zichang Tan, Chuangchuang Tan, Yunchao Wei, Jingdong Wang, and Yao Zhao. Forgery-aware adaptive transformer for generalizable synthetic image detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10770–10780, 2024.

Xingchao Liu, Chengyue Gong, and Qiang Liu. Flow straight and fast: Learning to generate and transfer data with rectified flow. *arXiv preprint arXiv:2209.03003*, 2022.

Yunpeng Luo, Junlong Du, Ke Yan, and Shouhong Ding. Lareˆ 2: Latent reconstruction error based method for diffusion-generated image detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 17006–17015, 2024.

Alex 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. *arXiv preprint arXiv:2112.10741*, 2021.

Alexander Quinn Nichol and Prafulla Dhariwal. Improved denoising diffusion probabilistic models. In *International conference on machine learning*, pp. 8162–8171. PMLR, 2021.

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*, pp. 24480–24489, 2023.

Mats L. Richter Christopher Pal Pablo Pernias, Dominic Rampas and Marc Aubreville. Wuerstchen: Efficient pretraining of text-to-image models. 2023.

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. *arXiv preprint arXiv:2307.01952*, 2023.

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*, pp. 8748–8763. PMLR, 2021.

Jonas Ricker, Denis Lukovnikov, and Asja Fischer. Aeroblade: Training-free detection of latent diffusion images using autoencoder reconstruction error. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9130–9140, 2024.

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*, pp. 10684–10695, 2022.

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.

Zeyang Sha, Zheng Li, Ning Yu, and Yang Zhang. De-fake: Detection and attribution of fake images generated by text-to-image generation models. In *Proceedings of the 2023 ACM SIGSAC conference on computer and communications security*, pp. 3418–3432, 2023.

Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*, 2020a.

Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. *arXiv preprint arXiv:2011.13456*, 2020b.

Yang Song, Conor Durkan, Iain Murray, and Stefano Ermon. Maximum likelihood training of score-based diffusion models. *Advances in neural information processing systems*, 34:1415–1428, 2021.

Chuangchuang Tan, Yao Zhao, Shikui Wei, Guanghua Gu, Ping Liu, and Yunchao Wei. Frequency-aware deepfake detection: Improving generalizability through frequency space domain learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 5052–5060, 2024a.

Chuangchuang Tan, 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. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 28130–28139, 2024b.

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

Bram Wallace, Meihua Dang, Rafael Rafailov, Linqi Zhou, Aaron Lou, Senthil Purushwalkam, Stefano Ermon, Caiming Xiong, Shafiq Joty, and Nikhil Naik. Diffusion model alignment using direct preference optimization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8228–8238, 2024.

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*, pp. 8695–8704, 2020.

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*, pp. 22445–22455, 2023.

Mingxuan Yi, Zhanxing Zhu, and Song Liu. Monoflow: Rethinking divergence gans via the perspective of wasserstein gradient flows. In *International Conference on Machine Learning*, pp. 39984–40000. PMLR, 2023.