# D³QE: Learning Discrete Distribution Discrepancy-aware Quantization Error for Autoregressive-Generated Image Detection

Yanran Zhang[1,*]    Bingyao Yu[1,*,†]    Yu Zheng[1]    Wenzhao Zheng[1]    Yueqi Duan[2]    Lei Chen[1]
Jie Zhou[1]    Jiwen Lu[1,†]

[1] Department of Automation, Tsinghua University, China
[2] Department of Electronic Engineering, Tsinghua University, China

{zhangyr21}@mails.tsinghua.edu.cn; {wenzhao.zheng}@outlook.com;
{yuby, yu-zheng, duanyueqi, leichenthu, jzhou, lujiwen}@tsinghua.edu.cn

## Abstract

*The emergence of visual autoregressive (AR) models has revolutionized image generation while presenting new challenges for synthetic image detection. Unlike previous GAN or diffusion-based methods, AR models generate images through discrete token prediction, exhibiting both marked improvements in image synthesis quality and unique characteristics in their vector-quantized representations. In this paper, we propose to leverage Discrete Distribution Discrepancy-aware Quantization Error (D³QE) for autoregressive-generated image detection that exploits the distinctive patterns and the frequency distribution bias of the codebook existing in real and fake images. We introduce a discrete distribution discrepancy-aware transformer that integrates dynamic codebook frequency statistics into its attention mechanism, fusing semantic features and quantization error latent. To evaluate our method, we construct a comprehensive dataset termed **ARForensics** covering 7 mainstream visual AR models. Experiments demonstrate superior detection accuracy and strong generalization of D³QE across different AR models, with robustness to real-world perturbations. Code is available at https://github.com/Zhangyr2022/D3QE.*

## 1. Introduction

With the advent of Generative Adversarial Networks (GANs) [12] and Variational AutoEncoders (VAEs) [20], significant advancements have been made in the realm of generative AI technology within computer vision. Following this, the emergence of innovative technologies such as Flow Models and Diffusion Models [41, 42] has further enhanced the fidelity and quality of image generation. Nowa-
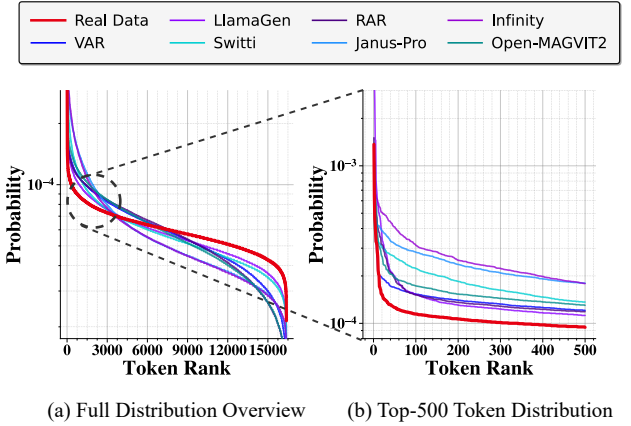


(a) Full Distribution Overview    (b) Top-500 Token Distribution

Figure 1. **Visualization of Discrete Distribution Discrepancy.** To elucidate the mechanism of **D³QE**, we analyze token probability distributions from LlamaGen's tokenizer using autoregressive sampling. (a) shows the full codebook vector probability distribution, while (b) displays the top-500 activation probabilities. The real data exhibits pronounced long-tail characteristics, whereas generated samples demonstrate concentrated probability mass in the peak regions, which **D³QE** leverages for detection.

days, Autoregressive Models [13, 45, 48] are capable of not only accurately capturing the structural features of images but also efficiently producing high-quality visual content. On one hand, visual generative models have the potential to drastically reduce the time expenditure associated with manual creation, thereby empowering industries such as art, film production, and education. On the other hand, while these models facilitate the easy acquisition of images that can be indistinguishable from reality to the human eye, they also usher in a host of potential social risks and ethical dilemmas.

In recent times, considerable efforts have been dedicated to the detection of generated images, with the goal of mitigating the trust crisis and addressing privacy risks that arise from the use of generative models. Existing detection meth-

---

ods have primarily focused on high-frequency artifacts in GANs [11, 36] or iterative noise patterns in diffusion models [56], overlooking the unique characteristics of autoregressive models' discrete encoding. Traditional detection methods based on superficial statistical features struggle to identify these samples because the artifacts manifest in the discrete latent space rather than in pixel-level patterns, making them particularly challenging to detect through conventional image analysis techniques. For the newly emerging generative models, the generalization capability of detectors is paramount, and the latest AR models present a significant challenge to their effectiveness.

Discrete coding enhances the efficiency of reasoning and fosters diverse outcomes in generative models, while also highlighting the variations in statistical distributions across different images. As illustrated in Figure 1, the discrete feature reveals distinct patterns in the utilization of codebook tokens among various generative models, with a more pronounced discrepancy between real and generated images. Motivated by this, we delve into the prior knowledge of codebooks to construct robust features and enhance their expressiveness. Furthermore, by integrating the frequency disparity between real and fake codebooks into the cross attention mechanism and aligning it with quantization error, we merge the features with the semantic features extracted by the backbone network. A classifier is then employed to predict the final outcome. For the first time, we have established a new benchmark termed **ARForensics** for the detection of images generated by AR models, encompassing the current top-performing mainstream AR models. Effectiveness and generalization of our method were rigorously tested in a challenging experimental setting that included GANs, diffusion models, and AR models, demonstrating its robust performance.

## 2. Related Work

**Visual Generation.** In recent years, visual generation models have experienced rapid development, with generated images and videos finding widespread applications in creative design and media production [4, 41]. Mainstream generative models encompass four paradigms: GANs, VAEs, Diffusion Models, and Autoregressive Models. GANs [12, 38] generate realistic images through adversarial training, with subsequent improvements [3, 18, 19] significantly enhancing generation quality. VAEs [14, 20] are based on latent space, while follow-up studies [21, 49, 52] addressed the blurry reconstruction issue through improved encoding structures. Diffusion models [15, 30, 35, 41–43] achieve high quality generation through iterative denoising processes. Recently, visual autoregressive models [50, 51] have demonstrated remarkable capabilities by discretizing visual content into sequences and progressively predicting conditional probabilities, offering advantages in training stability and generation speed. Related works include token-based autoregressive modeling [10, 17, 37, 45, 60] and scale-based autoregressive modeling [13, 48], both achieving significant progress in visual content generation. With the rapid development of autoregressive models, exploring effective detection methods for autoregressive-generated images has become particularly crucial.

**AI-generated Image Detection.** With the rapid advancement of generative models, AI-generated image detection techniques have become crucial for ensuring information security and maintaining digital media authenticity. Recent research has evolved from local feature analysis to global semantic mining. Early studies [27–29] focused mainly on handcrafted features, including color distribution anomalies, saturation differences, and texture co-occurrence patterns. However, these methods showed limited generalization to newer generative models.

The research community has proposed numerous detection methods targeting specific generative architectures. For GAN detection, studies have shifted towards frequency domain analysis. For instance, CNNSpot [55] enhanced cross-GAN architecture generalization through optimized data augmentation, while FreDect [11] revealed artifacts introduced by GAN upsampling operations in the frequency domain. Following the rise of diffusion models, UnivFD [32] leveraged ViT's pre-trained features to train universal linear classifiers, while DIRE [56] and AEROBLADE [40] achieved detection based on ADM [9] reconstruction errors and autoencoder reconstruction errors, respectively. NPR [47] designed a detection network that targets artifacts from common upsampling operations, and FatFormer [24] integrated local forgery traces through CLIP adapters.

However, existing approaches face two core challenges. First, while most research focuses on GANs and diffusion models, specific detection methods for autoregressive generative models remain underexplored. Artifacts from these models may exist in directional correlations or latent space discretization features. Second, current benchmark datasets lack samples from autoregressive models, limiting the validation of generalization capabilities of detection methods. Although the Chameleon benchmark [58] has improved in terms of diversity and realism, a more comprehensive evaluation framework is needed to support research on emerging models.

## 3. Methods

In this section, we present $\mathbf{D^3QE}$, a novel framework for detecting autoregressive generated images. Our method leverages the unique discretization characteristics of visual autoregressive models. We first analyze the theoretical foundations of autoregressive modeling. Then, we detail our detection approach that combines discrete distribution aware-
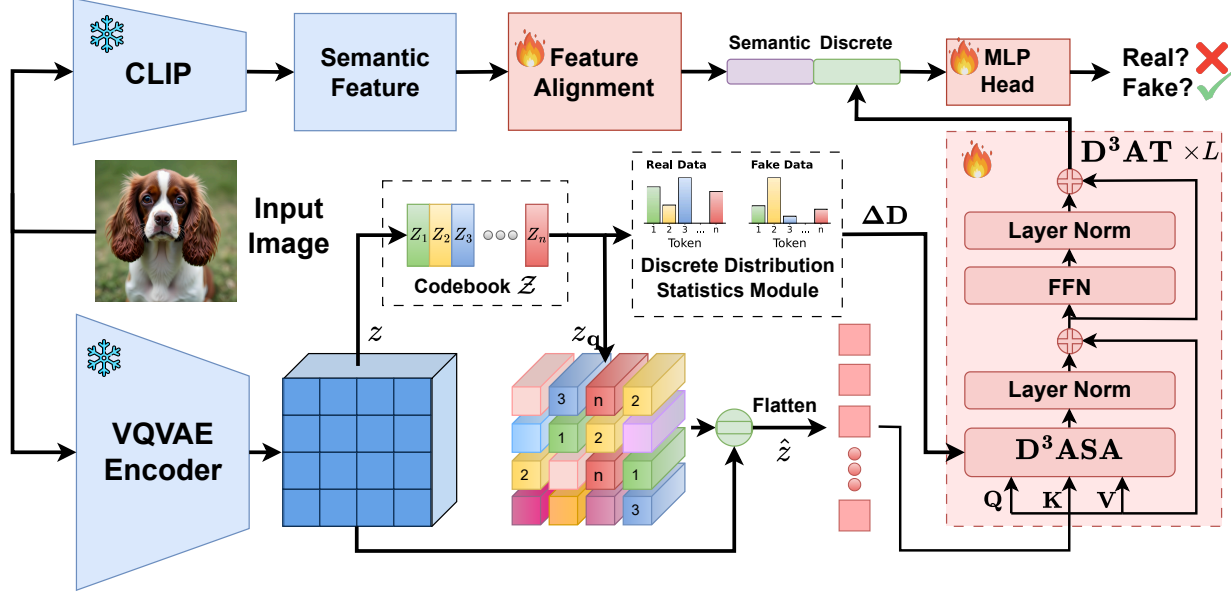
Figure 2. **$D^3QE$ pipeline.** Our approach first extracts quantized representations through a VQVAE encoder, computes the discrete distribution discrepancy between pre- and post-quantization features, and obtains discrete features via the **$D^3AT$** module. Semantic features are extracted using CLIP in parallel. The feature alignment module processes global semantic features, which then fuse with local discrete features for binary classification between generated and real samples. Blue snowflake symbols ❄ indicate frozen parameters, while red flame symbols 🔥 denote trainable modules.

ness with semantic understanding.

## 3.1. Preliminary

**Visual Autoregressive Modeling.** Visual autoregressive models generate visual content in a sequential manner. These models operate through two key processes: discrete quantization and autoregressive modeling. The approach first trains a discrete variational autoencoder to quantize vectors in the latent space. Then, it performs autoregressive prediction of subsequent elements. This methodology has proven highly effective in capturing complex visual dependencies and generating high-quality content. [57]

**Modeling via Next Token Prediction.** The next-token prediction methodology, borrowed from Natural Language Processing, has demonstrated remarkable generative capabilities in recent times. [10, 45] At its core, this approach employs vector quantization through a VQVAE-like [52] structure to compress continuous visual content into discrete sequences. The discretization process transforms input images into continuous latent representations, which are then quantized using a learnable codebook. After discretization, the model performs autoregressive prediction by estimating the probability of each subsequent token based on all preceding tokens. This sequential generation approach effectively captures both local patterns and global structural relationships in visual data.

**Modeling via Next Scale Prediction.** VAR [48] pioneered the Next Scale Prediction approach, which discretizes content into multi-scale sequences through a hierarchical struc-

ture. Unlike the token-by-token prediction, it models visual content from coarse to fine scales, where the autoregressive unit is a complete token map. The discretization process utilizes Residual Quantization from RQVAE [21], obtaining discrete token maps through coarse-to-fine residual estimation. This hierarchical strategy enables high-quality image reconstruction with compact codebook capacity. The subsequent autoregressive modeling predicts finer-scale representations conditioned on preceding coarser scales.

## 3.2. Motivation and Design Principles

**Design Insights.** From the above analysis, we observe that discretization serves as a crucial component in mainstream visual autoregressive modeling. This discretization process fundamentally distinguishes visual autoregressive models from continuous generative paradigms (e.g. diffusion models), a design choice that has been systematically validated in seminal works such as VQVAE [52], VQGAN [10], and VAR [48]. This architectural decision has profound implications for both model efficiency and generation quality.

Discretization has become a core feature of visual autoregressive models due to the following advantages. First, autoregressive models inherently decompose joint distributions through conditional probability chains. By transforming high-dimensional continuous visual data into discrete symbolic sequences, the model circumvents the curse of dimensionality while leveraging mature classification-based cross-entropy optimization paradigms from language modeling. Second, discrete distributions enable exact likeli-
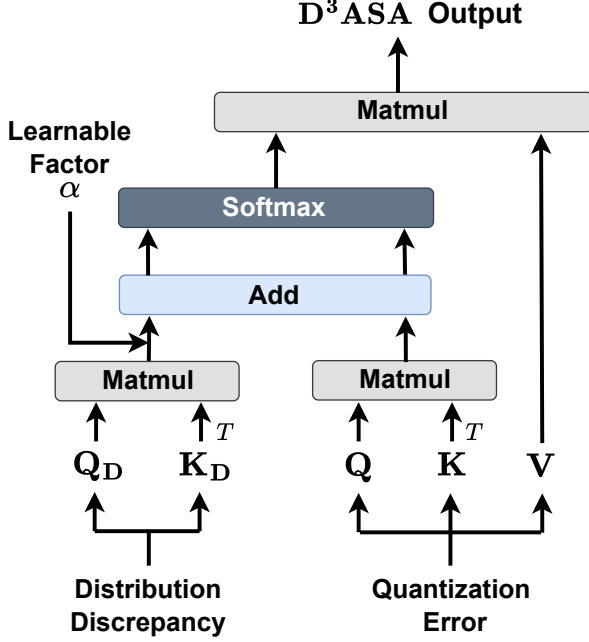
**D³ASA Output**

Figure 3. Illustration of **D³ASA** Module in Equation 8, which incorporates distribution discrepancy information into the attention mechanism.

hood computation through classification cross-entropy loss, avoiding the mode collapse issues prevalent in continuous models. While continuous autoregressive models like MAR [22] attempt to bypass discretization, they require diffusion losses for probability density estimation and often suffer from detail loss due to the smoothness of continuous latent spaces. Moreover, the structural homology between discrete tokens and NLP vocabularies enables visual autoregressive models to directly inherit architectural advantages from language models (as demonstrated in DALL·E [39] and Parti [59]), facilitating unified cross-modal modeling.

**Overview.** Based on these insights, we propose to detect autoregressive generated images by analyzing their distinctive discrete distribution patterns. And our architecture consists of three key components: a quantization error representation module, a discrete distribution discrepancy-aware transformer, and a semantic feature embedding module.

### 3.3. D³QE

Based on the above analysis, we present our detection framework **D³QE** that explicitly leverages the statistical signatures in autoregressive generated images. The discretization process in visual autoregressive models introduces distinctive statistical signatures that can be leveraged for detection. This phenomenon occurs primarily because the finite codebook capacity struggles to fully capture the long-tailed distribution of natural images. The training objective of discrete VAEs forces the encoder to map similar features to the same codebook entries, resulting in high-

frequency tokens corresponding to common local patterns. Rare patterns in real data such as specific object parts are compressed into high-frequency tokens due to their low occurrence rate, leading to reduced generation diversity. Furthermore, the explicit truncation introduced by top-p/top-k [16, 44, 46] sampling strategies directly results in the truncation of long-tail distributions. As shown in Figure 1, these effects create observable differences in codebook distribution statistics between real and generated images. To effectively capture these distinctive patterns between real and synthetic samples, we propose the following modules.

**Quantization Error Representation and Discrete Distribution Statistics Module.** We first employ a frozen discrete autoencoder to tokenize images into discrete representations. Specifically, given an input RGB image $I \in \mathbb{R}^{H \times W \times 3}$, where $H$ and $W$ denote the height and width respectively, a deep neural network $\mathcal{E}$ encodes it into a continuous latent map $z = \mathcal{E}(I) \in \mathbb{R}^{h \times w \times c}$, where $h$, $w$, and $c$ represent the height, width, and channel dimensions of the latent space. The latent vectors are then projected into a learnable finite codebook $\mathcal{Z} = \{z_k\}_{k=1}^{N} \subset \mathbb{R}^c$, which contains $N$ discrete vectors. The process of finding the nearest codebook vector for each latent vector can be formulated as:

$$z_q = \left( \arg\min_{z_k \in \mathcal{Z}} \|z_{ij} - z_k\| \right) \in \mathbb{R}^{h \times w \times c} \quad (1)$$

where $z_{ij}$ represents the latent vector at spatial position $(i, j)$ in the continuous latent map.

During training, we implement two discrete distribution tracking modules to monitor the distribution patterns of quantization indices for both real and synthetic images. These modules maintain frequency statistics for each codebook entry:

$$D_s^{(t+1)}[k] = D_s^{(t)}[k] + \sum_{i,j} \mathbf{1}[q(z_{ij}) = k], s \in \{\text{real}, \text{fake}\} \quad (2)$$

where $D_s^{(t)}[k]$ tracks the cumulative frequency of codebook index $k \in \{1, \ldots, N\}$ at training step $t$, $q(z_{ij})$ denotes the index of the nearest codebook entry for the latent vector at position $(i, j)$.

After obtaining the quantized representation $z_q$, we compute the quantization error features to capture the discrepancy between continuous and discrete representations. This quantization gap potentially encodes distinctive patterns that differentiate real from synthetic images:

$$\hat{z} = (z_q - z) \in \mathbb{R}^{h \times w \times c} \quad (3)$$

**Discrete Distribution Discrepancy-Aware Transformer (D³AT).** To effectively capture the distinctive patterns between real and synthetic samples, we propose a transformer-based module that explicitly incorporates codebook distribution information. Since the codebook usage patterns often differ significantly between real and synthetic images,

we first compute their distribution discrepancy:

$$\Delta \mathbf{D} = \text{normalize}(D_{\text{fake}} - D_{\text{real}}) \quad (4)$$

where $\Delta \mathbf{D} \in \mathbb{R}^N$ represents the normalized difference in codebook entry frequencies. The input features are reshaped into a sequence $\hat{\mathbf{X}} \in \mathbb{R}^{n \times c}$ following raster scan order, where $n$ denotes the sequence length and $c$ is the feature dimension. To effectively model both local and global dependencies while maintaining distribution awareness, our transformer architecture consists of $L$ layers. Each incorporating a novel Discrete Distribution Discrepancy-Aware Self-Attention ($\mathbf{D}^3\mathbf{ASA}$) mechanism. For the $\ell$-th layer:

$$\hat{\mathbf{X}}_\ell = \text{LN}(\mathbf{D}^3\mathbf{ASA}(\mathbf{X}_{\ell-1}, \Delta \mathbf{D})) + \mathbf{X}_{\ell-1} \quad (5)$$

$$\mathbf{X}_\ell = \text{LN}(\text{MLP}(\hat{\mathbf{X}}_\ell)) + \hat{\mathbf{X}}_\ell \quad (6)$$

where $\text{LN}(\cdot)$ denotes layer normalization and $\mathbf{X}_0 = \hat{\mathbf{X}}$. The $\mathbf{D}^3\mathbf{ASA}$ mechanism enhances traditional self-attention by incorporating codebook distribution information. The distribution-aware attention is formulated as:

$$\mathbf{Q_D} = \text{MLP}_q(\Delta \mathbf{D}), \quad \mathbf{K_D} = \text{MLP}_k(\Delta \mathbf{D}) \quad (7)$$

$$\mathbf{D}^3\mathbf{ASA}(\mathbf{X}, \Delta \mathbf{D}) = \text{softmax}\left(\frac{\mathbf{QK}^T}{\sqrt{d_k}} + \frac{\mathbf{Q_D K_D}^T}{\alpha}\right)\mathbf{V} \quad (8)$$

where $\{\mathbf{Q}, \mathbf{K}, \mathbf{V}\}$ are query, key, and value matrices projected from input features $\mathbf{X}$, $\text{MLP}_q$ and $\text{MLP}_k$ are learnable distribution projections, and $\alpha$ is a learnable scaling factor that balances the contribution of distribution information.

**Semantic Feature Embedding.** Beyond the distinctive patterns in local codebook token distributions, synthetic images often exhibit global semantic discrepancies compared to real images. To capture these high-level semantic differences, we leverage a pre-trained CLIP-ViT model to extract semantic features $\mathbf{F}_{\text{CLIP}}$. The CLIP features provide complementary global context information that helps identify subtle semantic inconsistencies in synthetic images.

**Classifier.** Finally, we construct our classifier by combining both global semantic features and local token distribution patterns. To reduce computational complexity while preserving discriminative information, we first apply average pooling to the $\mathbf{D}^3\mathbf{AT}$ output features to obtain a compact representation $\mathbf{F}_{\text{D}}$. The final prediction is computed by:

$$y = \text{MLP}(\text{concat}[\mathcal{A}_{\text{D}}(\mathbf{F}_{\text{D}}), \mathcal{A}_{\text{CLIP}}(\mathbf{F}_{\text{CLIP}})]) \quad (9)$$

where $\mathcal{A}_{\text{D}}$ and $\mathcal{A}_{\text{CLIP}}$ are feature alignment modules consisting of MLPs and layer normalization to project features into a shared embedding space before concatenation.

## 4. Experiments

In this section, we systematically constructed a comprehensive dataset of autoregressive model-generated images incorporating various generation strategies. Following dataset construction, we conducted systematic training and validation of our proposed model alongside existing SOTA baselines. Through comparative analysis of model performance, we validated the effectiveness of our proposed method and its critical components. We further evaluated the framework through robustness and generalization experiments.

### 4.1. Settings

**ARForensics: A Dataset of Images Generated by Autoregressive Models.** To validate the effectiveness of our method, we constructed the first benchmark dataset specifically designed for visual autoregressive models. We selected 7 representative autoregressive generative models — LlamaGen [45], VAR [48], Infinity [13], Janus-Pro [5], RAR [61], Switti [54], and Open-MAGVIT2 [26], covering diverse architectures (token-based and scale-based) and resolutions. These models exhibit significant variations in their discretization processes and key technical parameters such as codebook capacity.

The dataset comprises 152,000 real samples and 152,000 generated samples. The real data come from ImageNet [8], one of the most influential benchmarks in computer vision, which contains manually annotated images across 1,000 fine-grained categories. ImageNet's rigorous quality control system provides a reliable foundation for model evaluation. Our dataset consists of three splits: a training set of 100,000 LlamaGen-generated images paired with an equal number of randomly sampled ImageNet images (100 per category), a validation set of 10,000 image pairs, and a comprehensive test set incorporating 6,000 samples from each of the 7 autoregressive models, balanced with corresponding ImageNet test samples. It's worth noting that real images across all subsets are independently sampled to avoid evaluation bias from data overlap. This balanced design enables detection models to fully capture the characteristics of different generators while mitigating the impact of sample imbalance common in traditional datasets.

For image generation methodology, text-to-image models (Infinity, Janus-Pro, Switti) utilize a standard prompt template "A photo of [class]", where [class] corresponds to ImageNet labels. Other autoregressive models (LlamaGen, VAR, RAR, Open-MAGVIT2) directly employ their ImageNet pre-trained versions, generating images through category-conditional synthesis. This approach produces synthetic images with high variability and reasonableness.

**Cross-Paradigm Test Set.** To comprehensively validate our method's cross-domain generalization capability, we constructed a multi-modal test set of generated samples: 1) Based on the ForenSynths [55] dataset, we selected samples from representative GAN architectures including ProGAN [18], StyleGAN [19], StyleGAN2 [53], BigGAN [3], CycleGAN [7], StarGAN [6], and GauGAN [33]; 2) We incorporated samples from mainstream

Table 1. **Performance comparison on ARForensics dataset.** Detection accuracy (Acc.) and average precision (A.P.) of various detectors (rows) against real and AI-generated images from different generative models (columns).

| Method | LlamaGen | | VAR | | Infinity | | Janus-Pro | | RAR | | Switti | | Open-MAGVIT2 | | Mean | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Acc. | A.P. | Acc. | A.P. | Acc. | A.P. | Acc. | A.P. | Acc. | A.P. | Acc. | A.P. | Acc. | A.P. | Acc. | A.P. |
| CNNSpot[55] | 99.94 | 99.94 | 50.26 | 70.53 | 50.87 | 78.06 | 95.7 | 99.95 | 50.80 | 61.67 | 56.58 | 93.91 | 50.12 | 57.39 | 64.90 | 80.21 |
| FreDect [11] | 99.80 | 100.00 | 52.88 | 88.18 | 50.17 | 60.13 | 88.94 | 99.54 | 52.52 | 83.31 | 50.04 | 59.01 | 57.09 | 86.53 | 64.49 | 82.39 |
| Gram-Net [25] | 99.57 | 99.98 | 55.04 | 84.57 | 52.38 | 76.80 | 74.48 | 97.33 | 49.95 | 52.72 | 57.74 | 88.66 | 50.08 | 53.72 | 62.75 | 79.11 |
| LNP [23] | 99.48 | 99.99 | 49.64 | 55.42 | 49.76 | 49.94 | 99.53 | 99.98 | 49.69 | 55.61 | 70.28 | 94.16 | 49.63 | 54.92 | 66.86 | 72.86 |
| UnivFD [32] | 89.87 | 96.53 | 80.53 | 91.62 | 71.72 | 85.77 | 84.28 | 93.94 | 88.33 | 95.93 | 76.00 | 88.43 | 66.21 | 80.87 | 79.56 | 90.44 |
| NPR [47] | 99.96 | 100.00 | 56.87 | 88.68 | 88.48 | 97.98 | 93.67 | 99.18 | 52.30 | 74.99 | 51.97 | 87.04 | 63.00 | 92.11 | 72.32 | 91.43 |
| **$D^3QE$(ours)** | 97.19 | 99.43 | 85.33 | 95.30 | 62.88 | 79.39 | 92.28 | 97.53 | 91.69 | 97.77 | 75.31 | 89.09 | 70.08 | 85.98 | 82.11 | 92.07 |

diffusion models through the GenImage [62] dataset, including ADM [9], GLIDE [31], Midjourney [1], Stable Diffusion V1.4 [41], Stable Diffusion V1.5 [41], and Wukong [2], to evaluate adaptability of $D^3QE$ across different generative paradigms.

**Evaluation Metrics.** Our experiments strictly follow standard evaluation protocols in the field of generated image detection, employing Average Accuracy (Acc.) and Average Precision (A.P.) as core metrics. Acc is computed through binary classification with a fixed threshold of 0.5, while AP evaluates comprehensive performance of the classifier across different decision thresholds based on the area under the precision-recall curve.

**Baseline Methods.** We conducted comparative experiments with state-of-the-art detection methods spanning multiple technical approaches: CNNSpot [55], FreDect [11], Gram-Net [25], LNP [23], UnivFD [32], and NPR [47]. All baselines were evaluated using official source code and recommended parameter configurations to ensure fair comparison.

### 4.2. Implementation Details

Our VQVAE encoder adopts the visual tokenizer in LlamaGen with a 16× downsampling tokenizer and a codebook of size 16,384. The encoder processes input images at $256 \times 256$ resolution. When optimizing subsequent modules, we freeze the CLIP encoder, VQVAE backbone, and codebook, while dynamically learning codebook statistics during training. The two-layer $D^3AT$ module uses hidden dimension 512, while semantic features are extracted via CLIP-ViT from $224 \times 224$ preprocessed inputs. Experiments were conducted on an NVIDIA RTX 4090 GPU using PyTorch [34]. Training utilized AdamW via learning rate 0.0001, weight decay 0.01, batch size 32, for 10 epochs.

### 4.3. Quantitative Results

In this section, we have thoroughly validated the effectiveness of our proposed method across a variety of generative models through extensive experiments. We conducted systematic experiments on 7 autoregressive models, 7 GAN

models, and 6 diffusion models. The results demonstrate that our method achieves significant performance improvements compared to existing approaches across all models.

#### 4.3.1. Performance on ARForensics

As shown in Table 1, our proposed method demonstrates superior generalization capability across mainstream autoregressive models (including VAR, RAR, Open-MAGVIT2, etc.). Compared to traditional CNN-based detector CNNSpot, our method achieves significant improvements of 18.21% and 11.86% in average accuracy and average precision, respectively. In particular, for the latest scale-based autoregressive model VAR, our method achieves 85.33% accuracy and 95.30% AP, substantially outperforming the second-best method UnivFD at 80.53%. This advantage stems from our deep modeling of discrete codebook statistical characteristics in autoregressive generation systems: capturing information loss during VQVAE compression through quantized residual features, while revealing codebook distribution concentration phenomena through codebook frequency difference maps.

Notably, previous baselines typically perform well on models that are architecturally similar to the training set model LlamaGen (e.g., Janus-Pro), but degrade significantly on architecturally distinct models (e.g., VAR/Switti), indicating their inability to capture common characteristics across autoregressive models. In contrast, our method maintains high detection performance on traditional raster-order models while demonstrating unique advantages in detecting both novel scale-based paradigms like VAR and random-scan-order models like RAR. These results validate that our $D^3AT$ successfully captures cross-scale statistical biases through dynamic attention modulation. The features obtained from this module, when jointly optimized with CLIP semantic features, enable our codebook distribution-based framework to effectively handle rapidly evolving autoregressive architectures.

#### 4.3.2. Performance on Cross-Paradigm Models

To evaluate the generalization capability of our method, we directly apply our model trained on autoregressive samples

Table 2. **Performance comparison on GAN-based synthesis using ForenSynths [55] test set.** Detection accuracy (Acc.) and average precision (AP) of various detectors (rows) against real and AI-generated images from different generative models (columns).

| Method | ProGAN | | StyleGAN | | StyleGAN2 | | BigGAN | | CycleGAN | | StarGAN | | GauGAN | | Mean | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Acc. | A.P. | Acc. | A.P. | Acc. | A.P. | Acc. | A.P. | Acc. | A.P. | Acc. | A.P. | Acc. | A.P. | Acc. | A.P. |
| CNNSpot [55] | 50.26 | 47.83 | 49.97 | 43.89 | 49.99 | 46.49 | 50.03 | 41.16 | 49.74 | 50.56 | 50.00 | 44.66 | 50.00 | 52.73 | 50.00 | 46.76 |
| FreDect [11] | 50.25 | 66.83 | 50.97 | 71.46 | 49.92 | 56.13 | 50.48 | 55.12 | 50.68 | 53.87 | 50.93 | 98.44 | 49.94 | 33.03 | 50.45 | 62.12 |
| Gram-Net [25] | 49.78 | 45.85 | 50.04 | 50.27 | 49.77 | 45.98 | 49.78 | 38.00 | 48.07 | 54.19 | 50.00 | 83.00 | 50.00 | 50.65 | 49.64 | 52.56 |
| LNP [23] | 50.00 | 44.06 | 50.69 | 50.69 | 50.01 | 50.01 | 50.00 | 48.99 | 50.00 | 55.86 | 50.00 | 35.76 | 50.00 | 52.87 | 50.10 | 48.32 |
| UnivFD [32] | 88.17 | 94.12 | 72.98 | 80.90 | 72.23 | 81.14 | 88.78 | 95.60 | 71.23 | 73.74 | 79.99 | 79.99 | 91.52 | 97.33 | 80.70 | 86.12 |
| NPR [47] | 51.36 | 93.00 | 52.54 | 74.35 | 50.93 | 75.80 | 50.30 | 64.07 | 48.83 | 66.31 | 53.83 | 98.92 | 50.03 | 66.09 | 51.12 | 76.93 |
| **D³QE(ours)** | 95.20 | 97.68 | 77.67 | 88.65 | 75.83 | 88.61 | 86.03 | 94.79 | 82.44 | 92.31 | 74.64 | 85.65 | 94.31 | 97.94 | 83.73 | 92.23 |

Table 3. **Performance comparison on diffusion-based generation using GenImage [62] test set.** Detection accuracy (Acc.) and average precision (AP) of various detectors (rows) against real and AI-generated images from different generative models (columns).

| Method | ADM | | Glide | | Midjourney | | SDv1.4 | | SDv1.5 | | Wukong | | Mean | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Acc. | A.P. | Acc. | A.P. | Acc. | A.P. | Acc. | A.P. | Acc. | A.P. | Acc. | A.P. | Acc. | A.P. |
| CNNSpot [55] | 50.40 | 55.54 | 54.81 | 86.75 | 50.93 | 76.88 | 50.23 | 63.90 | 50.29 | 65.17 | 50.35 | 63.25 | 51.17 | 68.58 |
| FreDect [11] | 51.83 | 58.32 | 63.82 | 91.69 | 50.57 | 63.73 | 56.80 | 90.23 | 56.73 | 89.66 | 55.75 | 87.31 | 55.91 | 80.16 |
| Gram-Net [25] | 50.62 | 50.54 | 59.43 | 90.96 | 51.99 | 78.01 | 53.08 | 82.31 | 53.41 | 82.46 | 52.18 | 77.37 | 53.45 | 76.94 |
| LNP [23] | 49.61 | 55.52 | 49.66 | 54.10 | 50.00 | 51.08 | 59.37 | 88.02 | 59.72 | 88.45 | 58.87 | 87.51 | 54.54 | 70.78 |
| UnivFD [32] | 79.79 | 90.86 | 85.02 | 94.07 | 65.33 | 78.21 | 79.29 | 91.16 | 79.90 | 91.01 | 81.18 | 92.16 | 78.42 | 89.58 |
| NPR [47] | 59.47 | 69.62 | 89.89 | 98.39 | 55.74 | 97.38 | 55.33 | 89.98 | 55.51 | 90.38 | 55.67 | 75.19 | 61.94 | 86.82 |
| **D³QE(ours)** | 70.43 | 83.98 | 88.89 | 96.36 | 61.21 | 75.29 | 83.33 | 94.10 | 83.37 | 93.32 | 84.43 | 94.52 | 78.61 | 89.60 |

to detect GAN and diffusion generated images. As shown in Tables 2 and 3, our method demonstrates robust generalization across different generative architectures.

**Performance on GANs.** In GAN evaluation, our method achieves an average accuracy of 83.73% and AP of 92.23%, surpassing all baseline methods. Notably, we attain high AP of 97.68% and 97.94% on ProGAN and GauGAN, respectively. Despite GANs lacking explicit discretization, our codebook-based framework effectively captures distributional anomalies in GAN-generated images. This effectiveness likely stems from the hierarchical upsampling structure in GAN which imposes low-dimensional manifold constraints. The structure results in concentrated distribution patterns similar to discretization effects, which our **D³AT** successfully identifies.

**Performance on Diffusion Models.** Our method demonstrates exceptional generalization to diffusion models, achieving an average accuracy of 78.61% and AP of 89.60%, comparable to state-of-the-art approaches. Detection accuracy reaches 83.33%, 83.37%, and 84.43% on Stable Diffusion v1.4, v1.5, and Wukong respectively. While diffusion models' step-wise denoising fundamentally differs from autoregressive discrete generation, their iterative nature induces structured patterns in feature distributions. Our method's success in identifying these patterns can be attributed to the distribution-aware mechanism's sensitivity to similar feature patterns and the semantic embedding module's effective capture of global semantic inconsistencies.

### 4.3.3. Robustness to Unseen Perturbations

Images in real-world scenarios often undergo unpredictable perturbations, posing significant challenges for generated content detection. To validate the robustness of our method, we evaluated it on ARForensics datasets under JPEG compression (with quality $q \in [60, 95]$) and center cropping (with crop factor $f \in [0.5, 0.9]$ and subsequent resizing). As shown in Figure 4, experiments show that traditional methods generally suffer significant performance degradation under pixel-level perturbations, primarily due to the destruction of local artifact features left by generative models. In contrast, our approach demonstrates superior adaptability through discrete distribution awareness and feature fusion. Under JPEG compression, our method maintains detection AP above 85% even when the quality factor drops to 60, consistently showing greater robustness than previous approaches. When facing severe cropping with $f = 0.5$, our method still preserves over 80% detection AP. These results validate the stability of our proposed multi-granularity feature fusion strategy across various perturbation conditions.

### 4.4. Ablation Studies

To validate the effectiveness of model components, we systematically analyzed how different module configurations and parameter settings affect detection performance.

In the Table 4(a) of module analysis, the base model (Model ① ) achieves 79.56% accuracy using only CLIP semantic features. Incorporating VQVAE residual features
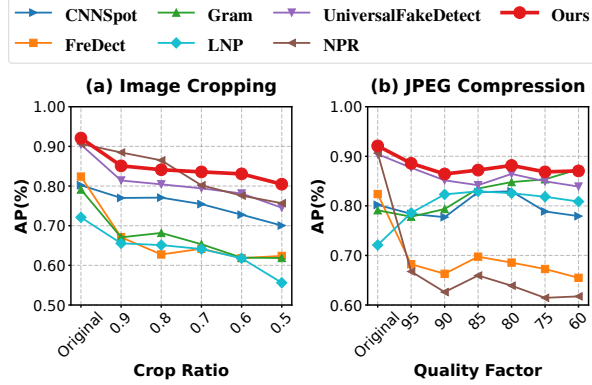
Figure 4. **Robustness Analysis.** Performance comparison under image cropping and JPEG compression. Our method maintains superior accuracy across different perturbation levels, demonstrating strong robustness against common image transformations.

Table 4. **Ablation studies on model components and parameter settings. (R: Residual, D: Discrete, V: Vanilla)**

(a) Module Analysis

| Model | Module Configuration | | | Acc. |
|---|---|---|---|---|
| | CLIP | Latent | Transformer | |
| ① | ✓ | ✗ | ✗ | 79.56 |
| ② | ✓ | R | ✗ | 79.92 |
| ③ | ✓ | D | V | 80.39 |
| ④ | ✓ | R | V | 80.72 |
| ⑤ | ✓ | R | $\mathbf{D^3AT}$ | 82.11 |

(b) Dimension Analysis

| $\mathbf{D^3AT}$ dim | Acc. |
|---|---|
| 128 | 80.83 |
| 256 | 81.57 |
| 384 | 80.95 |
| 512 | 82.11 |
| 1024 | 80.37 |

(Model ②) improves performance to 79.92%, demonstrating the effectiveness of discrete latent space modeling in enhancing feature representation. Comparing discrete features $z_q$ (Model ③) with residual features $\hat{z}$ (Model ④) reveals that residual quantization information more precisely captures distributional shifts in generated images, improving accuracy by 0.33%. Replacing the standard Transformer with our $\mathbf{D^3AT}$ (Model ⑤) further enhances detection accuracy to 82.11%, validating the effectiveness of our codebook statistical feature fusion mechanism.

In the Table 4(b) of dimension sensitivity tests, the $\mathbf{D^3AT}$ module achieves peak performance at 512 dimensions (82.11%). Lower dimensions restrict representational capacity (80.83% at 128 dimensions), while higher dimensions lead to overfitting (80.37% at 1024 dimensions), indicating significant model sensitivity to feature dimensionality. These results demonstrate that the synergistic design of residual quantization features and discrete distribution discrepancy-Aware self-attention attention mechanisms is crucial to improve detection performance.

### 4.5. Qualitative Results

To gain deeper insights into our model's intrinsic properties, we visualize the codebook activation distributions to reveal fundamental differences between real and generated samples. As illustrated in Figures 5(a) and (b), we analyze the normalized logarithmic activation frequencies of the first
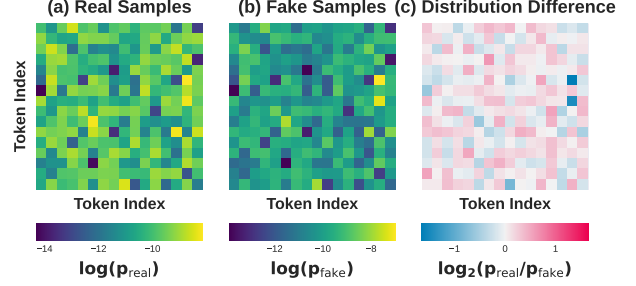


Figure 5. **Visualization of codebook activation patterns.** Heatmaps show normalized logarithmic activation frequencies of VQVAE codebook vectors for (a) real samples and (b) generated samples, with (c) their log-ratio difference. Real samples exhibit uniform activation patterns, while generated samples show significant polarization in high-frequency regions.

256 vectors in the VQVAE original codebook, based on the statistical distributions from our trained model.

The distributions reveal striking contrasts. Real samples exhibit balanced codebook utilization with uniform activation patterns. Generated samples, however, exhibit severe polarization. Their high-frequency codebook entries show anomalous peaks, with activation rates 3-5 times higher than real samples, while low-frequency regions show reduced coverage. The per-vector difference heatmap (Figure 5(c)) illustrates this disparity, revealing mode collapse characteristics in the discrete latent space. These distributional patterns reflect inherent limitations of autoregressive models in capturing complex real-world distributions, providing strong empirical support for our detection framework.

## 5. Conclusion

In this paper, we have proposed to learn Discrete Distribution Discrepancy-aware Quantization Error ($\mathbf{D^3QE}$) for autoregressive-generated image detection, which aims to exploit the distinctive patterns and the frequency distribution bias for various AR models. Further, we introduce a discrete distribution discrepancy-aware transformer to utilize dynamic codebook frequency statistics for combining semantic features with quantization error latent. Finally, we construct a comprehensive dataset covering mainstream visual AR models to evaluate our method, and experiments show that $\mathbf{D^3QE}$ achieves superior accuracy and strong generalization across different AR models, while maintaining robustness under various real-world perturbations.

## Acknowledgement

# References

[1] Midjourney. https://www.midjourney.com/home. 6

[2] Wukong. https://xihe.mindspore.cn/modelzoo/wukong, 2023. 6

[3] Andrew Brock, Jeff Donahue, and Karen Simonyan. Large scale GAN training for high fidelity natural image synthesis. In *ICLR*, 2019. 2, 5

[4] Tim Brooks, Bill Peebles, Connor Holmes, Will DePue, Yufei Guo, Li Jing, David Schnurr, Joe Taylor, Troy Luhman, Eric Luhman, Clarence Ng, Ricky Wang, and Aditya Ramesh. https://openai.com/index/sora/, 2024. Accessed: 2024-11. 2

[5] Xiaokang Chen, Zhiyu Wu, Xingchao Liu, Zizheng Pan, Wen Liu, Zhenda Xie, Xingkai Yu, and Chong Ruan. Janus-pro: Unified multimodal understanding and generation with data and model scaling. *arXiv preprint arXiv:2501.17811*, 2025. 5

[6] Yunjey Choi, Minje Choi, Munyoung Kim, Jung-Woo Ha, Sunghun Kim, and Jaegul Choo. StarGAN: Unified generative adversarial networks for multi-domain image-to-image translation. In *CVPR*, 2018. 5

[7] Casey Chu, Andrey Zhmoginov, and Mark Sandler. CycleGAN, a master of steganography. *arXiv preprint arXiv:1712.02950*, 2017. 5

[8] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *CVPR*, 2009. 5

[9] Prafulla Dhariwal and Alexander Nichol. Diffusion models beat GANs on image synthesis. In *NeurIPS*, 2021. 2, 6

[10] Patrick Esser, Robin Rombach, and Bjorn Ommer. Taming transformers for high-resolution image synthesis. In *CVPR*, 2021. 2, 3

[11] Joel Frank, Thorsten Eisenhofer, Lea Schönherr, Asja Fischer, Dorothea Kolossa, and Thorsten Holz. Leveraging frequency analysis for deep fake image recognition. In *ICML*, 2020. 2, 6, 7

[12] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *NeurIPS*, 2014. 1, 2

[13] Jian Han, Jinlai Liu, Yi Jiang, Bin Yan, Yuqi Zhang, Zehuan Yuan, Bingyue Peng, and Xiaobing Liu. Infinity: Scaling bitwise autoregressive modeling for high-resolution image synthesis. In *CVPR*, 2025. 1, 2, 5

[14] Irina Higgins, Loïc Matthey, Arka Pal, Christopher P. Burgess, Xavier Glorot, Matthew M. Botvinick, Shakir Mohamed, and Alexander Lerchner. beta-vae: Learning basic visual concepts with a constrained variational framework. In *ICLR*, 2017. 2

[15] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *NeurIPS*, 2020. 2

[16] Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. The curious case of neural text degeneration. In *ICLR*, 2020. 4

[17] Wenyi Hong, Ming Ding, Wendi Zheng, Xinghan Liu, and Jie Tang. Cogvideo: Large-scale pretraining for text-to-video generation via transformers. In *ICLR*, 2023. 2

[18] Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. Progressive growing of GANs for improved quality, stability, and variation. In *ICLR*, 2018. 2, 5

[19] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *CVPR*, 2019. 2, 5

[20] Diederik P. Kingma and Max Welling. Auto-encoding variational bayes. In *ICLR*, 2014. 1, 2

[21] Doyup Lee, Chiheon Kim, Saehoon Kim, Minsu Cho, and Wook-Shin Han. Autoregressive image generation using residual quantization. In *CVPR*, 2022. 2, 3

[22] Tianhong Li, Yonglong Tian, He Li, Mingyang Deng, and Kaiming He. Autoregressive image generation without vector quantization. In *NeurIPS*, 2024. 4

[23] Bo Liu, Fan Yang, Xiuli Bi, Bin Xiao, Weisheng Li, and Xinbo Gao. Detecting generated images by real images. In *ECCV*, 2022. 6, 7

[24] Huan Liu, Zichang Tan, Chuangchuang Tan, Yunchao Wei, Jingdong Wang, and Yao Zhao. Forgery-aware adaptive transformer for generalizable synthetic image detection. In *CVPR*, 2024. 2

[25] Zhengzhe Liu, Xiaojuan Qi, and Philip HS Torr. Global texture enhancement for fake face detection in the wild. In *CVPR*, 2020. 6, 7

[26] Zhuoyan Luo, Fengyuan Shi, Yixiao Ge, Yujiu Yang, Limin Wang, and Ying Shan. Open-magvit2: An open-source project toward democratizing auto-regressive visual generation. *arXiv preprint arXiv:2409.04410*, 2024. 5

[27] Scott McCloskey and Michael Albright. Detecting GAN-generated imagery using color cues. *arXiv preprint arXiv:1812.08247*, 2018. 2

[28] Scott McCloskey and Michael Albright. Detecting GAN-generated imagery using saturation cues. In *ICIP*, 2019.

[29] Lakshmanan Nataraj, Tajuddin Manhar Mohammed, Shivkumar Chandrasekaran, Arjuna Flenner, Jawadul H Bappy, Amit K Roy-Chowdhury, and BS Manjunath. Detecting GAN generated fake images using co-occurrence matrices. *arXiv preprint arXiv:1903.06836*, 2019. 2

[30] Alexander Quinn Nichol and Prafulla Dhariwal. Improved denoising diffusion probabilistic models. In *ICML*, 2021. 2

[31] Alexander Quinn Nichol, Prafulla Dhariwal, Aditya Ramesh, Pranav Shyam, Pamela Mishkin, Bob Mcgrew, Ilya Sutskever, and Mark Chen. Glide: Towards photorealistic image generation and editing with text-guided diffusion models. In *ICML*, 2022. 6

[32] Utkarsh Ojha, Yuheng Li, and Yong Jae Lee. Towards universal fake image detectors that generalize across generative models. In *CVPR*, 2023. 2, 6, 7

[33] Taesung Park, Ming-Yu Liu, Ting-Chun Wang, and Jun-Yan Zhu. GauGAN: semantic image synthesis with spatially adaptive normalization. In *ACM SIGGRAPH*, 2019. 5

[34] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An

imperative style, high-performance deep learning library. In *NeurIPS*, 2019. 6

[35] William Peebles and Saining Xie. Scalable diffusion models with transformers. In *ICCV*, 2023. 2

[36] Yuyang Qian, Guojun Yin, Lu Sheng, Zixuan Chen, and Jing Shao. Thinking in frequency: Face forgery detection by mining frequency-aware clues. In *ECCV*, 2020. 2

[37] Liao Qu, Huichao Zhang, Yiheng Liu, Xu Wang, Yi Jiang, Yiming Gao, Hu Ye, Daniel K Du, Zehuan Yuan, and Xinglong Wu. Tokenflow: Unified image tokenizer for multimodal understanding and generation. In *CVPR*, 2025. 2

[38] Alec Radford, Luke Metz, and Soumith Chintala. Unsupervised representation learning with deep convolutional generative adversarial networks. In *ICLR*, 2016. 2

[39] Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. Zero-shot text-to-image generation. In *ICML*, 2021. 4

[40] Jonas Ricker, Denis Lukovnikov, and Asja Fischer. AEROBLADE: training-free detection of latent diffusion images using autoencoder reconstruction error. In *CVPR*, 2024. 2

[41] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *CVPR*, 2022. 1, 2, 6

[42] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. In *ICLR*, 2021. 1

[43] Yang Song, Jascha Sohl-Dickstein, Diederik P. Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. In *ICLR*, 2021. 2

[44] Alessandro Sordoni, Michel Galley, Michael Auli, Chris Brockett, Yangfeng Ji, Margaret Mitchell, Jian-Yun Nie, Jianfeng Gao, and Bill Dolan. A neural network approach to context-sensitive generation of conversational responses. In *HLT-NAACL*, 2015. 4

[45] Peize Sun, Yi Jiang, Shoufa Chen, Shilong Zhang, Bingyue Peng, Ping Luo, and Zehuan Yuan. Autoregressive model beats diffusion: Llama for scalable image generation. *arXiv preprint arXiv:2406.06525*, 2024. 1, 2, 3, 5

[46] Ilya Sutskever, Oriol Vinyals, and Quoc V. Le. Sequence to sequence learning with neural networks. In *NeurIPS*, 2014. 4

[47] Chuangchuang Tan, Huan Liu, Yao Zhao, Shikui Wei, Guanghua Gu, Ping Liu, and Yunchao Wei. Rethinking the up-sampling operations in cnn-based generative network for generalizable deepfake detection. In *CVPR*, 2024. 2, 6, 7

[48] Keyu Tian, Yi Jiang, Zehuan Yuan, Bingyue Peng, and Liwei Wang. Visual autoregressive modeling: Scalable image generation via next-scale prediction. In *NeurIPS*, 2025. 1, 2, 3, 5

[49] Arash Vahdat and Jan Kautz. NVAE: A deep hierarchical variational autoencoder. In *NeurIPS*, 2020. 2

[50] Aäron van den Oord, Nal Kalchbrenner, Lasse Espeholt, Koray Kavukcuoglu, Oriol Vinyals, and Alex Graves. Conditional image generation with pixelcnn decoders. In *NeurIPS*, 2016. 2

[51] Aäron van den Oord, Nal Kalchbrenner, and Koray Kavukcuoglu. Pixel recurrent neural networks. In *ICML*, 2016. 2

[52] Aäron van den Oord, Oriol Vinyals, and Koray Kavukcuoglu. Neural discrete representation learning. In *NeurIPS*, 2017. 2, 3

[53] Yuri Viazovetskyi, Vladimir Ivashkin, and Evgeny Kashin. StyleGAN2 distillation for feed-forward image manipulation. In *ECCV*, 2020. 5

[54] Anton Voronov, Denis Kuznedelev, Mikhail Khoroshikh, Valentin Khrulkov, and Dmitry Baranchuk. Switti: Designing scale-wise transformers for text-to-image synthesis. *arXiv preprint arXiv:2412.01819*, 2024. 5

[55] Sheng-Yu Wang, Oliver Wang, Richard Zhang, Andrew Owens, and Alexei A Efros. Cnn-generated images are surprisingly easy to spot... for now. In *CVPR*, 2020. 2, 5, 6, 7

[56] Zhendong Wang, Jianmin Bao, Wengang Zhou, Weilun Wang, Hezhen Hu, Hong Chen, and Houqiang Li. Dire for diffusion-generated image detection. In *ICCV*, 2023. 2

[57] Jing Xiong, Gongye Liu, Lun Huang, Chengyue Wu, Taiqiang Wu, Yao Mu, Yuan Yao, Hui Shen, Zhongwei Wan, Jinfa Huang, et al. Autoregressive models in vision: A survey. *arXiv preprint arXiv:2411.05902*, 2024. 3

[58] Shilin Yan, Ouxiang Li, Jiayin Cai, Yanbin Hao, Xiaolong Jiang, Yao Hu, and Weidi Xie. A sanity check for ai-generated image detection. In *ICLR*, 2025. 2

[59] Jiahui Yu, Yuanzhong Xu, Jing Yu Koh, Thang Luong, Gunjan Baid, Zirui Wang, Vijay Vasudevan, Alexander Ku, Yinfei Yang, Burcu Karagol Ayan, Ben Hutchinson, Wei Han, Zarana Parekh, Xin Li, Han Zhang, Jason Baldridge, and Yonghui Wu. Scaling autoregressive models for content-rich text-to-image generation. *TMLR*, 2022. 4

[60] Lijun Yu, Yong Cheng, Kihyuk Sohn, José Lezama, Han Zhang, Huiwen Chang, Alexander G Hauptmann, Ming-Hsuan Yang, Yuan Hao, Irfan Essa, et al. Magvit: Masked generative video transformer. In *CVPR*, 2023. 2

[61] Qihang Yu, Ju He, Xueqing Deng, Xiaohui Shen, and Liang-Chieh Chen. Randomized autoregressive visual generation. *arXiv preprint arXiv:2411.00776*, 2024. 5

[62] Mingjian Zhu, Hanting Chen, Qiangyu Yan, Xudong Huang, Guanyu Lin, Wei Li, Zhijun Tu, Hailin Hu, Jie Hu, and Yunhe Wang. Genimage: A million-scale benchmark for detecting ai-generated image. In *NeurIPS*, 2023. 6, 7