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

Tsinghua University

ICCV HONOLULU HAWAII OCT 19-23, 2025

Introduction:

- **Challenge** Autoregressive (AR) models hide forgery artifacts in the discrete latent space, not pixels, challenging existing detectors.
- **Key Insight: Distribution Discrepancy** Real and AR-generated images show a stark Discrete Distribution Discrepancy. Real data has a long-tail token distribution; fakes show concentrated high-frequency usage.

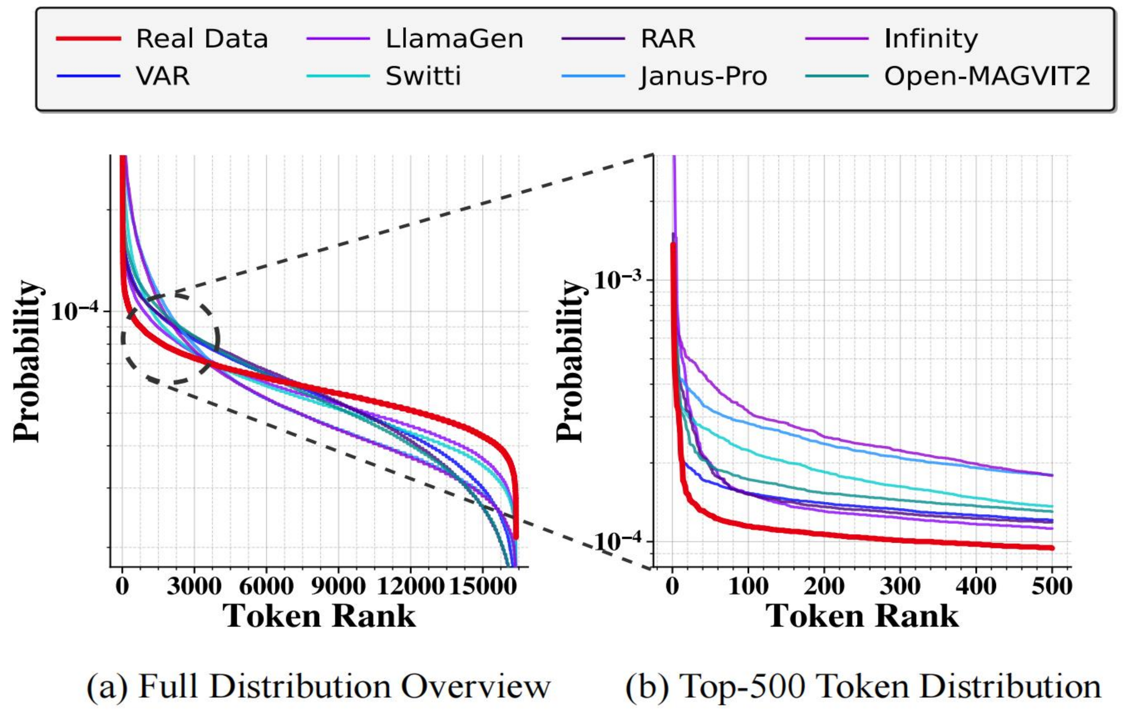


Figure 1: Token Distribution Bias. Real data: long-tail token usage. Fakes: concentrated, high-frequency bias.

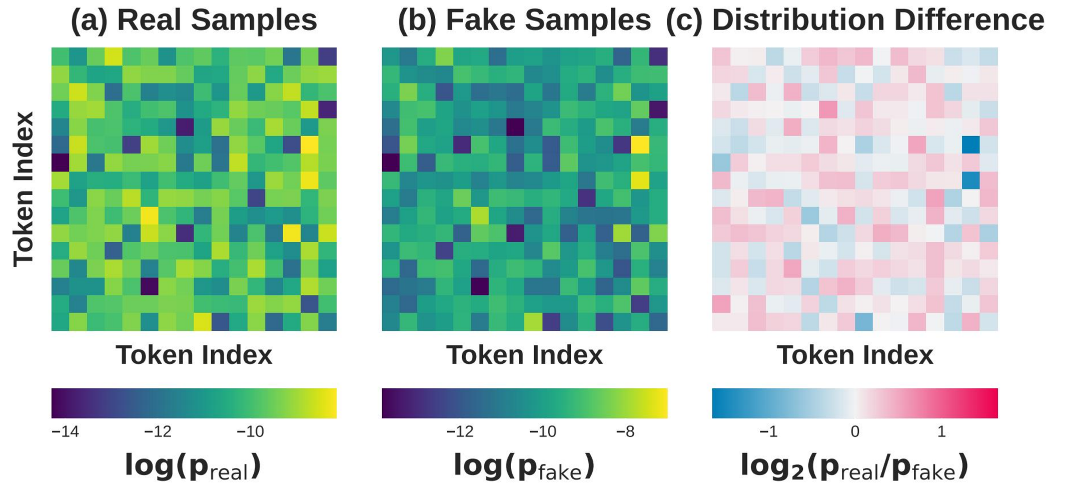
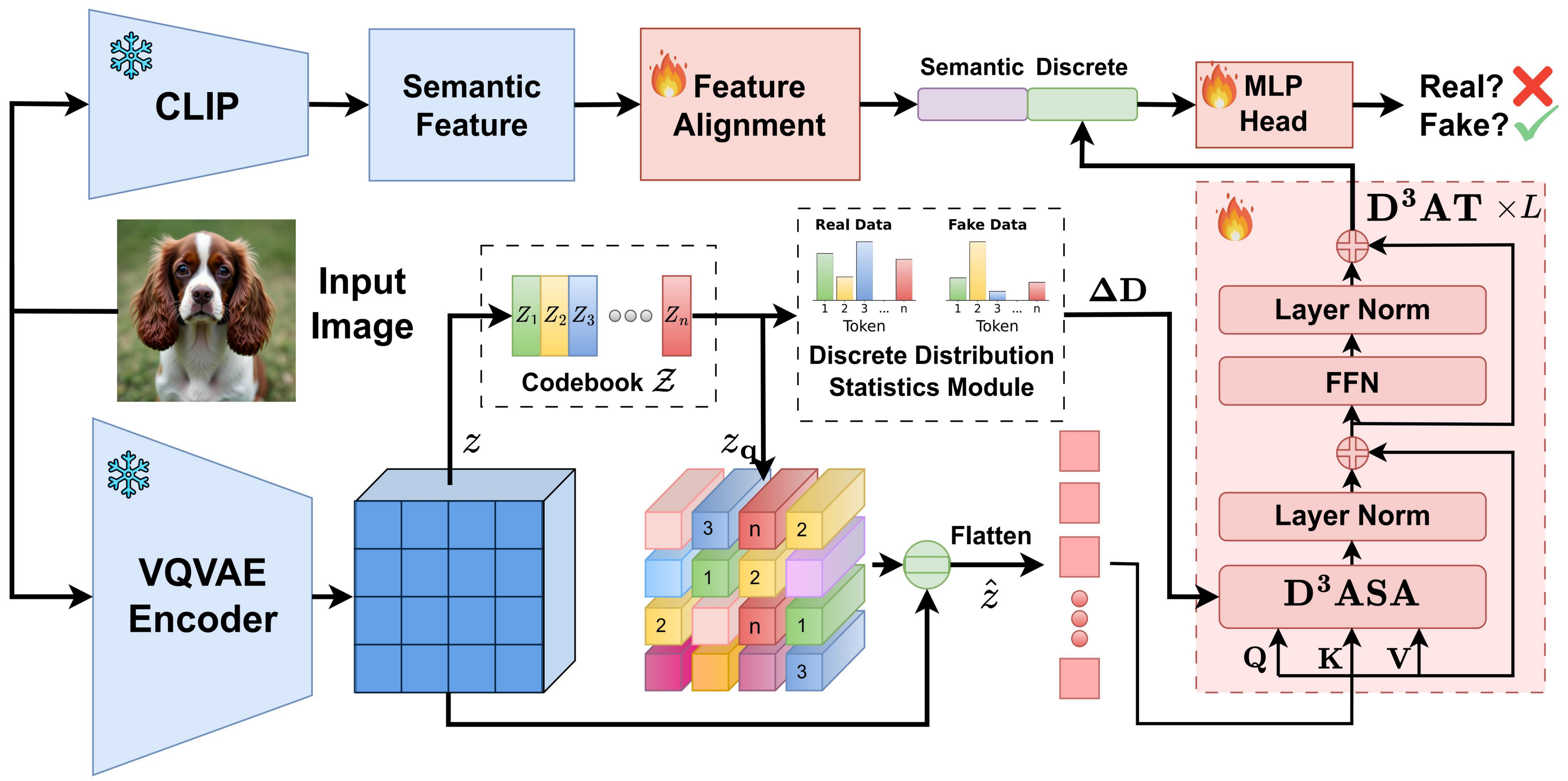


Figure 2: Codebook Activation Difference. Heatmaps: Real (a) balanced codebook. Fakes (b) polarized hotspots. (c) Core artifact for detection.

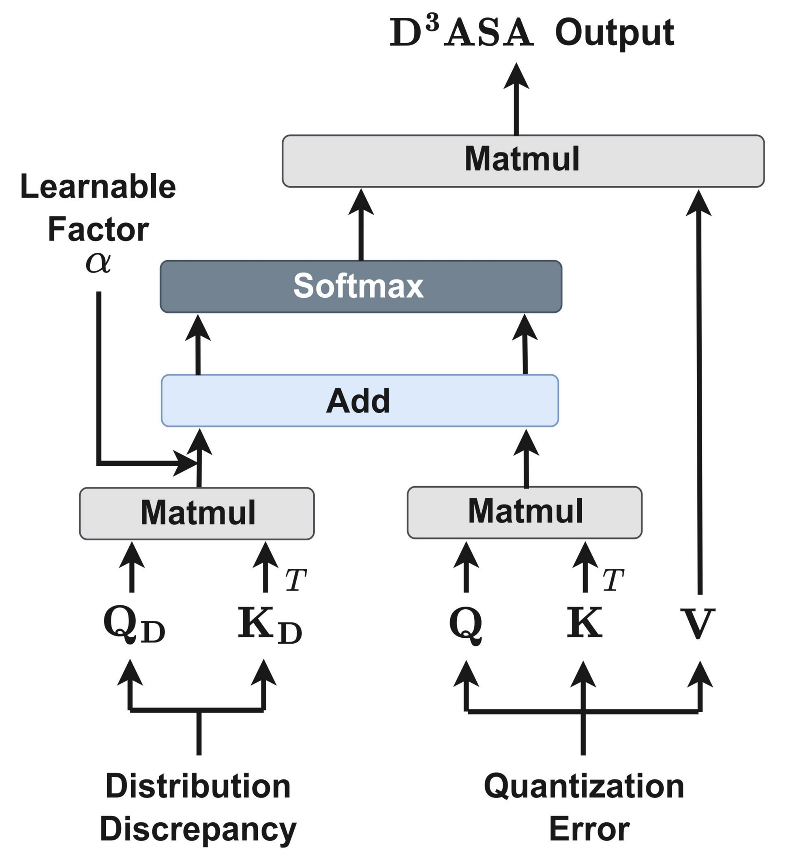
Our Contributions:

- **D³QE:** A framework analyzing codebook distribution bias & quantization error.
- **D³AT:** A transformer integrating distribution statistics into attention.
- **ARForensics:** The first large-scale benchmark for AR detection (7 models)

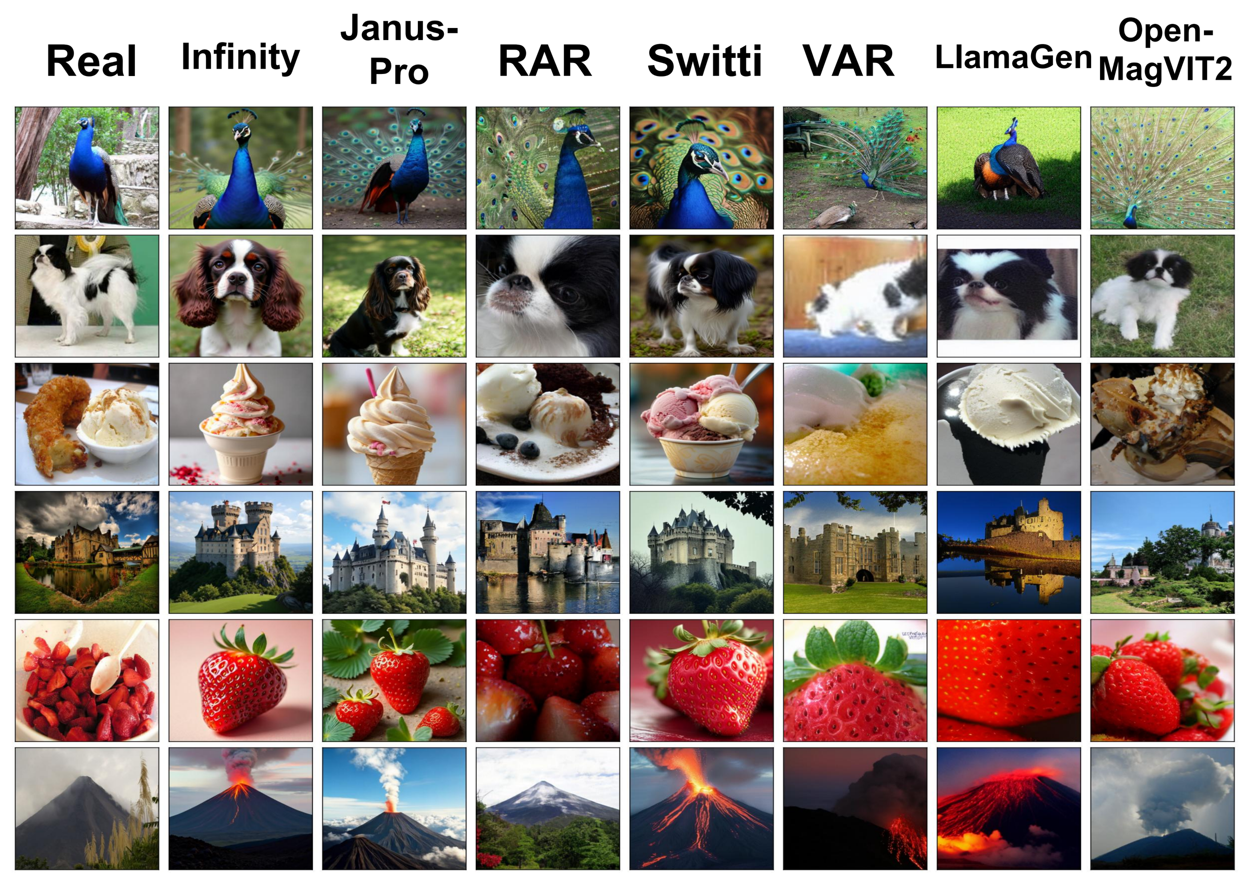
Method:



- The D³QE pipeline fuses discrete features (from VQVAE) and semantic features (from CLIP). Our D³AT and D³ASA module processes quantization error, guided by distribution discrepancy, for robust classification.



- The core D³ASA module enhances self-attention by incorporating global codebook usage statistics.



- **ARForensics:** The first large-scale benchmark for **AR image detection**. 152K diverse images from 7 mainstream AR models, providing a robust testbed for next-gen forgery forensics.

Experiments:

- Performance comparison on **ARForensics** dataset.

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
D³QE(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

- Performance comparison on GAN-based synthesis using **ForenSynths** test set.

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

- Performance comparison on Diffusion-based generation using **GenImage** test set.

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

Conclusion:

- **D³QE**: First leverages core discrete distribution bias + quantization error for AR detection.
- Dataset: Release **ARForensics** for future forgery research.
- Strong performance on ARForensics and cross-paradigm generalization.

Arxiv **Github** **Project**