CADMA

Capacity-Aware Coverage Decomposition for Generative Model Assessment

NeurIPS 2026
NEURAL INFORMATION PROCESSING SYSTEMS

Our task: evaluating the **diversity** of generative AI outputs

Debin Meng · Zheng Gao · Ziquan Liu · Yanran Li
Ioannis Patras · Georgios Tzimiropoulos

Queen Mary University of London

University of Bedfordshire

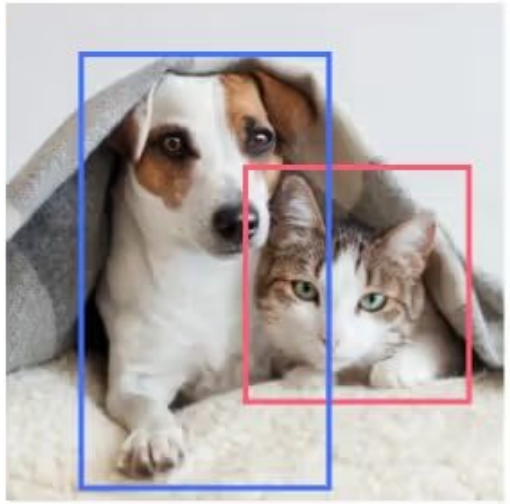
Output diversity is central to generative AI

Classification



Cat

Object Detection



Cat, Dog

Instance Segmentation



Cat, Dog

Consistent predictions

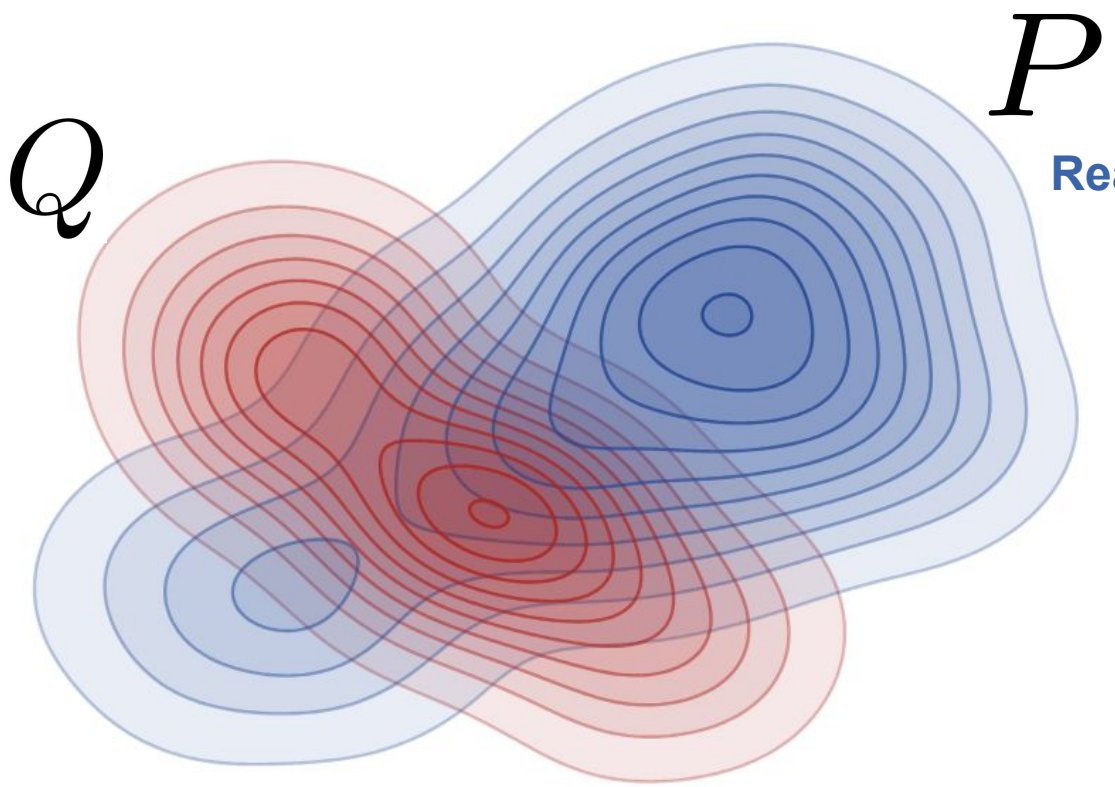
Image Generation



Dog

Diverse outputs

Recall and Coverage assess output diversity



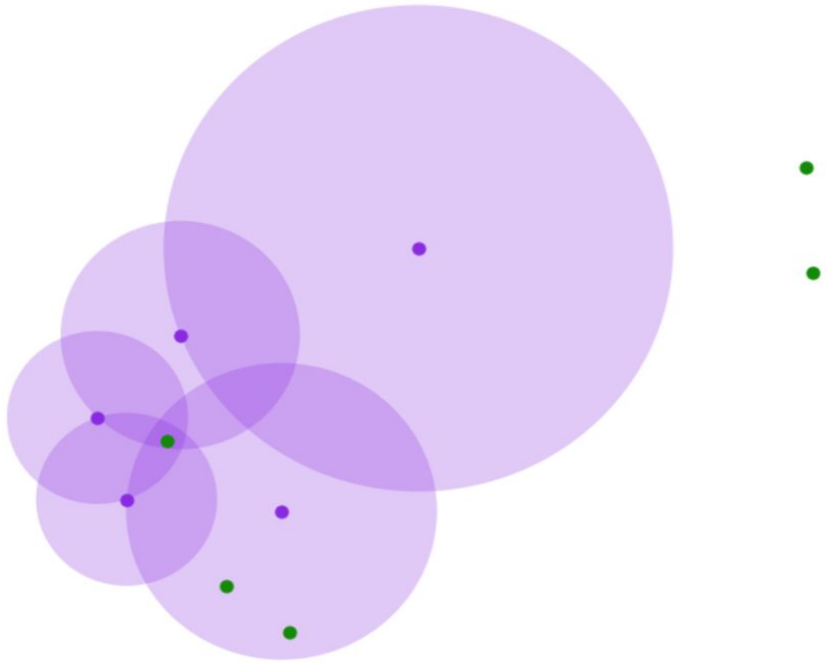
Real distribution

Generated distribution

$$\text{Recall}(P, Q) := \mathbb{E}_{X \sim P} \left[\mathbf{1}\{X \in \text{supp}(Q)\} \right], \tag{1}$$

How much of the real distribution is covered by generated samples?

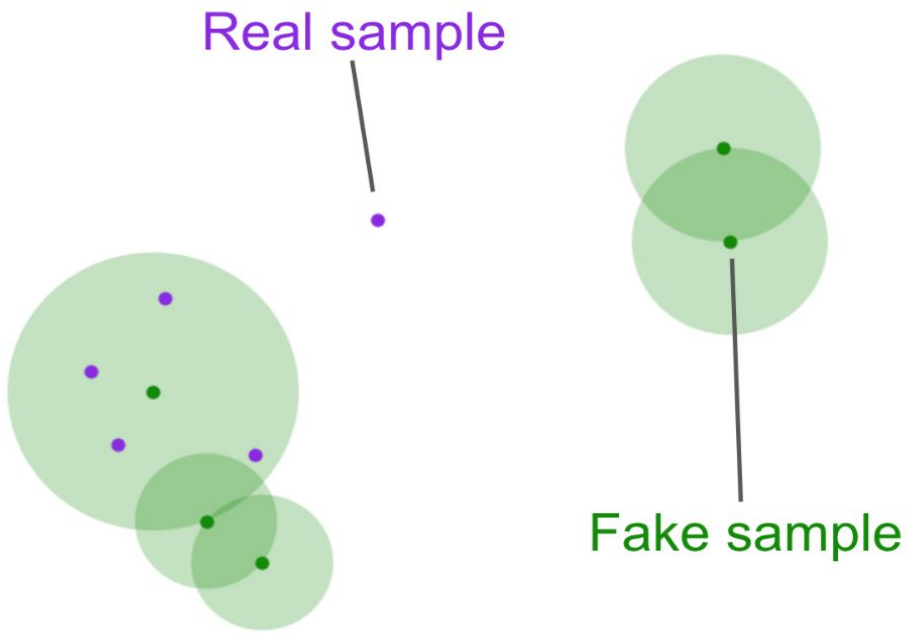
Limitation 1: geometry-based coverage



$$\text{Coverage} = \frac{1+1+1+1}{1+1+1+1+1} = 80\%$$

$$\text{Capacity-Coverage} = \frac{1+1}{1+1+1+1+1} = 40\%$$

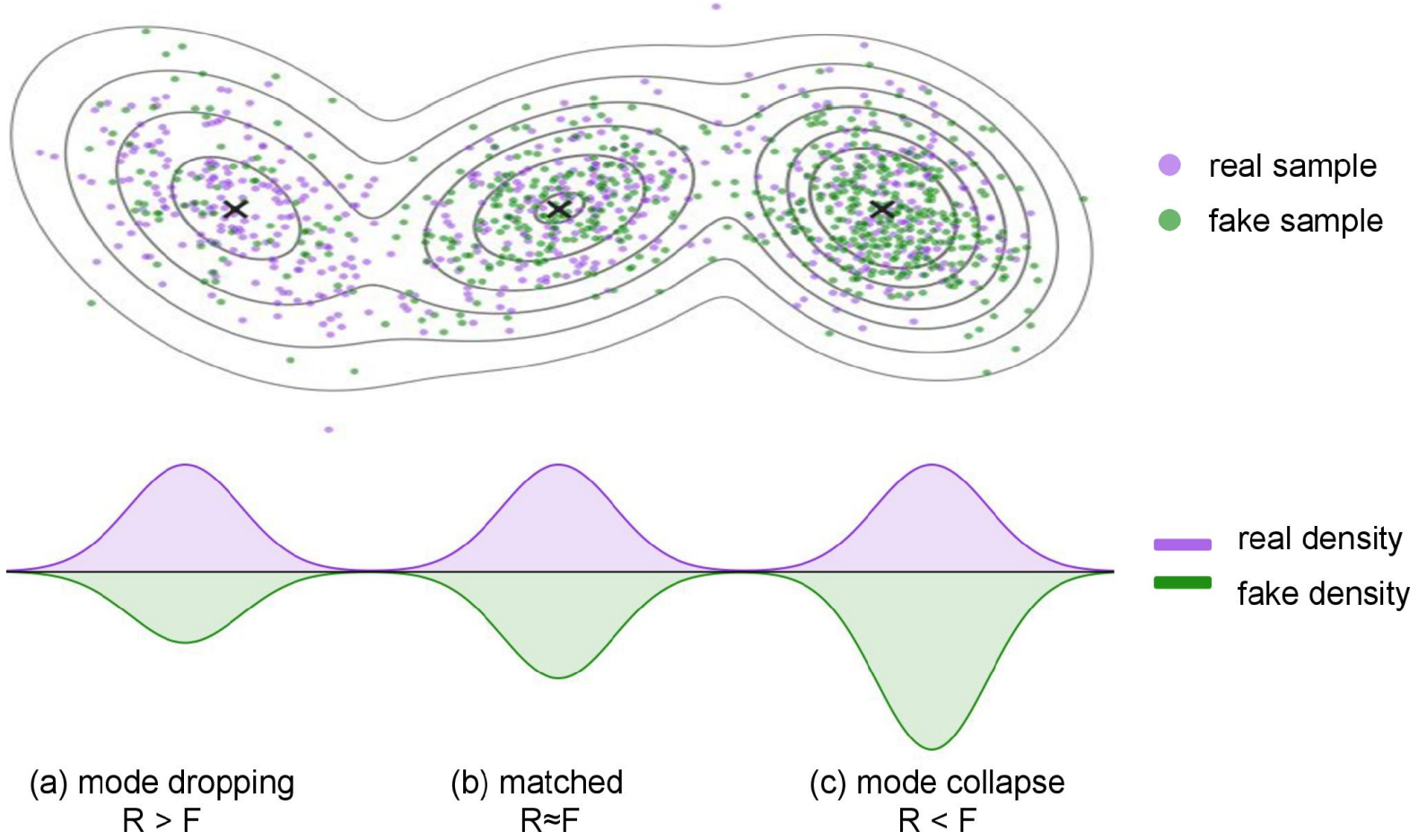
(a) Coverage versus Capacity-Coverage



$$\text{I-Recall} = \frac{1+1+1+1}{1+1+1+1+1} = 80\%$$

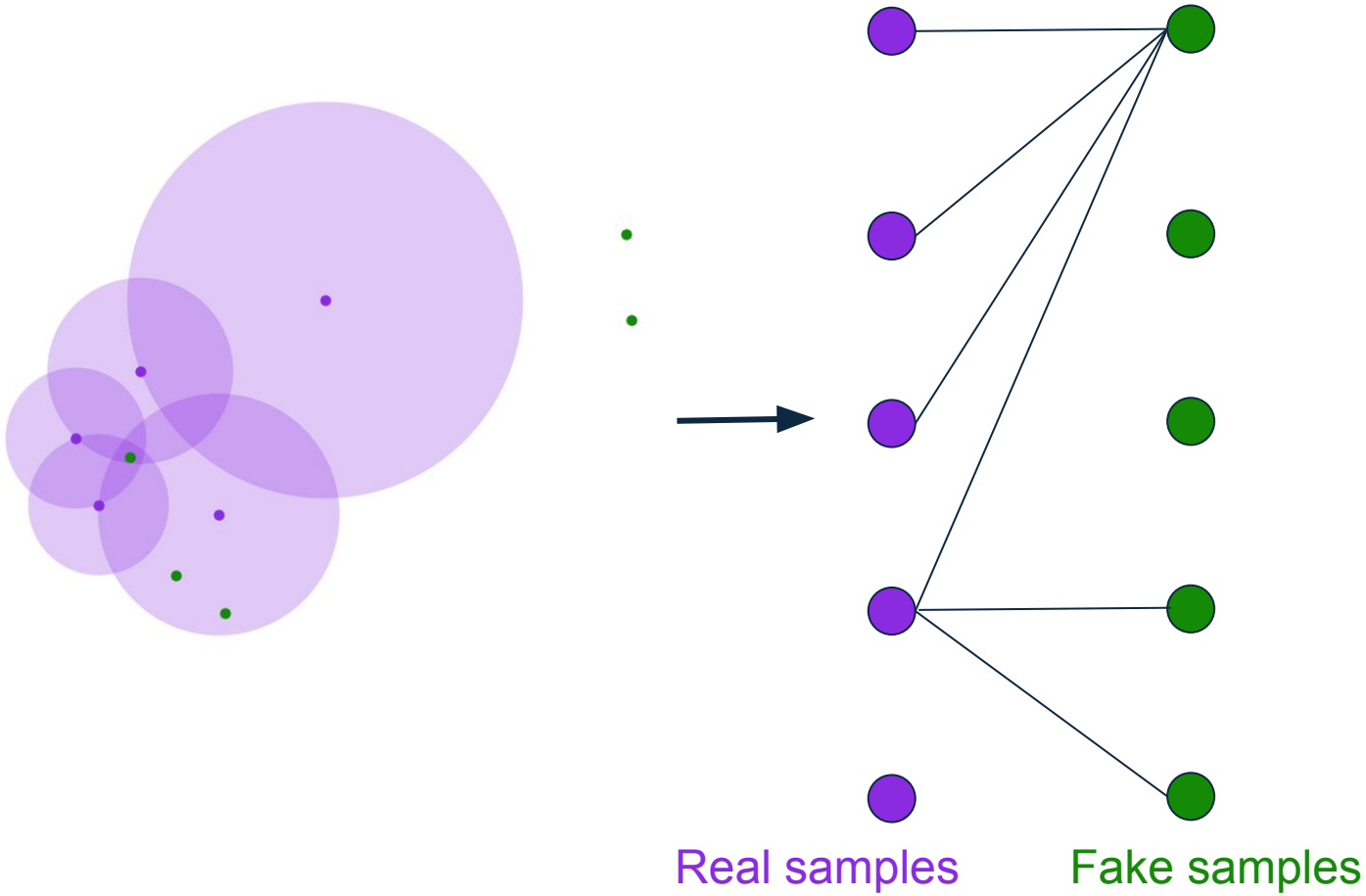
(b) I-Recall

Limitation 2: one score hides different failures



Mode dropping and mode collapse require separate diagnostics.

Geometry-based coverage as a bipartite graph

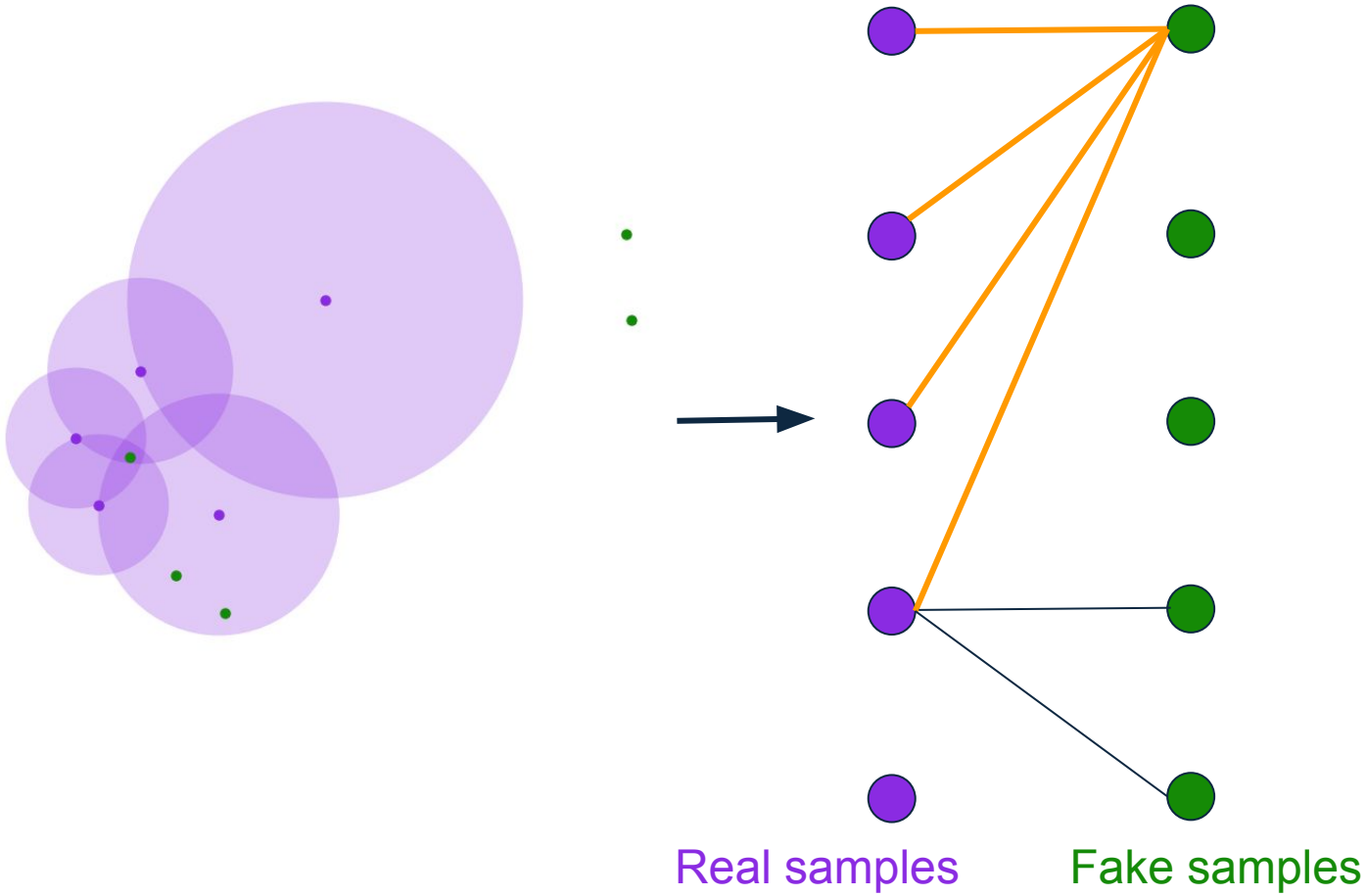


Represent coverage as a bipartite graph

Edges indicate generated samples inside real-sample neighbourhoods.

$$\text{Coverage} = \frac{1+1+1+1}{1+1+1+1+1} = 80\%$$

Why geometry can overestimate coverage

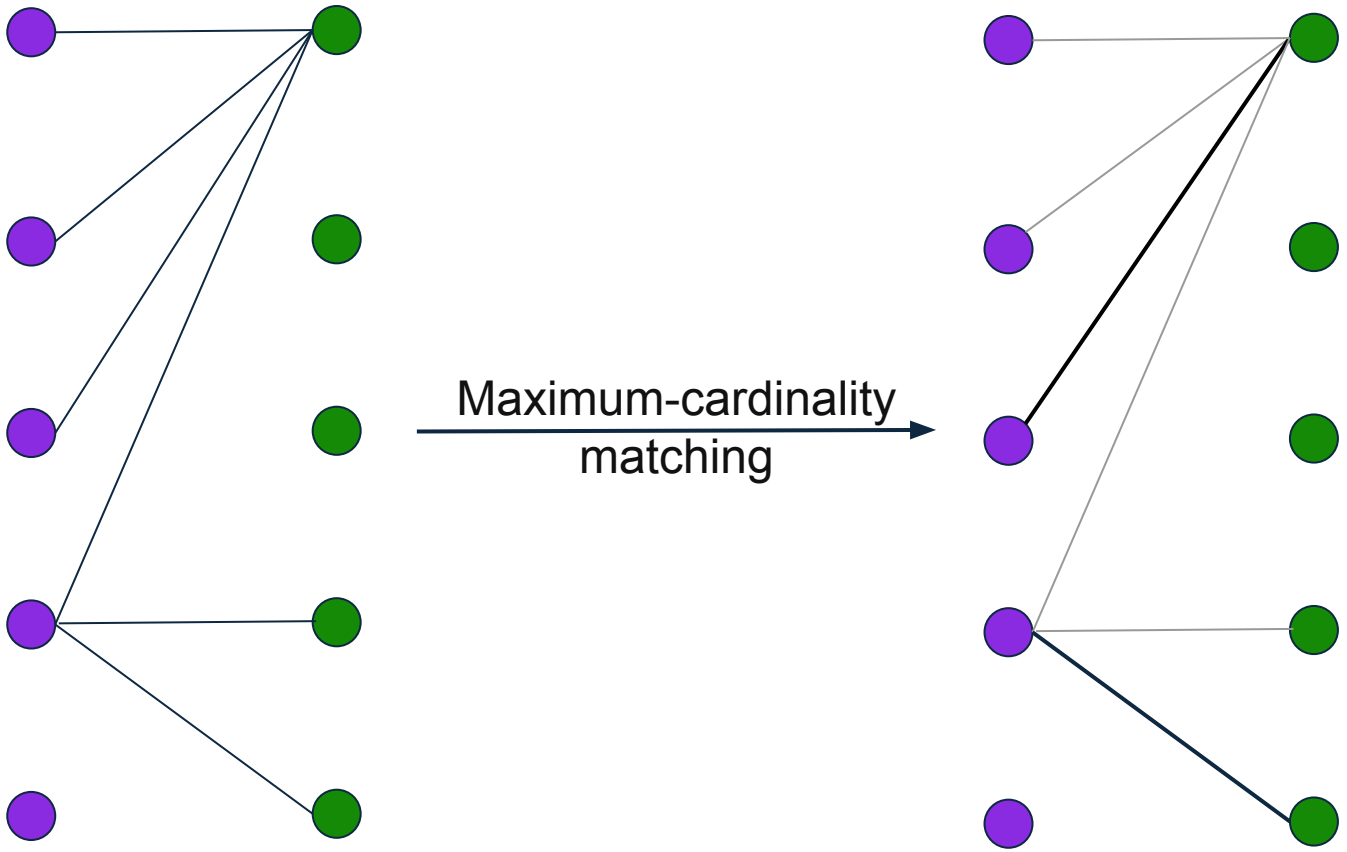


$$Coverage = \frac{N^{cov}}{N} = \frac{4}{5} = 0.8$$

A single generated sample can cover several real samples.

Removing the other four generated samples leaves Coverage unchanged.

Local capacity matching



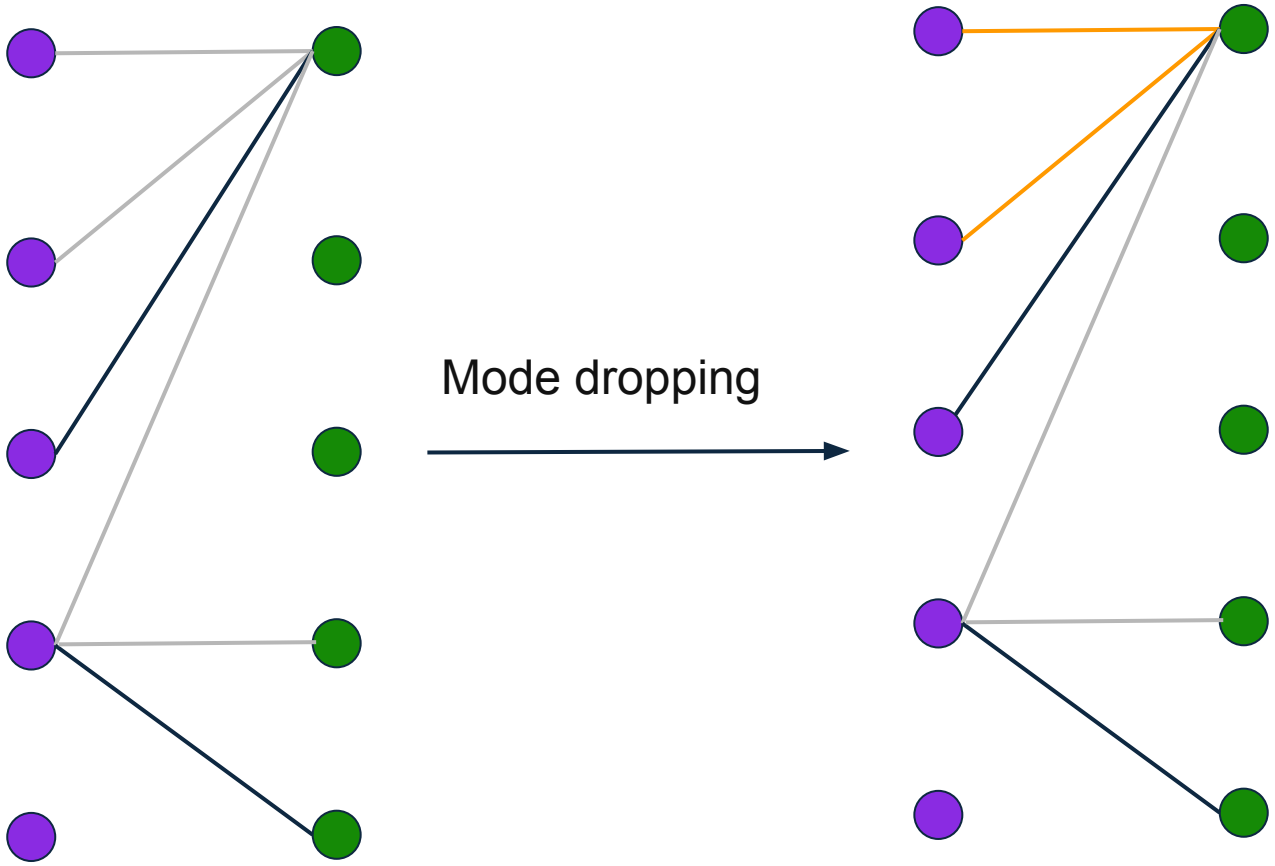
$$Coverage = \frac{N^{Cov}}{N} = \frac{4}{5} = 0.8$$

$$C_{cap} = \frac{N^{mat}}{N} = \frac{2}{5} = 0.4$$

Each sample can be matched at most once, yielding capacity-aware coverage.

Real samples Fake samples

Mode-dropping score



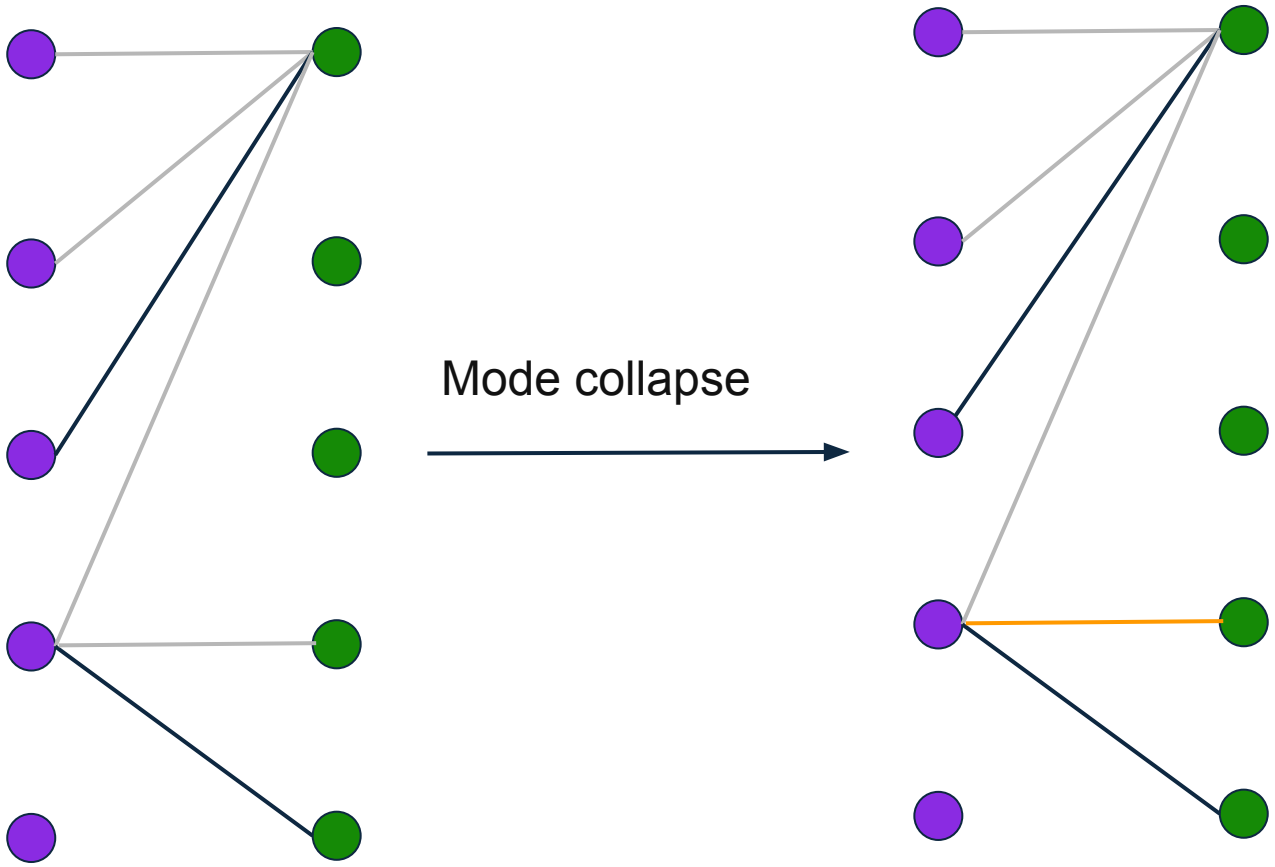
Real samples Fake samples

$$D_{drop} = Covearge - C_{cap}$$

$$= \frac{N^{cov} - N^{mat}}{N} = \frac{2}{5}$$

Coverage is overly optimistic because it includes the D_{drop} component.

Mode-collapse score



Real samples Fake samples

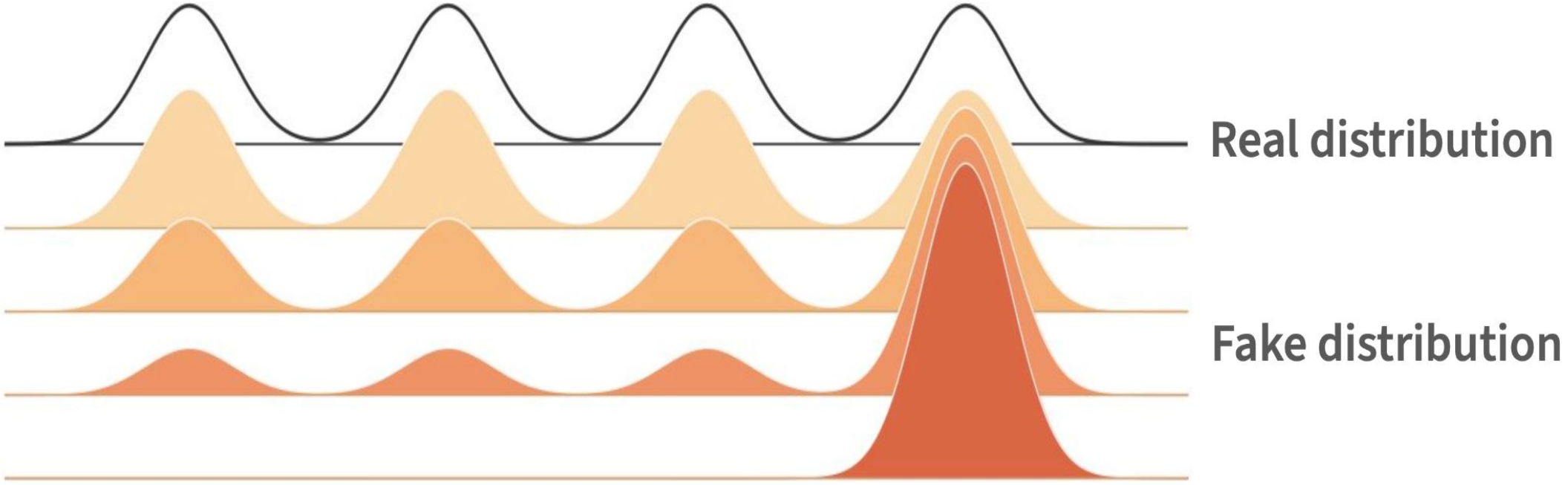
$$D_{collapse} = \frac{M^{cov} - N^{mat}}{N}$$

$$= \frac{3-2}{5} = \frac{1}{5}$$

Locally covered generated samples without distinct real counterparts

$D_{collapse}$

Controlled mode dropping and mode collapse

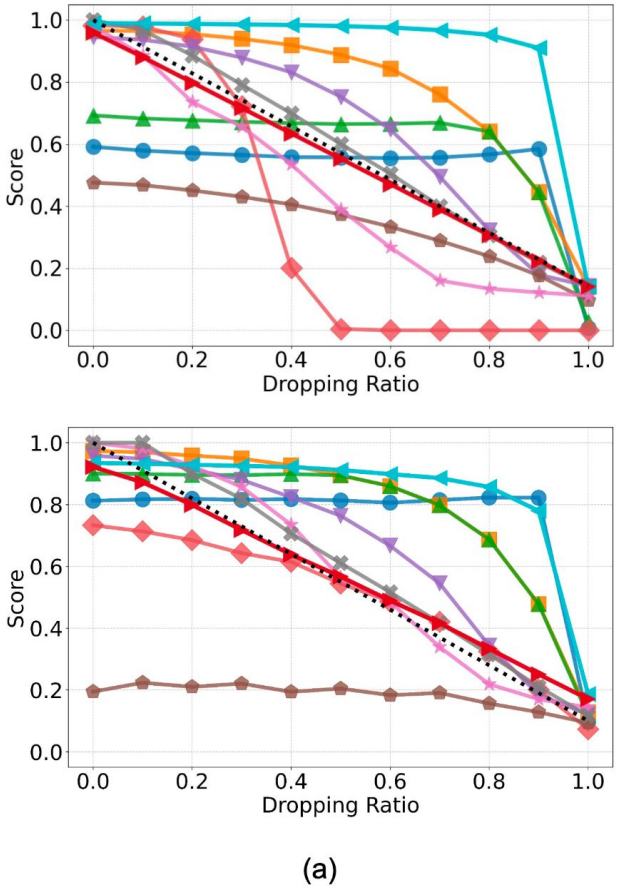


Gradually move generated samples from several modes into one remaining mode, creating under-representation and local concentration.

Mode Collapse and Dropping experiments on toy and real data

Gaussian

CIFAR-10

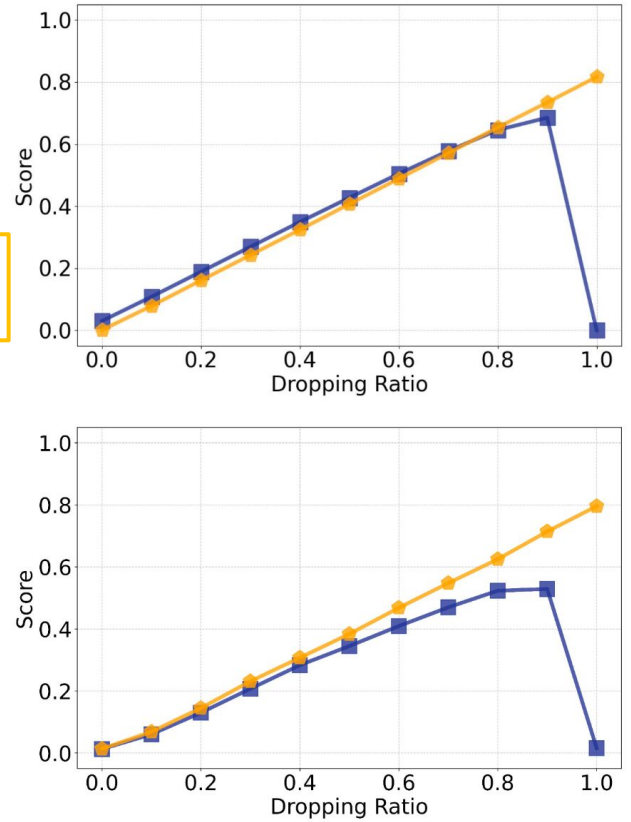


(a)

- Coverage$_{our}$
- Capacity Coverage ($\mathcal{C}_{cap}$)
- Recall
- Coverage
- symRecall
- P-recall
- RecallCover
- β-Recall
- TopR
- ClippedCoverage

About 25K real samples and 25K generated samples per setting

- Mode Dropping Score ($\mathcal{D}_{drop}$)
- Mode Collapse Score ($\mathcal{D}_{collapse}$)



(b)

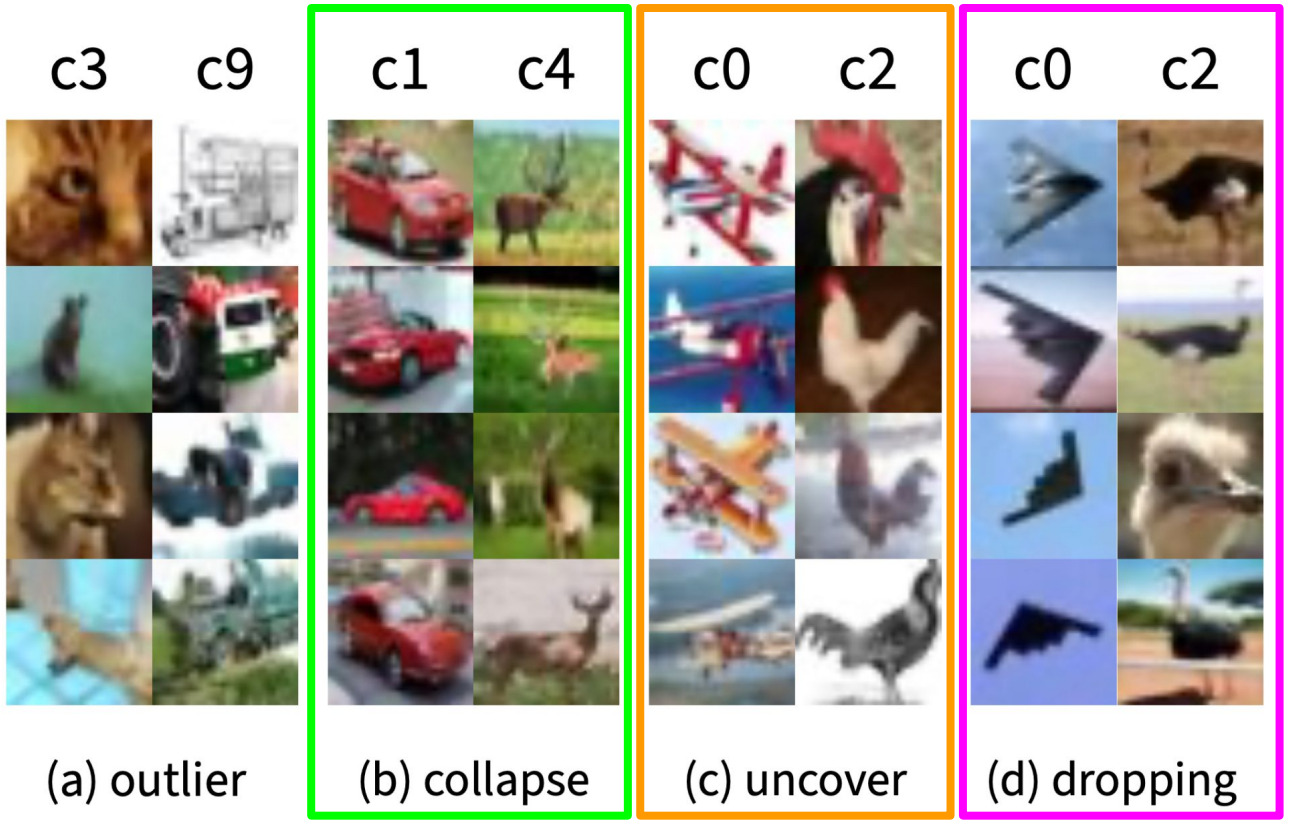
Controlled benchmarks: error to reference trends

Table 1: Mean absolute error. Best and second-best results are bolded and underlined.

Method	Gaussian		CIFAR-10		Avg.
	Mode Collapse (↓)	Outlier Pollution (↓)	Mode Collapse (↓)	Outlier Pollution (↓)	
Recall Kynkäänniemi et al. (2019)	0.206	0.214	0.252	0.179	0.213
Coverage Naeem et al. (2020)	0.204	0.203	0.239	0.189	0.209
symRecall Khayatkhoei and AbdAlmageed (2023)	0.180	0.183	0.227	0.175	0.191
P-recall Park and Kim (2023)	0.255	0.283	0.080	0.079	0.174
RecallCover Cheema and Urner (2023)	0.088	0.085	0.108	0.090	0.093
β-Recall Alaa et al. (2022)	0.230	0.260	0.368	0.560	0.355
TopR Kim et al. (2023)	0.119	0.127	0.054	0.240	0.135
ClippedCoverage Salvy et al. (2026)	<u>0.025</u>	**0.020**	<u>0.051</u>	<u>0.045</u>	<u>0.035</u>
CADMA ($\mathcal{C}_{\text{cap}}$)	**0.019**	<u>0.023</u>	**0.039**	**0.025**	**0.027**

Paper Table 1 · Controlled Gaussian and CIFAR-10 settings

Sample-level diagnosis on CIFAR-10



Mode collapse: the model over-produces red cars within the “car” class.

Uncovered: rooster-like birds have no generated counterparts.

Mode dropping: bat-like aircraft are under-represented within the “airplane” class.

Time complexity

Under the sparse-graph assumption:

$$T_{\mathrm{CADMA}} = O(dn^2 + kn^{3/2}).$$

Neighbourhood construction time

Matching time

For fixed k and d, the relative matching cost vanishes as n grows.

$$\rho_{\mathrm{HK}}^{\mathrm{time}} = \frac{T_{\mathrm{matching}}}{T_{\mathrm{neighborhood}}} = O(\frac{kn^{3/2}}{dn^2}) = O(\frac{k}{d\sqrt{n}})$$

Runtime across 1K–50K samples

Runtime in seconds

Table 6: Wall-clock runtime of CADMA and baseline metrics, measured in seconds using precomputed features. N denotes the number of samples in each of the real and generated sets.

N	CADMA (s)	T_{neigh} (s)	T_{add} (s)	$T_{\text{add}}/T_{\text{CADMA}}$	ClippedCoverage (s)	Coverage (s)
1K	0.225	0.223427	0.001189	0.529%	0.205	0.158
2.5K	0.662	0.658441	0.003792	0.572%	0.795	0.625
5K	2.166	2.157869	0.007753	0.358%	2.641	2.094
7.5K	4.792	4.779245	0.012223	0.255%	5.674	4.466
10K	8.266	8.230676	0.034747	0.420%	10.252	7.947
20K	31.757	31.645776	0.110717	0.349%	39.798	30.817
30K	68.509	68.328158	0.179587	0.262%	84.795	69.686
40K	123.505	122.035564	1.468471	1.189%	165.461	116.686
50K	193.375	191.806797	1.566866	0.810%	238.140	179.521

Memory usage and additional metric overhead

Table 7: Memory usage in MB. Percentages in parentheses are relative to the Feature Memory column.

N	Feature Memory	CADMA	ClippedCoverage	Coverage
1K	167.3	8.3 (5.0%)	3.5 (2.1%)	5.2 (3.1%)
2.5K	191.6	11.1 (5.8%)	3.7 (1.9%)	9.9 (5.2%)
5K	234.0	16.8 (7.3%)	6.1 (2.6%)	16.3 (6.9%)
7.5K	274.7	24.4 (8.9%)	4.1 (1.5%)	17.0 (6.2%)
10K	314.3	28.5 (9.1%)	3.9 (1.3%)	15.7 (5.0%)
20K	478.7	39.5 (8.3%)	10.2 (2.1%)	24.2 (5.1%)
30K	640.6	50.0 (7.8%)	12.9 (2.0%)	32.5 (5.1%)
40K	806.0	52.8 (6.6%)	19.2 (2.4%)	40.4 (5.0%)
50K	970.1	65.8 (6.8%)	21.2 (2.2%)	53.0 (5.4%)

Only ~13 MB more peak memory than Coverage at 50K samples.

Diagnostics remain practical at 50K samples

0.810%

of total runtime for matching
and score decomposition

1.566866 s additional operations
193.375 s total CADMA runtime

65.8 MB

additional peak memory
6.8% of the feature-loading baseline

970.1 MB after feature loading
50,000 samples in each set

Precomputed features · A100 40 GB · one CPU thread · CADMA k = 70

Paper Tables 6–7 · Exact reported values

Conclusion

- CADMA evaluates diversity more reliably.
- Local capacity matching reduces optimistic coverage.
- Separate scores diagnose mode dropping and collapse.

GitHub code

github.com/Open-Debin/CADMA_NeurIPS2026

Queen Mary University of London

University of Bedfordshire