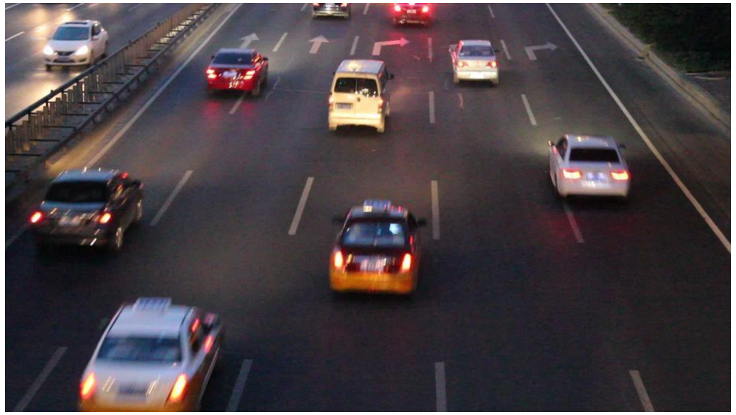
PTTA:
Purifying Malicious Samples for Test-Time Model Adaptation

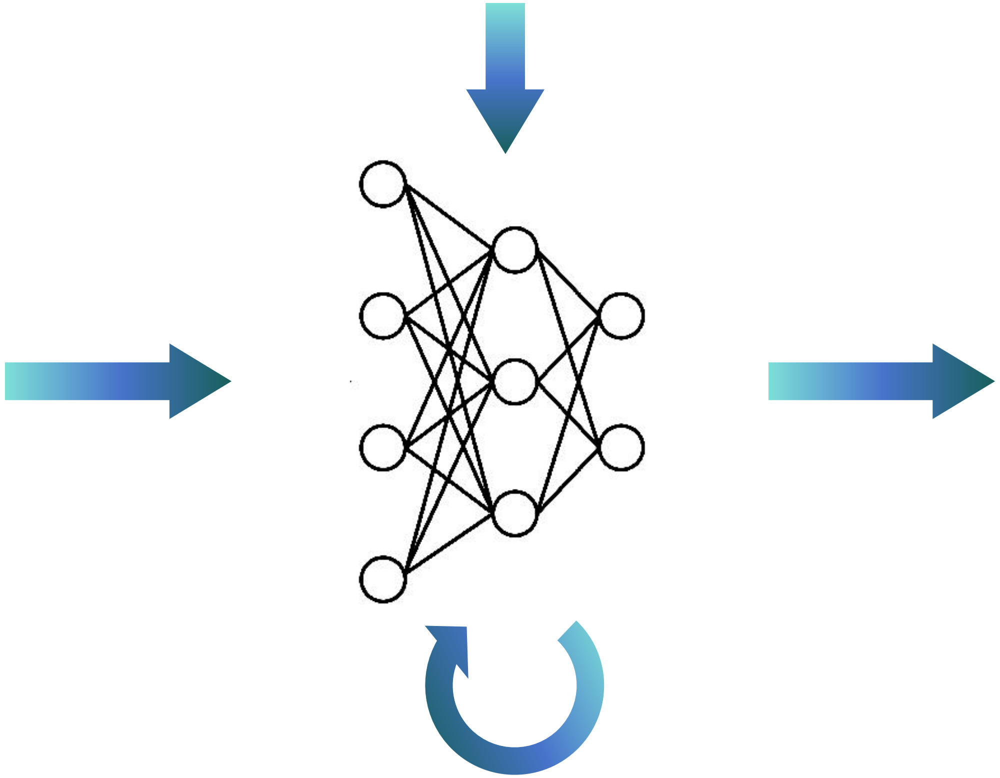
Jing Ma, Hanlin Li, Xiang Xiang

Introduction

Model Zoo (e.g., Hugging Face 🤗)

Test Samples



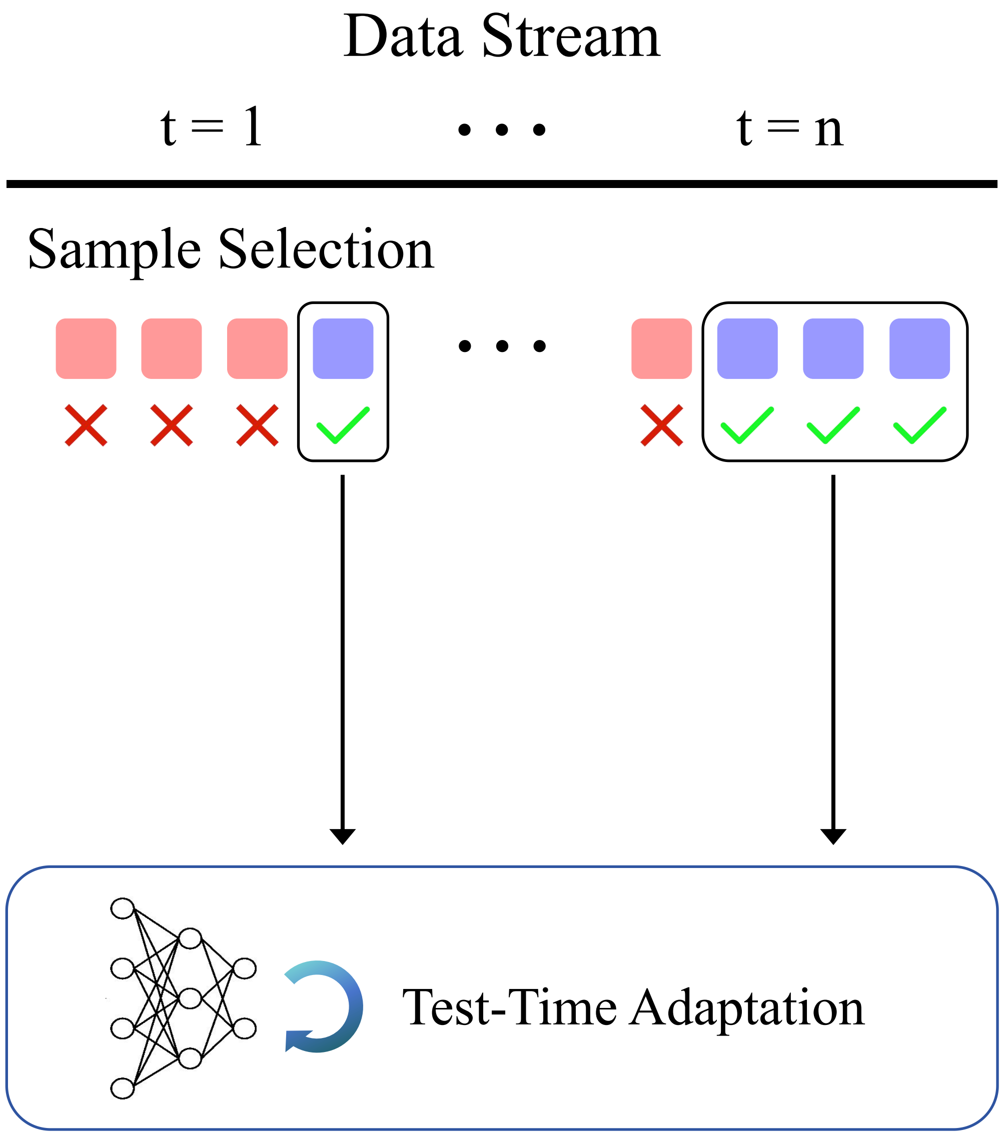


Online Prediction

Accuracy ↑

Test-Time Model Adaptation

Malicious Sample Hazards

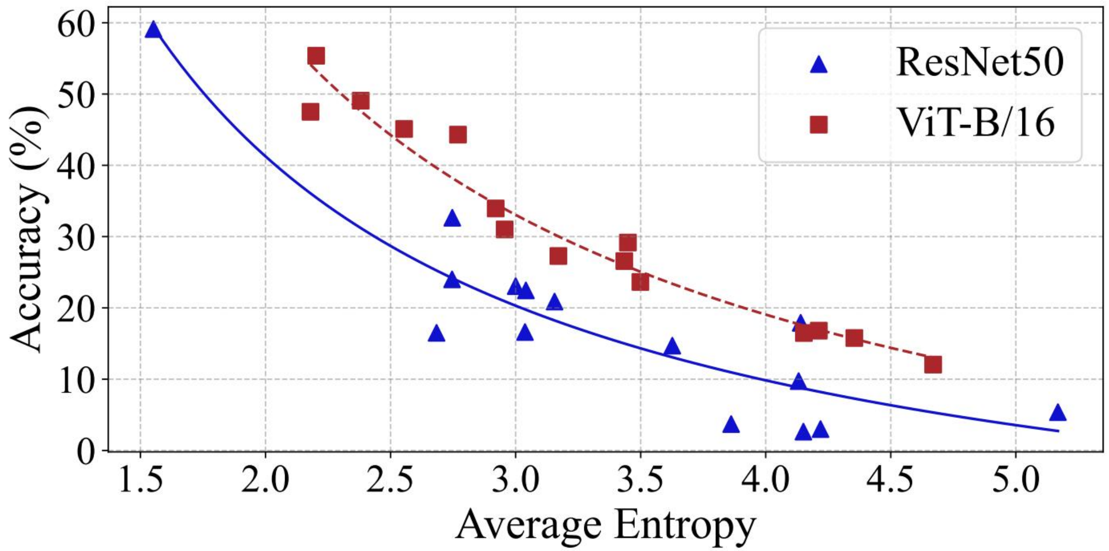


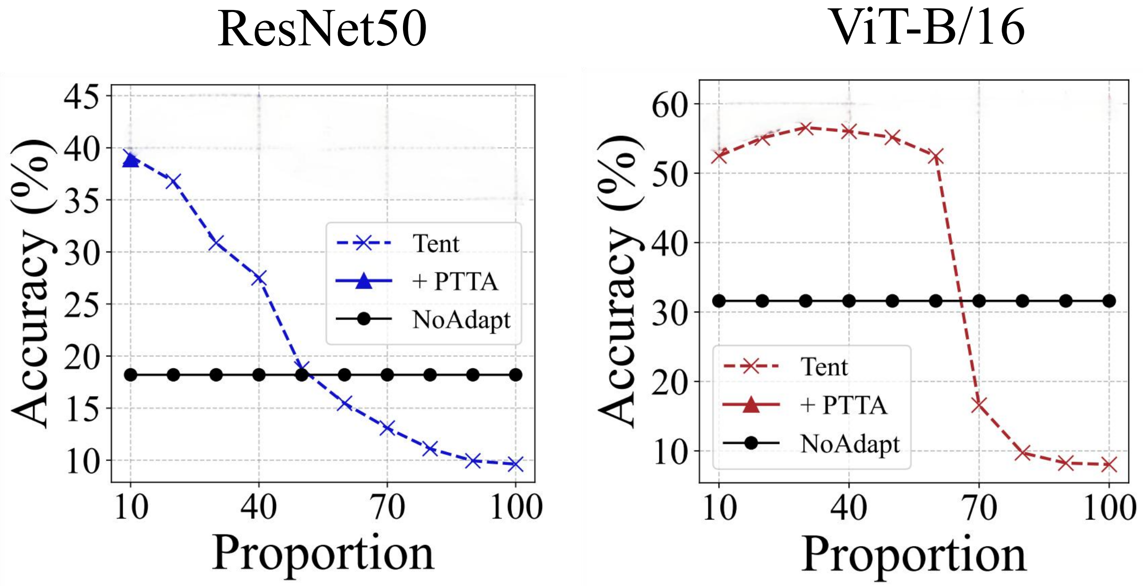
Data Stream

t = 1 • • • t = n

Sample Selection

Test-Time Adaptation





ResNet50

ViT-B/16

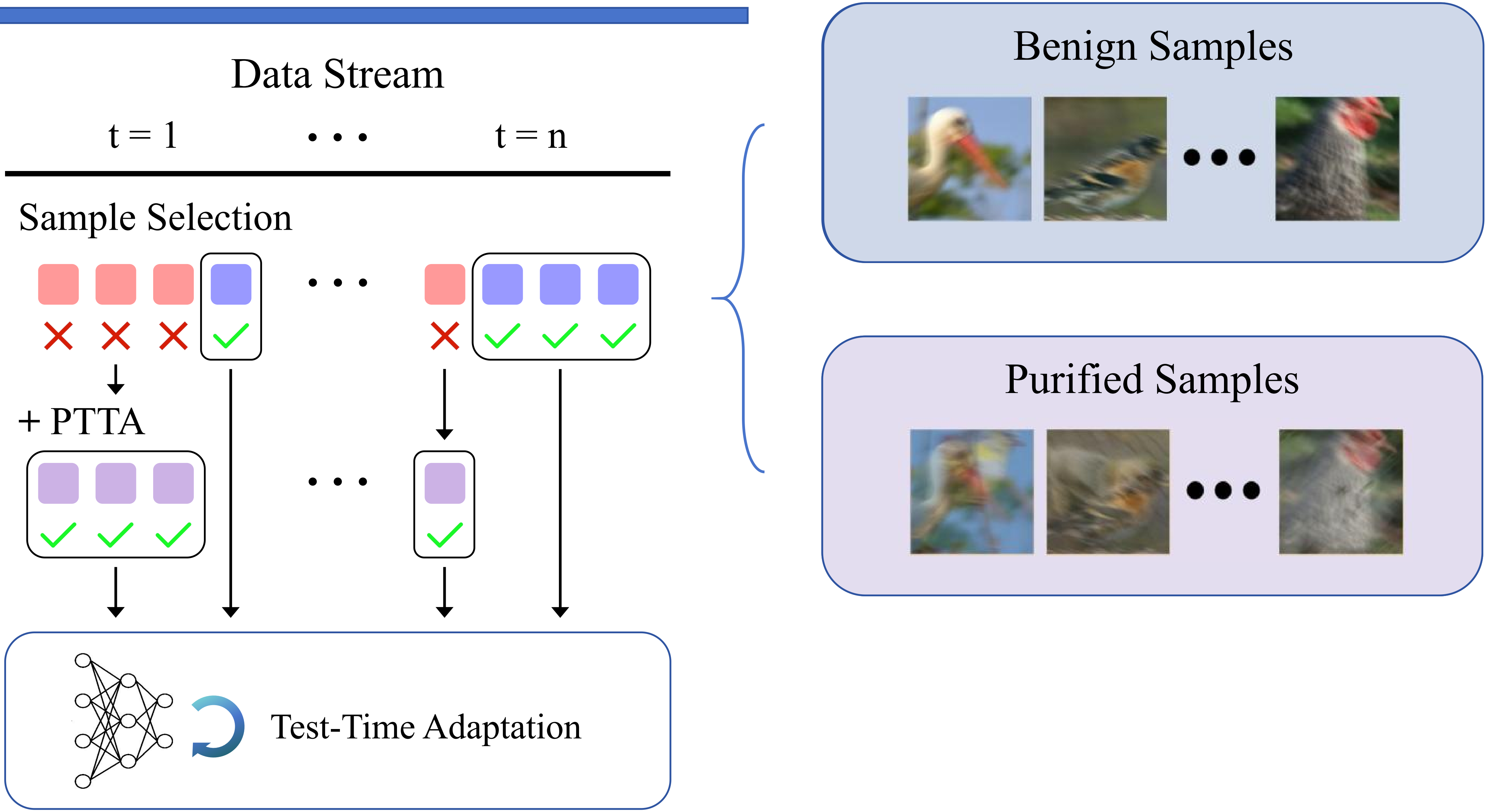
Introduction

Sample selection wastes our limited test samples.

Question:

Rather than selecting and discarding malicious samples, why not **purify** them into benign ones?

Introduction



Data Stream

t = 1 • • • t = n

Sample Selection

✗ ✗ ✗ ✓ • • • ✗ ✓ ✓ ✓

+ PTTA

✓ ✓ ✓ • • • ✓

Test-Time Adaptation

Benign Samples

Purified Samples

Method

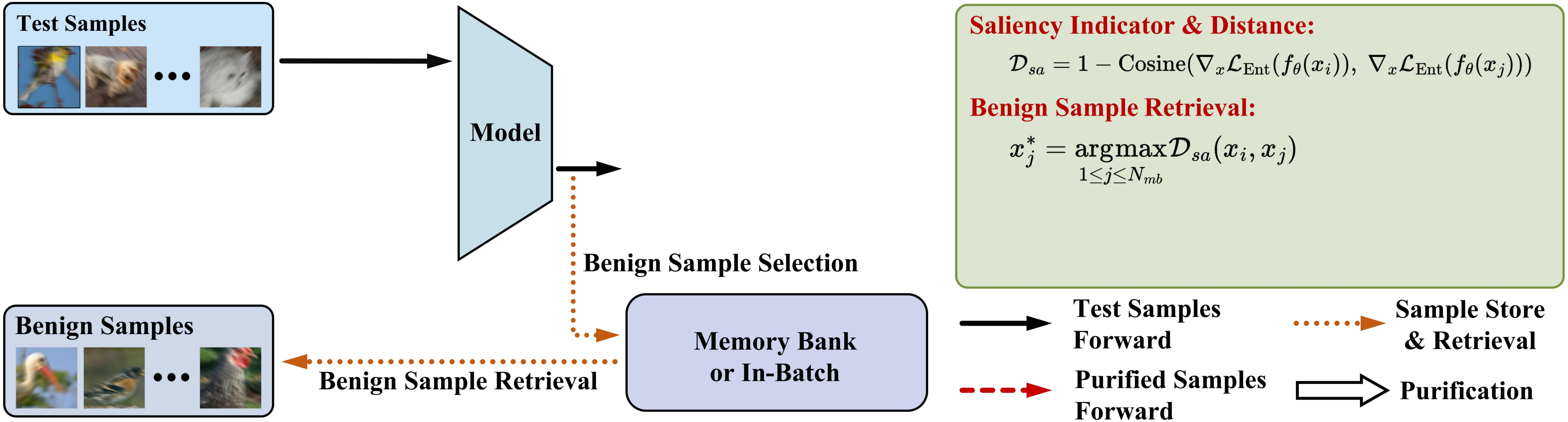
- 1. Logit-Saliency Indicator

$$\nabla_z \mathcal{L}_{\text{Ent}}(f_\theta(x)) = -(\mathbf{z} - \mathbf{p} \cdot \mathbf{z}) \odot \mathbf{p} \qquad (4)$$

f_θ	The model with parameters θ	$\mathbf{z}$	The output logits
x	A test sample	$\mathbf{p}$	The predicted probabilities
$\mathcal{L}_{\text{Ent}}$	Entropy minimization objective		

Method

- 2. Benign Sample Retrieval



Test Samples

Model

Benign Sample Selection

Memory Bank or In-Batch

Benign Samples

Benign Sample Retrieval

Saliency Indicator & Distance:

$$\mathcal{D}_{sa} = 1 - \text{Cosine}(\nabla_x \mathcal{L}_{\text{Ent}}(f_\theta(x_i)),\ \nabla_x \mathcal{L}_{\text{Ent}}(f_\theta(x_j)))$$

Benign Sample Retrieval:

$$x_j^* = \underset{1 \le j \le N_{mb}}{\text{argmax}} \mathcal{D}_{sa}(x_i, x_j)$$

Test Samples Forward

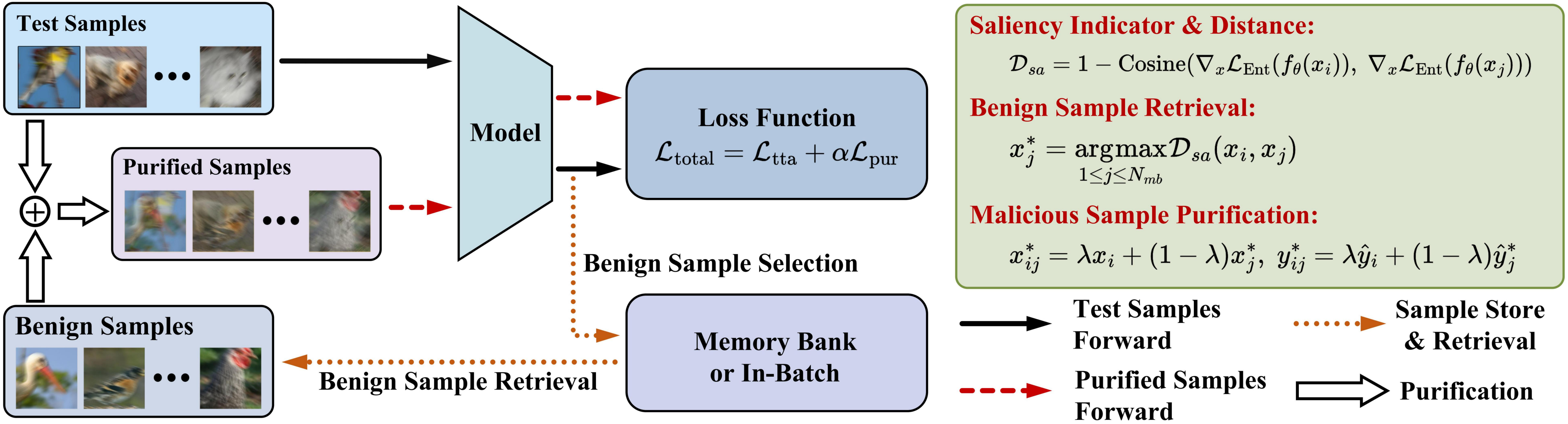
Sample Store & Retrieval

Purified Samples Forward

Purification

Method

- ➢ 3. Malicious Sample Purification



Saliency Indicator & Distance:

$$\mathcal{D}_{sa} = 1 - \text{Cosine}(\nabla_x \mathcal{L}_{\text{Ent}}(f_\theta(x_i)),\ \nabla_x \mathcal{L}_{\text{Ent}}(f_\theta(x_j)))$$

Benign Sample Retrieval:

$$x_j^* = \underset{1 \le j \le N_{mb}}{\text{argmax}} \mathcal{D}_{sa}(x_i, x_j)$$

Malicious Sample Purification:

$$x_{ij}^* = \lambda x_i + (1-\lambda)x_j^*,\ y_{ij}^* = \lambda \hat{y}_i + (1-\lambda)\hat{y}_j^*$$

Purification Loss: $$\mathcal{L}_{\text{pur}} = -\frac{1}{N_{bs}} \sum_{i}^{N_{bs}} \sum_{c}^{C} (y_{ij}^*)_c \log \sigma_c(f_\theta(x_{ij}^*))$$

Main Experiments

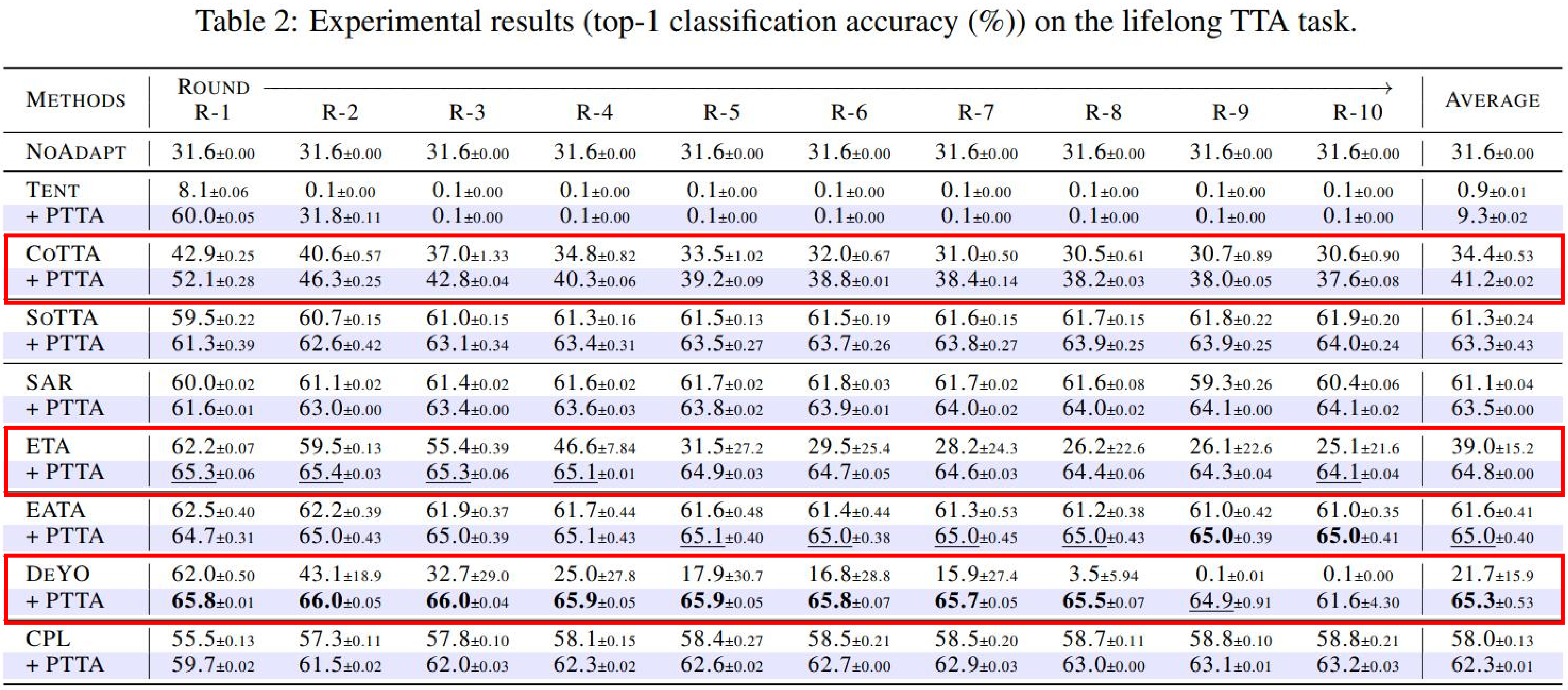
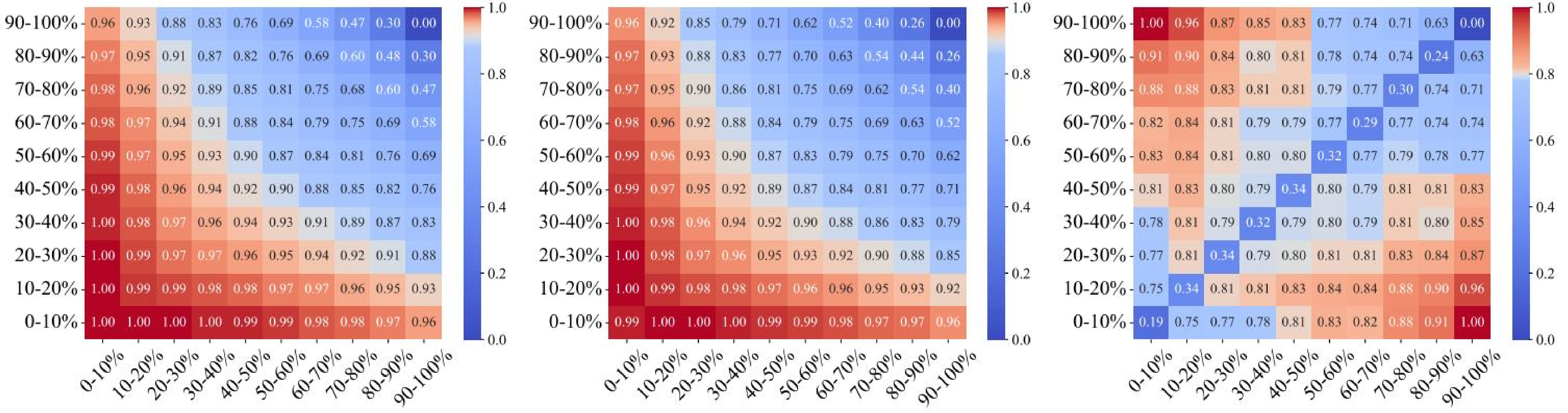


Table 2: Experimental results (top-1 classification accuracy (%)) on the lifelong TTA task.

Methods	Round R-1	R-2	R-3	R-4	R-5	R-6	R-7	R-8	R-9	R-10	Average
NoAdapt	31.6±0.00	31.6±0.00	31.6±0.00	31.6±0.00	31.6±0.00	31.6±0.00	31.6±0.00	31.6±0.00	31.6±0.00	31.6±0.00	31.6±0.00
Tent	8.1±0.06	0.1±0.00	0.1±0.00	0.1±0.00	0.1±0.00	0.1±0.00	0.1±0.00	0.1±0.00	0.1±0.00	0.1±0.00	0.9±0.01
+ PTTA	60.0±0.05	31.8±0.11	0.1±0.00	0.1±0.00	0.1±0.00	0.1±0.00	0.1±0.00	0.1±0.00	0.1±0.00	0.1±0.00	9.3±0.02
CoTTA	42.9±0.25	40.6±0.57	37.0±1.33	34.8±0.82	33.5±1.02	32.0±0.67	31.0±0.50	30.5±0.61	30.7±0.89	30.6±0.90	34.4±0.53
+ PTTA	52.1±0.28	46.3±0.25	42.8±0.04	40.3±0.06	39.2±0.09	38.8±0.01	38.4±0.14	38.2±0.03	38.0±0.05	37.6±0.08	41.2±0.02
SoTTA	59.5±0.22	60.7±0.15	61.0±0.15	61.3±0.16	61.5±0.13	61.5±0.19	61.6±0.15	61.7±0.15	61.8±0.22	61.9±0.20	61.3±0.24
+ PTTA	61.3±0.39	62.6±0.42	63.1±0.34	63.4±0.31	63.5±0.27	63.7±0.26	63.8±0.27	63.9±0.25	63.9±0.25	64.0±0.24	63.3±0.43
SAR	60.0±0.02	61.1±0.02	61.4±0.02	61.6±0.02	61.7±0.02	61.8±0.03	61.7±0.02	61.6±0.08	59.3±0.26	60.4±0.06	61.1±0.04
+ PTTA	61.6±0.01	63.0±0.00	63.4±0.00	63.6±0.03	63.8±0.02	63.9±0.01	64.0±0.02	64.0±0.02	64.1±0.00	64.1±0.02	63.5±0.00
ETA	62.2±0.07	59.5±0.13	55.4±0.39	46.6±7.84	31.5±27.2	29.5±25.4	28.2±24.3	26.2±22.6	26.1±22.6	25.1±21.6	39.0±15.2
+ PTTA	65.3±0.06	65.4±0.03	65.3±0.06	65.1±0.01	64.9±0.03	64.7±0.05	64.6±0.03	64.4±0.06	64.3±0.04	64.1±0.04	64.8±0.00
EATA	62.5±0.40	62.2±0.39	61.9±0.37	61.7±0.44	61.6±0.48	61.4±0.44	61.3±0.53	61.2±0.38	61.0±0.42	61.0±0.35	61.6±0.41
+ PTTA	64.7±0.31	65.0±0.43	65.0±0.39	65.1±0.43	65.1±0.40	65.0±0.38	65.0±0.45	65.0±0.43	**65.0**±0.39	**65.0**±0.41	65.0±0.40
DeYO	62.0±0.50	43.1±18.9	32.7±29.0	25.0±27.8	17.9±30.7	16.8±28.8	15.9±27.4	3.5±5.94	0.1±0.01	0.1±0.00	21.7±15.9
+ PTTA	**65.8**±0.01	**66.0**±0.05	**66.0**±0.04	**65.9**±0.05	**65.9**±0.05	**65.8**±0.07	**65.7**±0.05	**65.5**±0.07	64.9±0.91	61.6±4.30	**65.3**±0.53
CPL	55.5±0.13	57.3±0.11	57.8±0.10	58.1±0.15	58.4±0.27	58.5±0.21	58.5±0.20	58.7±0.11	58.8±0.10	58.8±0.21	58.0±0.13
+ PTTA	59.7±0.02	61.5±0.02	62.0±0.03	62.3±0.02	62.6±0.02	62.7±0.00	62.9±0.03	63.0±0.00	63.1±0.01	63.2±0.03	62.3±0.01

Main Experiments

- ➢ Why logit-saliency indicator works?



	0-10%	10-20%	20-30%	30-40%	40-50%	50-60%	60-70%	70-80%	80-90%	90-100%
90-100%	0.96	0.93	0.88	0.83	0.76	0.69	0.58	0.47	0.30	0.00
80-90%	0.97	0.95	0.91	0.87	0.82	0.76	0.69	0.60	0.48	0.30
70-80%	0.98	0.96	0.92	0.89	0.85	0.81	0.75	0.68	0.60	0.47
60-70%	0.98	0.97	0.94	0.91	0.88	0.84	0.79	0.75	0.69	0.58
50-60%	0.99	0.97	0.95	0.93	0.90	0.87	0.84	0.81	0.76	0.69
40-50%	0.99	0.98	0.96	0.94	0.92	0.90	0.88	0.85	0.82	0.76
30-40%	1.00	0.98	0.97	0.96	0.94	0.93	0.91	0.89	0.87	0.83
20-30%	1.00	0.99	0.97	0.97	0.96	0.95	0.94	0.92	0.91	0.88
10-20%	1.00	0.99	0.99	0.98	0.98	0.97	0.97	0.96	0.95	0.93
0-10%	1.00	1.00	1.00	1.00	0.99	0.99	0.98	0.98	0.97	0.96

(a) Logit-saliency indicator

	0-10%	10-20%	20-30%	30-40%	40-50%	50-60%	60-70%	70-80%	80-90%	90-100%
90-100%	0.96	0.92	0.85	0.79	0.71	0.62	0.52	0.40	0.26	0.00
80-90%	0.97	0.93	0.88	0.83	0.77	0.70	0.63	0.54	0.44	0.26
70-80%	0.97	0.95	0.90	0.86	0.81	0.75	0.69	0.62	0.54	0.40
60-70%	0.98	0.96	0.92	0.88	0.84	0.79	0.75	0.69	0.63	0.52
50-60%	0.99	0.96	0.93	0.90	0.87	0.83	0.79	0.75	0.70	0.62
40-50%	0.99	0.97	0.95	0.92	0.89	0.87	0.84	0.81	0.77	0.71
30-40%	1.00	0.98	0.96	0.94	0.92	0.90	0.88	0.86	0.83	0.79
20-30%	1.00	0.98	0.97	0.96	0.95	0.93	0.92	0.90	0.88	0.85
10-20%	1.00	0.99	0.98	0.98	0.97	0.96	0.96	0.95	0.93	0.92
0-10%	0.99	1.00	1.00	1.00	0.99	0.99	0.98	0.97	0.97	0.96

(b) Feature-saliency indicator

	0-10%	10-20%	20-30%	30-40%	40-50%	50-60%	60-70%	70-80%	80-90%	90-100%
90-100%	1.00	0.96	0.87	0.85	0.83	0.77	0.74	0.71	0.63	0.00
80-90%	0.91	0.90	0.84	0.80	0.81	0.78	0.74	0.74	0.24	0.63
70-80%	0.88	0.88	0.83	0.81	0.81	0.79	0.77	0.30	0.74	0.71
60-70%	0.82	0.84	0.81	0.79	0.79	0.77	0.29	0.77	0.74	0.74
50-60%	0.83	0.84	0.81	0.80	0.80	0.32	0.77	0.79	0.78	0.77
40-50%	0.81	0.83	0.80	0.79	0.34	0.80	0.79	0.81	0.81	0.83
30-40%	0.78	0.81	0.79	0.32	0.79	0.80	0.79	0.81	0.80	0.85
20-30%	0.77	0.81	0.34	0.79	0.80	0.81	0.81	0.83	0.84	0.87
10-20%	0.75	0.34	0.81	0.81	0.83	0.84	0.84	0.88	0.90	0.96
0-10%	0.19	0.75	0.77	0.78	0.81	0.83	0.82	0.88	0.91	1.00

(c) Pixel-saliency indicator

Figure 6: The Saliency Distance (normalized to the range of $0 \sim 1$) among test samples sorted in the ascending order of prediction entropy and split by percentages. A good indicator satisfies $\mathcal{D}_{sa}(x^+, x^+) > \mathcal{D}_{sa}(x^+, x^-) > \mathcal{D}_{sa}(x^-, x^-)$.

Main Experiments

➢ **Unnecessary** of Sample Selection

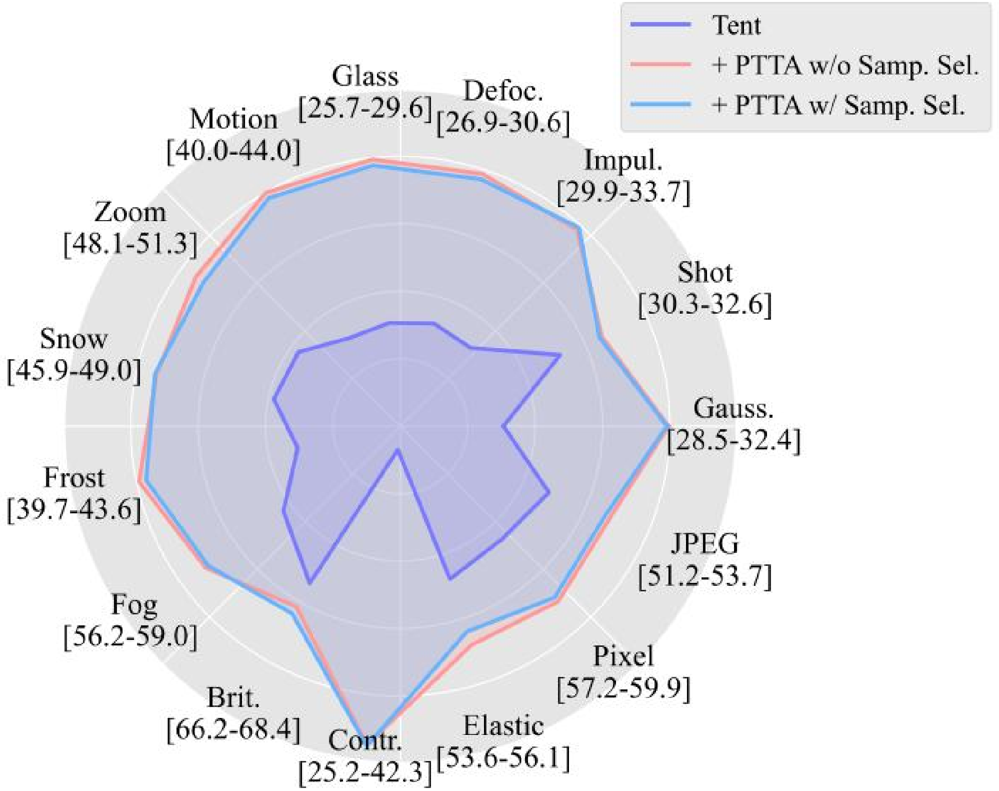
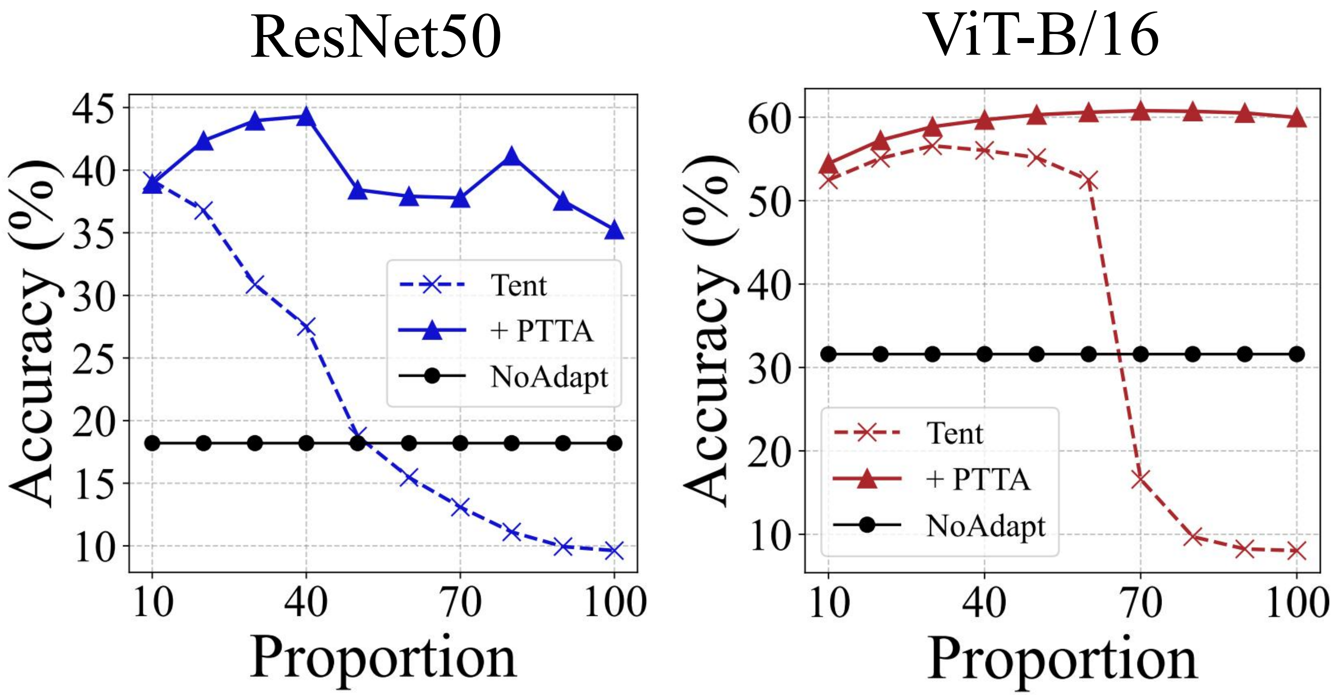


Figure 5: Comparison of selecting benign samples and using all samples as candidates for purifying malicious samples.

➢ **Insensitivity** to thresholds



Main Experiments

- Efficiency of PTTA

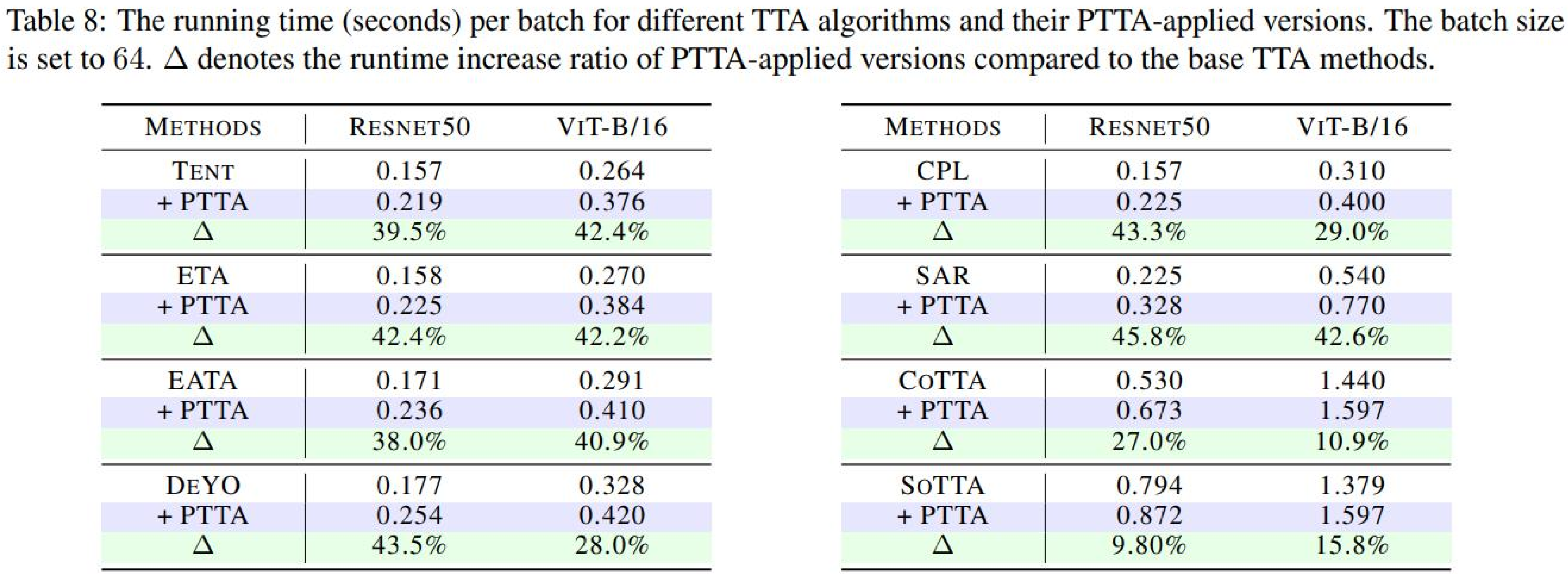


Table 8: The running time (seconds) per batch for different TTA algorithms and their PTTA-applied versions. The batch size is set to 64. Δ denotes the runtime increase ratio of PTTA-applied versions compared to the base TTA methods.

METHODS	RESNET50	VIT-B/16
TENT	0.157	0.264
+ PTTA	0.219	0.376
Δ	39.5%	42.4%
ETA	0.158	0.270
+ PTTA	0.225	0.384
Δ	42.4%	42.2%
EATA	0.171	0.291
+ PTTA	0.236	0.410
Δ	38.0%	40.9%
DEYO	0.177	0.328
+ PTTA	0.254	0.420
Δ	43.5%	28.0%

METHODS	RESNET50	VIT-B/16
CPL	0.157	0.310
+ PTTA	0.225	0.400
Δ	43.3%	29.0%
SAR	0.225	0.540
+ PTTA	0.328	0.770
Δ	45.8%	42.6%
COTTA	0.530	1.440
+ PTTA	0.673	1.597
Δ	27.0%	10.9%
SOTTA	0.794	1.379
+ PTTA	0.872	1.597
Δ	9.80%	15.8%

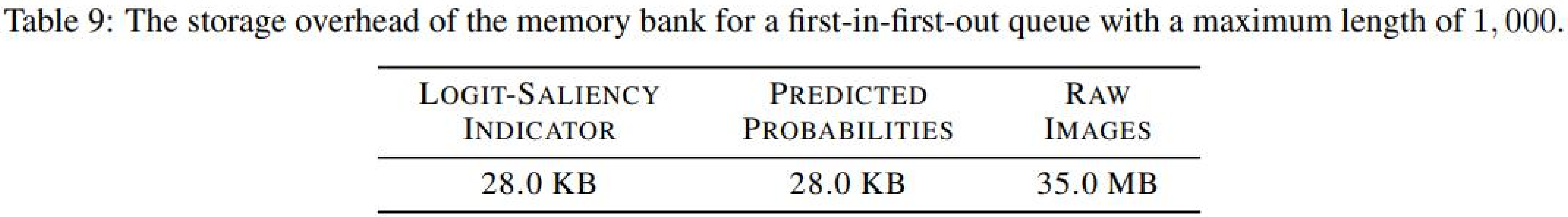


Table 9: The storage overhead of the memory bank for a first-in-first-out queue with a maximum length of $1,000$.

LOGIT-SALIENCY INDICATOR	PREDICTED PROBABILITIES	RAW IMAGES
28.0 KB	28.0 KB	35.0 MB

Thank You

More interesting analyses in our paper!

Code available on Github!