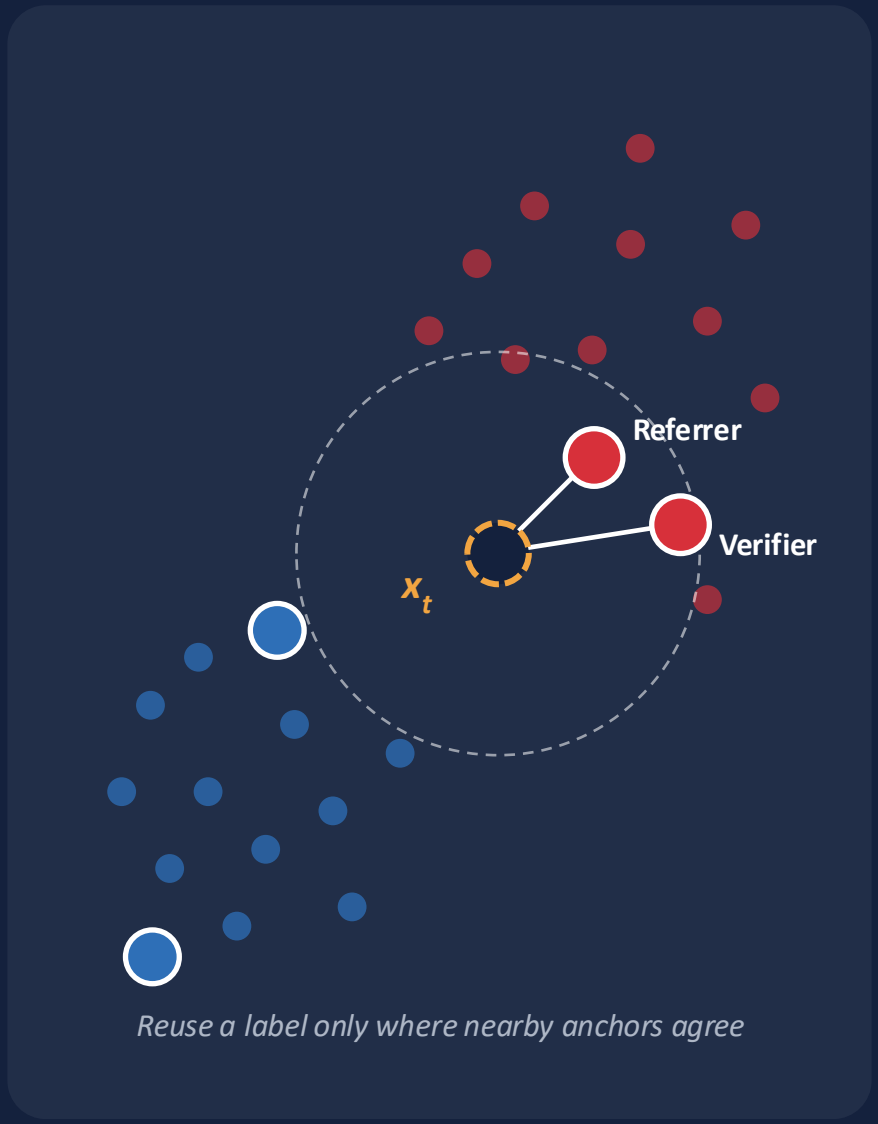
NeurIPS 2026 | KAIST

Training-Free Active Test-Time Adaptation for Vision-Language Models

ACTOR ACTive VLM-TTA for Online Retrieval

Jihwan Bang, Sumyeong Ahn, Hwanjun Song, Jae-Gil Lee

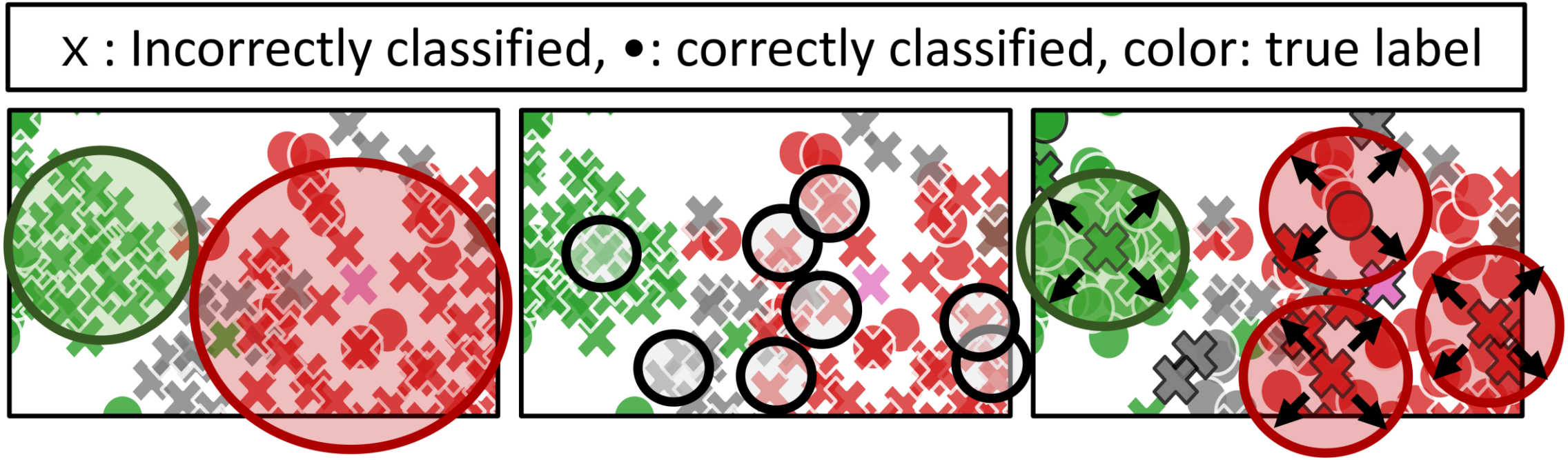
github.com/kaist-dmlab/ACTOR



MOTIVATION

Label-free TTA can't fix its own hard mistakes

Most VLM-TTA methods adapt from pseudo-labels, confidence, or entropy produced by the model itself. On hard, shifted samples, this self-generated signal can **reinforce errors instead of repairing them**.



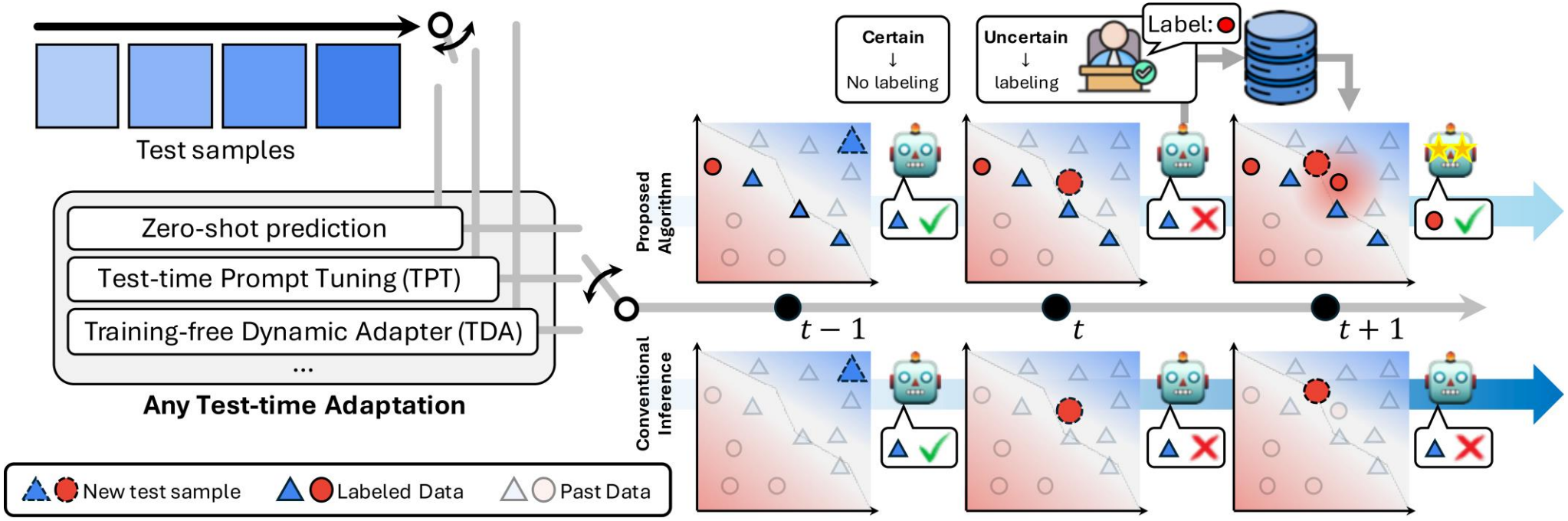
(a) Label-free TTA: errors persist on hard, shifted samples

(b) Query labels for a few uncertain samples → labeled anchors

(c) Anchors correct related future samples

PROBLEM SETTING · ACTIVE VLM-TTA

Labels are costed, delayed, and future-only



1 Predict first

x_t must be classified before its label is available

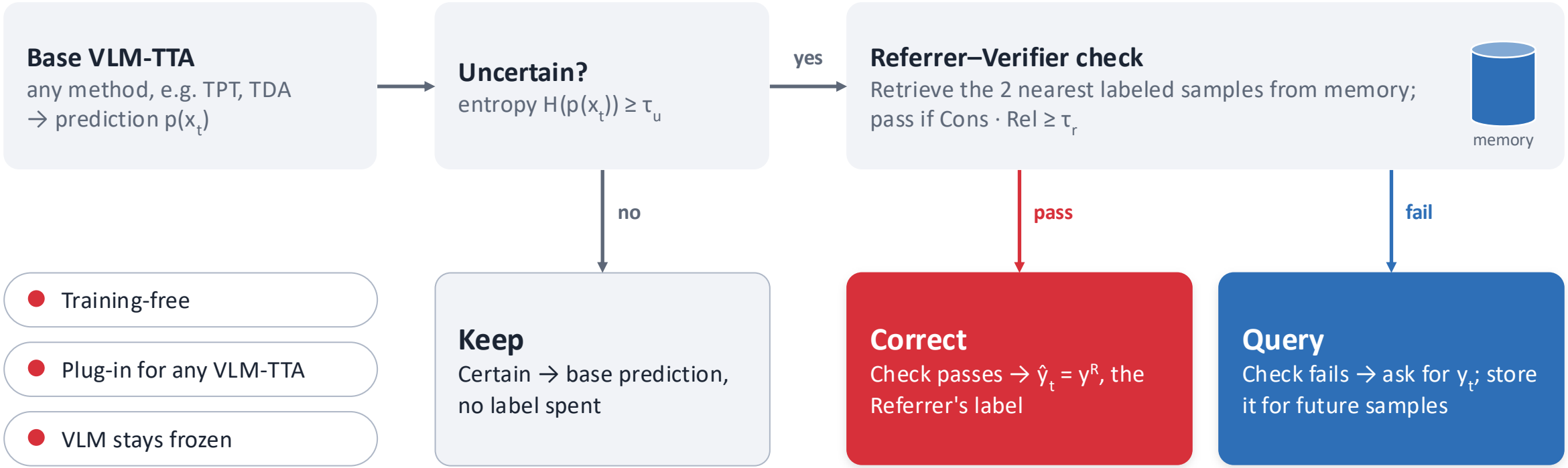
2 Query under a budget

Label at most B samples (10% of the stream)

3 Reuse only for the future

Labels help x_{t+1}, x_{t+2}, …; the VLM stays frozen

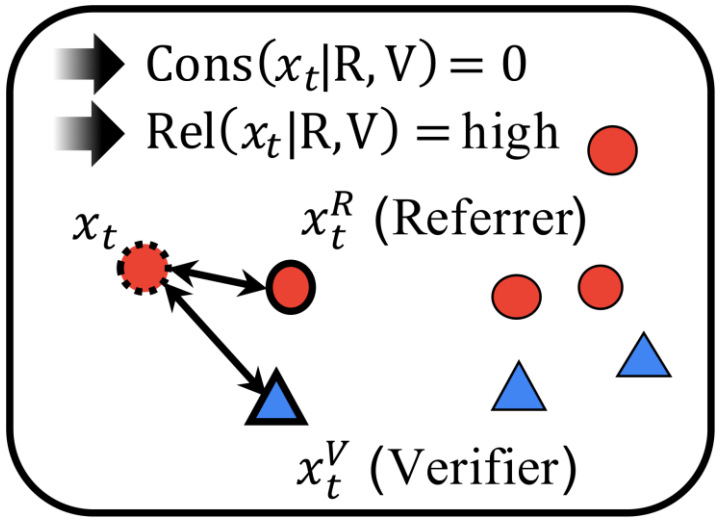
ACTOR: three-way routing, zero training



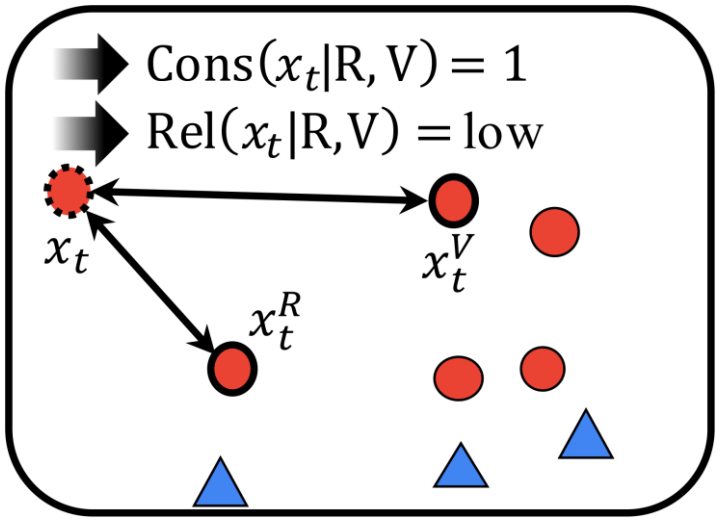
Budget spent? No new queries; accept the Referrer's label only if it is in the base model's Top-5 predictions.
One config for every dataset: $\tau_u = 0.2$, $\tau_r = 0.85$, $N = 5$ warm-up labels, budget $B = 10\%$ of the stream.

METHOD · RETRIEVAL-AUGMENTED CORRECTION

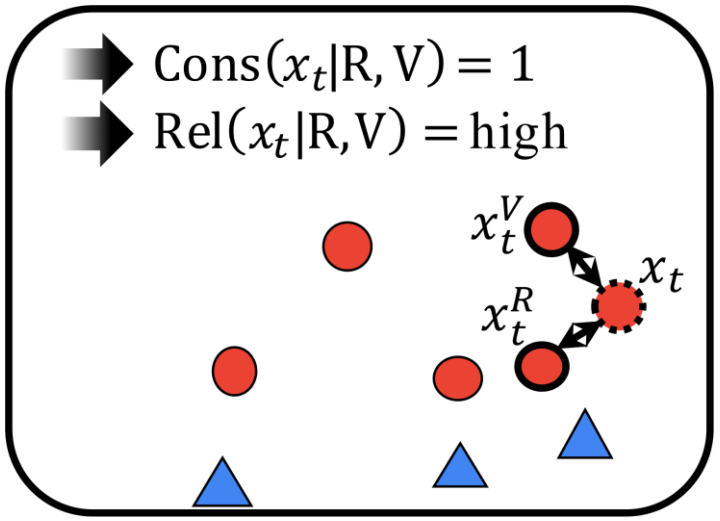
Reuse a label only when two close anchors agree



(a) Labels disagree → no correction



(b) Agree, but far → no correction



(c) Agree and close → correct

Consistency $\mathrm{Cons}(x_t \mid \mathrm{R}, \mathrm{V}) = \mathbb{1}\{\, y_t^{\mathrm{R}} = y_t^{\mathrm{V}} \,\}$

Reliability $\mathrm{Rel}(x_t \mid \mathrm{R}, \mathrm{V}) = \frac{1}{2}\Big[\cos\big(f(x_t), f(x_t^{\mathrm{R}})\big) + \cos\big(f(x_t), f(x_t^{\mathrm{V}})\big)\Big]$

Correct if $\mathrm{Cons} \cdot \mathrm{Rel} > \tau_r \implies \hat{y}_t = y_t^{\mathrm{R}}$

Product form: one label disagreement vetoes the correction.

Threshold τ_r = 0.85 keeps reuse inside trustworthy neighborhoods.

RESULTS · 15 BENCHMARKS, ONE FIXED CONFIGURATION

Consistent gains on every base method

+8.7 pp — **Zero-shot CLIP**
cross-domain avg.
61.94 → 70.67

+4.5 pp — **TDA (strongest base)**
cross-domain avg.
68.14 → 72.66

+4.1 pp — **TDA on ImageNet-S**
OOD benchmark
50.96 → 55.01

Method	Aircraft	Caltech101	Cars	DTD	EuroSAT	Flowers102	Food101	Pets	SUN397	UCF101	Average
					ViT-B/16						
Zero-shot	22.59	85.76	65.61	40.60	44.25	64.07	82.68	83.81	63.68	66.30	61.94
+ACTOR	27.33±0.71	90.31±0.12	67.77±0.25	47.94±0.42	**80.63**±0.54	80.53±0.20	83.57±0.06	87.10±0.13	65.30±0.03	76.28±0.54	70.67
TPT	23.43±0.08	93.21±1.07	66.52±0.20	46.33±1.10	42.69±0.16	68.25±0.93	83.94±1.18	84.42±0.98	65.19±0.47	67.62±0.41	64.16
+ACTOR	25.56±0.47	93.37±0.08	67.97±0.04	49.53±0.01	73.03±0.33	74.63±0.80	83.54±0.04	85.96±0.16	66.49±0.06	72.67±0.23	69.27
C-TPT	24.09±0.25	92.98±0.69	65.24±0.78	45.01±1.21	42.28±0.10	70.36±0.75	83.22±0.11	87.82±0.72	64.07±0.59	65.58±0.79	64.06
+ACTOR	27.78±0.47	93.27±0.18	66.88±0.37	50.59±0.17	74.72±0.79	77.49±0.43	83.73±0.05	87.70±0.02	66.22±0.05	74.65±0.64	70.30
TDA†	23.48±0.28	88.54±0.08	66.59±0.11	42.97±0.67	54.82±0.77	65.27±0.49	83.82±0.08	85.12±0.27	65.14±0.01	69.63±0.15	64.54
+ACTOR	25.74±0.21	90.41±0.43	67.94±0.04	49.41±0.34	75.40±0.54	79.20±0.49	84.20±0.06	87.37±0.10	66.18±0.01	76.01±0.62	70.18
TDA	25.29±0.30	**94.06**±0.08	66.39±1.72	45.48±0.79	63.95±0.92	71.62±0.23	**86.14**±0.06	89.85±0.13	67.70±0.04	70.94±0.28	68.14
+ACTOR	**27.84**±0.08	**94.06**±0.03	**68.76**±0.13	**51.63**±0.88	79.61±1.40	**82.28**±0.54	86.09±0.06	**90.75**±0.18	**68.20**±0.11	**77.38**±1.01	**72.66**

Table 2, ViT-B/16 rows: cross-domain benchmark, 10% label budget (gray rows = +ACTOR). ResNet-50 and the OOD benchmark (Table 1) follow the same trend.

RESULTS · LABEL-MATCHED COMPARISON & COST

Same labels, better use — at almost no cost

Prompt tuning on exactly the labels ACTOR queried

Table 3 · ViT-B/16 · same queried labels, future-only supervision

Method	Cross (10)	OOD (4)	OOD (5)	Avg. (15)
CLIP (Zero-shot)	61.94	58.47	59.48	61.12
CLIP + Online-CoOp	67.43	58.61	60.58	65.15
CLIP + Replay-CoOp	69.03	56.24	57.54	65.20
CLIP + ACTOR (Train-free)	**70.67**	**60.54**	**61.61**	**67.65**

+2.45 pp higher 15-dataset average than the best prompt-tuning baseline (67.65 vs. 65.20) — **best on every split, with no training**

Online-CoOp: one prompt update per queried label

Replay-CoOp: also reuses all accumulated labels

Per-sample inference time (ms)

Table 4 · rows: without / with ACTOR

ACTOR	Zero-shot	TPT	C-TPT	TDA
✗	10.07	509.87	511.33	10.84
✓	10.16	510.22	511.55	11.27
Rel.	+0.8%	+0.06%	+0.04%	+3.4%

+0.43 ms

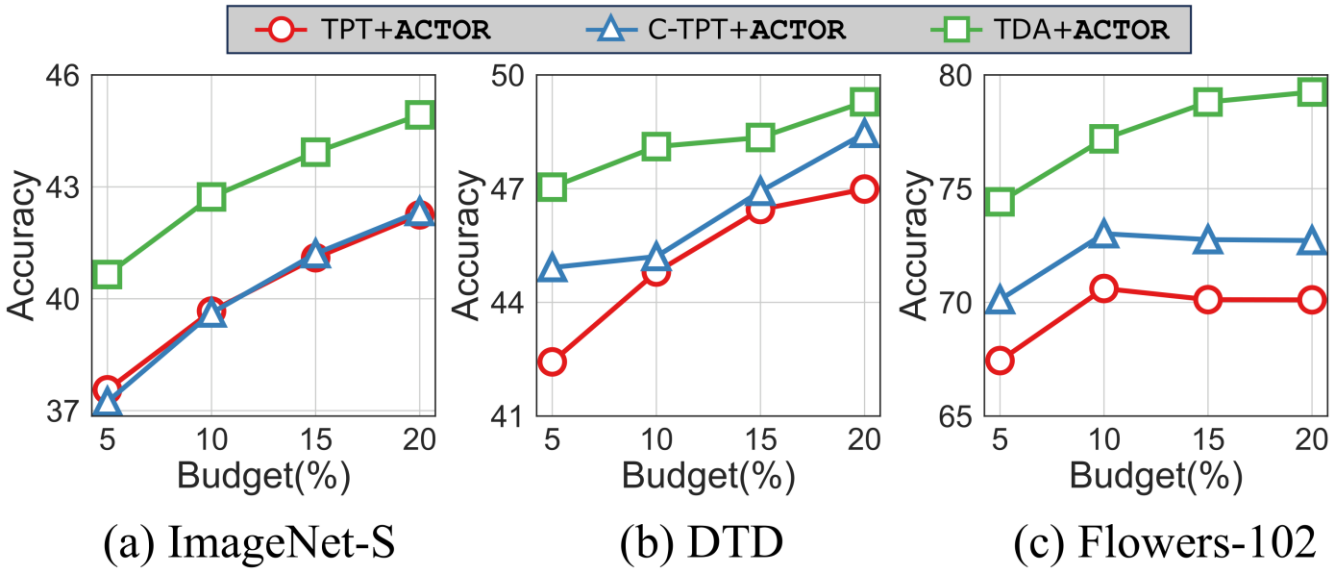
maximum added time per sample — no gradient updates

Prompt-tuning TTA (TPT, C-TPT): ~510 ms

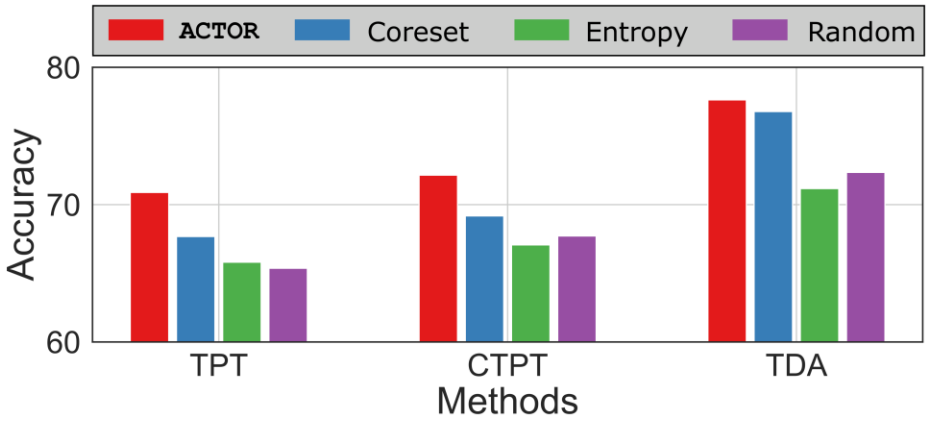
ANALYSIS

Robust to budget; tailored querying matters

Accuracy vs. label budget (5–20%)



Query rule vs. active learning



More labels, better retrieval. Accuracy generally rises with the budget, so the gains are not an artifact of the 10% setting.

Tailored querying wins. On Flowers-102, ACTOR beats Coreset, Entropy, and Random with every base TTA.

Takeaways

1. **A new setting: Active VLM-TTA**

   A frozen VLM spends a small label budget; labels only help future samples.

2. **A training-free method: ACTOR**

   Three-way routing + Referrer–Verifier check; plugs into any VLM-TTA.

3. **Consistent, cheap gains**

   Improves every base method · beats label-matched CoOp · ≤ 0.43 ms extra

CODE

github.com/kaist-dmlab/ACTOR

Thank you!