Is Less More?

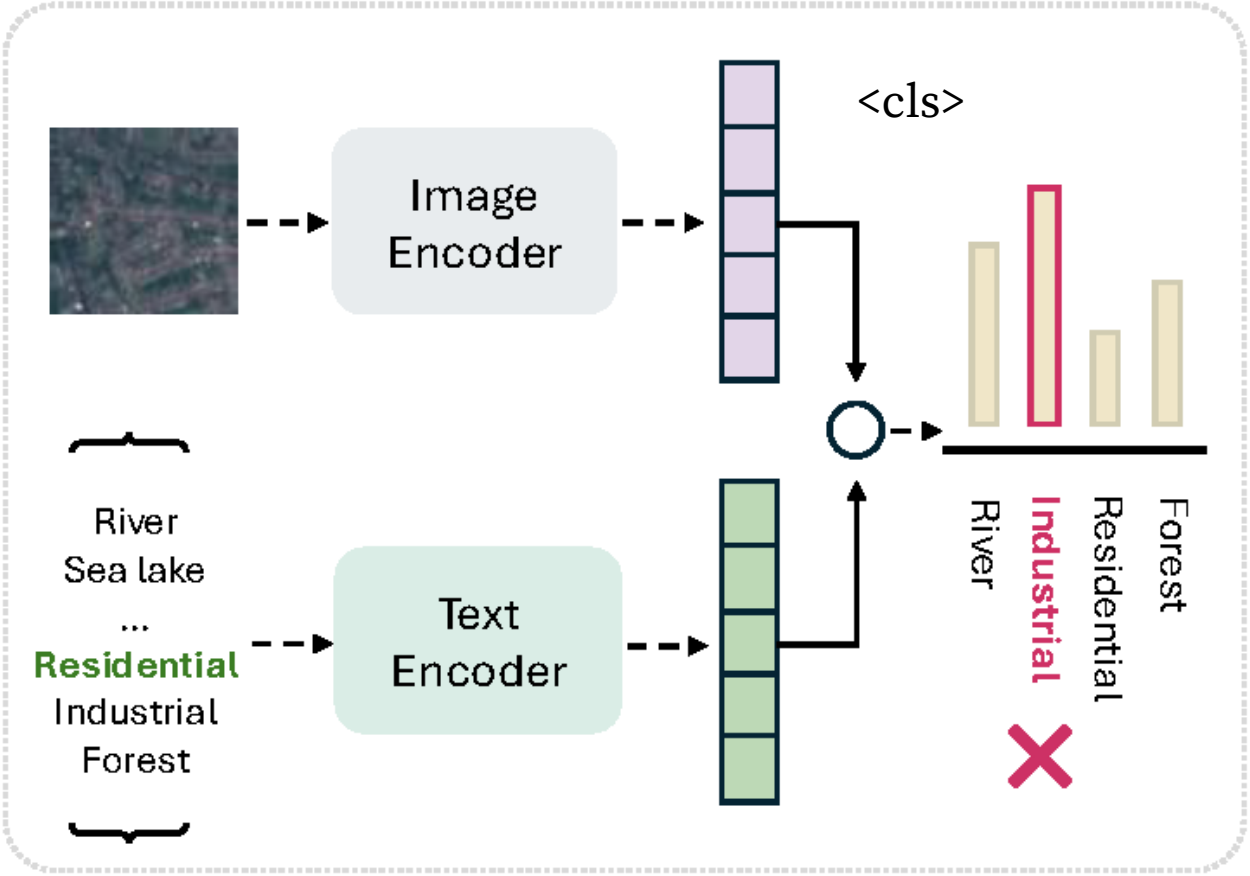
Exploring Token Condensation as Training-free Test-time Adaptation

Zixin Wang, Dong Gong*, Sen Wang, Zi Huang, Yadan Luo

*UQ, UNSW**

Australia

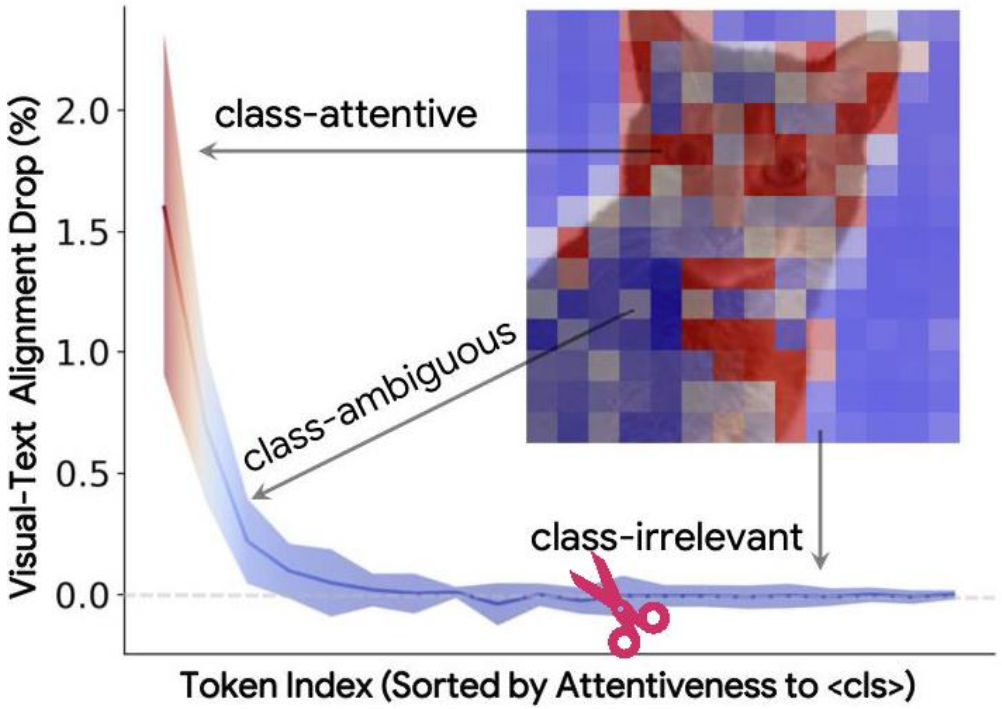
Why CLIP **Fails** on Unseen Data?



- CLIP is strong on zero-shot benchmarks.
- But on unseen datasets, visual–text alignment becomes unstable.
- Some visual information dilute alignment, leading to wrong predictions.

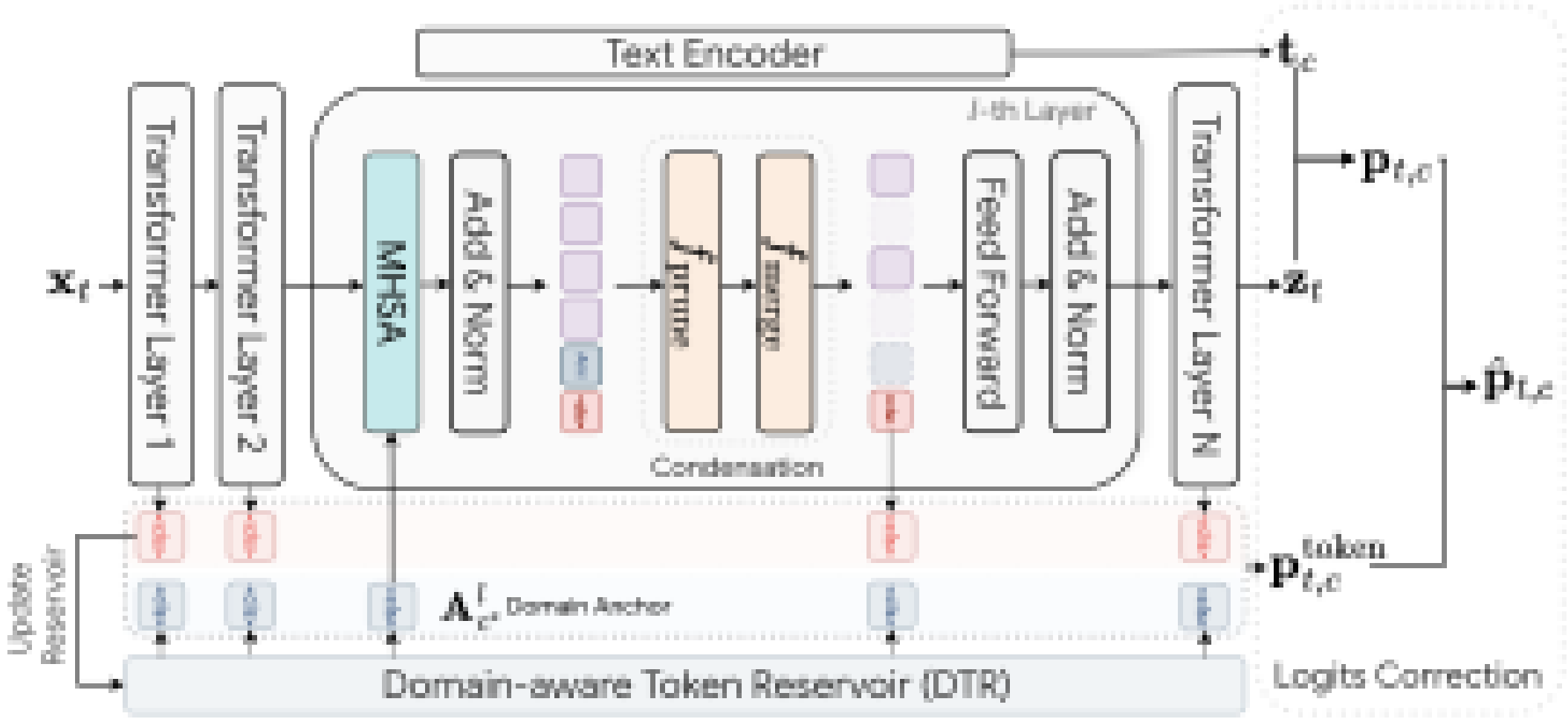
Root question: Is all visual information useful?

Not All Tokens Are Equal

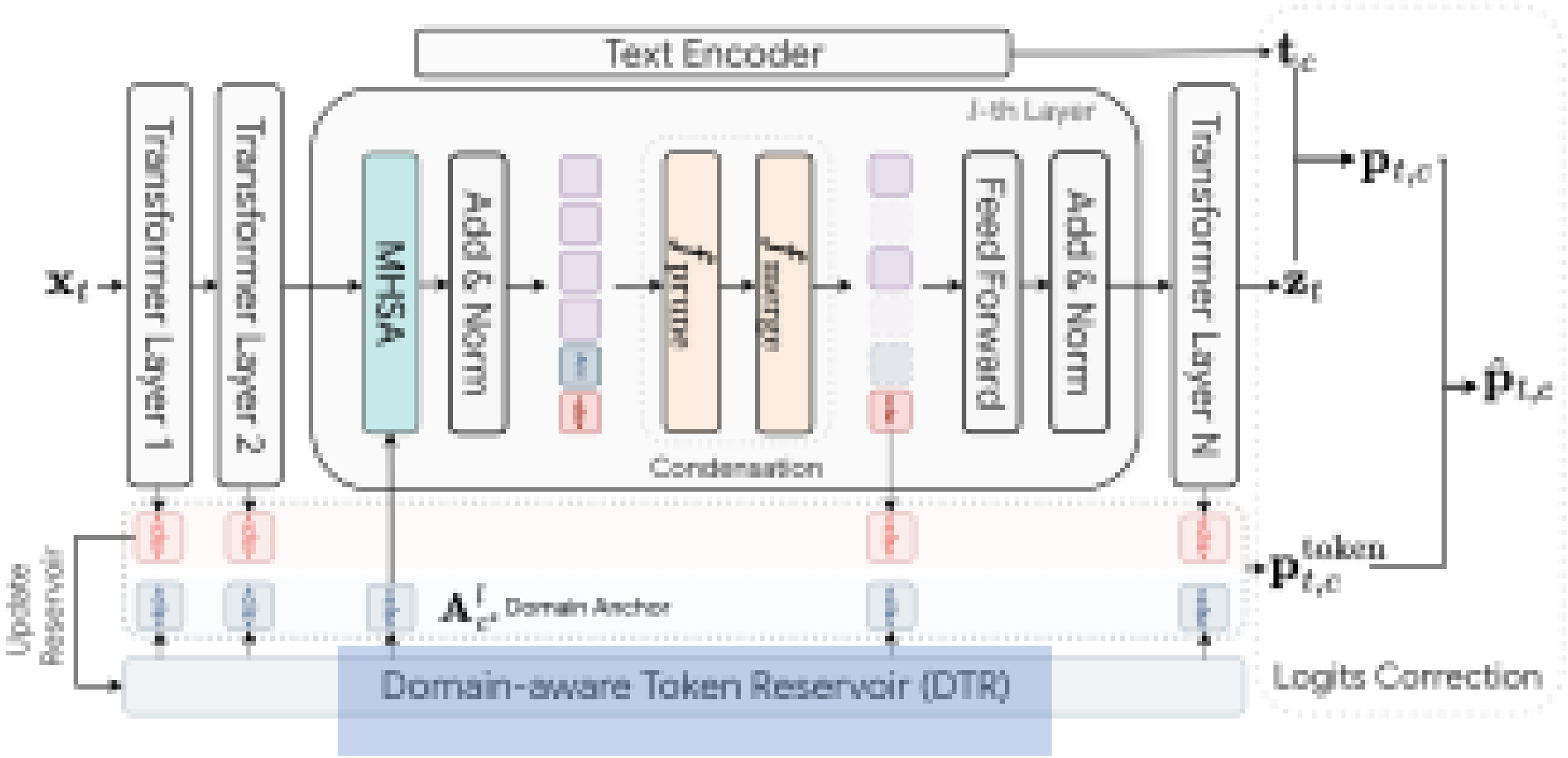


- Token influence is uneven: many low-attention tokens are class-irrelevant or ambiguous.
- Surprisingly, dropping them even improves visual–text alignment.

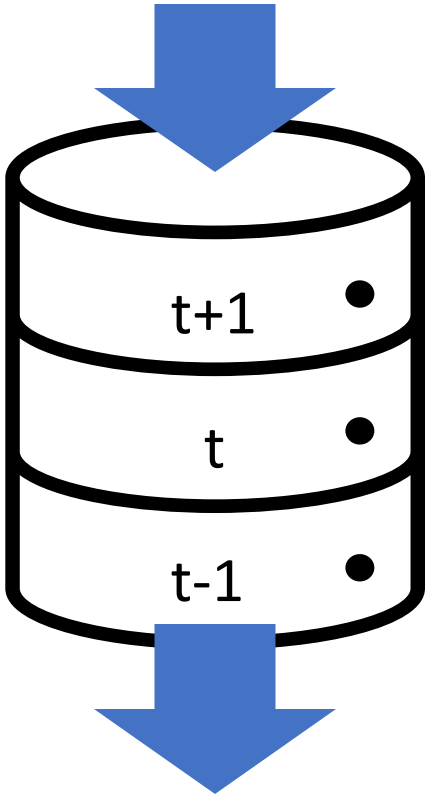
Token Condensation as Adaptation



Domain-aware Token Reservoir



Mining Good Tokens as Guidance

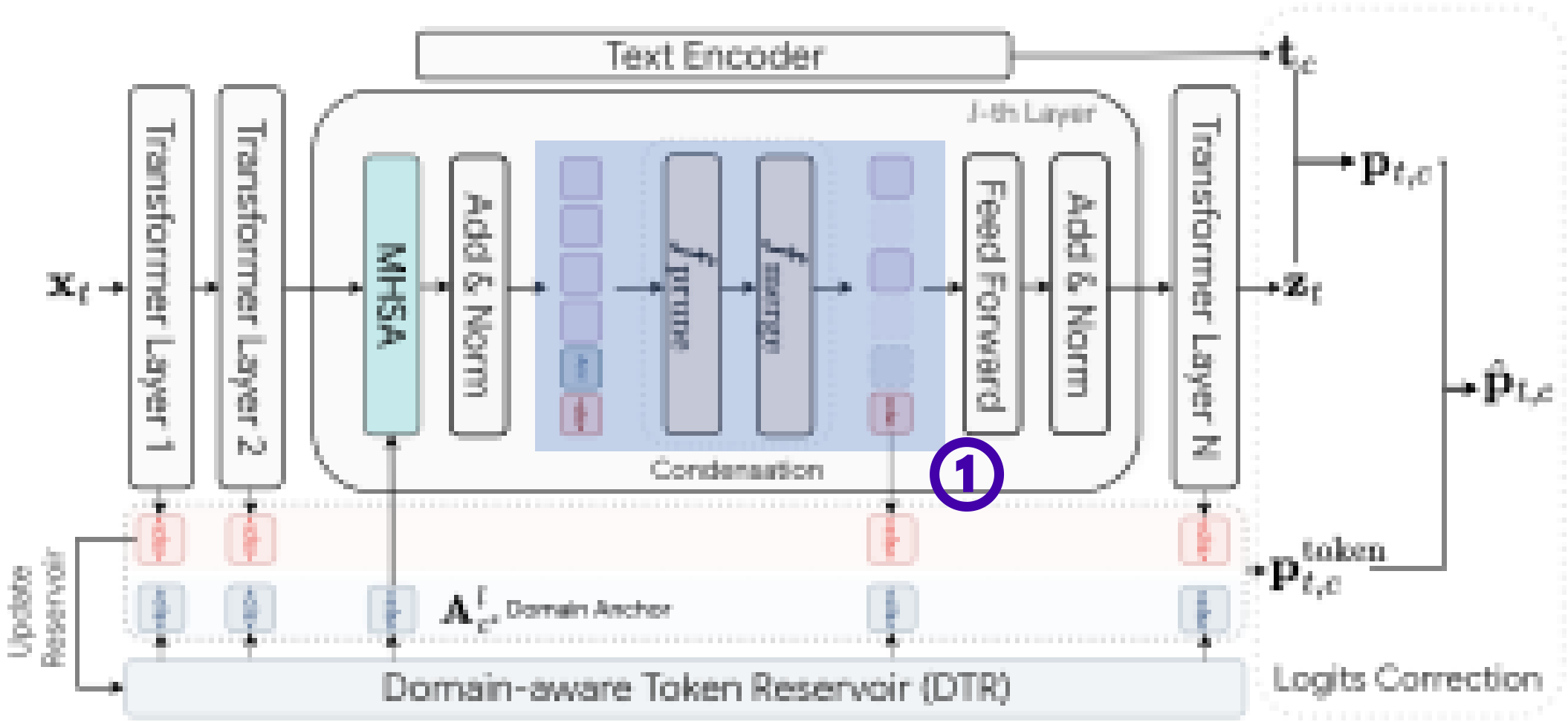


Each domain has tokens that align well.

- **What to store?** A per-class priority queue keeps the top-M most reliable domain anchors. Each item stores the entropy and the <cls> tokens as anchors.

- **How are anchors used?**
  - Guide token reduction. ①
  - Correct the distribution of logits. ②

Cross-head Token Reduction



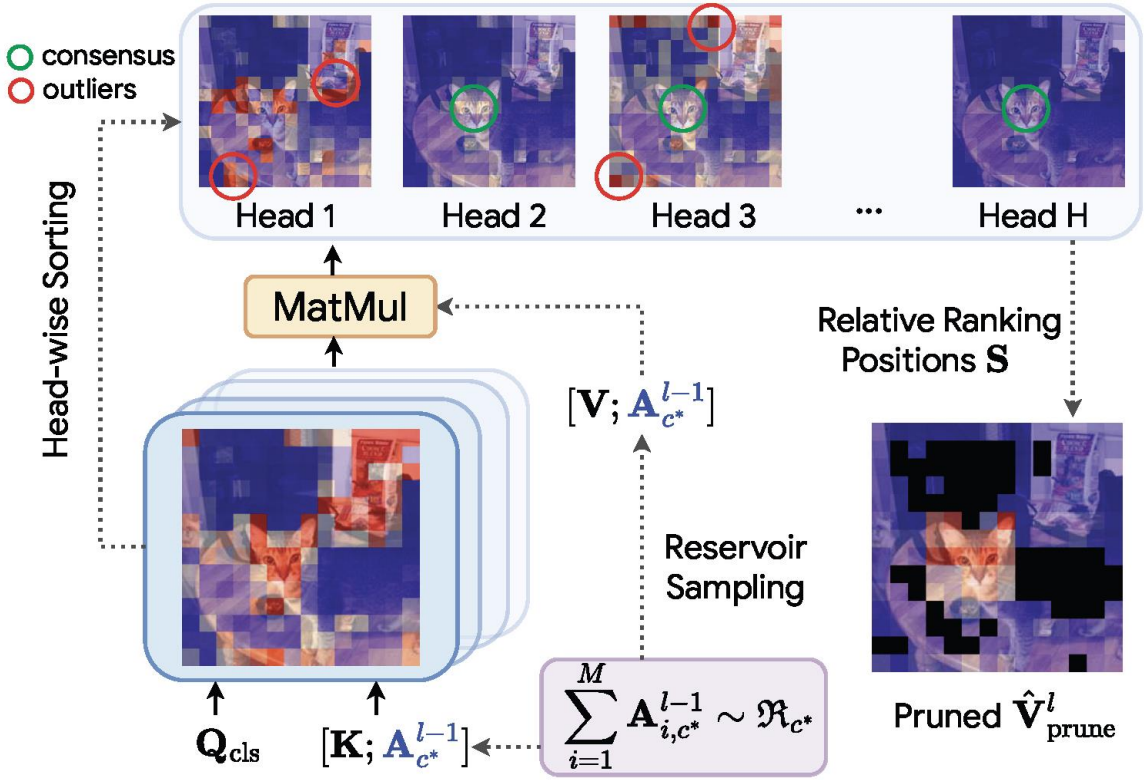
Select Visual Token Wisely

➢ **Our goal**

Training-free, efficient online adaptation → choose better tokens, not tune parameters.

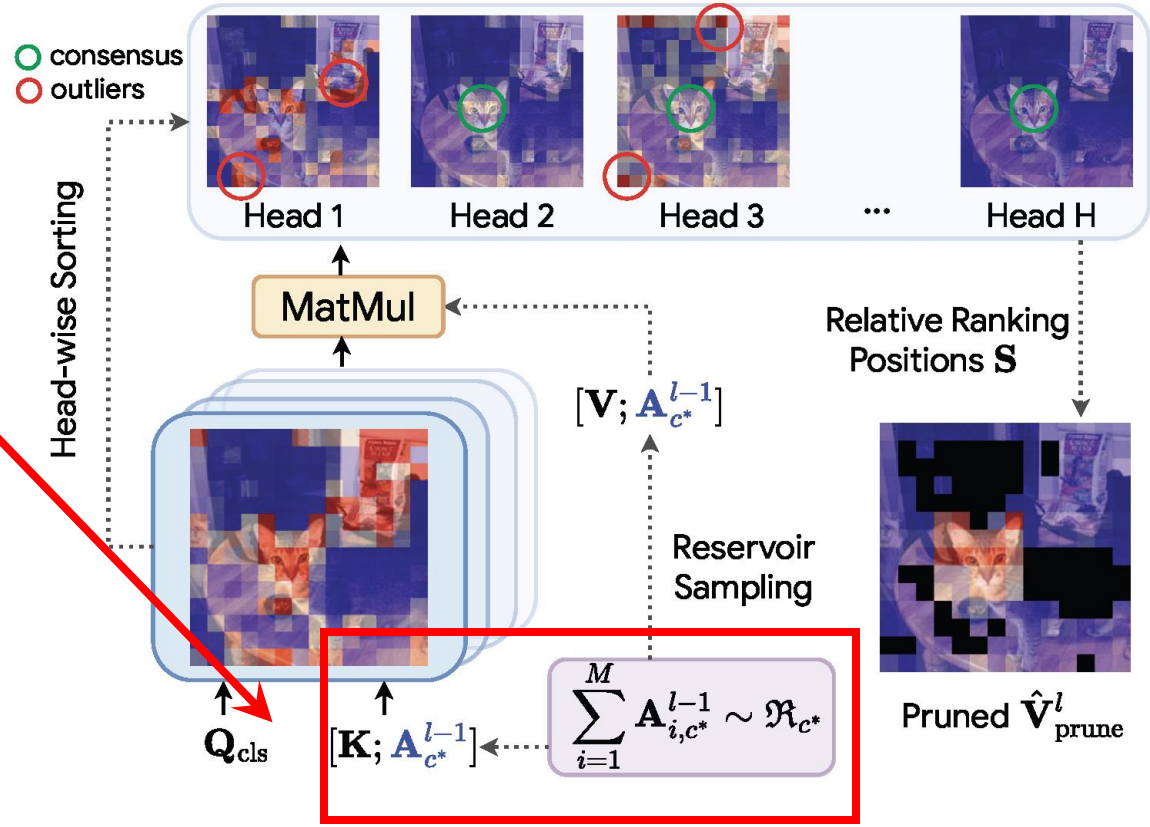
➢ **Vanilla pruning issues:**

1. <cls> may misalign → irrelevant tokens remain.
2. Averaging across heads → outlier heads dominate, hide useful info.

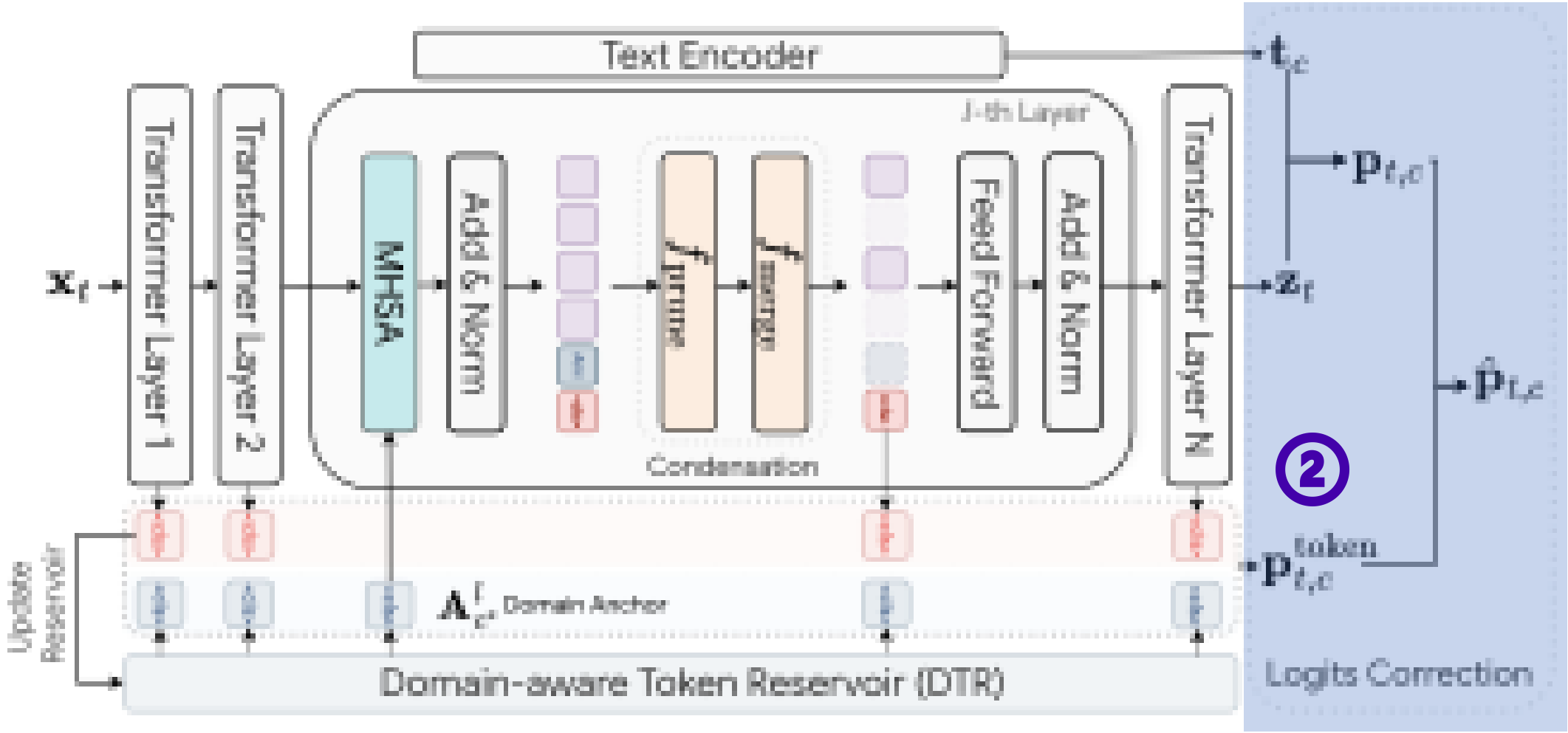


Select Visual Token Wisely

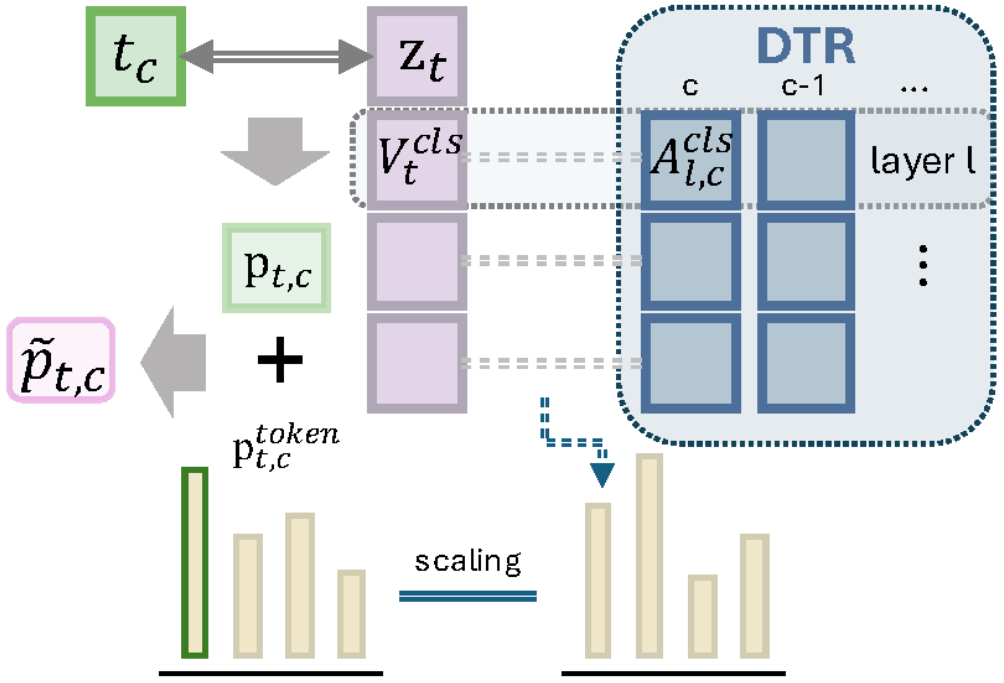
- **Step 1** — Domain-aware token evaluation: Use the sampled domain anchor to refine attention by
- **Step 2** — Cross-head consensus scoring: Compute $S_i^{head} = \frac{1}{H}\sum_h rank_h(i)$. Tokens that are consistently high across heads → robust to outliers.
- **Step 3** — Prune & Merge.



Anchor-Guided Logits Correction



Anchor-Guided Logits Correction



Basic idea: Use stored domain anchors as token-level classifiers to nudge predicted probabilities.

Mechanism:

- Compare current sample's <cls> tokens with stored anchors.
- Scale similarities across layers.
- Adjust logits accordingly.

Scaling coefficient $P = [\exp(\frac{l}{\beta})]_{l=1}^{L}$, β controls layer emphasis.
(small β → shallower layers; large β → deeper).

Experiments – Overall Performance

Table 1. Results on the cross-dataset benchmark using CLIP ViT-B/16, including the number of learnable parameters (L-Param.). * denotes the averaged GFLOPs. The best performance (aug-free) is **bolded**.

Method	Aug-free	Aircraft	Caltech101	Cars	DTD	EuroSAT	Flower102	Food101	Pets	SUN397	UCF101	Average	GFLOPs	L-Param.
CLIP	✓	23.22	93.55	66.11	45.04	50.42	66.99	82.86	86.92	65.63	65.16	64.59	17.59	0
CoOp	✗	18.47	93.70	64.51	41.92	46.39	68.71	85.30	89.14	64.15	66.55	63.88	17.59	2048
CoCoOp	✗	22.29	93.79	64.90	45.45	39.23	70.85	83.97	90.46	66.89	68.44	64.63	17.59	34,816
Tent	✓	8.97	93.39	62.69	39.78	20.85	61.23	83.70	87.76	65.30	66.93	59.06	17.59	40,960
SAR	✓	21.09	91.85	61.15	44.68	46.19	63.54	81.43	87.95	59.74	65.58	62.32	17.59	31,744
TPT	✗	24.78	94.16	66.87	47.75	42.44	68.98	84.67	87.79	65.50	68.04	65.10	1108.61	2048
Diff-TPT	✗	25.60	92.49	67.01	47.00	43.13	70.10	87.23	88.22	65.74	62.67	65.47	-	-
C-TPT	✗	23.90	94.10	66.70	46.80	48.70	69.90	84.50	87.40	66.00	66.70	65.47	1108.61	2048
MTA	✗	25.32	94.21	68.47	45.90	45.36	68.06	85.00	88.24	66.67	68.11	65.53	-	-
TDA	✓	23.91	**94.24**	**67.28**	**47.40**	58.00	71.42	86.14	88.63	**67.62**	70.66	67.53	17.59	0
EViT$_{R=0.9}$	✓	24.12	92.25	64.57	45.09	48.41	70.24	84.99	88.96	64.58	68.46	65.17	15.41	0
ToME$_{R=0.9}$	✓	24.66	92.49	63.10	44.92	48.64	69.22	85.04	87.90	64.22	68.62	64.88	15.31	0
ATS$_{R=0.9}$	✓	22.86	92.21	57.90	40.96	40.62	67.52	80.16	85.34	61.53	67.22	61.63	11.15*	0
EViT$_{R=0.7}$	✓	23.31	91.20	58.44	43.32	43.26	67.11	79.70	85.77	61.41	66.69	62.02	11.62	0
ToME$_{R=0.7}$	✓	22.26	90.79	55.48	42.32	40.12	64.11	79.36	84.19	60.66	63.97	60.33	11.45	0
ATS$_{R=0.7}$	✓	17.28	85.40	33.65	36.52	27.79	52.62	55.97	72.94	48.82	56.44	48.74	8.76*	0
TCA$_{R=0.9}$	✓	**24.87**	93.63	65.33	46.16	**70.43**	**73.33**	**85.31**	**89.53**	65.92	**72.38**	**68.69**	**15.45**$_{-12.2\%}$	0
TCA$_{R=0.7}$	✓	23.19	92.13	58.15	44.50	61.63	69.79	79.99	85.99	61.89	67.38	64.46	**11.69**$_{-33.5\%}$	0

Experiments – CLIP

Table 1. Results on the cross-dataset benchmark using CLIP ViT-B/16, including the number of learnable parameters (L-Param.). * denotes the averaged GFLOPs. The best performance (aug-free) is **bolded.**

Method	Aug-free	Aircraft	Caltech101	Cars	DTD	EuroSAT	Flower102	Food101	Pets	SUN397	UCF101	**Average**	**GFLOPs**	L-Param.
CLIP	✓	23.22	93.55	66.11	45.04	50.42	66.99	82.86	86.92	65.63	65.16	64.59	17.59	0
CoOp	✗	18.47	93.70	64.51	41.92	46.39	68.71	85.30	89.14	64.15	66.55	63.88	17.59	2048
CoCoOp	✗	22.29	93.79	64.90	45.45	39.23	70.85	83.97	90.46	66.89	68.44	64.63	17.59	34,816
Tent	✓	8.97	[illegible]	[illegible]	[illegible]	[illegible]	61.23	83.70	87.76	65.30	66.93	59.06	17.59	40,960
SAR	✓	21.09	91.85	61.15	44.68	46.19	63.54	81.43	87.95	59.74	65.58	62.32	17.59	31,744
TPT	✗	24.78	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	84.67	87.79	65.50	68.04	65.10	1108.61	2048
Diff-TPT	✗	25.60	92.49	67.01	47.00	43.13	70.10	87.23	88.22	65.74	62.67	65.47	-	-
C-TPT	✗	23.90	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	66.70	65.47	1108.61	2048
MTA	✗	25.32	94.21	68.47	45.90	45.36	68.06	85.00	88.24	66.67	68.11	65.53	-	-
TDA	✓	23.91	**94.24**	**67.28**	**47.40**	58.00	71.42	86.14	88.63	**67.62**	70.66	67.53	17.59	0
EViT$_{R=0.9}$	✓	24.12	92.25	64.57	45.09	48.41	70.24	84.99	88.96	64.58	68.46	65.17	15.41	0
ToME$_{R=0.9}$	✓	24.66	92.49	63.10	44.92	48.64	69.22	85.04	87.90	64.22	68.62	64.88	15.31	0
ATS$_{R=0.9}$	✓	22.86	92.21	57.90	40.96	40.62	67.52	80.16	85.34	61.53	67.22	61.63	11.15*	0
EViT$_{R=0.7}$	✓	23.31	91.20	58.44	43.32	43.26	67.11	79.70	85.77	61.41	66.69	62.02	11.62	0
ToME$_{R=0.7}$	✓	22.26	90.79	55.48	42.32	40.12	64.11	79.36	84.19	60.66	63.97	60.33	11.45	0
ATS$_{R=0.7}$	✓	17.28	85.40	33.65	36.52	27.79	52.62	55.97	72.94	48.82	56.44	48.74	8.76*	0
TCA$_{R=0.9}$	✓	**24.87**	93.63	65.33	46.16	**70.43**	**73.33**	**85.31**	**89.53**	65.92	**72.38**	**68.69**	15.45$_{-12.2\%}$	0
TCA$_{R=0.7}$	✓	23.19	92.13	58.15	44.50	61.63	69.79	79.99	85.99	61.89	67.38	64.46	11.69$_{-33.5\%}$	0

- ✓ Batch size = 1
- ✓ Lower FLOPs!
- ✓ No parameter updates required!

Experiments - SigLIP

Table 3. Improvements in Cross-Dataset benchmark over SigLIP and SigLIP v2 inference on ViT-B/16.

Method	Aircraft	Caltech101	Cars	DTD	EuroSAT	Flower102	Food101	Pets	SUN397	UCF101	**Average**
SigLIP	40.50	97.44	90.71	62.83	39.86	84.13	89.06	93.10	69.65	70.84	73.81
+TCA$_{R=0.9}$	+0.74%	+0.58%	+0.12%	+2.44%	+21.12%	+1.39%	-0.17%	+0.56%	+2.02%	+3.40%	+2.21%
SigLIP v2	50.05	97.61	93.36	62.41	42.09	85.30	90.14	94.88	73.34	72.09	76.13
+TCA$_{R=0.9}$	+0.42%	+0.25%	+0.11%	+2.76%	+11.17%	+0.39%	-0.04%	+0.02%	+0.64%	+3.22%	+1.32%

Input Image
Layer 3
Layer 6
Layer 9

Paper

GitHub

Thanks for listening ☺