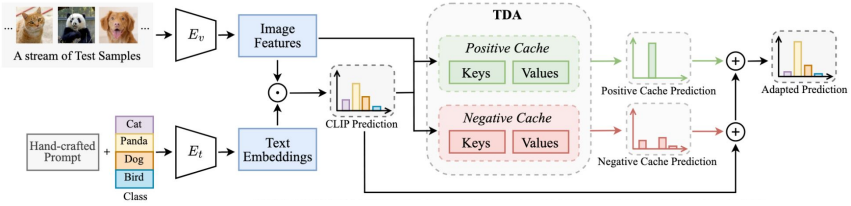
Efficient Test-Time Adaptation

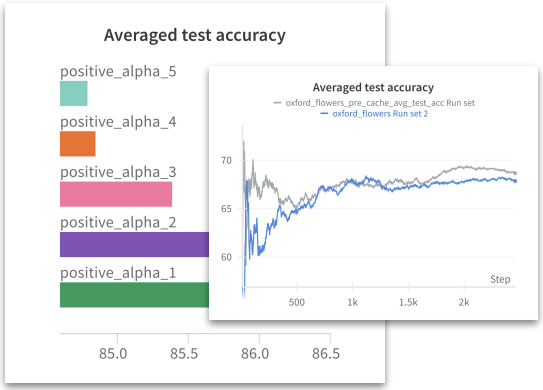
Final presentation

Juan Camacho, Andrea De Carlo, Samuele Bolotta

Outline



Efficient TTA



Experiments

Our proposal

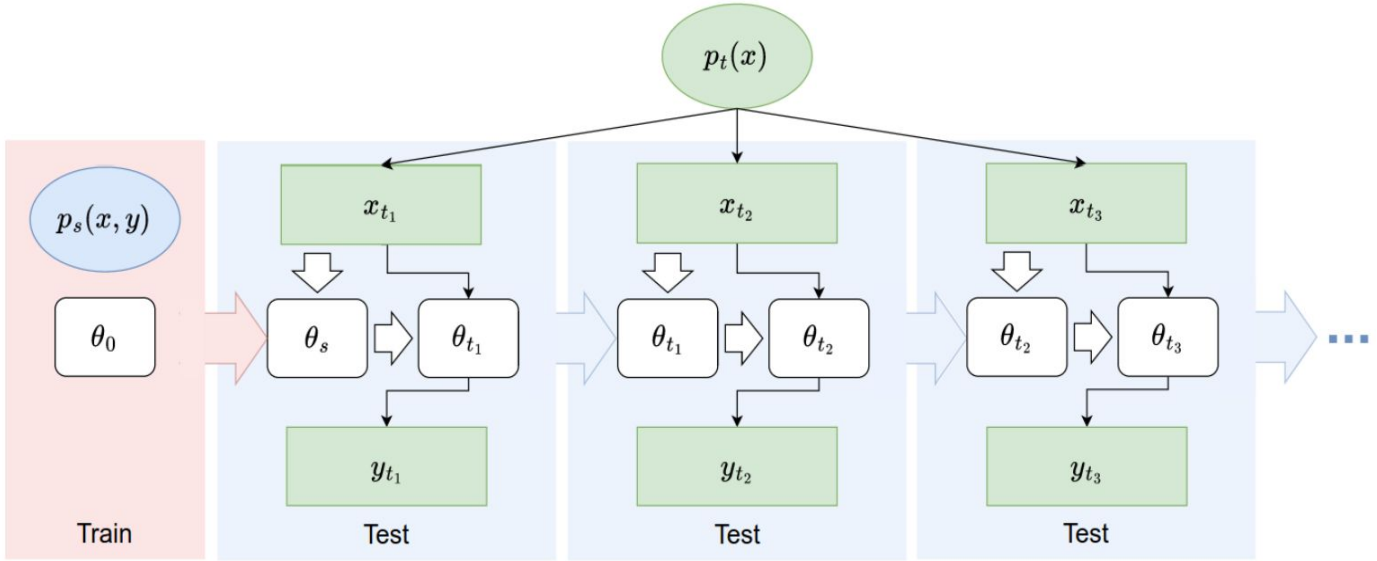
Background

Efficient Test-Time Adaptation

Models often face situations where test data differs significantly from training data

Background

Efficient Test-Time Adaptation

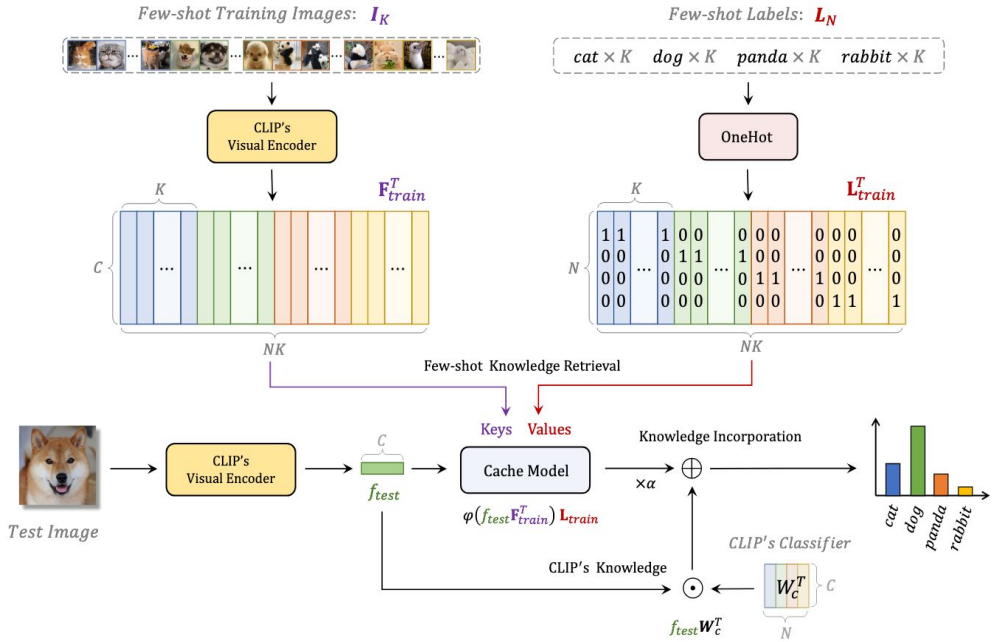


Test-Time Adaptation (TTA) addresses distribution shifts between training and test data.

Xiao e Snoek "Beyond Model Adaptation at Test Time: A Survey"

Background

TIP-Adapter

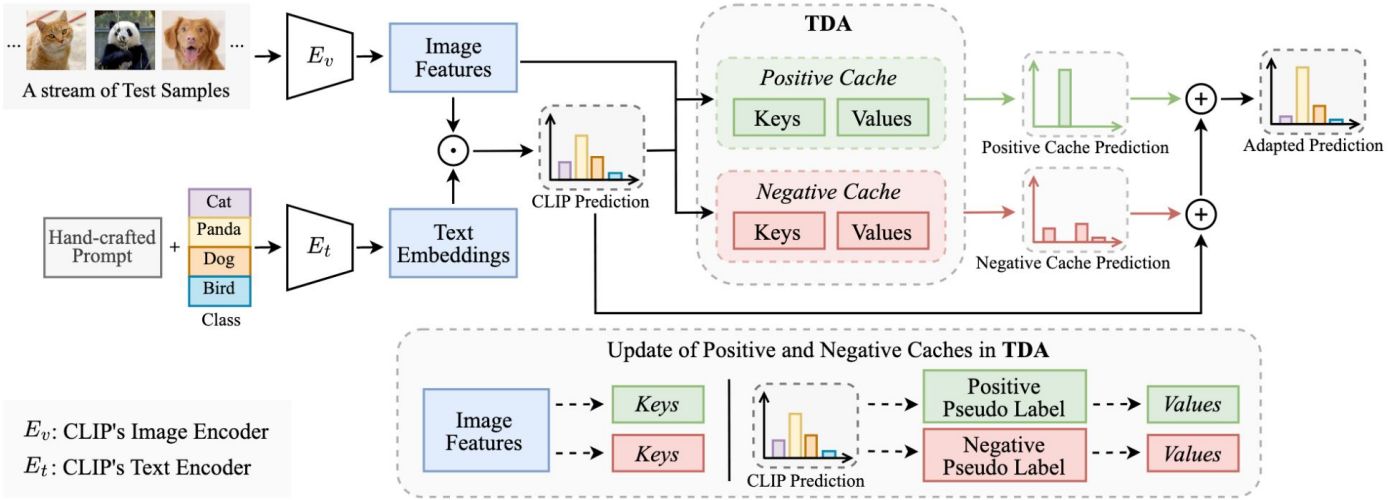


- Few-shot training data creates feature "keys" and one-hot label "values" in the cache
- Tip-Adapter skips backpropagation

Zhang et al. “Tip-Adapter: Training-free Adaption of CLIP for Few-shot Classification”

Background

Training-free Dynamic Adapter (TDA)



- The positive cache stores confident predictions and their features
- The negative cache identifies what a sample is not, reducing errors from uncertain predictions

Karmanov et al. “Efficient Test-Time Adaptation of Vision-Language Models”

Experiments

Dataset Statistics

Dataset	Classes	Test Size
ImageNet	1,000	50,000
ImageNet-V2	1,000	10,000
ImageNet-S	1,000	50,000
ImageNet-A	200	7,500
ImageNet-R	200	30,000
Aircraft	100	3,333
Caltech101	100	2,465
Cars	196	8,041
DTD	47	1,692
EuroSAT	10	8,100
Flowers102	102	2,463
Food101	101	30,300
Pets	37	3,669
SUN397	397	19,850
UCF101	101	3,783

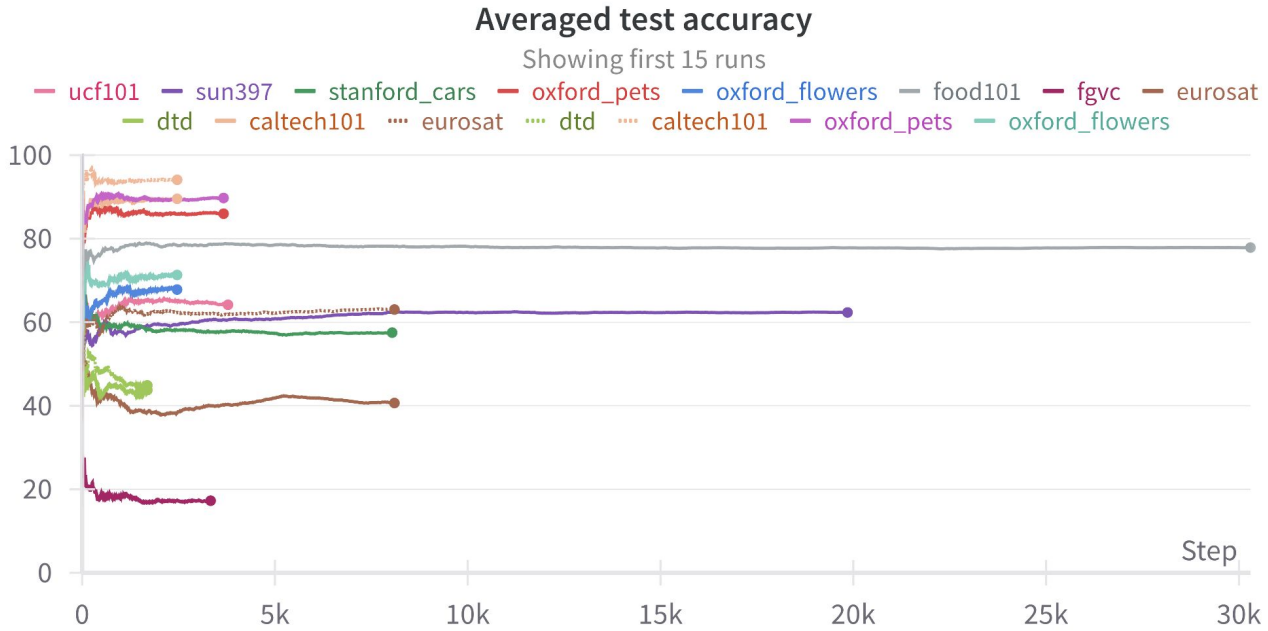
- **Out-of-Distribution (OOD) datasets**:
  These include ImageNet as the baseline and variants like ImageNet-A and ImageNet-R, which test robustness to adversarial and stylized shifts

—

- **Cross-Domain Datasets**:
  These explore adaptability to unique tasks, such as Aircraft and Cars, while datasets like Flowers102 focus on fine-grained classification, and broader ones like SUN397 test general scene adaptability

Benchmark CD

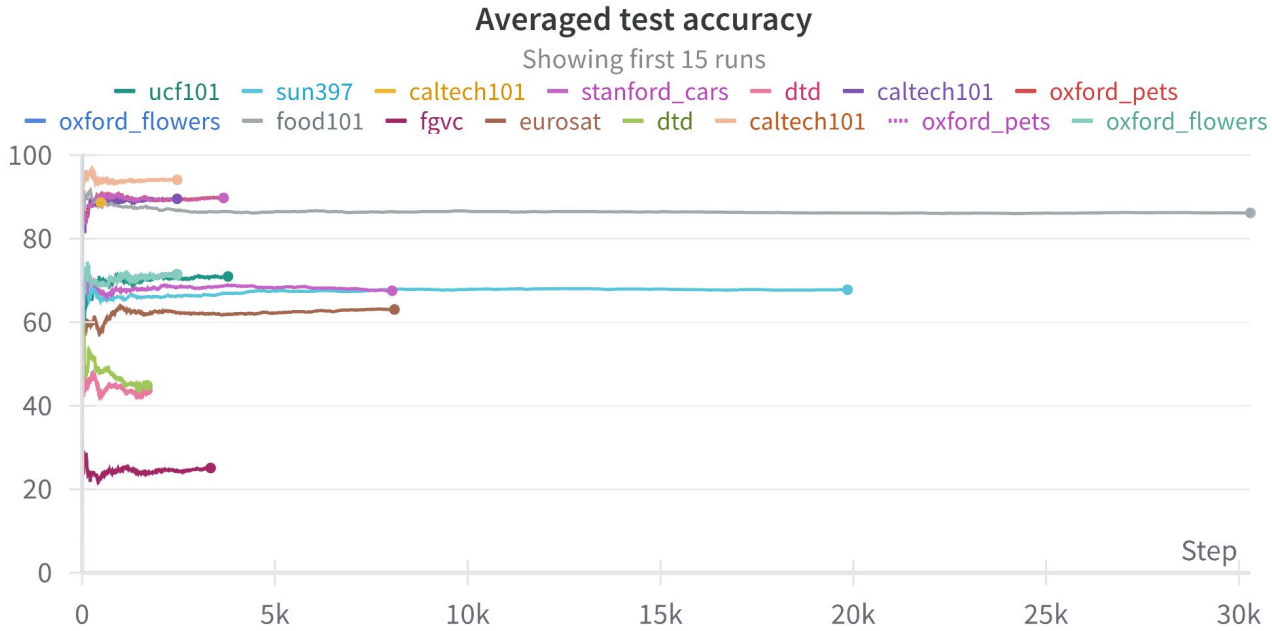
ResNet50



Method	Aircraft	Caltech101	Cars	DTD	EuroSAT	Flower102	Food101	Pets	SUN397	UCF101	Average
TDA (Ours)	**17.61**	**89.70**	57.78	**43.74**	**42.11**	**68.74**	77.75	86.18	62.53	**64.18**	**61.03**

Benchmark CD

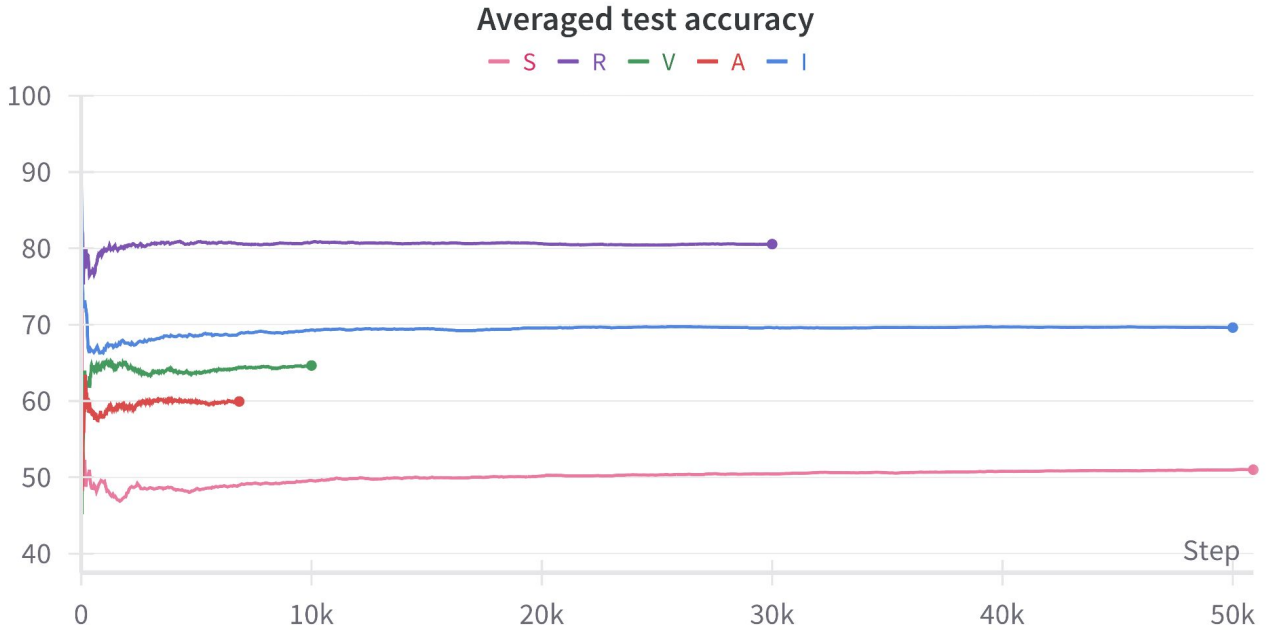
ViT-B/16



Method	Aircraft	Caltech101	Cars	DTD	EuroSAT	Flower102	Food101	Pets	SUN397	UCF101	Average
TDA (Ours)	23.91	**94.24**	**67.28**	47.40	**58.00**	**71.42**	86.14	88.63	**67.62**	**70.66**	**67.53**

Benchmark OOD

ResNet-50



Method	ImageNet	ImageNet-A	ImageNet-V2	ImageNet-R	ImageNet-S	Average	OOD Average
TDA (Ours)	61.35	30.29	55.54	**62.58**	**38.12**	**49.58**	**46.63**

i.i.d. vs non-i.i.d.

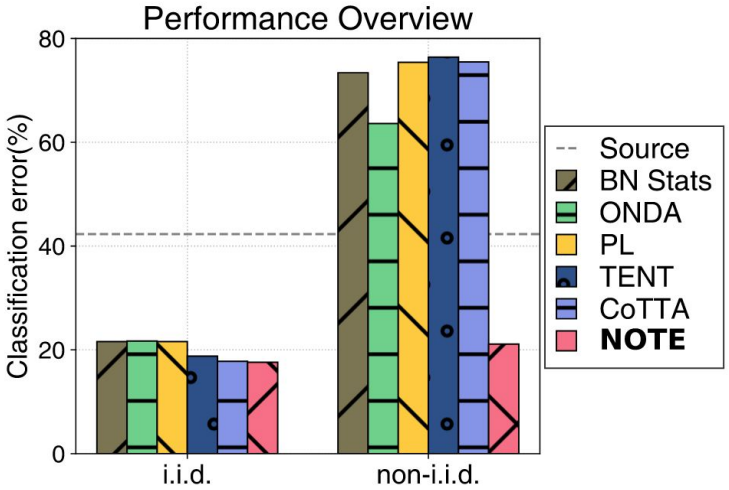
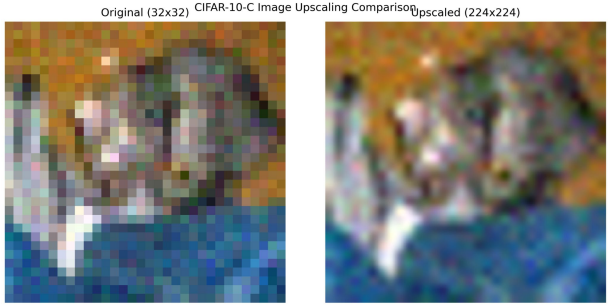


Figure 2: Average classification error (%) of existing TTA methods and our method (NOTE) on CIFAR10-C [13]. The error rates significantly increase under the non-i.i.d. setting compared with the i.i.d. setting. Lower is better.

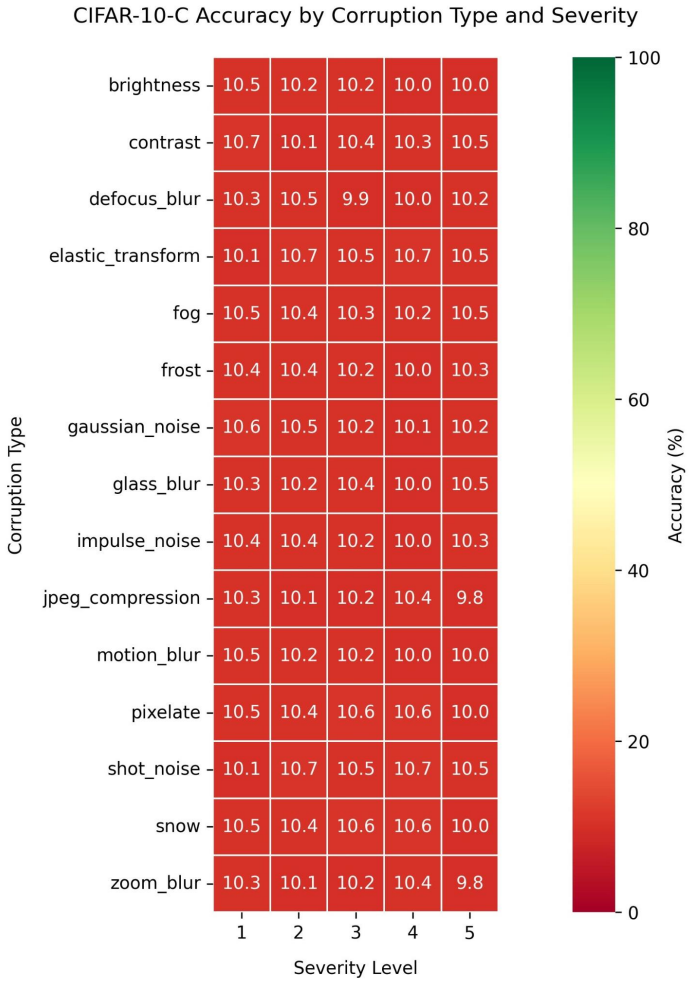
CIFAR10-C

ResNet-50

Gaussian Noise

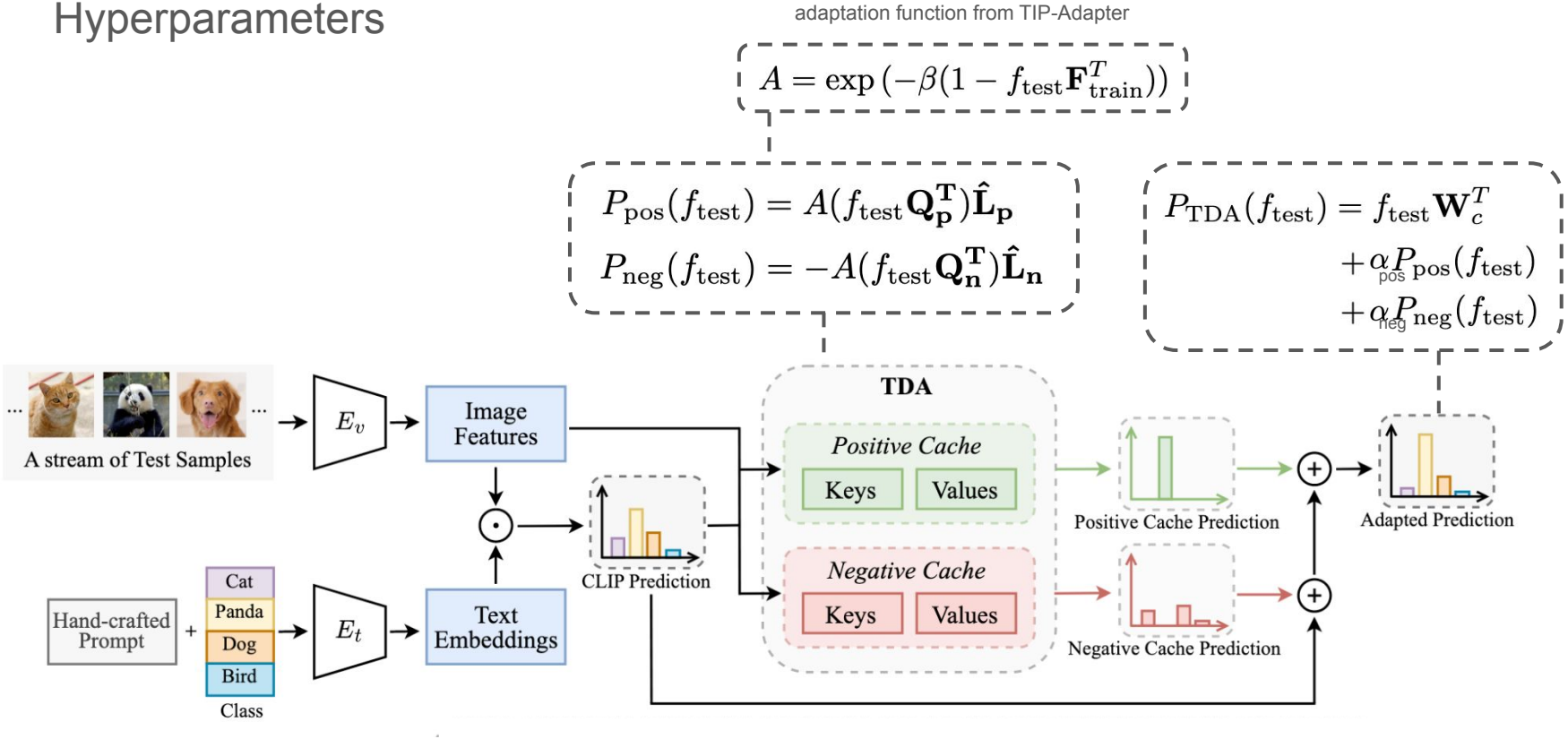


Defocus blur



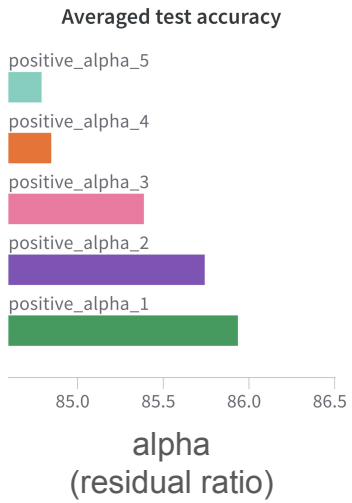
Ablation studies

Hyperparameters

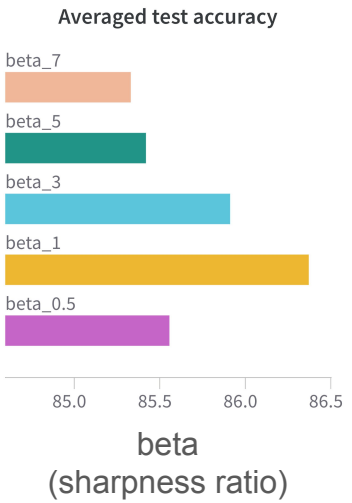


Ablation studies

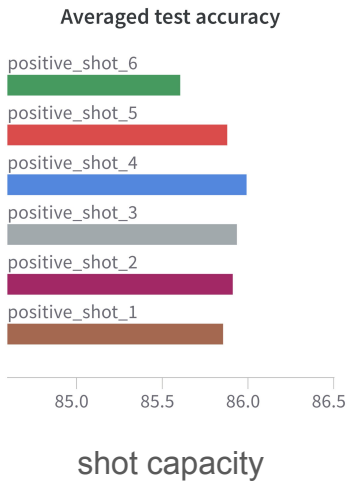
Positive cache parameters



alpha
(residual ratio)



beta
(sharpness ratio)



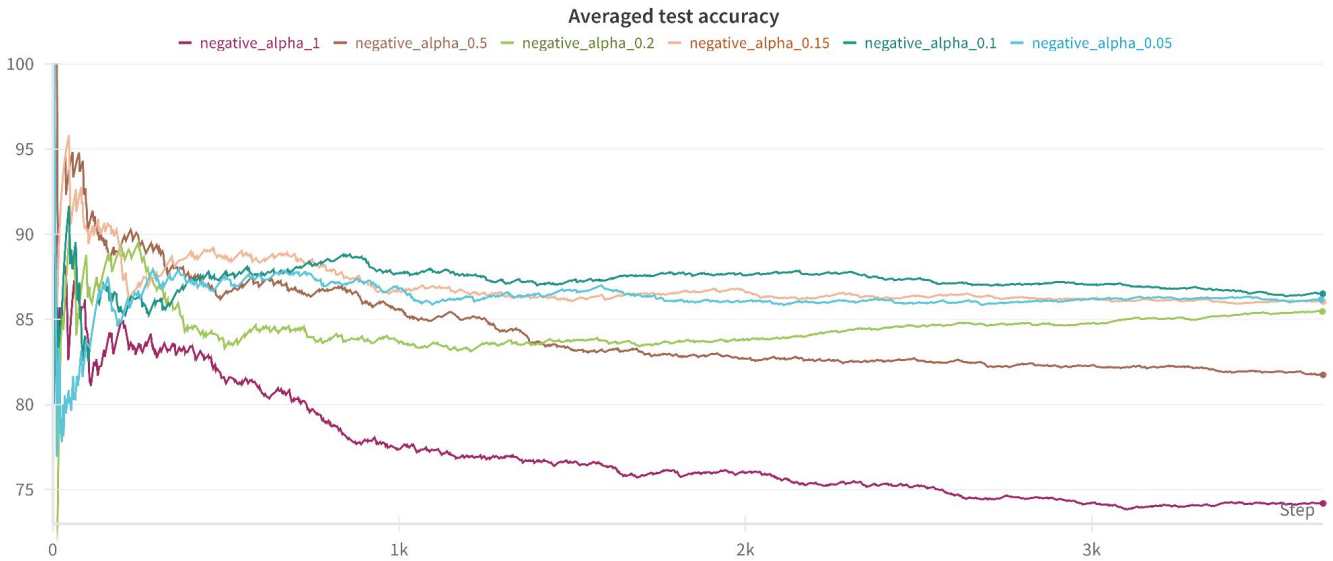
shot capacity

- accuracy improves with ***alpha*** going toward 1, means that **prior knowledge from CLIP** and **few-shot knowledge from cache** models are equally important
- ***beta* follows a non linear pattern**, it also has limited impact w.r.t. the other parameters
- larger ***size of cache*** intuitively should lead to better results, but this is not the case, probably due to the amount of classes and the distribution of the dataset considered

tested on Oxford_pets dataset

Ablation studies

Negative cache parameters

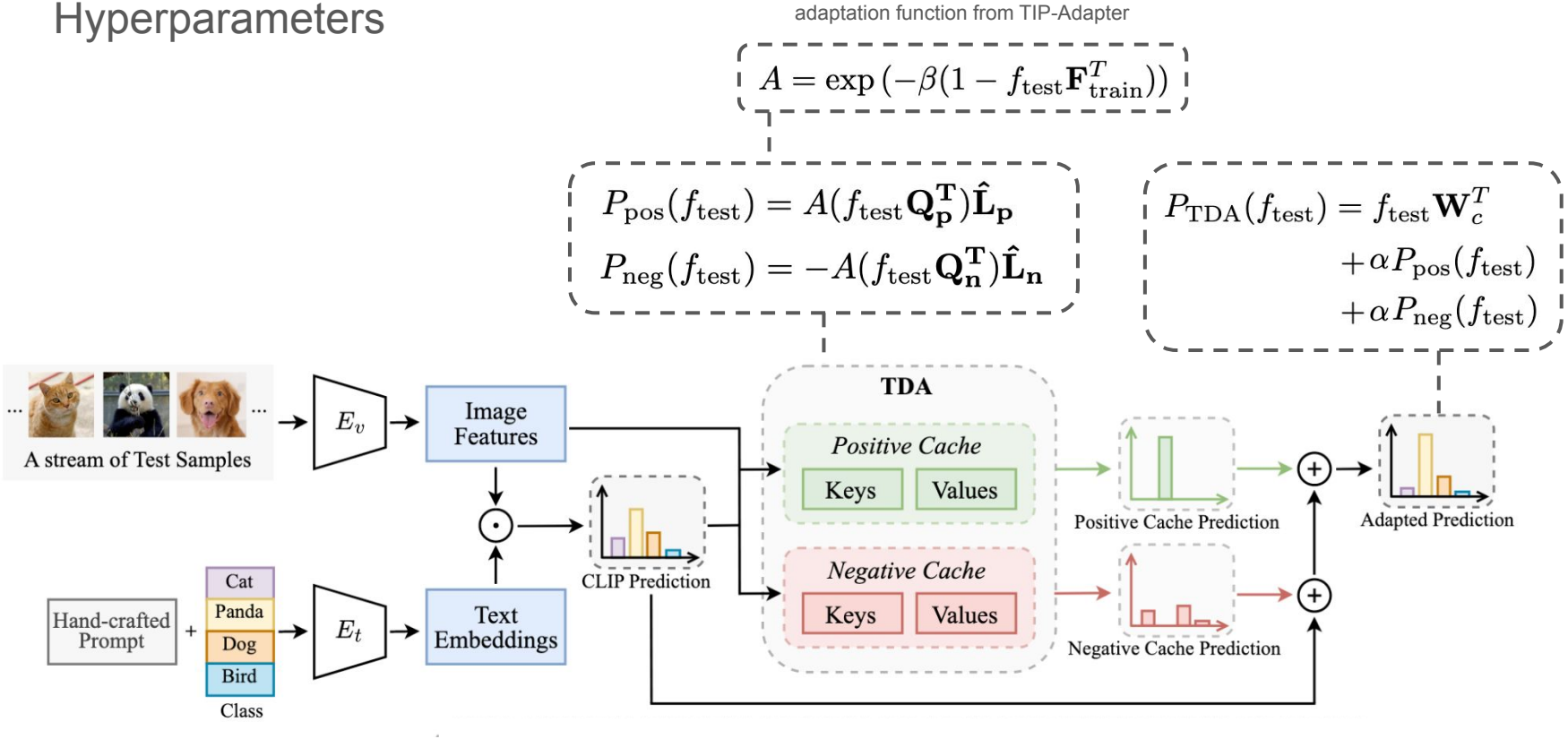


this is a plot showing how a strong presence of the negative cache (*negative_alpha* = 1) can worsen the performance

tested on Oxford_pets dataset

Ablation studies

Hyperparameters



Ablation studies

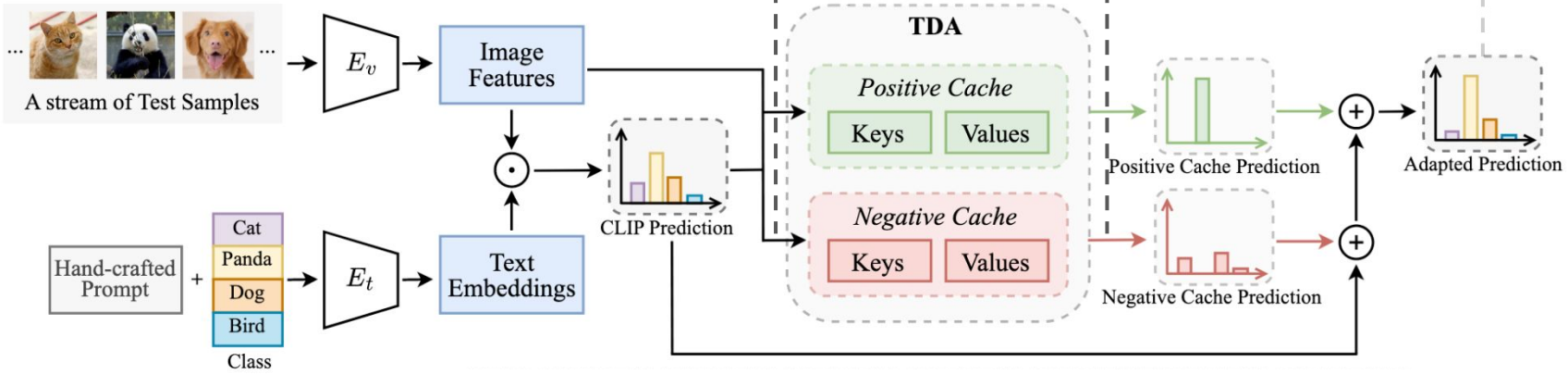
Hyperparameters

Entropy thresholds

$$\gamma(f_{\text{test}}) : \tau_l < \mathrm{H}(f_{\text{test}}\mathbf{W}_c^T) < \tau_h$$

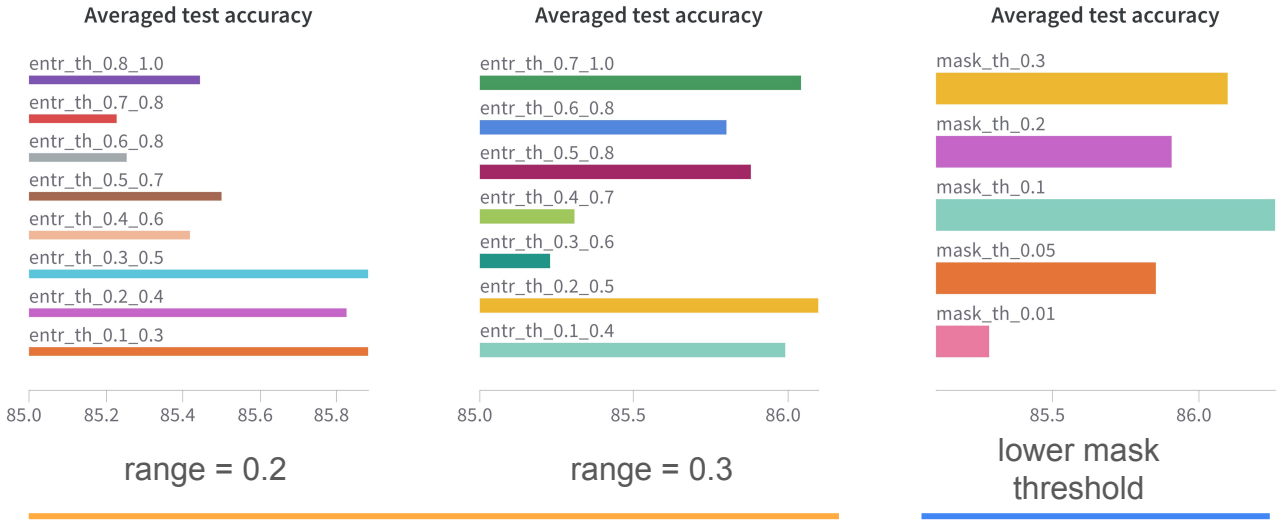
Mask lower threshold

$$\hat{\mathbf{L}}_{\mathbf{n}} = -\mathbb{1}[p_l < P(\mathbf{Q_n})]$$



Ablation studies

Negative cache parameters

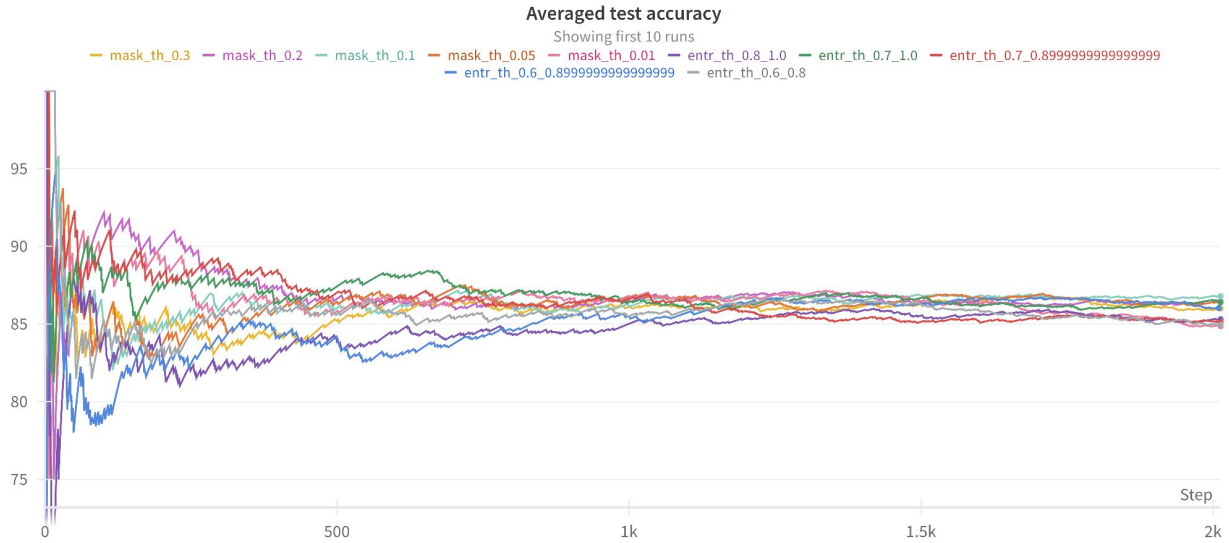


- different **entropy ranges** show that both using **low or high confidence range can improve performances**
- for the **mask threshold** we note that the performance deteriorates when it is less than 0.1, because the **confident classes with low probability should not be included in negative pseudo labels**

tested on Oxford_pets dataset

Ablation studies

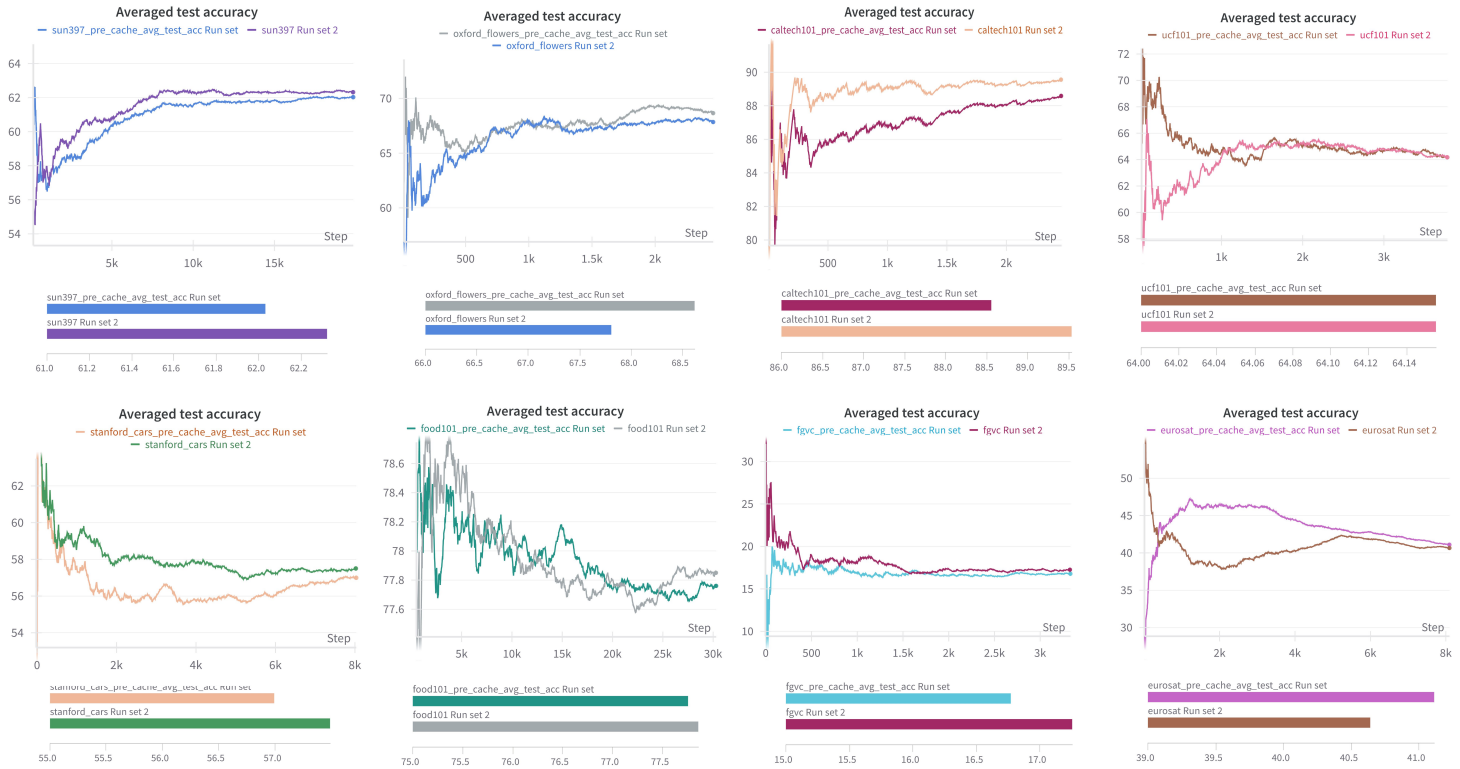
Number of seen samples



- average accuracy **oscillates because of few samples seen**
- however it converges after some iterations, **showing also less cache updates**

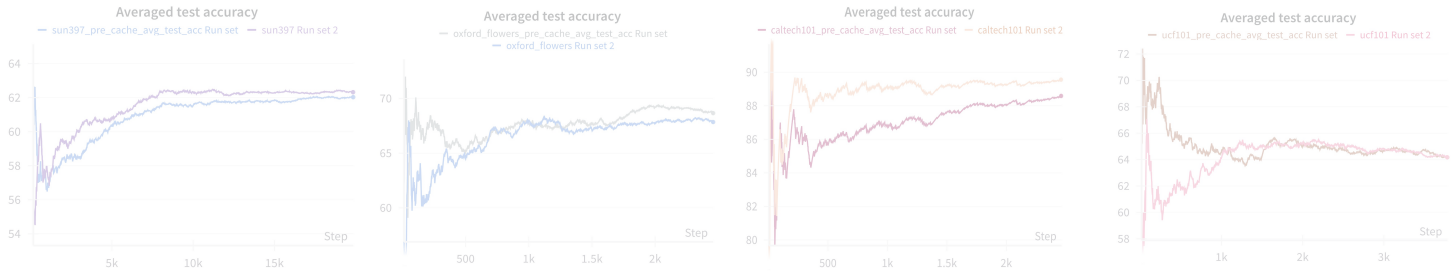
tested on Oxford_pets dataset

Ablation studies: predictions made **before vs after** cache update



upper bar with the cache update after the prediction, lower bar is the standard approach

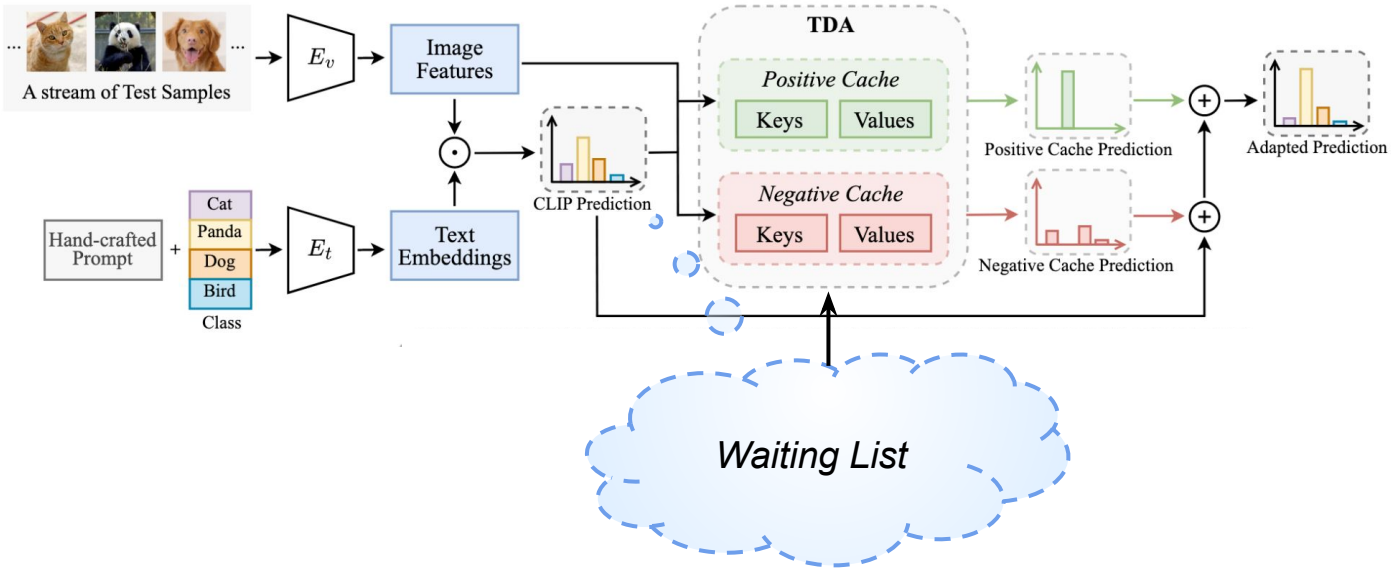
Ablation studies: predictions made **before vs after** cache update



this leads mostly to equal or less performances,
probably **due to the amount of samples and classes of the specific dataset**

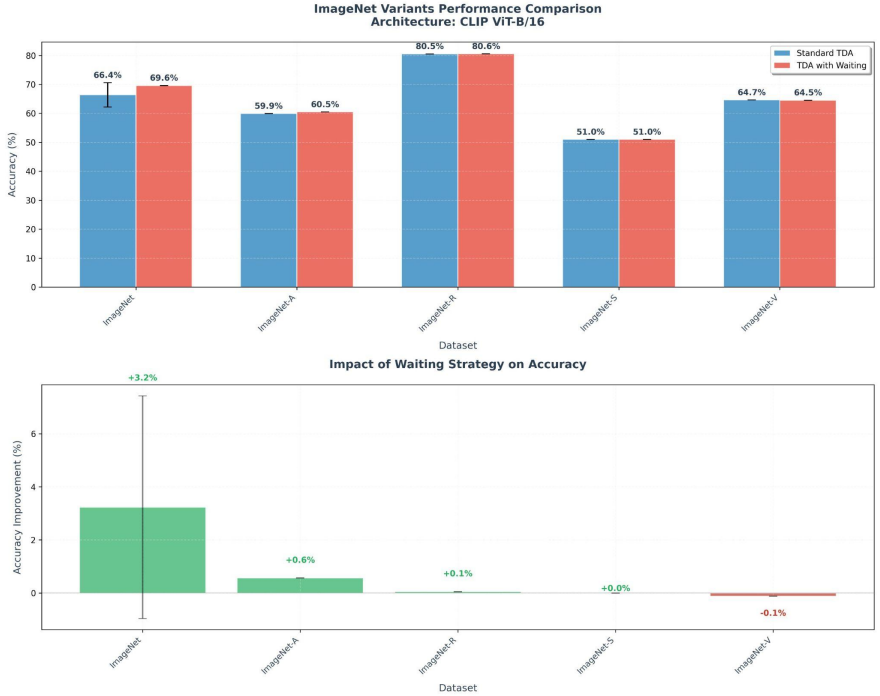
upper bar with the cache update after the prediction, lower bar is the standard approach

Our idea: Waiting List



- Confident (low-entropy) samples are processed first, while high-entropy samples are deferred to a waiting list for later refinement
- **Hypothesis**: Revisiting uncertain samples after refining the cache with confident predictions leads to more informed decisions.

Our idea: Waiting List

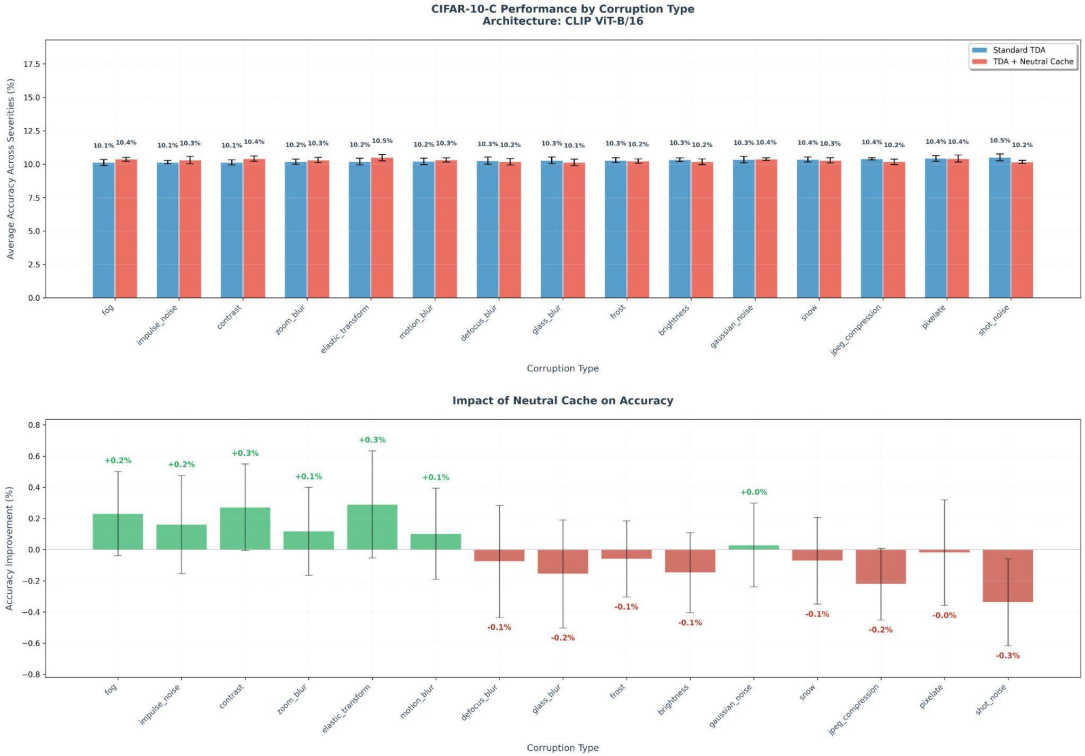


- **Data shift** between datasets
- **CLIP overfitting** to ImageNet
- For ImageNet the cache captures more relevant features that help the samples in the waiting list

tested on ImageNet dataset

CIFAR10-C

ResNet-50



- **Minimal improvements** over some corruptions
- Delaying classification of samples encourages a **better distribution of classes** in the cache for some corruptions

Conclusions

and TakeAways

what we did:

- **Benchmark** on different datasets, both OOD and CD with failure cases on CIFAR-10-C (non-iid data stream)
- We evaluated **how the performance changed** w.r.t. changing **hyperparameters** and the orders of data presented considering **budget-aware constraints**
- We tried to mitigate the issues adding a **Waiting List** to the model, which improved performance on ImageNet but didn't help on more challenging dataset like as CIFAR10-C

Future work:

- Further experiments to test if the waiting list is useful or not:
  - Creating a **synthetic dataset** that contains elements representative of basic concepts which are in relation with harder to understand concepts. Same as when we start learning basic and general concepts, and evolve to more fine grained understanding of the details.

Thanks for
your attention!