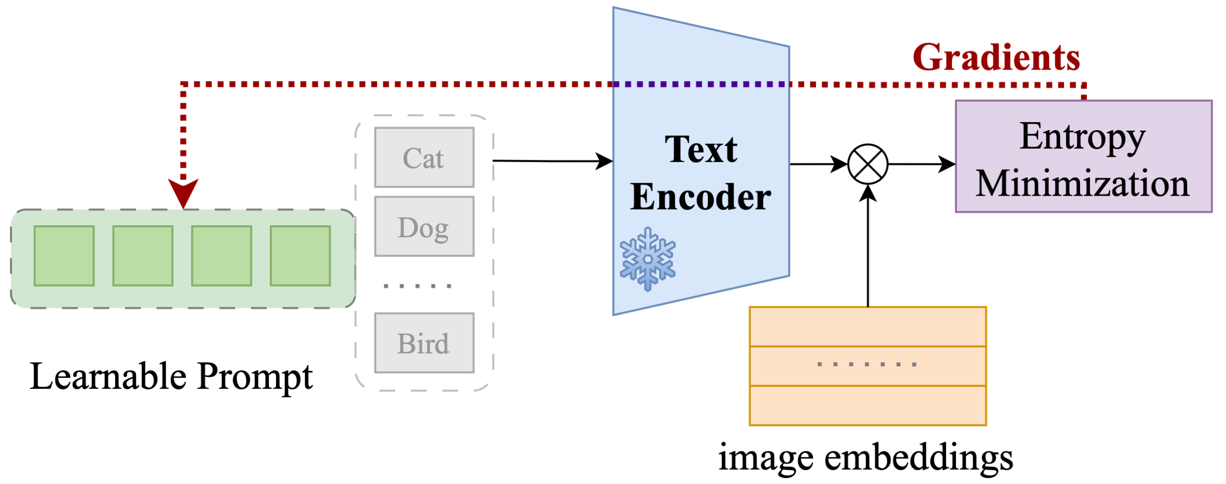
Just Shift It: Test-Time Prototype Shifting for Zero-Shot Generalization with Vision-Language Models

Elaine Sui, Xiaohan Wang, Serena Yeung-Levy

We introduce the **Test-Time Prototype Shifting** (**TPS**) framework, a test-time adaptation method using VLMs by dynamically learning <u>shift vectors</u> for each prototype based solely on the given test sample.

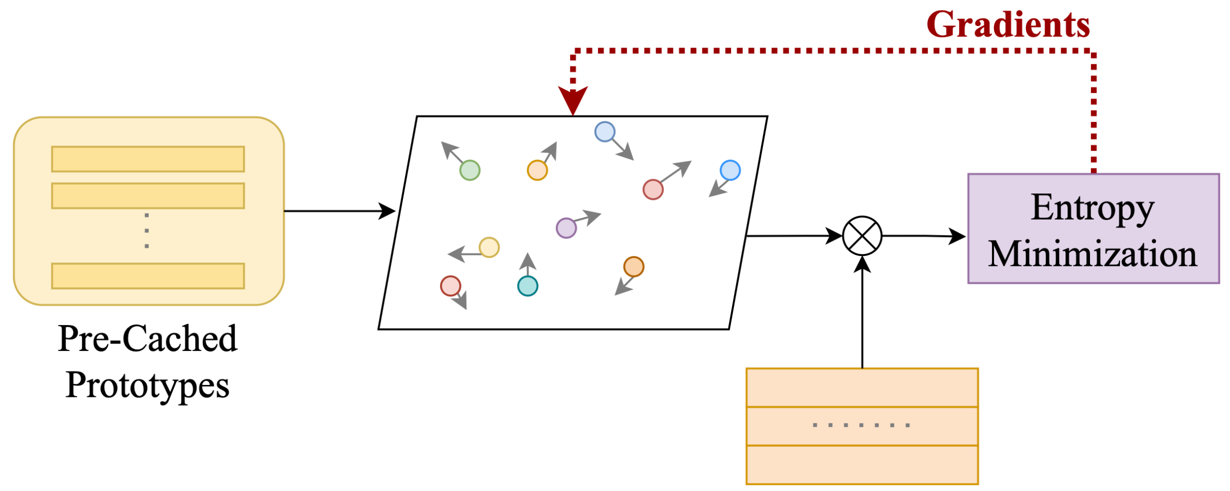
Introduction

Many state-of-the-art TTA methods require backpropagation through the text encoder, which incurs high computational and memory demands.



Test-Time Prompt Tuning (TPT)

Ultimately, TTA methods indirectly perturb the prototypes in the feature space. We propose to **learn this shift directly**.

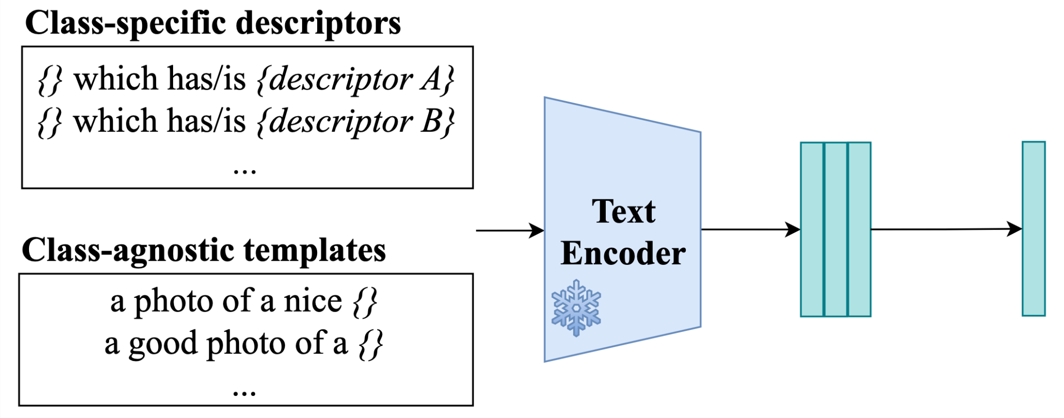


Test-Time Prototype Shifting (Ours)

Method

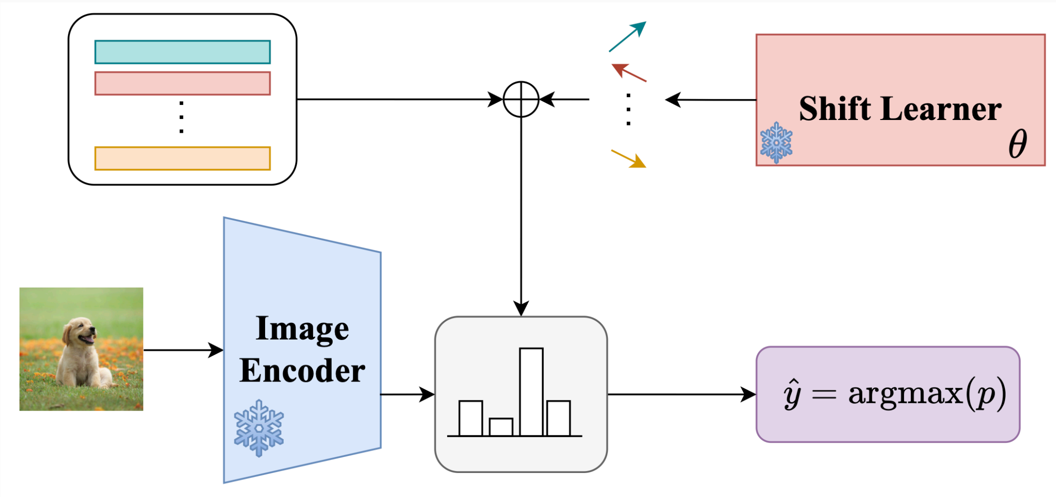
1. Prototype Generation

Generate class-conditioned descriptors via GPT-4 and compute the mean of the descriptor and CLIP ImageNet templated prompt embeddings.

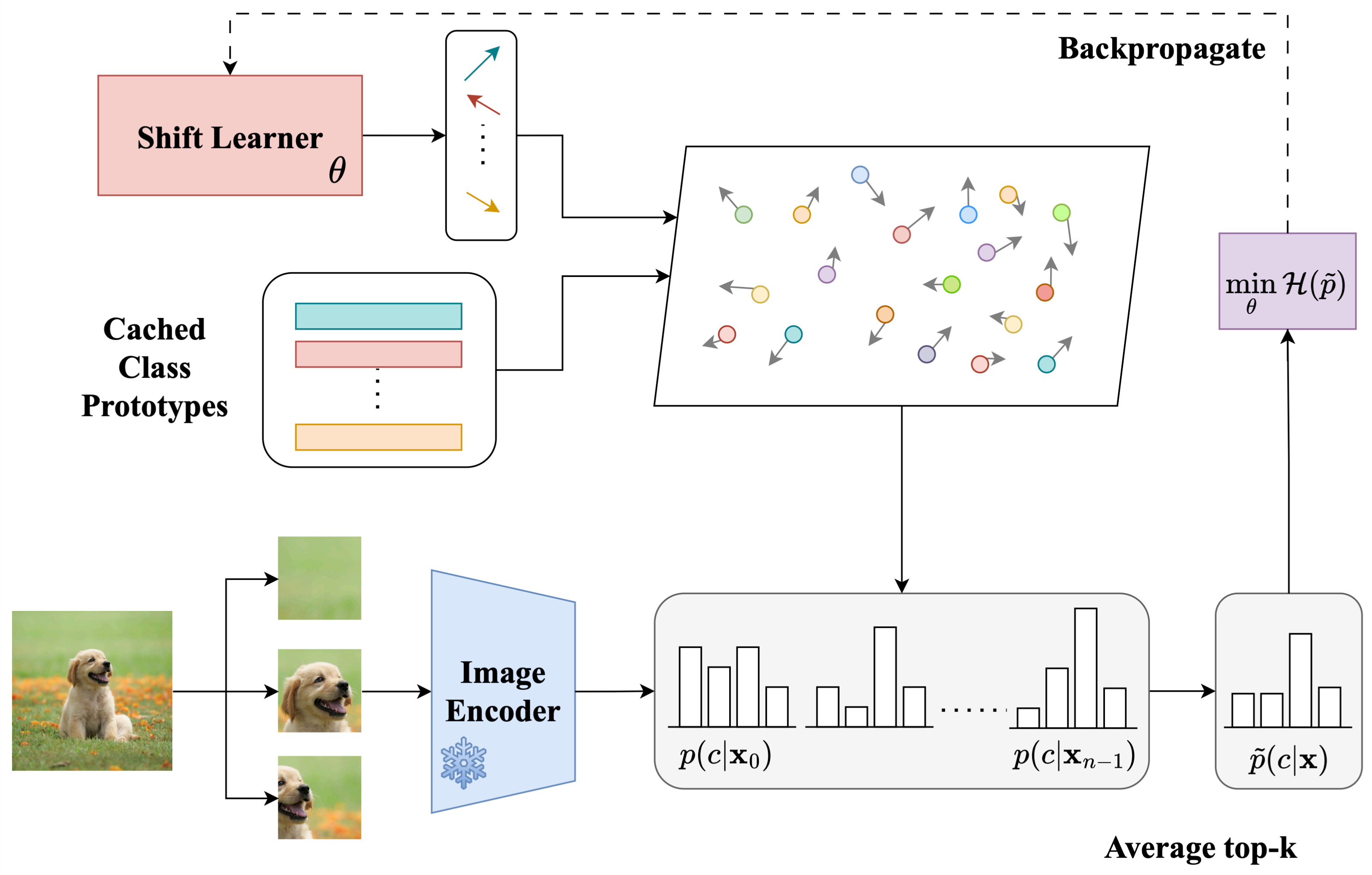


3. Test-Time Inference

Compute the final prediction for the shifted class prototypes and the original image embedding with CLIP similarity.



2. Test-Time Shift Tuning



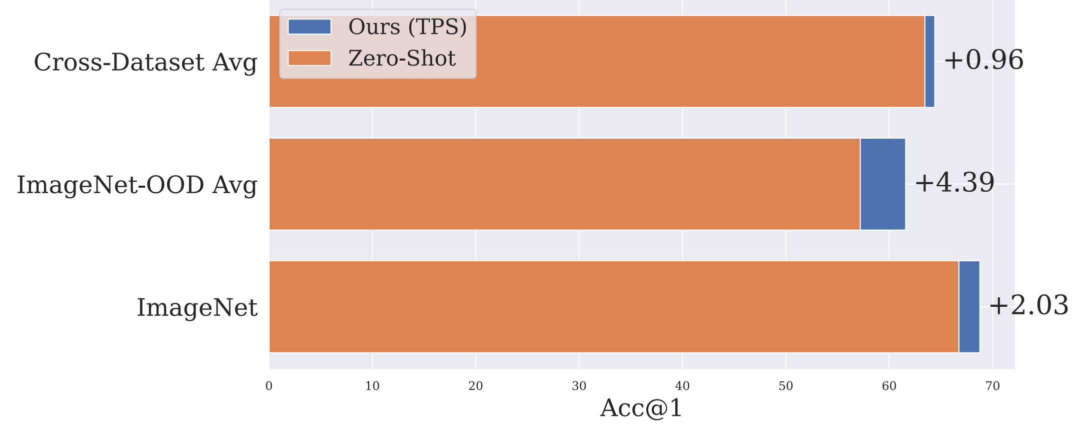
Tune the Shift Learner to generate small perturbations to the class prototypes to close the gap between source and target distributions.

Marginal entropy of the CLIP similarities of the shifted prototypes and augmented image embeddings is minimized.

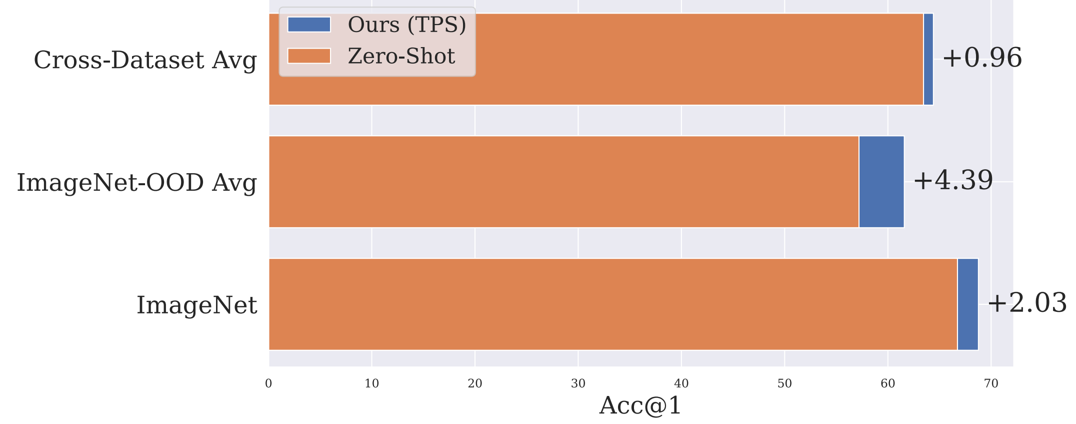
Main Results

Test-Time Prototype Shifting significantly improves zero-shot classification in natural distribution shift and cross-dataset generalization benchmarks.

Prototypes initialized with ***"a photo of a {classname}"***



Prototypes initialized with ***mean pool of CLIP templates and descriptors***



Efficiency Analysis

Our method (TPS) runs **more than 10x as fast and uses 10x less memory** than TPT on an A6000 GPU.