STAR: Test-Time Adaptation Can Enhance Universal Prompt Learning for Vision-Language Models

Supplementary Material

A. Preliminaries

CLIP-based OOD Detection for VLMs. Vision-language models (VLMs) have demonstrated promising performance in image classification compared to traditional vision-only representation methods. Among them, CLIP [6] is widely recognized as a pioneering VLM, leveraging self-supervised contrastive learning to jointly train an image encoder $\mathcal{I} : \mathbf{x} \to \mathbb{R}^d$ (e.g., ViT [1]) and a text encoder $\mathcal{T} : \mathbf{t} \to \mathbb{R}^d$ (e.g., Transformer [7]) in a shared representation space. After pretraining on millions of image-text pairs, CLIP acquires a unified image-text alignment space, thereby enabling strong zero-shot capabilities and making it particularly effective for OOD detection. Building upon CLIP, MCM [4] introduces a concept matching score by maximizing the similarity between an image feature $\mathcal{I}(\mathbf{x})$ and a text concept $\mathcal{T}(\mathbf{t}_k)$:

$$S_{\text{MCM}}(\mathbf{x}) = \max_k \frac{e^{s(\mathbf{x},\mathbf{t}_k)/\tau}}{\sum_k e^{s(\mathbf{x},\mathbf{t}_k)/\tau}}, \tag{1}$$

where $s(\mathbf{x}, \mathbf{t}_k) = \frac{\mathcal{I}(\mathbf{x})\cdot\mathcal{T}(\mathbf{t}_k)}{\|\mathcal{I}(\mathbf{x})\|\cdot\|\mathcal{T}(\mathbf{t}_k)\|}$ measures cosine similarity between the image feature $\mathcal{I}(\mathbf{x})$ and the text concept $\mathcal{T}(\mathbf{t}_k)$ and τ is the temperature. Subsequently, we could detect OOD samples by

$$G(x;\lambda) = \begin{cases} 1, & S_{\text{MCM}}(\mathbf{x}) \geq \lambda \\ 0, & S_{\text{MCM}}(\mathbf{x}) < \lambda \end{cases}, \tag{2}$$

in which λ is the threshold. To better exploit both global and local visual features and handle multi-object scenarios, GL-MCM [5] incorporates local patch-level scores to calibrate the global detection score:

$$S_{\text{GL-MCM}}(\mathbf{x}) = S_{\text{MCM}}(\mathbf{x}) + \max_{i,k} \frac{e^{s(\mathbf{x}_i,\mathbf{t}_k)/\tau}}{\sum_k e^{s(\mathbf{x}_i,\mathbf{t}_k)/\tau}}, \tag{3}$$

where $\mathbf{x}_i$ corresponds to the ith patch of the image $\mathbf{x}$. MCM and GL-MCM provide two distinct score functions for OOD detection, and selecting an appropriate threshold λ remains a critical concern. Based on the observation that OOD scores often exhibit a bimodal distribution, OWTTT [3] proposes an adaptive thresholding strategy by optimizing the intra-class variance for a better separation.

Prompt Learning for OOD Detection. Recalling that CLIP provides a unified representation space for both images and texts, the textual prompt can be represented as $\mathbf{t} = \{\mathbf{w}_1, \cdots, \mathbf{w}_N, \mathbf{c}\}$, where $\{\mathbf{w}_i\}_{i=1}^N$ denote learnable context vectors optimized during prompt learning, and $\mathbf{c}$ corresponds to its class name. LoCoOp [5] performs prompt learning by extracting ID-irrelevant nuisances $\tilde{\mathbf{x}}$ from ID images for OOD regularization, whose optimization objective is defined as:

$$\mathcal{L}_{\text{LoCoOp}} = \mathbb{E}_{\mathcal{X},\mathcal{Y}} \left[\mathcal{L}_{\text{CE}}(p(\mathbf{y}|\mathbf{x},\mathbf{w}),\mathbf{y}) - \lambda_L \, \text{H}(p(\mathbf{y}|\tilde{\mathbf{x}},\mathbf{w}))\right], \tag{4}$$

in which $\mathcal{L}_{\text{CE}}$ represents the cross-entropy loss and $\text{H}(p(\mathbf{y}|\tilde{\mathbf{x}},\mathbf{w}))$ measures the entropy of the predictive distribution induced by spurious OOD features extracted from ID images. λ_L is the balancing hyperparameter. Notably, the decomposition of foreground and background in images is often imperfect, while predictions with low uncertainty tend to have an even stronger positive impact on OOD detection. Motivated by this insight, SCT [8] introduces an uncertainty-aware calibration for LoCoOp, adjusting the loss function as follows:

$$\mathcal{L}_{\text{SCT}} = \mathbb{E}_{\mathcal{X},\mathcal{Y}} \left[\phi \cdot \mathcal{L}_{\text{CE}}(p(\mathbf{y}|\mathbf{x},\mathbf{w}),\mathbf{y}) - \psi \cdot \lambda_S \, \text{H}(p(\mathbf{y}|\tilde{\mathbf{x}},\mathbf{w}))\right], \tag{5}$$

where $\phi = 1 - p(\mathbf{y}|\mathbf{x},\mathbf{w})$ and $\psi = p(\mathbf{y}|\mathbf{x},\mathbf{w})$ are adaptive modulating factors that calibrate prompt learning according to confidence estimation. λ_S is the balancing hyperparameter. Furthermore, AMCN [2] generates multiple adaptive ID and OOD prompts to facilitate the separation of ID and OOD samples under prompt guidance, thereby improving OOD detection performance.

B. Theoretical Analysis

In Section 3.1, we introduce a separability-aware conjugate optimization framework equipped with a soft gating mechanism. We begin by constructing an entropy-guided divided learning strategy that enables the model to adapt with explicit out-of-distribution (OOD) awareness during inference. To stabilize optimization, we then incorporate a soft gating mechanism (Lemma 3.1) that smoothly approximates a hard indicator while remaining differentiable. Given that the loss function exhibits a Legendre–Fenchel structure, the optimization can be reformulated into a convex conjugate problem. The resulting problem is solved in two stages: (i) establishing the local invertibility of g (Lemma 3.2), and (ii) deriving conjugate-optimized pseudo-labels.

In this section, we present the detailed proofs of the lemmas and the theorem provided in the main manuscript, offering a comprehensive theoretical analysis that substantiates the validity of our STAR.

Proof of Lemma 3.1. We consider a soft gating function that converges pointwise to the hard indicator, providing an efficient and differentiable approximation.on.

Proof. Consider three cases for $\mathrm{H}(\mathbf{p})$:

If $\mathrm{H}(\mathbf{p}) < \theta$, then $\mathrm{H}(\mathbf{p}) - \theta < 0$. As $n \to \infty$, $n(\mathrm{H}(\mathbf{p}) - \theta) \to -\infty$, so

$$\phi_n(\mathbf{p}) = \frac{1}{1 + \exp(n(\mathrm{H}(\mathbf{p}) - \theta))} \to \frac{1}{1+0} = 1. \quad (6)$$

If $\mathrm{H}(\mathbf{p}) > \theta$, then $\mathrm{H}(\mathbf{p}) - \theta > 0$. As $n \to \infty$, $n(\mathrm{H}(\mathbf{p}) - \theta) \to +\infty$, so

$$\phi_n(\mathbf{p}) = \frac{1}{1 + \exp(n(\mathrm{H}(\mathbf{p}) - \theta))} \to \frac{1}{1+\infty} = 0. \quad (7)$$

If $\mathrm{H}(\mathbf{p}) = \theta$, then

$$\phi_n(\mathbf{p}) = \frac{1}{1 + \exp(0)} = \frac{1}{2}. \quad (8)$$

Hence, $\phi_n(\mathbf{p}) \to \mathbf{1}(\mathrm{H}(\mathbf{p}) < \theta)$ pointwise except at $\mathrm{H}(\mathbf{p}) = \theta$. □

Proof of Lemma 3.2. Here, let $g(\mathbf{h}) = \phi_n \log \mathbf{p}$, where $\phi_n(\mathbf{p}) = \frac{1}{1+\exp(n(\mathrm{H}(\mathbf{p})-\theta))}$. Define the subspace $\mathcal{S} := \{\mathbf{v} \in \mathbb{R}^K \mid \mathbf{1}^\top \mathbf{v} = 0\}$. Our goal is to prove that g is locally invertible on the subspace $\mathcal{S}$.

Proof. Specifically, we first derive the explicit form of $\nabla_{\mathbf{h}} g(\mathbf{h})$, and subsequently prove that it is non-singular when restricted to $\mathcal{S}$, thereby establishing the local invertibility of g with respect to the subspace $\mathcal{S}$.

(i) Calculate $\nabla_{\mathbf{h}} g(\mathbf{h})$. We first calculate $\nabla_{\mathbf{h}} \mathbf{p}$, $\nabla_{\mathbf{p}} \mathrm{H}(\mathbf{p})$, $\nabla_{\mathbf{h}} \log \mathbf{p}$, and $\nabla_{\mathbf{p}} \phi_n(\mathbf{p})$:

$$\nabla_{\mathbf{h}} \mathbf{p} = \mathrm{diag}(\mathbf{p}) - \mathbf{p} \cdot \mathbf{p}^\top, \quad (9)$$

$$\nabla_{\mathbf{p}} \mathrm{H}(\mathbf{p}) = -(\log \mathbf{p} + \mathbf{1}), \quad (10)$$

$$\nabla_{\mathbf{h}} \log \mathbf{p} = \mathbf{I} - \mathbf{1}\mathbf{p}^\top, \quad (11)$$

$$\nabla_{\mathbf{p}} \phi_n = n\phi_n(1-\phi_n)(\log \mathbf{p} + \mathbf{1}). \quad (12)$$

where $\mathbf{1}$ is an all-one vector whose shape is same as $\mathbf{p}$. And then $\nabla_{\mathbf{h}} g(\mathbf{h})$ can be derived by:

$$\begin{aligned}\nabla_{\mathbf{h}} g(\mathbf{h}) =& \nabla_{\mathbf{h}}(\phi_n \log \mathbf{p}) \\ =& \log \mathbf{p} \cdot (\nabla_{\mathbf{h}}(\phi_n))^\top + \phi_n \nabla_{\mathbf{h}}(\log \mathbf{p}) \\ =& \log \mathbf{p} \cdot \left(\nabla_{\mathbf{h}} \mathbf{p} \cdot \nabla_{\mathbf{p}} \phi_n\right)^\top \\ &+ \phi_n \cdot \nabla_{\mathbf{p}}(\log \mathbf{p}) \cdot \nabla_{\mathbf{h}} \mathbf{p}. \end{aligned} \quad (13)$$

Furthermore, we define $\mathbf{u}$ as the following formula:

$$\begin{aligned}\mathbf{u} \triangleq \nabla_{\mathbf{h}} \phi_n &= \nabla_{\mathbf{h}} \mathbf{p}\ \nabla_{\mathbf{p}} \phi_n \quad (14) \\ &= n\phi_n(1-\phi_n)(\mathrm{diag}(\mathbf{p}) - \mathbf{p}\mathbf{p}^\top)(\log \mathbf{p} + \mathbf{1}).\end{aligned}$$

Then, $\nabla_{\mathbf{h}} g(\mathbf{h})$ satisfies:

$$\nabla_{\mathbf{h}} g(\mathbf{h}) = \log \mathbf{p} \cdot \mathbf{u}^T + \phi_n(\mathbf{I} - \mathbf{1}\mathbf{p}^\top) \quad (15)$$

.

(ii) Null space of $\nabla_{\mathbf{h}} g(\mathbf{h})$ and invertibility on $\mathcal{S}$.

Let $\mathbf{J} \triangleq \nabla_{\mathbf{h}} g(\mathbf{h})$ and assume $\mathbf{J}\mathbf{v} = \mathbf{0}$ for some $\mathbf{v} \in \mathbb{R}^K$. Define scalars $s := \mathbf{u}^\top \mathbf{v}$ and $b := \mathbf{p}^\top \mathbf{v}$. Using $\mathbf{J} = \log \mathbf{p} \cdot \mathbf{u}^\top + \phi_n(\mathbf{I} - \mathbf{1}\mathbf{p}^\top)$, we obtain

$$(\log \mathbf{p})\, s + \phi_n(\mathbf{v} - \mathbf{1}b) = \mathbf{0}, \quad (16)$$

that is,

$$\mathbf{v} = \mathbf{1}b - \frac{s}{\phi_n} \log \mathbf{p}. \quad (17)$$

The equality holds since $\phi_n \in (0,1)$, which is guaranteed by the Sigmoid function. Subsequently, we multiply both sides by $\mathbf{p}^\top$ and derive the following equation:

$$b = b - \frac{s}{\phi_n} \mathbf{p}^\top \log \mathbf{p} \quad \Longrightarrow \quad s \cdot \mathbf{p}^\top \log \mathbf{p} = 0. \quad (18)$$

Since $\mathbf{p} = \mathrm{softmax}(\mathbf{h})$ has strictly positive entries and $\sum_i p_i = 1$, we have $\mathbf{p}^\top \log \mathbf{p} = \sum_i p_i \log p_i < 0$, hence $s = 0$. Therefore $\mathbf{v} = \mathbf{1}b$, implying

$$\mathrm{Ker}(\mathbf{J}) = \mathrm{span}\{\mathbf{1}\} \quad \text{and} \quad \mathrm{rank}(\mathbf{J}) = K - 1. \quad (19)$$

Restricting $\mathbf{J}$ to $\mathcal{S}$ removes the null direction $\mathbf{1}$, so $\mathbf{J}|_{\mathcal{S}}$ is non-singular. By the inverse function theorem, g is locally invertible on $\mathcal{S}$. □

Derivation of Conjugate-Derived Pseudo-Labels During test-time, the ground-truth label $\mathbf{y}$ is not available. Inspired by convex optimization theory, we derive conjugate pseudo-labels for estimating $\mathbf{y}$.

Proof. Since the loss function satisfies a Legendre-Fenchel structure, we can derive the optimal logits $\mathbf{h}$ by minimizing the empirical loss. The corresponding objective can be written as

$$\min_{\mathbf{h} \in \mathcal{S}} \left\{ f(\mathbf{h}) - \mathbf{y}^\top g(\mathbf{h}) \right\} = \min_{\mathbf{z} = g(\mathbf{h})} \left\{ (f \circ g^{-1})(\mathbf{z}) - \mathbf{y}^\top \mathbf{z} \right\}. \quad (20)$$

Given the common assumption that, after training on a large-scale dataset, the representation $\mathbf{h}$ has converged close to the optimal solution, we estimate the pseudo label $\mathbf{y}$ by the conjugate relation

$$\tilde{\mathbf{y}} = \nabla_{\mathbf{z}} (f \circ g^{-1})(\mathbf{z}) \Big|_{\mathbf{z} = g(\mathbf{h})} = \left(\left(\nabla_{\mathbf{h}} g(\mathbf{h}) \right) \big|_{\mathcal{S}} \right)^{-\top} \nabla_{\mathbf{h}} f(\mathbf{h}), \quad (21)$$

where $\mathcal{S} = \{\mathbf{v} \in \mathbb{R}^K : \mathbf{1}^\top \mathbf{v} = 0\}$ and the restriction to $\mathcal{S}$ is necessary since $\nabla_{\mathbf{h}} g(\mathbf{h})$ has a one-dimensional null space $\mathrm{span}\{\mathbf{1}\}$ (Lemma 3.2).

The proof proceeds in two steps. First, we establish that $\nabla_{\mathbf{h}} g(\mathbf{h})$ is non-singular when restricted to $\mathcal{S}$, which guarantees the existence of the corresponding inverse mapping on $\mathcal{S}$. Second, we derive the explicit form of $\tilde{\mathbf{y}}$ by applying the chain rule.

Specifically, we compute $\nabla_{\mathbf{h}} g(\mathbf{h})$ and $\nabla_{\mathbf{h}} f(\mathbf{h})$ as follows.

(i) By Lemma 3.2,

$$\begin{aligned}\nabla_{\mathbf{h}} g(\mathbf{h}) =& \nabla_{\mathbf{h}}(\phi_n \log \mathbf{p}) \\ =& \log \mathbf{p} \cdot (\nabla_{\mathbf{h}}(\phi_n))^\top + \phi_n \nabla_{\mathbf{h}}(\log \mathbf{p}) \\ =& \log \mathbf{p} \cdot \left(\nabla_{\mathbf{h}} \mathbf{p} \cdot \nabla_{\mathbf{p}} \phi_n\right)^\top \\ &+ \phi_n \cdot \nabla_{\mathbf{p}}(\log \mathbf{p}) \cdot \nabla_{\mathbf{h}} \mathbf{p} \\ \triangleq& A + B, \end{aligned} \tag{22}$$

where

$$A = n\phi_n(1-\phi_n) \log \mathbf{p} \cdot (\log \mathbf{p} + \mathbf{1})^\top (\mathrm{diag}(\mathbf{p}) - \mathbf{p}\mathbf{p}^\top), \tag{23}$$

and

$$B = \phi_n(\mathbf{I} - \mathbf{1}\mathbf{p}^\top). \tag{24}$$

Moreover, Lemma 3.2 implies that $\nabla_{\mathbf{h}} g(\mathbf{h})$ has rank $K-1$ and is nonsingular when restricted to $\mathcal{S}$.

(ii) Calculate $\nabla_{\mathbf{h}} f(\mathbf{h})$:

$$\begin{aligned}\nabla_{\mathbf{h}} f(\mathbf{h}) &= \alpha \cdot \nabla_{\mathbf{h}}\big((1-\phi_n)(\log K - \mathrm{H}(\mathbf{p}))\big) \\ &= -\alpha \cdot \nabla_{\mathbf{h}} \phi_n \cdot (\log K - \mathrm{H}(\mathbf{p})) \\ &\quad - \alpha(1-\phi_n) \cdot \nabla_{\mathbf{h}} \mathrm{H}(\mathbf{p}) \\ &= -\alpha \cdot n\phi_n(1-\phi_n)(\log K - \mathrm{H}(\mathbf{p})) \\ &\quad \cdot (\mathrm{diag}(\mathbf{p}) - \mathbf{p}\mathbf{p}^\top)(\log \mathbf{p} + \mathbf{1}) \\ &\quad + \alpha(1-\phi_n)(\mathrm{diag}(\mathbf{p}) - \mathbf{p}\mathbf{p}^\top)(\log \mathbf{p} + \mathbf{1}) \\ &= \alpha \cdot [1 - n\phi_n(\log K - \mathrm{H})]\,(1-\phi_n)\Psi, \end{aligned} \tag{25}$$

where $\Psi = (\mathrm{diag}(\mathbf{p}) - \mathbf{p}\mathbf{p}^\top)(\log \mathbf{p} + \mathbf{1})$. Therefore, the estimation of $\tilde{\mathbf{y}}$ is given by

$$\begin{aligned}\tilde{\mathbf{y}} = \alpha \cdot \left(\left[n\phi_n(1-\phi_n) \log \mathbf{p} \cdot \Psi^\top + \phi_n(\mathbf{I} - \mathbf{1}\mathbf{p}^\top)\right]\Big|_{\mathcal{S}}\right)^{-\top} \\ \cdot [1 - n\phi_n(\log K - \mathrm{H})]\,(1-\phi_n)\Psi. \end{aligned} \tag{26}$$

□

C. Influence on ViT backbone on Performance

In this section, we evaluate the harmonic mean of accuracy on iNaturalist and SUN using different ViT backbones, namely ViT-B/16 and ViT-B/32. Furthermore, we observe that the model's performance decreases when moving from ViT-B/16 to ViT-B/32 across both 1-shot and 16-shot settings on both datasets.This finding indicates that STAR is robust to variations in backbone architecture, maintaining competitive performance across different patch configurations.

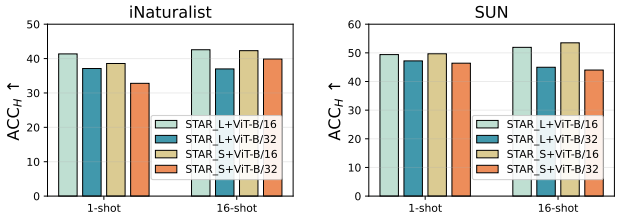

Figure 1. $\mathrm{ACC_H}$ comparison of STAR using two different image encoders, ViT-B/16 and ViT-B/32.

D. Algorithm

Algorithm 1 describes the detailed pipeline of STAR for enhancing universal prompt learning in test-time adaptation for VLMs.

Algorithm 1: STAR for Test-time Adaptation

Input: Test batch containing mixed samples $\mathcal{D} = \{\mathbf{x}_i^{\mathrm{ID}}\}_{i=1}^{N} \cup \{\mathbf{x}_j^{\mathrm{OOD}}\}_{j=1}^{M} = \{\mathbf{x}_i\}_{i=1}^{N_B}$

for $x_i \leftarrow 1$ **to** N_B **do**
  // Conjugate Optimization
  Compute the gating threshold θ by Eq. 2 ;
  Generate pseudo-labels $\hat{\mathbf{y}}$ via Eq. 5 ;
  Optimize the prompt learner using Eq. 14 ;

for $x_i \leftarrow 1$ **to** N_B **do**
  // Prototypical Retrieval
  Initialize class prototypes with $\mathcal{T}(\mathbf{t}_k)$;
  Populate the memory bank $\mathcal{B}$ using Eq. 16 ;
  Update prototypes $\mathcal{C} = \{\mathbf{c}_k\}_{k=1}^{K}$ via Eq. 17 ;
  Compute the similarity score $s_{i,k}^c$ using Eq. 18 ;
  Derive the class-aware threshold τ_k by Eq. 20 ;

for $x_i \leftarrow 1$ **to** N_B **do**
  // Context Adaptation
  Compute convex conjugate loss $\mathcal{L}_n$ via Eq. 5 ;
  Update parameters $\Theta \leftarrow \Theta - \mathrm{lr} \cdot \nabla \mathcal{L}_n$;

References

[1] Alexey Dosovitskiy. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 1

[2] Xiang Fang, Arvind Easwaran, and Blaise Genest. Adaptive multi-prompt contrastive network for few-shot out-of-distribution detection. *arXiv preprint arXiv:2506.17633*, 2025. 1

[3] Yushu Li, Xun Xu, Yongyi Su, and Kui Jia. On the robustness of open-world test-time training: Self-training with dynamic prototype expansion. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11836–11846, 2023. 1

[4] Yifei Ming, Ziyang Cai, Jiuxiang Gu, Yiyou Sun, Wei Li, and Yixuan Li. Delving into out-of-distribution detection with

vision-language representations. *Advances in neural information processing systems*, 35:35087–35102, 2022. 1

[5] Atsuyuki Miyai, Qing Yu, Go Irie, and Kiyoharu Aizawa. Zero-shot in-distribution detection in multi-object settings using vision-language foundation models. *arXiv preprint arXiv:2304.04521*, 2023. 1

[6] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PmLR, 2021. 1

[7] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017. 1

[8] Geng Yu, Jianing Zhu, Jiangchao Yao, and Bo Han. Self-calibrated tuning of vision-language models for out-of-distribution detection. *Advances in Neural Information Processing Systems*, 37:56322–56348, 2024. 1