
Align Your Prompts: Test-Time Prompting with Distribution Alignment for Zero-Shot Generalization

Jameel Hassan^{1*} Hanan Gani^{1*} Noor Hussein¹ Muhammad Uzair Khattak¹

Muzammal Naseer¹ Fahad Shahbaz Khan^{1,2} Salman Khan^{1,3}

¹Mohamed Bin Zayed University of AI ²Linköping University ³Australian National University

{jameel.hassan, hanan.ghani, noor.hussein, uzair.khattak
muzammal.naseer, fahad.khan, salman.khan} @mbzuai.ac.ae

Abstract

The promising zero-shot generalization of vision-language models such as CLIP has led to their adoption using prompt learning for numerous downstream tasks. Previous works have shown test-time prompt tuning using entropy minimization to adapt text prompts for unseen domains. While effective, this overlooks the key cause for performance degradation to unseen domains – distribution shift. In this work, we explicitly handle this problem by aligning the out-of-distribution (OOD) test sample statistics to those of the source data using prompt tuning. We use a single test sample to adapt multi-modal prompts at test time by minimizing the feature distribution shift to bridge the gap in the test domain. Evaluating against the domain generalization benchmark, our method improves zero-shot top-1 accuracy beyond existing prompt-learning techniques, with a 3.08% improvement over the baseline MaPLe. In cross-dataset generalization with unseen categories across 10 datasets, our method improves consistently across all datasets compared to the existing state-of-the-art. Our source code and models are available at <https://jameelhassan.github.io/promptalign/>.

1 Introduction

Deep neural networks (DNNs) have outperformed humans in numerous visual recognition tasks [12, 8]. However, their impressive performance generally holds in the case when test data originate from the same distribution as the training data. In most real-world applications, the train and test data distributions can significantly differ due to factors such as natural variations or changes in sensing equipment. The sensitivity of models to such unseen distributional shifts during inference results in their performance degradation [14, 31, 29]. A plethora of previous works [36, 39, 37] have explored test-time adaptation as a mechanism to overcome the crucial problem of distribution shift in test data. However, test-time adaptation has been minimally explored for the increasingly popular foundation models, which are large DNNs trained on massive vision-language datasets [30, 17, 44, 40, 41].

Foundation models emerging at the intersection of various modalities such as vision, language, and audio, are proving to be effective in numerous downstream applications. Among these models, Vision-Language (V-L) model CLIP (Contrastive Language-Image Pretraining) [30] has been pre-trained on large-scale image-text pairs from the web, and can generalize well to zero-shot recognition tasks. The CLIP model has parallel vision and language encoding branches and the similarities between the embeddings obtained from both the branches are used to classify an input image. At inference, a handcrafted prompt such as ‘a photo of a <CLS>’ is used as a query for the text encoder. However, adapting CLIP efficiently to specific downstream tasks is still a challenging problem. Naively fine-tuning such models poses the risk of losing their inherent generalization abilities [21]. Instead, recent

*Equal contribution

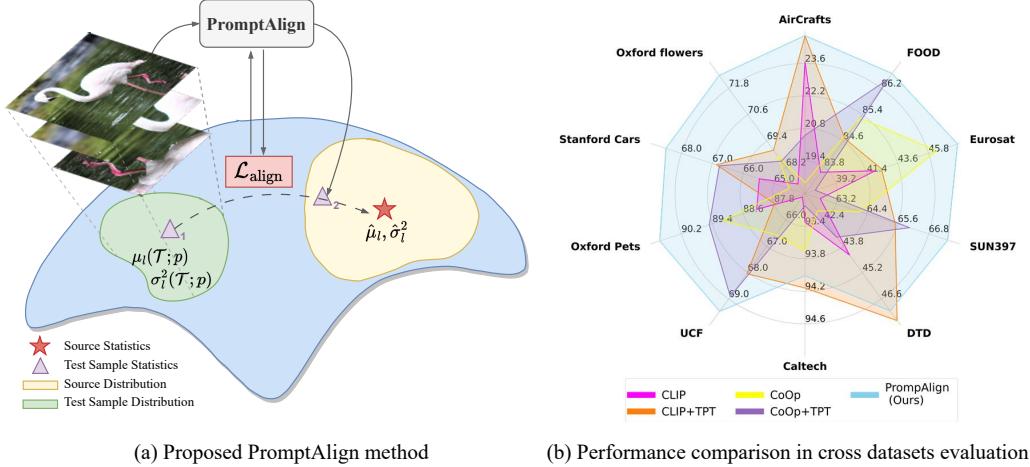


Figure 1: (a) PromptAlign matches the distribution statistics $\mu_l(\mathcal{T}; p)$, $\sigma_l^2(\mathcal{T}; p)$, obtained from multiple augmented views of a single test sample, with the source data distribution statistics $\hat{\mu}_l$, $\hat{\sigma}_l^2$. This effectively brings the test sample closer to the distribution of the source data, where the domain shift is denoted by $\blacktriangle_1 \rightarrow \blacktriangle_2$. \mathcal{T} denotes the distribution of the test sample, p represents the prompts that are updated and l refers to the vision-backbone layers. (b) Owing to the distribution matching via prompts, PromptAlign surpasses the existing state-of-the-art prompt learning approaches on 8 out of 10 datasets in cross-dataset generalization benchmarks.

methods have shown the promise of prompt learning on training data [48, 47, 18, 19, 42] as opposed to handcrafted prompts, allowing the model to better adapt to the training data distribution.

Existing prompt learning approaches are deployed at the training phase to learn representative prompts based on the training data for the downstream task. This conventional approach does not explicitly handle the distribution shift in the test set. Leveraging the capability of prompt learning, a recent approach TPT [34] performs test-time adaptation by tuning the text prompts on the fly to adapt the model to the test sample. For image classification, the model updates the prompts by minimizing the entropy of the top confident samples which are obtained using different augmented views. However, TPT does not explicitly align the pre-trained CLIP to become aware of the test sample distribution.

For the effective test-time adaptation of V-L foundation models, it is crucial to bridge the distribution gap between the pre-training dataset and the downstream evaluation set for high zero-shot generalization. To this end, we propose PromptAlign, a test-time token distribution alignment strategy using prompt learning. The TPT setting, which tunes the prompts on the text encoder branch alone poses an architectural limitation in performing distribution alignment of tokens. Further, at test time, there is no knowledge transfer between the text and vision branches in the intermediate stages. Since the text encoder features will be static given a dataset (as the input is the same class labels), distribution alignment of tokens can only be performed on the vision branch. Given these constraints and to further extend the strength of prompts for test time adaptation, we propose to employ a multi-modal prompt learning model (MaPLE) [18] for distribution alignment. PromptAlign explicitly aligns the mean and variances of the image token embeddings of a proxy source dataset, computed offline, with the image token embeddings of the test sample. We extend TPT with the token alignment strategy, which enforces to bridge the distribution shift in the test data (Fig. 1 a). For each input test sample, we obtain randomly augmented views that are fed into the model to obtain the token embedding statistics. Without any noticeable compute overhead, we update the prompts on both the text and vision branches of CLIP to minimize jointly the feature distribution shift and entropy of predictions.

We extensively demonstrate the effectiveness of PromptAlign by evaluating the zero-shot generalization on two representative benchmarks: domain generalization and cross-dataset generalization. In the domain generalization setting, our method improves the baseline model by 3.08% on average across the four datasets and has the highest average Top-1 accuracy compared to existing state-of-the-art methods. In the cross-dataset setting, our method achieves an absolute average improvement of 1.82% over the existing state-of-the-art method which uses test-time prompt tuning, while attaining the best Top-1 accuracy in 8 out of the 10 datasets. Our contributions can be summarized as follows:

- Given only a single test sample, we introduce a distribution alignment strategy for V-L models to improve test-time adaptation. The distribution-aware pre-trained CLIP effectively narrows the distribution gap on the test domains. To the best of our knowledge, this is the first study to explore the potential of distribution alignment in V-L models at test time.
- The proposed strategy formulates a distribution alignment loss that utilizes offline computed source data statistics to encourage the test sample token distributions to be aligned with the source data token distributions. We harmonically combine the benefits of token distribution alignment and entropy minimization using a multi-modal prompt learning approach.
- Since CLIP-pre-training data is not publicly released, we study the statistics of ImageNet as a possible candidate for the source distribution, and our empirical results show that ImageNet is an effective proxy source dataset for large-scale V-L models such as CLIP.
- We validate our method PromptAlign through extensive experiments in domain generalization and cross-dataset benchmarks. PromptAlign improves the generalization of CLIP at test time beyond existing prompt-tuning methods, achieving state-of-the-art results.

2 Related Work

Prompting for Vision-Language models. Vision Language (V-L) foundation models [30, 17, 44, 40, 41] have emerged with the convergence of image and text modalities. Having been pre-trained on massive image-text pairs from the web in a contrastive self-supervised manner, these models like CLIP [30] and ALIGN [17] have shown strong generalizability towards downstream zero-shot recognition tasks. However, adapting them efficiently to specific downstream tasks with limited data is still a challenging problem. Prompting in CLIP-like models has been explored in the form of a text query that is usually given to the text encoder to instruct it to perform a specific task. Recent methods propose to learn these prompts by treating them as continuous learnable vectors and training them in an end-to-end manner while keeping the model parameters frozen. CoOp [48] proposes to fine-tune CLIP by learning a set of prompts in the text encoder. CoCoOp [47] highlights the inferior generalization capability of CoOp and conditions the text prompt tokens on image embeddings on the fly. MaPLe [18] proposes to jointly learn deep prompts at both vision and text encoders of CLIP. The vision prompts are further conditioned on text prompts via a V-L coupling function. While all these approaches promote the effective transfer of CLIP, they require training data to learn the prompts which restricts such adaptation on novel datasets at test time. A recent method TPT [34], attempts to learn prompts solely at test time with the objective of enforcing consistency regularization between multiple views of a test sample by minimizing their averaged entropy. However, this approach struggles to explicitly address the distribution misalignment between the pre-training data of CLIP and the downstream test data. Our method builds on multi-modal prompting variant [18] and explicitly enforces to match the distribution via a distribution alignment technique using a proxy dataset that serves as a replacement for unavailable pre-training data of CLIP. To the best of our knowledge, ours is the first work to explicitly align the distributions learned by V-L foundational models with test time data distribution.

Test-Time Adaptation (TTA). TTA [36, 37, 23] aims to bridge the distribution gap between the train and test data distributions at test time. These methods can broadly be categorized into two streams. The first stream of approaches [11, 36] typically utilizes a self-supervised proxy task such as predicting image rotations and thus alters the training phase. For example, Test-Time Training (TTT) [36] jointly trains the model for rotation prediction and image classification during training. TTT++ [23] adopts a self-supervised contrastive learning approach as an auxiliary task during training and test-time feature alignment module to adapt the model. By incorporating additional auxiliary tasks, these techniques aim to mitigate over-fitting and improve the generalization performance of the model. The second stream of approaches solely adopts the pre-trained model without altering the training phase. These methods improve model generalization at test time by imposing self-consistency regularization constraints or aligning statistics of training and test samples in a model. Approaches with self-consistency regularization constraints typically utilize entropy minimization within samples of a batch or multiple views of a single test sample. TENT [37] proposes entropy minimization of batch-wise prediction probability distributions. The need to have a batch of samples by generating multiple views via augmentations at test time is eliminated in [45]. Other methods, such as NORM [32] and DUA [25] adapt the model’s normalization layers by matching the batch normalization statistics computed during training with the test set distribution. Both [22] and ActMAD

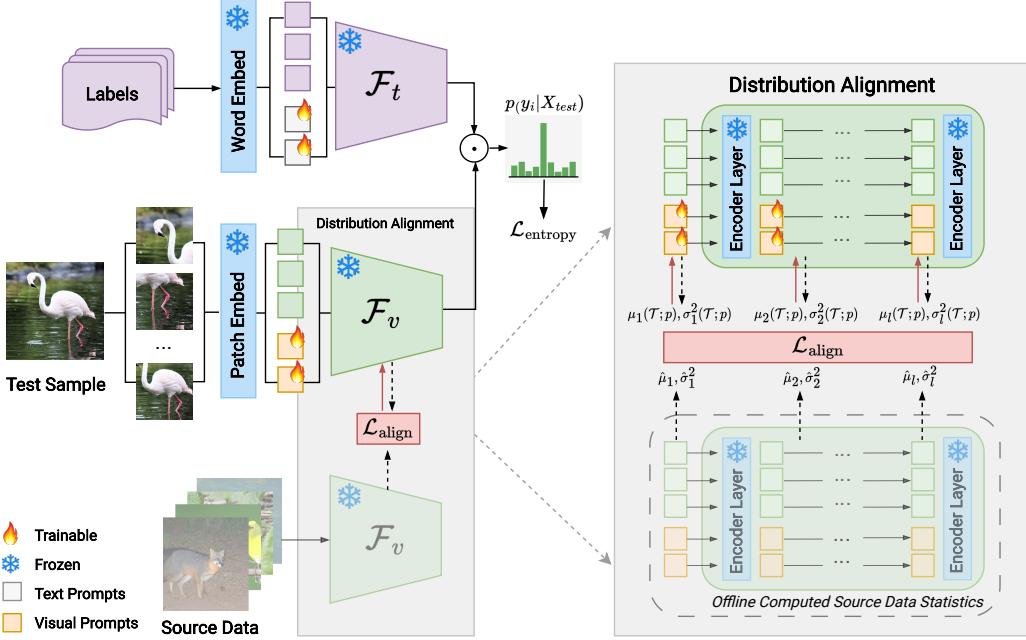


Figure 2: Overview of our proposed PromptAlign method for zero-shot image classification. At test time, a single test sample along with its augmented views is passed through the CLIP image encoder, and the text labels are passed to the CLIP text encoder. The token distribution statistics – mean and variance – of the test sample are aligned with the offline computed source data statistics using a distribution alignment loss. The resulting alignment loss from the distribution shift is combined with the entropy loss to update the multi-modal prompts.

[26] updates all model parameters using a batch of samples in a continuous manner, with a distribution matching loss between the source statistics and the batch statistics of activations across the network. However, it is unclear how these statistics alignment techniques can be adopted for foundational V-L models in a lightweight manner for a single sample. In this paper, we propose a multi-modal test time prompting approach without altering the pre-training phase of V-L models along with a strategy to obtain training distribution statistics. This effectively paves the way for distribution alignment together with entropy minimization for improved test-time optimization.

Test-time prompt tuning. As mentioned earlier, a plethora of works has been introduced on learning prompts to adapt V-L models using the downstream task training data. However, the learning of prompts at test time remains largely unexplored in the literature. Very recently, TPT [34] proposed a test time prompting approach by learning prompts at the text side with an entropy minimization objective. While effective, TPT struggles with addressing the explicit alignment of the pre-training data distribution of CLIP with the test data distribution. Further, such an alignment is not trivial as access to the pre-training data of CLIP is not available. Our work addresses these limitations by extending TPT and introducing a distribution alignment objective for CLIP. This objective leverages token distribution statistics from proxy source data and a *single* test sample using multi-modal prompt learning, and explicitly aligns the train and test sample distributions to mitigate the domain shift.

3 Methodology

We provide an overview of CLIP [30] and prompt learning and their application for zero-shot generalization to downstream tasks in Section 3.1. We then discuss test-time prompt tuning and briefly review previous approaches. Finally, we introduce PromptAlign in Section 3.2 and provide a detailed explanation of how we apply it to achieve improved zero-shot image classification.

3.1 Preliminaries

Contrastive Language-Image Pre-training (CLIP). CLIP consists of two parallel encoders, one for mapping the text input into a feature vector and the other for mapping the visual input. We denote the CLIP image and text encoders by \mathcal{F}_v and \mathcal{F}_t respectively, and their pre-trained parameters are represented by $\theta_{\text{CLIP}} = \{\theta_v, \theta_t\}$ respectively. The input image X is divided into M patches, which are projected to produce patch tokens, and a class token CLS is prepended to it, resulting in $\tilde{X}_0 = \{\text{CLS}, e_1, e_2, \dots, e_M\}$ where e_i is the embedding of the i^{th} patch. The image encoder \mathcal{F}_v encodes the input patches via transformer blocks to produce a latent visual feature representation $\tilde{f}_v = \mathcal{F}_v(\tilde{X}_0, \theta_v)$. The corresponding class label y is embedded within a text template, such as ‘a photo of a <CLS>’, which is formulated as $\tilde{Y}_0 = \{\text{sos}, t_1, t_2, \dots, t_L, c_k, \text{EOS}\}$. Here, $t_l|_{l=1}^L$ and c_k are the word embeddings corresponding to the text template and the class label, respectively, while sos and EOS are the start and end token embeddings. The text encoder \mathcal{F}_t encodes \tilde{Y} via transformer blocks to produce the latent textual feature as $\tilde{f}_t = \mathcal{F}_t(\tilde{Y}_0, \theta_t)$. For zero-shot inference, each text feature with class labels $y = \{1, 2, \dots, C\}$ is paired with the image feature to compute a similarity score $s_i = \text{sim}(\tilde{f}_t \cdot \tilde{f}_v)$, where $\text{sim}(\cdot)$ denotes the cosine similarity. The prediction probability on X can be denoted by $p(y_i|X) = \frac{\exp(\text{sim}(\tilde{f}_t \cdot \tilde{f}_v)\tau)}{\sum_{i=1}^C \exp(\text{sim}(\tilde{f}_t \cdot \tilde{f}_v)\tau)}$, where τ is the temperature of the softmax.

Prompt tuning on downstream tasks. CLIP contains a plethora of knowledge leveraged from training on millions of noisy image-text pairs. To effectively extract the rich features learned by the CLIP model, recent approaches [48, 47, 18, 2, 42] append extra learnable prompts while keeping image and text encoders frozen. These prompts modify the context of the model input without distorting the pre-trained CLIP features. Prompts are appended either at the image or text encoder side and learn contextual information tailored towards a specific task. In our work, we use a recently introduced multi-modal prompting baseline [18] that learns prompt tokens on both the text and image encoders.

Specifically, we append learnable V visual and T text prompts given as $\mathbf{p}_v = \{\mathbf{p}_v^1, \mathbf{p}_v^2, \dots, \mathbf{p}_v^V\}$, and $\mathbf{p}_t = \{\mathbf{p}_t^1, \mathbf{p}_t^2, \dots, \mathbf{p}_t^T\}$ with the visual and textual input tokens respectively. The image encoder processes the input $\tilde{X}_0^p = \{\text{CLS}, \mathbf{p}_v, e_1, e_2, \dots, e_M\}$ to generate a prompted visual feature representation denoted as \tilde{f}_v^p . Similarly, the text encoder processes the input $\tilde{Y}_0^p = \{\text{sos}, \mathbf{p}_t, t_1, t_2, \dots, t_L, c_k, \text{EOS}\}$ producing the textual feature representation \tilde{f}_t^p . Our approach uses deep prompting as utilized in [18], along with text prompts and conditional image prompts at subsequent transformer blocks. We jointly represent the visual and textual prompts by \mathbf{p} . We refer the readers to [18] for more details on the baseline architecture.

Test-time prompt tuning. Test-time prompt tuning introduced by [34] aims to benefit from the rich knowledge of CLIP to boost its generalization in a zero-shot manner. TPT can be viewed as a means to provide the model with a context that is customized for each individual test sample in order to more accurately recall the knowledge contained within CLIP. During inference, several randomly augmented views are generated from the given test sample X_{test} . The predictions having entropy below a certain threshold are kept, while other views are discarded using a confidence selection filter. The averaged entropy of the filtered predictions is then used to update the prompts \mathbf{p} in an unsupervised fashion using the following objective function.

$$\mathcal{L}_{\text{entropy}} = \arg \min_{\mathbf{p}} - \sum_{i=1}^C \tilde{p}_{\mathbf{p}}(y_i|X_{\text{test}}) \log \tilde{p}_{\mathbf{p}}(y_i|X_{\text{test}}), \quad (1)$$

where $\tilde{p}_{\mathbf{p}}(y_i|X_{\text{test}})$ represents the mean of vector class probabilities produced by the model across the different augmented views preserved after the confidence selection filter.

3.2 PromptAlign

Unimodal Test-time Prompt Tuning (TPT) [34] updates the text prompts on the fly at inference by minimizing the entropy loss. This approach does not explicitly handle distribution shift that arises in the test set which is sub-optimal. One solution to this problem is to align the source and target distributions by bringing the test sample into the domain of the source data distribution. However, TPT updates the prompts only on the text encoder branch with static labels which poses an architectural limitation in performing the distribution alignment of tokens. Hence, distribution alignment of tokens

can only be performed on the vision branch. We, therefore, utilize multi-modal prompt learning VL models [18] to explicitly handle the distribution shift of each test sample from the source distribution.

Given a test sample X_{test} , we take multiple augmented views and pass them through the visual encoder with deep learnable prompts as shown in Fig. 2. At each layer of the visual encoder, we compute the token alignment loss between the means and variances of the test sample with that of the means and variances of a proxy dataset which mimics the behavior of the source data distribution. Our final objective combines the entropy and alignment losses to update the prompts for a given test sample.

Proxy source dataset. In order to compute the token embedding statistics on the source dataset, we require the pre-training dataset of the CLIP model. The CLIP model was trained on over 400 million image-text pairs, which is not publicly available. However, previous works have shown that LAION400M [33] can be used as a training dataset to achieve the performance of CLIP [5], leading to subsets of LAION400M being used as a proxy dataset for the CLIP training dataset. In addition to this, CLIP has also been heavily tuned to achieve excellent zero-shot performance on ImageNet [1]. Therefore, we use ImageNet as the proxy source dataset for computing the mean and variance of token distributions. Source dataset statistics are computed offline and utilized directly at test time.

Token Distribution Alignment via multi-modal prompting. We generate N_k random views of the test sample using a set of augmentations \mathcal{H} . The mean and variance statistics of token embeddings of the test sample are computed at the output of each transformer layer of the CLIP model’s visual encoder, across the N_k views. Similarly, the source data statistics are pre-computed in an offline manner. We represent test sample distribution by (\mathcal{T}) and source distribution by (\mathcal{D}) . Specifically, we compute the token means and variances for the alignment as follows.

$$\boldsymbol{\mu}_l(\mathcal{T}; \mathbf{p}) = \frac{1}{N_k} \sum_{x \in \mathcal{H}(X)} \tilde{\mathbf{X}}_{l,x}^{\mathbf{p}} , \quad (2)$$

$$\boldsymbol{\sigma}_l^2(\mathcal{T}; \mathbf{p}) = \frac{1}{N_k} \sum_{x \in \mathcal{H}(X)} \left(\tilde{\mathbf{X}}_{l,x}^{\mathbf{p}} - \boldsymbol{\mu}_l(\mathcal{T}; \mathbf{p}) \right)^2 , \quad (3)$$

where $\boldsymbol{\mu}_l(\mathcal{T}; \mathbf{p})$ and $\boldsymbol{\sigma}_l^2(\mathcal{T}; \mathbf{p})$ are the vector means and variances of the test sample tokens at the layer l in the visual encoder and $\tilde{\mathbf{X}}_{l,x}^{\mathbf{p}}$ represents the prompted token embeddings at layer l for the augmented view input x . Similarly for each layer l in the visual encoder, we pre-compute the source data statistics as,

$$\hat{\boldsymbol{\mu}}_l = \boldsymbol{\mu}_l(\mathcal{D}, \theta_v) \quad \text{and} \quad \hat{\boldsymbol{\sigma}}_l^2 = \boldsymbol{\sigma}_l^2(\mathcal{D}, \theta_v) , \quad (4)$$

where θ_v denotes the parameters of the visual encoder from the pre-trained CLIP model. We compute the token distribution alignment loss between the mean and variances of the test sample and the source dataset statistics as follows,

$$\mathcal{L}_{\text{align}} = \frac{1}{L} \sum_{l=1}^L \left(\|\boldsymbol{\mu}_l(\mathcal{T}; \mathbf{p}) - \hat{\boldsymbol{\mu}}_l\|_1 + \|\boldsymbol{\sigma}_l^2(\mathcal{T}; \mathbf{p}) - \hat{\boldsymbol{\sigma}}_l^2\|_1 \right) . \quad (5)$$

As shown above, we use L_1 loss to enforce the distribution alignment of the test sample with the source distribution. The alignment loss $\mathcal{L}_{\text{align}}$ is added to the entropy loss (Eq. 1) to obtain the final objective $\mathcal{L}_{\text{final}}$. Finally, the combined objective is optimized to update the prompts \mathbf{p} .

$$\mathcal{L}_{\text{final}} = \mathcal{L}_{\text{entropy}} + \beta \mathcal{L}_{\text{align}} , \quad (6)$$

where the hyperparameter β controls the contribution of alignment loss to the overall objective function. Our alignment loss in Eq. 6 bridges the gap between source and token distributions by updating the prompts to align them with the source distribution.

Discussion on $\mathcal{L}_{\text{final}}$. As derived above, our test-time loss combines the entropy minimization ($\mathcal{L}_{\text{entropy}}$) and distribution alignment objectives ($\mathcal{L}_{\text{align}}$) together. The $\mathcal{L}_{\text{entropy}}$ objective enforces prediction consistency among different views of a sample, leading to robustness against various variations that could possibly occur at test time. On the other hand, $\mathcal{L}_{\text{align}}$ effectively narrows the domain shift and brings the test sample distribution closer to the pre-trained CLIP distribution space, which enforces CLIP to better understand the test sample. The combination of these loss objectives harmonically adapts CLIP to be robust to different sample variations and at the same time, enhances CLIP’s understanding of the underlying test sample domain for better generalization.

Table 1: **Effect of token distribution alignment strategy for domain generalization.** The base model MaPLe is trained on ImageNet and evaluated on datasets with domain shifts.

	Imagenet V2	Imagenet Sketch	Imagenet A	Imagenet R	OOD Avg.
MaPLe	64.07	49.15	50.90	76.98	60.28
MaPLe+TPT	64.87	48.16	58.08	78.12	62.31
PromptAlign	65.29	50.23	59.37	79.33	63.55

4 Experiments

Datasets. For the domain generalization setting, we evaluate the four out-of-distribution (OOD) variants of ImageNet [7]; ImageNetV2 [31], ImageNet-Sketch [38], ImageNet-A [16] and ImageNet-R [15]. We also evaluate the domain generalization setting on the recent challenging Photorealistic Unreal Graphics (PUG) dataset [3], comprising different textures, backgrounds, sizes and orientations. In the cross-dataset evaluation, we follow TPT [34] and evaluate the performance of methods on 10 image classification datasets covering a wide range of visual recognition tasks. This includes one generic-objects dataset Caltech101 [10]; five fine-grained datasets OxfordPets [28], StanfordCars [20], Flowers102 [27], Food101 [4] and FGVC-Aircraft [24], which contain images of animals, flowers and transportation; and four datasets of scenes, textures, satellite imagery and human actions – SUN397 [36], DTD [6], EUROSAT [13] and UCF101 [35] respectively.

Baselines. We evaluate PromptAlign with existing few-shot prompt learning methods for adapting CLIP including CoOp [48] and CoCoOp [47], and TPT [34] method. MaPLe [18] is a multi-modal prompt learning baseline, which adapts CLIP by learning deep prompts on both the text and vision branches. TPT is a test-time prompt tuning method, that tunes the prompt at test time per input sample achieving state-of-the-art performance in prompt learning when combined with CoOp.

Implementation details. Following MaPLe [18], we train on ImageNet using 16-shot training data with 2 prompt tokens for a depth of 3 layers. We optimize the prompts on both the text and vision branches using a single test image. We obtain 63 augmented views using random resized crops and horizontal flip augmentations to construct a batch of 64 images including the original image to mimic the setting of TPT. From the 64 predictions, we obtain top 10% confident predictions based on the lowest entropy and compute the average prediction probability. We compute the token distribution alignment loss between the tokens of all 64 images. We optimize the prompts to minimize the combined loss of average prediction entropy and the token distribution alignment loss using the AdamW optimizer. We use a learning rate of $5e^{-4}$ for the fine-grained datasets Flowers102, OxfordPets, Food101, SUN397, FGVC-Aircraft, and EuroSAT and a learning rate of 0.04 for the rest of the datasets, and set the loss scale factor β equal to 100. Further details are listed in Appendix A.

4.1 Token distribution alignment in V-L models

We first evaluate the performance of the token distribution alignment strategy over the baseline for test-time prompt tuning. Since the alignment can only be performed on the vision branch, we explicitly compare PromptAlign to vanilla MaPLe, and MaPLe with TPT using entropy minimization. Table 1 presents the results in domain generalization across ImageNet variants. TPT improves the performance of MaPLe from 60.28% to 62.31% on average, while the incorporation of token distribution alignment further improves the performance by 1.24%. This indicates that explicitly aligning the distribution of training and testing data enhances CLIP’s generalization. Moreover, TPT is intended to improve test-time adaptation as a plug-in module, but its effectiveness is not uniform, as observed from the drop in performance on the ImageNet-Sketch dataset. On the other hand, PromptAlign does not show any performance degradation and improves beyond vanilla MaPLe consistently on all datasets. We further present qualitative results of PromptAlign in Appendix C.

4.2 Domain Generalization

Compared to CLIP, all test-time adaptation methods demonstrate better performance, indicating that V-L models benefit from test-time adaptation approaches (Table 2). PromptAlign achieves the highest Top-1 accuracy averaged across all ImageNet variants. In comparison to the previous state-of-the-art CoOp+TPT, PromptAlign achieves an average improvement of 0.71%. Out of the four different

Method	Entropy loss	Distribution alignment	Top-1 Acc.
MaPLe	✗	✗	50.90
MaPLe+TPT	✓	✗	58.08
PromptAlign [†]	✗	✓	50.85
PromptAlign	✓	✓	59.37

Table 5: **Analysis on the alignment and entropy minimization loss.** The average of Top-1 accuracy (%) across three seeds is reported. PromptAlign[†] denotes our method excluding the entropy minimization loss.

5 Ablation

We perform ablative analysis on various components of PromptAlign. Unless stated otherwise, we show ablations on ImageNet-A dataset, the smallest domain generalization variant for simplicity. Detailed ablation trends across other ImageNet variants and datasets are presented in Appendix D.

Token Distribution Alignment. Table 5 summarizes the effect of token distribution alignment in PromptAlign. It can be observed that removing the entropy loss results in almost the same performance as vanilla MaPLe. Whereas combining alignment and entropy losses improves the performance remarkably. This is because the distribution alignment loss is a regularizer that functions together with the self-entropy loss. Since distribution alignment does not promote any discriminative learning as opposed to entropy loss, it is not expected to improve test-time adaptation on its own.

Loss variants for distribution alignment. We show the variation in performance with different loss choices for distribution alignment objective. Figure 4 compares the three loss choices L_1 , L_2 , and KL divergence. We observe that L_1 constraint performs best for the distribution alignment objective. Further analysis is shown in Table E of supplementary material.

Loss balancing factor β . We ablate the loss scale factor β to analyze the significance of the distribution alignment loss. The performance improves over the baseline model with TPT as depicted in Figure 3, even with a smaller weight scale and gradually plateaus close to a scale of 100. We use ImageNet validation set for this ablation and choose $\beta = 100$ across all experiments. We find a similar trend of the scaling factor on other datasets, which is discussed in Appendix D. For the MaPLe+TPT model, the value of β is 0 as there is no alignment objective in this case.

Higher Order Statistics. Since we align distributions computed using a small set of samples, we use relatively simpler (mean and variance) measures for our alignment strategy over higher order statistical measures. We evaluate this choice for the distribution alignment in PromptAlign using the 5th order statistics of Central Moment Discrepancy measure [43] in Table 6. Our observation reveals that the utilization of higher order statistics results in only a slight improvement implies that there is minimal distinction between opting for higher order statistics validating mean and variance based approach as a suitable measure for alignment in our case.

Prompt Regularization. We compare PromptAlign to a naive prompt regularization method in a continuous test-time adaptation setting, in which the model gradually adapts to the incoming distribution. We regularize the prompts in relation to the previous prompts, denoted by PromptReg in Table 7. We also lower the learning rate to $1e^{-5}$ to prevent diverging in the continuous setting. We

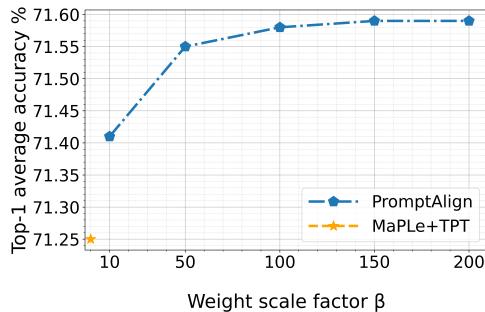


Figure 3: Effect of the loss scaling factor β on ImageNet. Scale factor plateaus after $\beta = 100$.

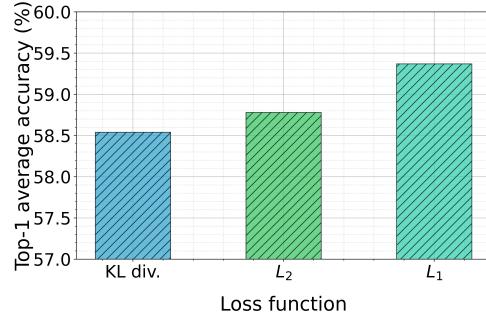


Figure 4: Effect of performance on the choice of loss function for distribution alignment.

Acknowledgements

The computational resources were provided by the National Academic Infrastructure for Supercomputing in Sweden (NAIIS), partially funded by the Swedish Research Council through grant agreement No. 2022-06725, and by the Berzelius resource, provided by Knut and Alice Wallenberg Foundation at the National Supercomputer Center.

References

- [1] Hyojin Bahng, Ali Jahanian, Swami Sankaranarayanan, and Phillip Isola. Exploring visual prompts for adapting large-scale models. *arXiv preprint arXiv:2203.17274*, 1(3):4, 2022.
- [2] Hyojin Bahng, Ali Jahanian, Swami Sankaranarayanan, and Phillip Isola. Visual prompting: Modifying pixel space to adapt pre-trained models. *arXiv preprint arXiv:2203.17274*, 2022.
- [3] Florian Bordes, Shashank Shekhar, Mark Ibrahim, Diane Bouchacourt, Pascal Vincent, and Ari S Morcos. Pug: Photorealistic and semantically controllable synthetic data for representation learning. *arXiv preprint arXiv:2308.03977*, 2023.
- [4] Lukas Bossard, Matthieu Guillaumin, and Luc Van Gool. Food-101—mining discriminative components with random forests. In *Computer Vision—ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6–12, 2014, Proceedings, Part VI 13*, pages 446–461. Springer, 2014.
- [5] Mehdi Cherti, Romain Beaumont, Ross Wightman, Mitchell Wortsman, Gabriel Ilharco, Cade Gordon, Christoph Schuhmann, Ludwig Schmidt, and Jenia Jitsev. Reproducible scaling laws for contrastive language-image learning. *arXiv preprint arXiv:2212.07143*, 2022.
- [6] Mircea Cimpoi, Subhransu Maji, Iasonas Kokkinos, Sammy Mohamed, and Andrea Vedaldi. Describing textures in the wild. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3606–3613, 2014.
- [7] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255. IEEE, 2009.
- [8] Andre Esteva, Brett Kuprel, Roberto A Novoa, Justin Ko, Susan M Swetter, Helen M Blau, and Sebastian Thrun. Dermatologist-level classification of skin cancer with deep neural networks. *Nature*, 542(7639):115–118, 2017.
- [9] Yuxin Fang, Wen Wang, Binhui Xie, Quan Sun, Ledell Wu, Xinggang Wang, Tiejun Huang, Xinlong Wang, and Yue Cao. Eva: Exploring the limits of masked visual representation learning at scale. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19358–19369, 2023.
- [10] Li Fei-Fei, Rob Fergus, and Pietro Perona. Learning generative visual models from few training examples: An incremental bayesian approach tested on 101 object categories. In *2004 Conference on Computer Vision and Pattern Recognition Workshop*, pages 178–178. IEEE, 2004.
- [11] Yossi Gandelsman, Yu Sun, Xinlei Chen, and Alexei Efros. Test-time training with masked autoencoders. *Advances in Neural Information Processing Systems*, 35:29374–29385, 2022.
- [12] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1026–1034, 2015.
- [13] Patrick Helber, Benjamin Bischke, Andreas Dengel, and Damian Borth. Eurosat: A novel dataset and deep learning benchmark for land use and land cover classification. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 12(7):2217–2226, 2019.
- [14] Dan Hendrycks and Thomas Dietterich. Benchmarking neural network robustness to common corruptions and perturbations. *arXiv preprint arXiv:1903.12261*, 2019.
- [15] Dan Hendrycks, Steven Basart, Norman Mu, Saurav Kadavath, Frank Wang, Evan Dorundo, Rahul Desai, Tyler Zhu, Samyak Parajuli, Mike Guo, et al. The many faces of robustness: A critical analysis of out-of-distribution generalization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8340–8349, 2021.

- [16] Dan Hendrycks, Kevin Zhao, Steven Basart, Jacob Steinhardt, and Dawn Song. Natural adversarial examples. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15262–15271, 2021.
- [17] Chao Jia, Yafei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In *ICML*, pages 4904–4916. PMLR, 2021.
- [18] Muhammad Uzair Khattak, Hanoona Rasheed, Muhammad Maaz, Salman Khan, and Fahad Shahbaz Khan. Maple: Multi-modal prompt learning. In *The IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023.
- [19] Muhammad Uzair Khattak, Syed Talal Wasim, Muzammal Naseer, Salman Khan, Ming-Hsuan Yang, and Fahad Shahbaz Khan. Self-regulating prompts: Foundational model adaptation without forgetting. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15190–15200, 2023.
- [20] Jonathan Krause, Michael Stark, Jia Deng, and Li Fei-Fei. 3d object representations for fine-grained categorization. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*, pages 554–561, 2013.
- [21] Ananya Kumar, Aditi Raghunathan, Robbie Jones, Tengyu Ma, and Percy Liang. Fine-tuning can distort pretrained features and underperform out-of-distribution. In *International Conference on Learning Representations*, 2022.
- [22] Wei Lin, Muhammad Jehanzeb Mirza, Mateusz Kozinski, Horst Possegger, Hilde Kuehne, and Horst Bischof. Video test-time adaptation for action recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22952–22961, 2023.
- [23] Yuejiang Liu, Parth Kothari, Bastien van Delft, Baptiste Bellot-Gurlet, Taylor Mordan, and Alexandre Alahi. Ttt++: When does self-supervised test-time training fail or thrive? In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, volume 34, pages 21808–21820. Curran Associates, Inc., 2021. URL https://proceedings.neurips.cc/paper_2021/file/b618c3210e934362ac261db280128c22-Paper.pdf.
- [24] Subhransu Maji, Esa Rahtu, Juho Kannala, Matthew Blaschko, and Andrea Vedaldi. Fine-grained visual classification of aircraft. *arXiv preprint arXiv:1306.5151*, 2013.
- [25] M Jehanzeb Mirza, Jakub Micorek, Horst Possegger, and Horst Bischof. The norm must go on: Dynamic unsupervised domain adaptation by normalization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14765–14775, 2022.
- [26] Muhammad Jehanzeb Mirza, Pol Jané Soneira, Wei Lin, Mateusz Kozinski, Horst Possegger, and Horst Bischof. Actmad: Activation matching to align distributions for test-time-training. *arXiv preprint arXiv:2211.12870*, 2022.
- [27] Maria-Elena Nilsback and Andrew Zisserman. Automated flower classification over a large number of classes. In *2008 Sixth Indian Conference on Computer Vision, Graphics & Image Processing*, pages 722–729. IEEE, 2008.
- [28] Omkar M Parkhi, Andrea Vedaldi, Andrew Zisserman, and CV Jawahar. Cats and dogs. In *2012 IEEE Conference on Computer Vision and Pattern Recognition*, pages 3498–3505. IEEE, 2012.
- [29] Joaquin Quinonero-Candela, Masashi Sugiyama, Anton Schwaighofer, and Neil D Lawrence. *Dataset Shift in Machine Learning*. MIT Press, 2008.
- [30] 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.
- [31] Benjamin Recht, Rebecca Roelofs, Ludwig Schmidt, and Vaishaal Shankar. Do imagenet classifiers generalize to imagenet? In *International Conference on Machine Learning*, pages 5389–5400. PMLR, 2019.
- [32] Steffen Schneider, Evgenia Rusak, Luisa Eck, Oliver Bringmann, Wieland Brendel, and Matthias Bethge. Improving robustness against common corruptions by covariate shift adaptation. *Advances in Neural Information Processing Systems*, 33:11539–11551, 2020.

- [33] Christoph Schuhmann, Richard Vencu, Romain Beaumont, Robert Kaczmarczyk, Clayton Mullis, Aarush Katta, Theo Coombes, Jenia Jitsev, and Aran Komatsuzaki. Laion-400m: Open dataset of clip-filtered 400 million image-text pairs. *arXiv preprint arXiv:2111.02114*, 2021.
- [34] Manli Shu, Weili Nie, De-An Huang, Zhiding Yu, Tom Goldstein, Anima Anandkumar, and Chaowei Xiao. Test-time prompt tuning for zero-shot generalization in vision-language models. *Advances in Neural Information Processing Systems*, 35:14274–14289, 2022.
- [35] Khurram Soomro, Amir Roshan Zamir, and Mubarak Shah. A dataset of 101 human action classes from videos in the wild. *Center for Research in Computer Vision*, 2(11), 2012.
- [36] Yu Sun, Xiaolong Wang, Zhuang Liu, John Miller, Alexei Efros, and Moritz Hardt. Test-time training with self-supervision for generalization under distribution shifts. In *International Conference on Machine Learning*, pages 9229–9248. PMLR, 2020.
- [37] Dequan Wang, Evan Shelhamer, Shaoteng Liu, Bruno Olshausen, and Trevor Darrell. Tent: Fully test-time adaptation by entropy minimization. *arXiv preprint arXiv:2006.10726*, 2020.
- [38] Haohan Wang, Songwei Ge, Zachary Lipton, and Eric P Xing. Learning robust global representations by penalizing local predictive power. *Advances in Neural Information Processing Systems*, 32, 2019.
- [39] Qin Wang, Olga Fink, Luc Van Gool, and Dengxin Dai. Continual test-time domain adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7201–7211, 2022.
- [40] Lewei Yao, Runhui Huang, Lu Hou, Guansong Lu, Minzhe Niu, Hang Xu, Xiaodan Liang, Zhenguo Li, Xin Jiang, and Chunjing Xu. Filip: Fine-grained interactive language-image pre-training. *arXiv preprint arXiv:2111.07783*, 2021.
- [41] Lu Yuan, Dongdong Chen, Yi-Ling Chen, Noel Codella, Xiyang Dai, Jianfeng Gao, Houdong Hu, Xuedong Huang, Boxin Li, Chunyuan Li, et al. Florence: A new foundation model for computer vision. *arXiv preprint arXiv:2111.11432*, 2021.
- [42] Yuhang Zang, Wei Li, Kaiyang Zhou, Chen Huang, and Chen Change Loy. Unified vision and language prompt learning. *arXiv preprint arXiv:2210.07225*, 2022.
- [43] Werner Zellinger, Thomas Grubinger, Edwin Lughofer, Thomas Natschläger, and Susanne Saminger-Platz. Central moment discrepancy (cmd) for domain-invariant representation learning. *arXiv preprint arXiv:1702.08811*, 2017.
- [44] Xiaohua Zhai, Xiao Wang, Basil Mustafa, Andreas Steiner, Daniel Keysers, Alexander Kolesnikov, and Lucas Beyer. Lit: Zero-shot transfer with locked-image text tuning. In *CVPR*, pages 18123–18133, 2022.
- [45] Marvin Zhang, Sergey Levine, and Chelsea Finn. Memo: Test time robustness via adaptation and augmentation. *Advances in Neural Information Processing Systems*, 35:38629–38642, 2022.
- [46] Pan Zhang, Bo Zhang, Ting Zhang, Dong Chen, Yong Wang, and Fang Wen. Prototypical pseudo label denoising and target structure learning for domain adaptive semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12414–12424, 2021.
- [47] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Conditional prompt learning for vision-language models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.
- [48] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Learning to prompt for vision-language models. *International Journal of Computer Vision (IJCV)*, 2022.

Appendix

In the following, we provide further details on our implementation under Section A. In Section B, we provide additional results of our proposed PromptAlign method on cross-dataset generalization, base-to-novel generalization and for different CLIP backbone architectures; and Section C shows additional qualitative results. Finally, Section D presents ablation trends across diverse datasets for the effect of the proxy dataset, the effect of test sample statistics and different parameter choices.

A Additional Implementation details

A.1 Hardware and Software details

We implement our method with the MaPLE [18] multi-modal prompting model with the CLIP ViT-B/16 backbone architecture. Our models were implemented on a single NVIDIA A100 40GB GPU using the PyTorch framework.

A.2 Effect of layer depth for token distribution alignment loss

The token distribution alignment loss is computed from all the corresponding tokens of 64 crops spanning the first 3 layers of the visual encoder. This is chosen in line with the baseline multimodal prompt learning model [18] where the deep prompts are present in layers 1-3. In order to analyze the impact of different layer depths on the alignment loss, we conduct an ablation study. Figure 6 displays the performance comparison of alignment losses obtained from various layer combinations using the ImageNet-A dataset. As observed in the Figure 6, enforcing the token alignment loss across the first three layers of the vision encoder of CLIP provides the best performance.

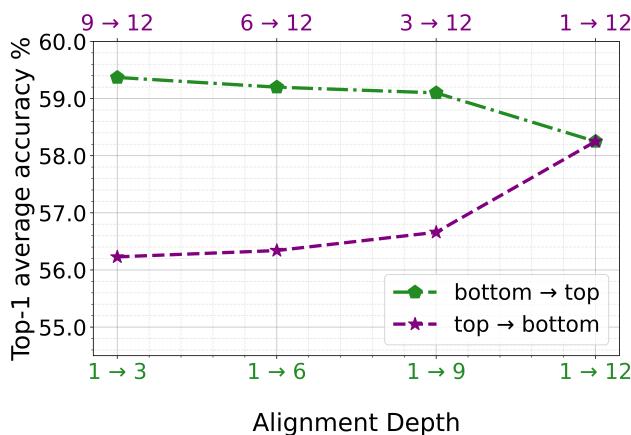


Figure 6: **Effect of alignment depth on performance:** We examine how the performance of PromptAlign is affected by altering the number of alignment layers that contribute to the alignment loss. Our findings indicate that the token distribution alignment loss is most impactful when applied to the lower layers (specifically, layer 1 → 3). We hypothesize that the effectiveness of the token distribution alignment loss in these layers can be attributed to the presence of our prompts, which are located within the first three layers of the visual encoder.

B Additional results

B.1 Cross dataset evaluations

In Table 4 of the main paper, we compared the performance of PromptAlign with the current state-of-the-art prompt learning methods, CoOp and CoCoOp, for cross-dataset generalization. Additionally, in Table 8, we evaluate existing prompt learning methods that have been integrated with TPT and present further comparisons of our method with them. As shown in Table 8, PromptAlign outperforms both CoOp+TPT and CoCoOp+TPT in 9 out of 10 datasets. It is worth noting that when integrated with TPT, the performance of CoOp and CoCoOp decreases for the majority of cases. In contrast, PromptAlign consistently improves performance across different datasets.

B.2 Base to Novel Generalization

We also evaluate the performance of our method on the base-to-novel generalization benchmark setting. On average, our distribution alignment strategy improves the model performance by nearly 1% in the base classes and by 0.79% on novel classes. We observe that TPT alone drops the performance of the model in some instances such as for OxfordPets, Eurosat and UCF101. On such instances, PromptAlign regulates the drop in accuracy cause by TPT, while in most cases achieves better performance in novel classes of these datasets.

The loss scaling factor β plateaus close to 100, whereas the performance increases with the increasing number of augmented views and prompt update steps. As discussed in Section 5, the token distribution alignment scaling factor β is set to 100 using the ImageNet validation set, which is used for all experiments.

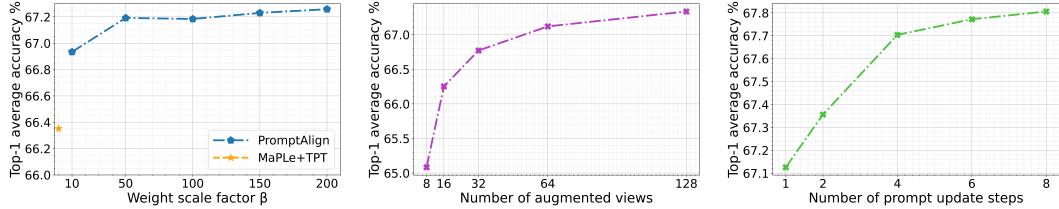


Figure 8: Effect of loss scaling factor β (left), number of augmented views (center), and number of prompt update steps (right) averaged across the 5 datasets. The loss scaling factor trend is similar to that on ImageNet validation set, and the average accuracy increases consistently with the number of augmented views and the number of prompt update steps.

E Analysis of losses for distribution alignment

The comparison of different alignment loss functions is presented in Fig 4 of the main paper. All three losses improve on top of the test-time prompting approach. In terms of a test-time setting, the L1 loss appears better suited to equally valuing all errors as opposed to L2 which primarily penalizes outliers. Moreover, the KL divergence loss does not enforce strict distribution alignment hence leading to sub-par performance. We further show a comparative analysis of alignment loss gradients in Table 15. Each row indicates the ratio of the average gradient in comparison to the gradient when using each of the other losses. In our experiments we notice that the shift in the statistical measures are small, which would lead to very small gradients in the case of L2 and KL divergence losses. This can be observed from Table 15. Specifically, the gradient of the L1 loss is almost 80 times greater than that of the L2 loss and approximately 4 times greater than that of the KL divergence loss. It is evident that the L1 Loss results in improved alignment, primarily due to its higher gradient values, indicating that prompt updates with an L1 penalty can yield better alignment.

Table 15: Comparison of different alignment losses. L1 Loss yields better alignment due smaller deviation in statistical measures, thus higher gradient values.

	L1-Loss	L2-Loss	KL-Divergence
L1-Loss	1	81.40	3.95
L2-Loss	0.0123	1	0.048
KL-Divergence	0.25	20.5	1

F Limitations and Future Directions

As a test-time adaptation strategy, PromptAlign improves CLIP performance but relies on the proxy source dataset statistics for aligning the train data distribution with the test sample statistics. Further, it involves a single update step during runtime to optimize the prompts. A potential future direction would be to explore the adaptation of PromptAlign to other downstream tasks such as object detection and segmentation.