FLATNESS GUIDED TEST-TIME ADAPTATION FOR VISION-LANGUAGE MODELS

Anonymous authors
Paper under double-blind review

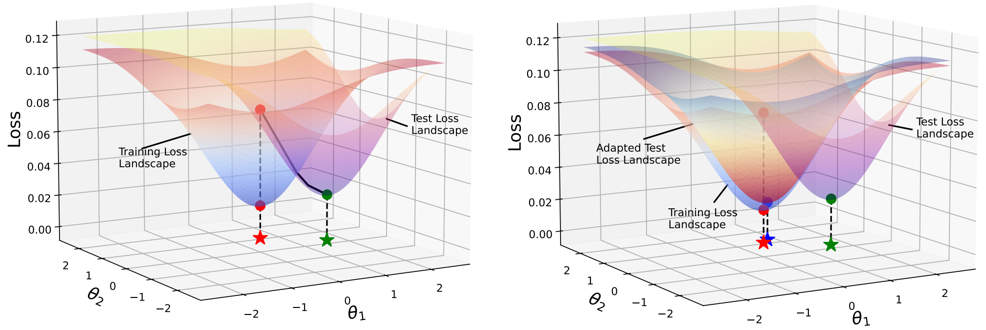
ABSTRACT

Test-time adaptation (TTA) of Vision-Language Models (VLMs) has emerged as a technique for tackling distribution shifts during the test time. Recent research indicates that the test-time adaptation is intrinsically linked to the model's training history. However, existing TTA methods, such as Test-time Prompt Tuning, often design adaptation strategies in isolation from the models' training characteristics, which degrade their performance. This paper argues that the flatness acquired via sharpness-aware training is an efficient clue for the test-time adaptation of VLMs. Built on this insight, this paper proposes a novel Flatness-Guided Adaptation framework (FGA) for VLMs to cohesively unify training and test-time procedures. Its core idea is to leverage the alignment between the training minimum and test loss flat regions to guide the adaptation process. Specifically, our FGA consists of a prompt-tuning stage and a test-time adaptation stage. In the tuning stage, a Sharpness-Aware Prompt Tuning method is utilized to identify the training flat minimum, offering a geometric clue of flatness for subsequent adaptation. In the test stage, a Sharpness-based Test Sample Selection approach is proposed to ensure the alignment of flat minima between the training and each augmented test sample's loss landscape. In comparison to existing TTA methods, our FGA avoids the expensive prompt parameter updates during test time, and substantially reduces the computation overhead. Extensive experiments on both domain generalization and cross-dataset benchmarks demonstrate that our FGA achieves superior performance over prevalent TTA methods. Notably, FGA even surpasses SOTA performance by 4.55% on ImageNet-A, when using a ViT-B/16 image encoder. Our code will be available soon.

1 INTRODUCTION

Recent advancements in vision-language pretraining, such as CLIP (Radford et al., 2021), have generated new opportunities for developing foundational models in vision tasks (Jia et al., 2021; Yang et al., 2022). These models, trained on extensive collections of image-text pairs, can learn and represent a diverse range of visual concepts. By means of well-designed prompts, they can be applied to downstream tasks in a zero-shot manner without requiring task-specific data (Li et al., 2022; Ramesh et al., 2022; Patashnik et al., 2021). Consequently, various prompt tuning methods (Zhou et al., 2022b;a) are proposed to directly learn prompts using training data from downstream tasks. Though these methods find better prompts compared to hand-crafted ones, the learned prompts are limited to the training distribution and may have limited generalization beyond that.

To address this issue, several studies (Chen et al., 2022; Boudiaf et al., 2022; Wang et al., 2020) have attempted to develop test-time adaptation (TTA) methods, which aim to rapidly adjust pre-trained models to unlabeled test data streams during inference. Among various TTA strategies, methods based on Test-Time Prompt Tuning (TPT) (Shu et al., 2022), which optimize a set of learnable prompts via entropy minimization on augmented test views, have demonstrated promising performance and gained significant attraction (Shu et al., 2022; Feng et al., 2023; Yoon et al., 2024). Recent research indicates that the test-time adaptation is intrinsically influenced by the model's training history (Goyal et al., 2022). However, most existing TTA methods, including TPT-based methods, design adaptation strategies in isolation, treating the test phase as a standalone optimization problem disconnected from the model's training history. This isolation from the training phase may fail to



(a) **Traditional TTA Methods**

(b) **Our FGA**

Figure 1: **Comparison of conventional TTA methods and Flatness-Guided Adaptation (FGA).** (a) Traditional TTA methods treat the test landscape as static, aiming to optimize parameters to achieve the test flat minimum (★), using the training minimum (★) as an initialization. (b) Our FGA keeps parameters unchanged during testing. Instead, it adjusts the test landscape to a position where the training minimum (★) is already very close to the minimum (★) of the adapted test landscape.

exploit valuable geometric and representational properties inherent in the pre-trained model, leading to suboptimal test-time adaptation.

To improve model generalization, seeking flat minima within the training loss landscape has emerged as an effective training strategy over the past few years (Foret et al., 2020; Kwon et al., 2021; Kim et al., 2022). It is widely observed that parameters residing in flat minima tend to generalize better to out-of-distribution data (Cha et al., 2021; Zhang et al., 2024b; Li et al., 2025; Zou et al., 2024) than those sharp ones. Nonetheless, conventional TTA methods often ignore the influence of sharpness-aware training on the test-time adaptation. While Sharpness-Aware Minimization (SAM) (Foret et al., 2020) seeks flat regions during training, its principle is rarely extended to guide test-time adaptation in a unified framework. This disconnection leads to computationally expensive test-time optimizations (e.g., backpropagation in TPT (Shu et al., 2022)) that are agnostic to loss geometric structure and often yield suboptimal generalization. This paper argues that the flatness is not merely a desirable property during training but a powerful clue that can dictate test-time adaptation.

Inspired by this insight, this paper proposes a novel *Flatness-Guided Adaptation* (FGA) framework, which cohesively unifies training and test-time procedures from the perspective of loss landscape geometry. It mainly leverages the alignment between the training minimum and test loss flat regions to guide the adaptation process (see Figure 1). Specifically, our FGA framework consists of two synergistic stages: (1) In the prompt tuning stage, a *Sharpness-Aware Prompt Tuning* (SAPT) method is utilized to fine-tune the prompts on the downstream training dataset, aiming at seeking the training flat minimum. Since flatter minima generally indicate better model generalization than sharper ones (Keskar et al., 2016; Dziugaite & Roy, 2017; Jiang et al., 2019; Foret et al., 2020), the minimization of sharpness not only improves model generalization but also provides a test-time criterion to measure the alignment of flat minima within the training and test loss landscapes. (2) In the test-time stage, FGA leverages the geometric clue of flatness acquired via SAPT. For a given test sample, a *Sharpness-based Test Sample Selection* (STSS) method is proposed to intelligently select its augmented views based on the sharpness score of their loss landscapes around the training flat minimum. This ensures that the final prediction is derived from a test-time loss landscape whose flat minima align with those identified during training. During this process, loss landscapes are efficiently altered through data augmentations. In comparison with existing TTA methods, our FGA avoids the expensive prompt parameter updates during test time, eliminating the computational overhead of adaptation and offering a more plausible adaptation strategy. Theoretical analysis suggests that using the sharpness-based metric will help distinguish the proximity of test samples to the training distribution. The closer an augmented sample is to the training distribution, the smaller its sharpness-based score is likely to be. Since models tend to generate more reliable results for data closer to the training distribution, FGA significantly improves the generalization ability of vision-language models. Extensive experiments on domain generalization (Hendrycks et al., 2021b) and cross-dataset (Zhou et al., 2022a) benchmarks demonstrate the superior performance of FGA over prevailing TTA methods.

Our main contributions can be summarized as follows:

- A novel Flatness-Guided Adaptation (FGA) framework is proposed to cohesively unify training and test-time procedures for vision-language models. By ensuring the alignment of model's training flat minimum with flat regions in test loss landscapes, it significantly enhances the generalization capabilities of VLMs under distribution shifts.
- Theoretical analysis is presented to offer a clearer insight into how sample selection at test time improves the reliability of predictions.
- Extensive experiments on domain generalization and cross-dataset benchmarks demonstrate the superior performance of FGA over other prevalent TTA methods, while significantly eliminating the computational overhead.

2 Related Work

Test-time adaptation (TTA) of vision-language models. Vision-language models like CLIP have shown strong performance in various tasks. To enhance CLIP's transfer learning for downstream classification tasks, methods like text prompt learners (e.g., CoOp (Zhou et al., 2022b) and Co-CoOp (Zhou et al., 2022a)) and visual adapters (e.g., Tip-Adapter (Zhang et al., 2022)) have been proposed. However, these methods struggle with distribution misalignment between pre-training and test data. Test-time adaptation (TTA) methods address this by adjusting models during testing, with two main streams (Abdul Samadh et al., 2024): the first modifies the training process using a self-supervised proxy task, such as image rotation prediction, and uses it to guide test-time optimization (e.g., Test-Time Training (Sun et al., 2020) and TTT++ (Liu et al., 2021)); the second adapts models without altering the training process (e.g., TPT (Shu et al., 2022), which uses entropy minimization to learn adaptive parameters during testing). DiffTPT (Feng et al., 2023) introduces a diffusion model to generate diverse augmentations for further improvements. PromptAlign (Abdul Samadh et al., 2024) adds an explicit term to align the learned distributions with that of test data. Meanwhile, online methods like TDA (Karmanov et al., 2024) and DPE (Zhang et al., 2024a) use a key-value cache or prototype set to adapt progressively to test data. They benefit from information aggregated during testing but are unsuitable for single test-sample scenarios, unlike TPT-based methods. This paper proposes a novel Flatness-Guided Adaptation (FGA) framework that leverages the geometry of loss landscapes to enhance CLIP's generalization and inference efficiency in single test-sample adaptation scenarios. By avoiding backpropagation and parameter updates during testing, FGA significantly reduces computational overhead while achieving robust out-of-domain performance.

Generalization from a loss landscape view. In recent years, optimization techniques aimed at flat minima in loss landscapes have surged to improve the generalization of deep models (Keskar et al., 2016; Dziugaite & Roy, 2017; Jiang et al., 2019). Among them, SAM (Foret et al., 2020), which focuses on finding parameters located in regions of the loss landscape with consistently low loss values, has gained significant attention for its effectiveness and scalability. To seek flatter minima, numerous SAM variants, such as ASAM (Kwon et al., 2021) and FisherSAM (Kim et al., 2022), have already been developed over the past few years. This concept of flat minima has also been extended to improve the out-of-domain generalization of deep models (Zou et al., 2024; Cha et al., 2021; Li et al., 2025). However, most of them focus on the training stage. SAR (Niu et al., 2023) and SoTTA (Gong et al., 2023), two online test-time adaptation (TTA) methods, both utilize sharpness-aware minimization at test time to improve robustness by seeking flat minima. Yet, they operate solely during testing without accounting for training-testing sharpness interactions. In contrast, our FGA applies sharpness-aware minimization during training to establish flatness as a criterion for subsequent alignment, then at test time adapts by adjusting loss landscapes through augmentation selection—without updating model parameters—and preserving the pre-trained flat minimum's optimality on adapted test loss landscapes.

3 Methodology

3.1 Preliminaries

Contrastive Language-Image Pre-training. CLIP (Radford et al., 2021) primarily comprises a Text Encoder $\boldsymbol{E}_t$ and an Image Encoder $\boldsymbol{E}_v$. The Image Encoder is available in two architectures:

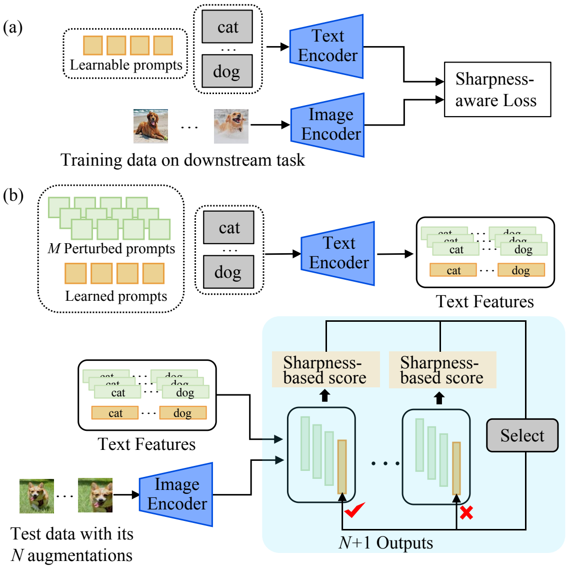


Figure 2: **Overview of our Flatness-Guided Adaptation (FGA).** It consists of two synergistic mechanisms: (a) Sharpness-aware Prompt Tuning: It optimizes the model parameters to reduce the loss value and sharpness, enabling stable and effective adaptations during test time without direct access to training data. (b) Sharpness-based Test Sample Selection: It introduces a selection mechanism to identify augmented test samples that ensure the training flat minimum aligns with those in their loss landscapes, enabling more confident predictions.

one based on ResNet (He et al., 2016) and the other using the popular Vision Transformer (ViT) (Dosovitskiy et al., 2020). This encoder transforms an input image $\boldsymbol{x}$ into its feature representation, i.e., $\boldsymbol{e}_i = \boldsymbol{E}_v(\boldsymbol{x})$. For a classification task with K classes, the corresponding class labels are formatted into a text template, "a photo of a [cls]", which is then mapped to tokens $\boldsymbol{y}_k = (\text{SOS}, t_1, t_2, \ldots, t_L, c_k, \text{EOS})$. Here, SOS and EOS represent the embeddings of the start and end tokens, while $t_1, t_2, \ldots, t_L$ corresponds to the phrase "a photo of a", and the token c_k denotes the specific description of the k-th class. The text encoder of CLIP, designed as a Transformer architecture, processes these tokens to generate text features: $\boldsymbol{e}_{t,k} = \boldsymbol{E}_t(\boldsymbol{y}_k)$. During the pre-training stage, CLIP is trained on the WIT dataset (Radford et al., 2021) through a contrastive learning approach. In this setup, each image is paired with its corresponding text sentence as a positive sample, while all other image-text combinations are treated as negative samples. The goal of the contrastive learning objective is to enhance the cosine similarity of positive pairs while reducing that of negative pairs. In the classification stage, all classes in the dataset are converted to text, and the cosine similarity between image embeddings and text embeddings is computed to determine the probability of an image belonging to each category:

$$p(y_k \mid \boldsymbol{x}) = \frac{\exp\left(\text{sim}\left(\boldsymbol{e}_{t,k} \cdot \boldsymbol{e}_i\right) \tau\right)}{\sum_{j=1}^{K} \exp\left(\text{sim}\left(\boldsymbol{e}_{t,j} \cdot \boldsymbol{e}_i\right) \tau\right)}, \tag{1}$$

where τ is the temperature of the softmax.

Prompt tuning. Prompt tuning has emerged as a popular tuning method for Transformer-based models in downstream tasks. This approach does not modify the model parameters; rather, it changes the input to the model, making it highly efficient. Specifically, instead of using the template "a photo of a [cls]", it replaces the tokens associated with the hand-crafted prompts ("a photo of a") with learnable parameters $\boldsymbol{p} = (p_1, \ldots, p_L)$, which are then updated based on the dataset used for downstream tasks.

Test-time prompt tuning. To prevent overfitting that may arise from prompts learned on the downstream training set—which may not perform effectively on test data with distribution shifts—test-time prompt tuning (TPT) (Shu et al., 2022) fine-tunes a specific prompt for each test sample. During testing, multiple augmented views of the test samples are generated. Then, predictions with entropy below a predetermined threshold are kept, while others are discarded using a confidence filter. The averaged entropy of selected predictions is then used as a loss function to update the prompts.

3.2 Flatness-Guided Adaptation

Our proposed Flatness-Guided Adaptation (FGA) framework fundamentally offers a unified, loss landscape-centric methodology that seamlessly bridges the training and test phases. This approach leverages the flatness as a universal guiding principle to enhance both generalization during training and robust adaptation during inference under distribution shifts. As illustrated in Figure 2, FGA integrates two complementary mechanisms that operate in concert: sharpness-aware prompt tuning (SAPT) during training and sharpness-guided test sample selection (STSS) during inference. This section will elaborates on the key idea and technical details of the framework.

3.2.1 Sharpness-Aware Prompt Tuning

FGA mainly exploits the alignment between training and test flat minima for the efficient adaptation of VLMs. However, a major challenge in achieving this alignment arises from inherent data constraints. Test samples remain unknown during training, and at test time, training data becomes inaccessible due to storage limitations and privacy requirements. To overcome this, FGA focuses on the intrinsic properties of the training minimum—ensuring that the test loss landscape shares these desirable characteristics: (1) the training minimum should correspond to a low loss value, indicating that the model is effectively learning from the data; (2) the training minimum may display implicit biases, such as reduced sharpness, which are often beneficial for improved generalization.

Traditional training methods for optimizing the prompts, such as CoOp (Zhou et al., 2022b), typically use cross-entropy loss to fine-tune the prompt $\boldsymbol{p}$:

$$\ell_{\mathrm{CE}}(\boldsymbol{p}) = -\sum_{i=1}^{n} \log p_{\boldsymbol{p}}(y_i|\boldsymbol{x}_i), \tag{2}$$

where $p_{\boldsymbol{p}}(y_i|\boldsymbol{x}_i)$ represents the predictive probability that $\boldsymbol{x}_i$ belongs to the its true label class. While standard SGD tends to find flat minima, methods such as SAM (Foret et al., 2020) can enhance this implicit bias through explicit perturbation-aware optimization. To enable more precise alignment during testing, we adopt Sharpness-aware Prompt Tuning (SAPT) during training, which jointly minimizes both the loss and its "sharpness":

$$\ell_{\mathrm{SAPT}}(\boldsymbol{p}) = \ell_{\mathrm{CE}}(\boldsymbol{p}) + \lambda \max_{||\boldsymbol{\epsilon}|| \le \rho} \left[\ell_{\mathrm{CE}}(\boldsymbol{p}+\boldsymbol{\epsilon}) - \ell_{\mathrm{CE}}(\boldsymbol{p})\right]. \tag{3}$$

The first and second terms above represent the loss value and loss sharpness, with λ acting as a hyperparameter to balance them. Similar to previous studies (Foret et al., 2020; Kwon et al., 2021), sharpness is defined as the sensitivity of the training loss to small perturbations $\boldsymbol{\epsilon}$ (with a norm less than ρ) added to the prompts $\boldsymbol{p}$. Since the perturbation strength ρ is small enough, we can apply a Taylor expansion to approximately solve for the optimal perturbation $\boldsymbol{\epsilon}^{\star}$:

$$\boldsymbol{\epsilon}^{\star} = \arg\max_{\|\boldsymbol{\epsilon}\| \le \rho} \ell_{\mathrm{CE}}(\boldsymbol{p}+\boldsymbol{\epsilon}) - \ell_{\mathrm{CE}}(\boldsymbol{p}) \approx \arg\max_{\|\boldsymbol{\epsilon}\| \le \rho} \boldsymbol{\epsilon}^{T} \nabla_{\boldsymbol{p}} \ell_{\mathrm{CE}}(\boldsymbol{p}) = \rho \frac{\nabla_{\boldsymbol{p}} \ell_{\mathrm{CE}}(\boldsymbol{p})}{\|\nabla_{\boldsymbol{p}} \ell_{\mathrm{CE}}(\boldsymbol{p})\|}. \tag{4}$$

During training via (stochastic) gradient descent, the contribution from $\nabla_{\boldsymbol{p}} \boldsymbol{\epsilon}^{\star}$ can be disregarded due to the minor perturbation strength ρ.

In this way, SAPT not only yields robust prompts that enhance generalization but also provides the sharpness measure as additional information for adaptation during testing.

3.2.2 Sharpness-based Test Sample Selection

Through sharpness-aware prompt tuning, the prompts are positioned at a flat minimum within the training loss landscape. To avoid computationally expensive gradient descent during inference, we keep the pre-trained prompt fixed and instead adapt the test loss landscapes such that the well-trained prompt from the downstream training dataset (i.e., the training flat minimum) coincides with the flat minimum in the adapted test landscape, as illustrated in Figure 1.

To achieve this alignment, we propose a Sharpness-based Test Sample Selection (STSS) method. STSS utilizes data augmentations to create multiple test loss landscapes for each sample. By selecting augmented samples that align the training minimum with flat minima in their respective loss landscapes, we ensure the training minimum remain optimal. Given that such alignment typically corresponds to small loss values and reduced loss sharpness in these test landscapes, STSS introduces a sharpness-based score as a metric. To mitigate the computational burden of backpropagation in calculating sharpness, we redefine it as the maximum variation in the loss resulting from M random perturbations:

$$\ell_{\mathrm{STSS}}(\boldsymbol{p}) = \ell_{\mathrm{SRG}}(\boldsymbol{p}) + \lambda \max_{m=1,\ldots,M;\boldsymbol{\epsilon_m}\sim\mathcal{N}} \left[\ell_{\mathrm{SRG}}\left(\boldsymbol{p} + \rho' \frac{\boldsymbol{\epsilon_m}}{\|\boldsymbol{\epsilon_m}\|}\right) - \ell_{\mathrm{SRG}}(\boldsymbol{p})\right]. \tag{5}$$

Table 1: **Results on datasets with natural distribution shifts.** We report top-1 accuracy (%) for each method across five datasets, using the CLIP-ViT-B/16 backbone. We highlight the best results in **bold** and underline the second best results. The abbreviation "IN" means the ImageNet dataset. † denotes results reproduced by adapting the method to the single-sample setting.

Algorithm	IN	IN-A	IN-V2	IN-R	IN-Sketch	Avg.	OOD Avg.
CLIP-ViT-B/16 (Radford et al., 2021)	68.34	49.89	61.88	77.65	48.24	61.20	59.42
CoOp (Zhou et al., 2022b)	71.51	49.71	64.20	75.21	47.99	61.72	59.28
CoCoOp (Zhou et al., 2022a)	71.02	50.63	64.07	76.18	48.75	62.13	59.91
Tip-Adapter (Zhang et al., 2022)	70.75	51.04	63.41	77.76	48.88	62.37	60.27
TPT (Shu et al., 2022)	69.70	53.67	64.30	73.90	46.40	61.59	59.57
DiffTPT (Feng et al., 2023)	70.30	55.68	65.10	75.00	46.80	62.58	60.64
C-TPT (Yoon et al., 2024)	-	52.90	63.40	78.00	48.50	-	60.70
ZERO (Farina et al., 2024)	69.06	61.35	64.13	77.28	48.29	64.02	62.76
MTA (Zanella & Ben Ayed, 2024)	69.29	57.41	63.61	76.92	48.58	63.16	
PromptAlign* (Abdul Samadh et al., 2024)	-	59.37	65.29	79.33	50.23	-	63.55
TDA (Karmanov et al., 2024)	69.51	60.11	64.67	80.24	50.54	65.01	63.89
DPE (Zhang et al., 2024a)	71.91	59.63	65.44	80.40	**52.26**	65.93	64.43
TPT (Shu et al., 2022)+CoOp	73.30	56.88	66.60	73.80	49.40	64.00	61.67
DiffTPT (Feng et al., 2023)+CoOp	**75.00**	58.09	66.80	73.90	49.50	64.66	62.07
C-TPT (Yoon et al., 2024)+CoOp	72.90	52.73	65.61	76.46	48.63	63.27	60.86
ZERO (Farina et al., 2024)+CoOp	73.61	63.17	66.82	77.71	48.52	65.97	64.05
MTA (Zanella & Ben Ayed, 2024)+CoOp	73.99	59.29	66.97	78.20	49.96	65.68	63.61
SAR† (Niu et al., 2023)+CoOp	73.03	55.35	65.89	77.09	48.65	64.00	61.75
FGA(SAPT only + CoOp)	70.79	51.04	64.41	77.66	49.31	62.64	60.61
FGA(STSS only + CoOp)	73.99	64.00	67.11	77.92	49.36	66.48	64.60
FGA (Ours)	74.01	**65.90**	**67.23**	**81.24**	51.81	**68.04**	**66.55**

Here, ℓ_{STSS} represents the sharpness-based score used to select the most reliable augmented test samples, and ℓ_{SRG} denotes a surrogate loss function when test labels are unavailable, such as entropy (Wang et al., 2020; Goyal et al., 2022). The perturbation direction is expressed by $\boldsymbol{\epsilon_m}/\|\boldsymbol{\epsilon_m}\|$, where $\boldsymbol{\epsilon_m}$ is drawn from the standard normal distribution $\mathcal{N}$. The term ρ' controls the magnitude of perturbations during testing. To obtain $\ell_{\mathrm{SRG}}\left(\boldsymbol{p} + \rho' \frac{\boldsymbol{\epsilon_m}}{\|\boldsymbol{\epsilon_m}\|}\right)$, we first obtain text features for each category ($\boldsymbol{e}_{t,k,m}$) through the forward pass of the text encoder:

$$[\boldsymbol{e}_{t,k,1}, \ldots, \boldsymbol{e}_{t,k,M}] = \boldsymbol{E}_t([\boldsymbol{y}_{k,1}, \ldots, \boldsymbol{y}_{k,M}]), \tag{6}$$

where the input sequence $\boldsymbol{y}_{k,m}$ for the k-th category and m-th perturbation consists of the tokens: $\boldsymbol{y}_{k,m} = (\mathrm{SOS}, \boldsymbol{p} + \rho' \boldsymbol{\epsilon_m}/\|\boldsymbol{\epsilon_m}\|, c_k, \mathrm{EOS})$. Notably, the additional computational cost of this step is minimal, as text features only need to be computed once per test category. Then, the surrogate loss for perturbed prompts is:

$$\ell_{\mathrm{SRG}}\left(\boldsymbol{p} + \rho' \frac{\boldsymbol{\epsilon_m}}{\|\boldsymbol{\epsilon_m}\|}\right) = -\sum_{k=1}^{K} p_m(y_k|\boldsymbol{x}) \log p_m(y_k|\boldsymbol{x}), \tag{7}$$

with probabilities derived from cosine similarity:

$$p_m(y_k|\boldsymbol{x}) = \frac{\exp\left(\mathrm{sim}\left(\boldsymbol{e}_{t,k,m} \cdot \boldsymbol{e}_i\right) \tau\right)}{\sum_{j=1}^{K} \exp\left(\mathrm{sim}\left(\boldsymbol{e}_{t,j,m} \cdot \boldsymbol{e}_i\right) \tau\right)}. \tag{8}$$

Finally, the final prediction aggregates votes from the top r augmented samples with the lowest sharpness-based scores, which are more reliable predictions according to the theoretical analysis in the next section.

4 Theoretical Analysis

This section provides a theoretical explanation of how our method improves test-time classification. Let's begin with the following problem: During training, the model learns from data sampled independently and identically from distribution $\mathcal{S}$; During testing, however, data is drawn from two distinct distributions $\mathcal{T}_1$ and $\mathcal{T}_2$. Then, the question is: *How can we distinguish between these test distributions and determine on which one the model will perform more reliably?*

Table 2: **Cross-dataset generalization from ImageNet to fine-grained classification datasets.** During the prompt tuning stage, the prompts are tuned on ImageNet with 16-shot training data per category, using a ViT-B/16 image encoder.

Method	Caltech101	Pets	Cars	Flowers102	Aircraft	SUN397	DTD	Eurosat	Food101	UCF101	Avg.
CLIP-ViT-B/16 (Radford et al., 2021)	93.35	88.25	65.48	67.44	23.67	62.59	44.27	42.01	83.65	65.13	63.58
CoOp (Zhou et al., 2022b)	93.70	89.14	64.51	68.71	18.47	64.15	41.92	46.39	85.30	66.55	63.88
CoCoOp (Zhou et al., 2022a)	94.43	90.14	65.32	71.88	22.94	67.36	45.73	39.23	83.97	68.44	64.94
TPT (Shu et al., 2022)	94.16	87.79	66.87	68.98	24.78	65.50	47.75	42.44	84.67	68.04	65.10
DiffTPT (Feng et al., 2023)	92.49	88.22	67.01	70.10	25.60	65.74	47.00	43.13	**87.23**	62.67	64.92
C-TPT (Yoon et al., 2024)	93.60	88.20	65.80	69.80	24.00	64.80	46.00	43.20	83.70	65.70	64.48
ZERO (Farina et al., 2024)	93.66	87.75	68.04	67.68	25.21	65.03	46.12	34.33	86.53	67.77	64.21
PromptAlign (Abdul Samadh et al., 2024)	94.01	90.76	68.50	**72.39**	24.80	67.54	47.24	47.86	86.65	69.47	66.92
TDA (Karmanov et al., 2024)	94.24	88.63	67.28	71.42	23.91	67.54	47.40	**58.00**	86.14	**70.66**	67.53
TPT+CoOp (Zhou et al., 2022b)	93.75	88.93	67.06	68.25	25.89	66.40	47.15	48.78	83.82	66.53	65.66
TPT+MaPLe (Khattak et al., 2023)	93.59	90.72	66.50	72.37	24.70	67.54	45.87	47.80	86.64	69.19	66.50
ZERO (Farina et al., 2024)+CoOp	93.85	88.36	64.90	67.23	19.14	64.73	43.62	33.53	82.67	66.61	62.46
ZERO (Farina et al., 2024)+MaPLe	94.48	90.60	68.58	71.62	26.25	68.20	45.86	42.17	86.77	69.87	66.42
FGA (Ours)	**96.96**	**91.28**	**68.93**	72.11	**26.97**	**69.29**	**49.76**	47.58	84.95	68.17	**67.60**

To address this, we first derive an upper bound for the generalization error, which quantifies the model's performance on unseen data from $\mathcal{T}_1$ and $\mathcal{T}_2$. We will then explore how, when the test distributions are sufficiently distinguishable, FGA can effectively distinguish between them. This is crucial because, as we will show, when the test distribution closely resembles the training distribution, the generalization error bound decreases, leading to more accurate predictions.

Theorem 1 (Generalization Bound) *Consider real-valued function class $\mathcal{F} = \{f_{\boldsymbol{\theta}}(\cdot)\}$, and a bounded loss function $\ell : \mathbb{R} \times \mathbb{R} \to [0, M]$. Define ℓ^ρ as:*

$$\ell^\rho(f_{\boldsymbol{\theta}}(\boldsymbol{x}), y) = \max_{\|\boldsymbol{\epsilon}\|_2 \le \rho} \ell(f_{\boldsymbol{\theta}+\boldsymbol{\epsilon}}(\boldsymbol{x}), y). \tag{9}$$

Assume that ℓ^ρ is μ-Lipschitz with respect to f :

$$|\ell^\rho(f, y) - \ell^\rho(f', y)| \le \mu |f - f'| . \tag{10}$$

Denote the training and test distribution as $\mathcal{S}$ and $\mathcal{T}$, respectively. Then, with probability at least $1 - \delta$, the following inequality holds:

$$\mathbb{E}_{\mathcal{T}}[\ell^\rho(f_{\boldsymbol{\theta}}(\boldsymbol{X}_{\mathcal{T}}), Y_{\mathcal{T}})] \le \frac{M}{2} d_{\mathcal{F}\Delta\mathcal{F}}(\mathcal{S}; \mathcal{T}) + \hat{\ell}^\rho_{\mathcal{S}}(f_{\boldsymbol{\theta}}) + 2\mu R_n(\mathcal{F}, \mathcal{S}) + M\sqrt{\frac{\log(1/\delta)}{2n}}. \tag{11}$$

Here, $(\boldsymbol{X}_{\mathcal{T}}, Y_{\mathcal{T}})$ represents the random vector that follows the distribution $\mathcal{T}$. The term $d_{\mathcal{F}\Delta\mathcal{F}}(\mathcal{S}; \mathcal{T})$ quantifies the discrepancy between distributions $\mathcal{S}$ and $\mathcal{T}$, whose formal definition is provided in Appendix A. $R_n(\mathcal{F}, \mathcal{S})$ represents the Rademacher complexity (Zhang, 2023).

In the following, we will show that when the two test distributions are sufficiently distinguishable—compared with the tightness of the above upper bound—we can effectively differentiate between them. To proceed with this analysis, we first introduce the concepts of bound tightness and distribution separability.

Definition 2 (β-tightness) *Let α be an upper bound for the variable x such that $\Pr\{x \le \alpha\} \ge 1-\delta$. If there exists an oracle upper bound $\alpha^\star$ for which $\Pr\{x \le \alpha^\star\} = 1 - \delta$, we say that the upper bound is β-tight, where $\beta = |\alpha - \alpha^\star|$.*

Definition 3 (γ-separability) *Let $\mathcal{T}_1$ and $\mathcal{T}_2$ be two test distributions. We say that they are γ-separable if the condition $|d_{\mathcal{F}\Delta\mathcal{F}}(\mathcal{T}_1; \mathcal{S}) - d_{\mathcal{F}\Delta\mathcal{F}}(\mathcal{T}_2; \mathcal{S})| > \gamma$ holds. Here, $\mathcal{S}$ represents the training distribution.*

Theorem 4 *Consider a function class $\mathcal{F} = \{f_{\boldsymbol{\theta}}(\cdot)\}$, where the parameters $\boldsymbol{\theta}$ lie in a set such that the loss function is bounded within $[0, M]$. Let $p = (p_1, \ldots, p_K)$ and $q = (q_1, \ldots, q_K)$ be probability distributions over a finite set $\{1, \ldots, K\}$, with $p_i, q_i \ge \eta > 0$ for all i. Denote by $H(q)$ the entropy and by $H(p, q)$ the cross-entropy. Given a training distribution $\mathcal{S}$ and two γ-separable test distributions $\mathcal{T}_1$ and $\mathcal{T}_2$, assume $d_{\mathcal{F}\Delta\mathcal{F}}(\mathcal{S}, \mathcal{T}_1) < d_{\mathcal{F}\Delta\mathcal{F}}(\mathcal{S}, \mathcal{T}_2)$. Define the quantile function*

$Q_i(\delta)$ for the entropy loss of $f_{\boldsymbol{\theta}}$ on $\mathcal{T}_i$ such that $\Pr\{H^\rho < \mathbb{E}[H^\rho] + Q_i(\delta)\} = 1 - \delta$, and let $Q(\delta) = \sup\{Q_1(\delta), Q_2(\delta)\}$ be the supremum quantile. Then, with probability at least $1 - \delta$:

$$\begin{aligned}\mathbb{E}_{\mathcal{T}_i}\left[H^\rho(f_{\boldsymbol{\theta}}(X_{\mathcal{T}_i}))\right] \le & \frac{M}{2} d_{\mathcal{F}\Delta\mathcal{F}}\left(\mathcal{S};\mathcal{T}_i\right) + \hat{H}^\rho(Y_\mathcal{S}, f_\theta(X_\mathcal{S})) + 2\mu R_n(\mathcal{F},\mathcal{S}) + M\sqrt{\frac{\log(1/\delta)}{2n}} \\ & + \mathbb{E}_\mathcal{S}\|Y_\mathcal{S} - f_{\widetilde{\boldsymbol{\theta}}}(X_\mathcal{S})\|_1 + \frac{1}{\eta}\mathbb{E}_\mathcal{S}\|Y_\mathcal{S} - f_{\widetilde{\boldsymbol{\theta}}}(X_\mathcal{S})\|_1^2. \end{aligned} \tag{12}$$

Here, the notation $\widetilde{\boldsymbol{\theta}}$ is defined as $\widetilde{\boldsymbol{\theta}} := \boldsymbol{\theta} + \arg\max_{\|\boldsymbol{\epsilon}\| \le \rho} \max\{H(\boldsymbol{y}, f_{\boldsymbol{\theta}+\boldsymbol{\epsilon}}(\boldsymbol{x})), H(f_{\boldsymbol{\theta}+\boldsymbol{\epsilon}}(\boldsymbol{x}))\}$. Furthermore, if this bound is β_i-tight for $\mathcal{T}_1$ and $\mathcal{T}_2$ with $\beta_i < \gamma$, then there exists a threshold ξ such that:

$$\Pr\left\{H^\rho(f_\theta(X_{\mathcal{T}_1})) < \xi\right\} > \Pr\left\{H^\rho(f_\theta(X_{\mathcal{T}_2})) < \xi\right\}. \tag{13}$$

This inequality indicates that a test distribution further from the training distribution tends to exhibit a higher sharpness score. By comparing sharpness scores across test distributions, we can identify which one is closer to the training distribution, thus yielding more reliable predictions. Notably, the tunable parameter ρ controls the tightness of the upper bound, facilitating a precise differentiation between test distributions and improving the model performance. It is important to note that in the theoretical analysis presented in this section, we do not distinguish between ρ and ρ' (which are utilized to calculate the sharpness of the training and test loss landscapes, respectively). However, in practical implementation, we may opt to use different values for ρ and ρ' for better performance. Due to space limitations, detailed proofs and further discussions are provided in the appendix.

5 Experiments

5.1 Experimental Setup

Datasets. We conduct two types of experiments to evaluate the model's robustness to natural distribution shifts and its cross-dataset generalization capabilities, following previous research such as TPT (Shu et al., 2022). To assess the model's robustness to natural distribution shifts, we apply prompt tuning on the ImageNet (Deng et al., 2009) dataset, and evaluate its performance on four ImageNet variants—ImageNet-A (Hendrycks et al., 2021c), ImageNet-V2 (Recht et al., 2019), ImageNet-R (Hendrycks et al., 2021a) and ImageNet-Sketch (Wang et al., 2019)—which is also known as the domain generalization task. In addition, we perform cross-dataset evaluations for image classification across 10 datasets, each from a distinct domain with different classes: including Caltech101 (Fei-Fei et al., 2004), OxfordPets (Parkhi et al., 2012), StanfordCars (Krause et al., 2013), Flower102 (Nilsback & Zisserman, 2008), Aircraft (Maji et al., 2013), SUN397 (Xiao et al., 2010), DTD (Cimpoi et al., 2014), Food101 (Bossard et al., 2014), UCF101 (Soomro, 2012) and Eurosat (Helber et al., 2019). In this experiment, ImageNet serves as the source dataset, while the remaining fine-grained datasets are used as target datasets for evaluation.

Implementation details. Our experiments are based on pretrained CLIP (Radford et al., 2021) models, specifically CLIP-ResNet50 (using a ResNet50 image encoder) and CLIP-ViT-B/16 (using a Vision Transformer image encoder). Due to space limits, we focus on reporting the experimental results of CLIP-ViT-B/16, deferring those of CLIP-ResNet50 to the appendix. In the prompt tuning stage, our experiments are built on the CoOp (Zhou et al., 2022b) framework. The prompts are trained in a 16-shot manner on the ImageNet dataset. We set the number of prompts to 4 and utilize the SGD optimizer, with a learning rate of 0.002. For cross-dataset and domain generalization tasks, the prompts were trained for 5 and 50 epochs, with batch sizes of 4 and 32, respectively. The key hyperparameters ρ are determined through a grid search, with the values ranging from [0.05, 0.1, 0.3, 0.5, 0.7]. During testing, existing TPT-based methods usually leverage the input image along with its 63 augmented views. To ensure a fair comparison, we apply the same data augmentation strategy across all experiments. To avoid tuning hyperparameters on test data, we just set $\lambda = 1$ and $\rho' = 0.5$ for all experiments. Please refer to the Appendix for more discussions about other hyperparameters.

5.2 Main Results

Robustness to natural distribution shifts. We first compare the proposed FGA with prevalent TTA techniques on ImageNet and its variant OOD datasets. The results, presented in Table 1, highlight the

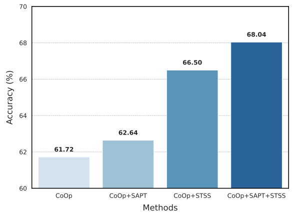


Figure 3: **Ablation study on main components of the proposed FGA.**

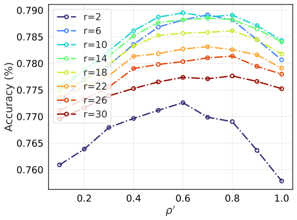
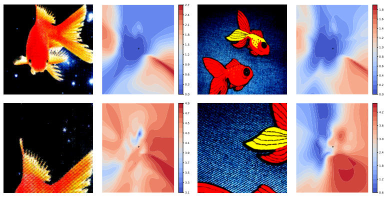


Figure 4: **The influence of the key hyperparameter ρ' on the test accuracy.**



(a) **Case 1**

(b) **Case 2**

Figure 5: **2D Visualization of loss landscapes associated with different augmented test samples.**

superior performance of FGA across several ImageNet-based OOD datasets. Notably, even the ablated version of our method, FGA (STSS only + CoOp), exhibits strong performance, surpassing all previous approaches with an OOD average of 64.60% and an overall average of 66.48%. We attribute this robustness to the fact that standard SGD training already imbues models with an implicit bias toward flatter minima. By explicitly enhancing this geometric property through SAPT, the full FGA algorithm achieves a substantial leap in generalization performance. Specifically, when compared to TPT+CoOp, FGA shows an average accuracy improvement of 4.88% (61.67% → 66.55%) on the OOD benchmark. Furthermore, our FGA also consistently surpasses other powerful TTA methods (e.g., DiffTPT, C-TPT, ZERO, MTA, and SAR) when they are combined with CoOp. It is critical to note that while online TTA methods like SAR benefit from aggregating information across a test data stream, FGA operates in a more challenging single-sample adaptation setting. For a fair comparison, we have adapted SAR to this setting. These superior results of FGA strongly demonstrate its effectiveness in enhancing CLIP's out-of-domain generalization across diverse datasets.

Cross-dataset generalization. We also observe superior performance of the FGA in evaluating cross-dataset generalization from ImageNet to various fine-grained classification benchmarks. Based on the comprehensive results presented in Table 2, the proposed FGA method demonstrates superior overall performance, achieving the highest average accuracy of 68.09% and attaining top-tier results on 6 out of 10 datasets, including a notably strong performance on Caltech101 (96.96%). Furthermore, FGA (67.60%) achieves an average accuracy improvement of 1.94% over the powerful baseline TPT+CoOp (65.66%). It also exhibits superior performance over the combinations of TTA methods (like TPT and ZERO) and different tuning methods (CoOp and MaPLe). These results further validate its effectiveness in adapting to diverse datasets during testing. It is important to note that due to the significant difference in the amount of target domain information available to different TTA settings, it is not imperative to expect single-sample TTA methods to surpass online TTA methods like TDA. It is particularly valuable for VLMs like CLIP, as it enables models to recognize more fine-grained categories in image classification without the need for additional training.

Runtime and Memory Efficiency. To quantify the computational advantage of FGA, we report runtime per test image and peak GPU memory usage, all measured on a single NVIDIA Tesla V100. Specifically, FGA achieves 22× faster inference than DiffTPT (0.07s vs 1.67s) and 9× speedup over TPT (0.07s vs 0.62s). Additionally, FGA's memory usage is 15× lower than TPT (1.28GB vs 19.24GB). These results validate FGA's computational efficiency, delivering high-performance test-time adaptation with minimal resource overhead. It is important to note that this work primarily focuses on the single test sample adaptation setting. Therefore, comparative analysis with online TTA methods that utilize aggregated test data falls outside the scope of the present investigation.

5.3 Ablation Study

Main components analysis. Our ablation study on the domain generalization benchmark (using CLIP-ViT-B/16 architecture, shown in Figure 3) validates the necessity of each FGA component: (1) Sharpness-aware prompt tuning (SAPT) enhances generalization, boosting CoOp's average accuracy by 0.92% (61.72%→62.64%) on ImageNet and OOD datasets; (2) Test-time sharpness selection (STSS) drives major gains, with CoOp+STSS outperforming CoOp by 4.82% (61.72%→66.54%); (3) SAPT synergistically enhances STSS, where full FGA (CoOp+SAPT+STSS) achieves a 5.40% gain over CoOp+SAPT (62.64%→68.04%)—exceeding standalone STSS improvements (4.82%). This confirms that flatter minima from SAPT intrinsically improve test-time sample selection.

Ablative analysis on key parameter ρ'. Theoretical analysis (Section 4) establishes ρ and ρ' as key generalization controllers, playing significant roles during the training and testing stages, respectively. Since ρ's role has been well explored in previous research (Foret et al., 2020), we focus on ρ' for test-time adaptation: as mentioned earlier, it governs distribution distinguishability, with proper values enhancing prediction reliability through sensitive discrimination. Empirical validation on ImageNet-R (Figure 4) shows non-monotonic accuracy dependence on ρ'—initially rising then falling. It is because extreme values ($\rho' \to 0$ or $\rho' \gg 0$) may yield uninformative sharpness measures and degrade performance. Crucially, $\rho' = 0$ degenerates to entropy maximization, and its comparison with non-zero cases also demonstrates sharpness's necessity. Notably, all experiments fix $\rho' = 0.5$ without test-data tuning, and this analysis solely aims to demonstrate the control effect of ρ' on generalization. Sample retention follows a similar trend: accuracy peaks then declines with increased retention. This reflects the probabilistic correlation: lower sharpness typically means greater proximity to the training distribution, meaning performance degrades when retaining excessively high-sharpness samples.

5.4 Visualization of Loss Landscapes

To intuitively validate FGA's effectiveness, we visualize the test data's loss surface using a 2D technique (Li et al., 2018) in Figure 5, revealing how sample selection enhances prediction reliability. The visualization demonstrates critical relationships: when parameters reside in flat minima (Figure 5, top), augmented samples maintain semantic integrity and enable reliable predictions. Conversely, parameters outside flat minima (bottom) yield distorted semantic representations that degrade generalization. This contrast directly demonstrates FGA's core mechanism—filtering unreliable test samples to prevent their negative impact, thereby boosting VLMs' generalization capacity.

6 Conclusion

This paper demonstrates that flatness operates not just as a beneficial training characteristic but as a key geometric clue for test-time adaptation. This understanding motivates the proposal of a novel framework, Flatness-Guided Adaptation (FGA), which utilizes the principle of loss landscape flatness as a unified guide to improve both training and test generalization against distribution shifts. Different from previous TTA methods that often fine-tune prompts per sample, it directs adaptation by leveraging the geometric relationship between training minima and test-time loss landscapes. Specifically, it first identifies flat minima during prompt tuning and then ensures the alignment across training and test landscapes via a selective mechanism. Comprehensive experiments and theoretical analysis confirm FGA's effectiveness and superior performance. We anticipate this work will advance the understanding of loss landscapes and inspire future TTA technologies.

Future Work. Our proposed FGA is grounded in a general analysis of loss landscape geometry, a foundational concept that is broadly applicable across diverse model architectures and learning paradigms. This foundation makes FGA a flexible component that could be integrated into modern visual-language models, advanced prompt tuning methods, or new types of test-time adaptation objectives to potentially enhance their performance. In this work, we intentionally followed the experimental setup introduced in the TPT paper. This choice allows a controlled and fair comparison with prior TTA methods, helping us to clearly demonstrate the contribution of FGA itself. The extension of FGA to other experimental configurations would require extensive engineering efforts to conduct large-scale experiments across diverse settings, and thus remains our future work.

References

Jameel Abdul Samadh, Mohammad Hanan Gani, Noor Hussein, Muhammad Uzair Khattak, Muhammad Muzammal Naseer, Fahad Shahbaz Khan, and Salman H Khan. Align your prompts: Test-time prompting with distribution alignment for zero-shot generalization. *Advances in Neural Information Processing Systems*, 36, 2024.

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*, pp. 446–461. Springer, 2014.

Malik Boudiaf, Romain Mueller, Ismail Ben Ayed, and Luca Bertinetto. Parameter-free online test-time adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8344–8353, 2022.

Junbum Cha, Sanghyuk Chun, Kyungjae Lee, Han-Cheol Cho, Seunghyun Park, Yunsung Lee, and Sungrae Park. Swad: Domain generalization by seeking flat minima. *Advances in Neural Information Processing Systems*, 34:22405–22418, 2021.

Dian Chen, Dequan Wang, Trevor Darrell, and Sayna Ebrahimi. Contrastive test-time adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 295–305, 2022.

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*, pp. 3606–3613, 2014.

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*, pp. 248–255. Ieee, 2009.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

Gintare Karolina Dziugaite and Daniel M Roy. Computing nonvacuous generalization bounds for deep (stochastic) neural networks with many more parameters than training data. *arXiv preprint arXiv:1703.11008*, 2017.

Matteo Farina, Gianni Franchi, Giovanni Iacca, Massimiliano Mancini, and Elisa Ricci. Frustratingly easy test-time adaptation of vision-language models. *arXiv preprint arXiv:2405.18330*, 2024.

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*, pp. 178–178. IEEE, 2004.

Chun-Mei Feng, Kai Yu, Yong Liu, Salman Khan, and Wangmeng Zuo. Diverse data augmentation with diffusions for effective test-time prompt tuning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 2704–2714, 2023.

Pierre Foret, Ariel Kleiner, Hossein Mobahi, and Behnam Neyshabur. Sharpness-aware minimization for efficiently improving generalization. *arXiv preprint arXiv:2010.01412*, 2020.

Taesik Gong, Yewon Kim, Taeckyung Lee, Sorn Chottananurak, and Sung-Ju Lee. Sotta: Robust test-time adaptation on noisy data streams. *Advances in Neural Information Processing Systems*, 36:14070–14093, 2023.

Sachin Goyal, Mingjie Sun, Aditi Raghunathan, and J Zico Kolter. Test time adaptation via conjugate pseudo-labels. *Advances in Neural Information Processing Systems*, 35:6204–6218, 2022.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

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.

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*, pp. 8340–8349, 2021a.

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*, pp. 8340–8349, 2021b.

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*, pp. 15262–15271, 2021c.

Chao Jia, Yinfei 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 *International conference on machine learning*, pp. 4904–4916. PMLR, 2021.

Yiding Jiang, Behnam Neyshabur, Hossein Mobahi, Dilip Krishnan, and Samy Bengio. Fantastic generalization measures and where to find them. *arXiv preprint arXiv:1912.02178*, 2019.

Adilbek Karmanov, Dayan Guan, Shijian Lu, Abdulmotaleb El Saddik, and Eric Xing. Efficient test-time adaptation of vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14162–14171, 2024.

Nitish Shirish Keskar, Dheevatsa Mudigere, Jorge Nocedal, Mikhail Smelyanskiy, and Ping Tak Peter Tang. On large-batch training for deep learning: Generalization gap and sharp minima. *arXiv preprint arXiv:1609.04836*, 2016.

Muhammad Uzair Khattak, Hanoona Rasheed, Muhammad Maaz, Salman Khan, and Fahad Shahbaz Khan. Maple: Multi-modal prompt learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 19113–19122, 2023.

Minyoung Kim, Da Li, Shell X Hu, and Timothy Hospedales. Fisher sam: Information geometry and sharpness aware minimisation. In *International Conference on Machine Learning*, pp. 11148–11161. PMLR, 2022.

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*, pp. 554–561, 2013.

Jungmin Kwon, Jeongseop Kim, Hyunseo Park, and In Kwon Choi. Asam: Adaptive sharpness-aware minimization for scale-invariant learning of deep neural networks. In *International conference on machine learning*, pp. 5905–5914. PMLR, 2021.

Aodi Li, Liansheng Zhuang, Xiao Long, Minghong Yao, and Shafei Wang. Seeking consistent flat minima for better domain generalization via refining loss landscapes. In *Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR)*, pp. 15349–15359, June 2025.

Hao Li, Zheng Xu, Gavin Taylor, Christoph Studer, and Tom Goldstein. Visualizing the loss landscape of neural nets. *Advances in neural information processing systems*, 31, 2018.

Liunian Harold Li, Pengchuan Zhang, Haotian Zhang, Jianwei Yang, Chunyuan Li, Yiwu Zhong, Lijuan Wang, Lu Yuan, Lei Zhang, Jenq-Neng Hwang, et al. Grounded language-image pre-training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10965–10975, 2022.

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? *Advances in Neural Information Processing Systems*, 34:21808–21820, 2021.

Subhransu Maji, Esa Rahtu, Juho Kannala, Matthew Blaschko, and Andrea Vedaldi. Fine-grained visual classification of aircraft. *arXiv preprint arXiv:1306.5151*, 2013.

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*, pp. 722–729. IEEE, 2008.

Shuaicheng Niu, Jiaxiang Wu, Yifan Zhang, Zhiquan Wen, Yaofo Chen, Peilin Zhao, and Mingkui Tan. Towards stable test-time adaptation in dynamic wild world. *arXiv preprint arXiv:2302.12400*, 2023.

Omkar M Parkhi, Andrea Vedaldi, Andrew Zisserman, and CV Jawahar. Cats and dogs. In *2012 IEEE conference on computer vision and pattern recognition*, pp. 3498–3505. IEEE, 2012.

Or Patashnik, Zongze Wu, Eli Shechtman, Daniel Cohen-Or, and Dani Lischinski. Styleclip: Text-driven manipulation of stylegan imagery. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 2085–2094, 2021.

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*, pp. 8748–8763. PMLR, 2021.

Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 1(2):3, 2022.

Benjamin Recht, Rebecca Roelofs, Ludwig Schmidt, and Vaishaal Shankar. Do imagenet classifiers generalize to imagenet? In *International conference on machine learning*, pp. 5389–5400. PMLR, 2019.

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.

K Soomro. Ucf101: A dataset of 101 human actions classes from videos in the wild. *arXiv preprint arXiv:1212.0402*, 2012.

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*, pp. 9229–9248. PMLR, 2020.

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.

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.

Jianxiong Xiao, James Hays, Krista A Ehinger, Aude Oliva, and Antonio Torralba. Sun database: Large-scale scene recognition from abbey to zoo. In *2010 IEEE computer society conference on computer vision and pattern recognition*, pp. 3485–3492. IEEE, 2010.

Jinyu Yang, Jiali Duan, Son Tran, Yi Xu, Sampath Chanda, Liqun Chen, Belinda Zeng, Trishul Chilimbi, and Junzhou Huang. Vision-language pre-training with triple contrastive learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 15671–15680, 2022.

Hee Suk Yoon, Eunseop Yoon, Joshua Tian Jin Tee, Mark Hasegawa-Johnson, Yingzhen Li, and Chang D Yoo. C-tpt: Calibrated test-time prompt tuning for vision-language models via text feature dispersion. *arXiv preprint arXiv:2403.14119*, 2024.

Maxime Zanella and Ismail Ben Ayed. On the test-time zero-shot generalization of vision-language models: Do we really need prompt learning? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 23783–23793, 2024.

Ce Zhang, Simon Stepputtis, Katia Sycara, and Yaqi Xie. Dual prototype evolving for test-time generalization of vision-language models. *Advances in Neural Information Processing Systems*, 37:32111–32136, 2024a.

Renrui Zhang, Wei Zhang, Rongyao Fang, Peng Gao, Kunchang Li, Jifeng Dai, Yu Qiao, and Hongsheng Li. Tip-adapter: Training-free adaption of clip for few-shot classification. In *European conference on computer vision*, pp. 493–510. Springer, 2022.

Ruipeng Zhang, Ziqing Fan, Jiangchao Yao, Ya Zhang, and Yanfeng Wang. Domain-inspired sharpness-aware minimization under domain shifts. *arXiv preprint arXiv:2405.18861*, 2024b.

Tong Zhang. *Mathematical analysis of machine learning algorithms*. Cambridge University Press, 2023.

Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Conditional prompt learning for vision-language models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 16816–16825, 2022a.

Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Learning to prompt for vision-language models. *International Journal of Computer Vision*, 130(9):2337–2348, 2022b.

Yingtian Zou, Kenji Kawaguchi, Yingnan Liu, Jiashuo Liu, Mong-Li Lee, and Wynne Hsu. Towards robust out-of-distribution generalization bounds via sharpness. *arXiv preprint arXiv:2403.06392*, 2024.