Spurious Feature Eraser: Stabilizing Test-Time Adaptation for Vision-Language Foundation Model

Huan Ma[1*†], **Yan Zhu**[1*], **Changqing Zhang**[1‡], **Peilin Zhao**[2], **Baoyuan Wu**[2], **Long-Kai Huang**[2], **Qinghua Hu**[1], **Bingzhe Wu**[2†]

[1]College of Intelligence and Computing, Tianjin University, Tianjin, China
[2]AI Lab, Tencent, Shenzhen, China

Abstract

Vision-language foundation models have exhibited remarkable success across a multitude of downstream tasks due to their scalability on extensive image-text paired data. However, these models also display significant limitations when applied to downstream tasks, such as fine-grained image classification, as a result of "decision shortcuts" that hinder their generalization capabilities. In this work, we find that the CLIP model possesses a rich set of features, encompassing both *desired invariant causal features* and *undesired decision shortcuts*. Moreover, the underperformance of CLIP on downstream tasks originates from its inability to effectively utilize pre-trained features in accordance with specific task requirements. To address this challenge, we propose a simple yet effective method, Spurious Feature Eraser (SEraser), to alleviate the decision shortcuts by erasing the spurious features. Specifically, we introduce a test-time prompt tuning paradigm that optimizes a learnable prompt, thereby compelling the model to exploit invariant features while disregarding decision shortcuts during the inference phase. The proposed method effectively alleviates excessive dependence on potentially misleading spurious information. We conduct comparative analysis of the proposed method against various approaches which validates the significant superiority.

Introduction

Vision-**L**anguage **F**oundation **M**odels (VLFMs) such as CLIP (Radford et al. 2021), have achieved remarkable success across a diverse range of downstream tasks (Chen et al. 2023a; Gu et al. 2021; Xu et al. 2023). This success can largely be attributed to their scalability on massive image-text pair data, enabling the model to capture high-quality representation of both image and text. Such a capability further facilitates zero-shot learning on out-of-distribution (OOD) data, not limited to the original training set (Kamath et al. 2021; Patashnik et al. 2021; Li et al. 2022; Ma et al. 2023a). However, despite the powerful zero-shot learning abilities of CLIP and its variants (Agarwal et al. 2021; Zhou et al. 2023b; Cao et al. 2015; Bai et al. 2024a), their application to certain downstream tasks, such as fine-grained image classification, reveals significant limitations (Yang et al. 2022; Shi et al. 2023; Ma et al. 2023b). The presence of "decision shortcuts" — the model's tendency to rely on simple, potentially superficial features for decision-making — severely hinders their generalization capabilities (Fan et al. 2023; Li et al. 2023). These shortcuts, often emerging as a byproduct of the training process, lead to a model that, while robust in familiar scenarios, fails to effectively adapt to more fine-grained or less common tasks. For example, as shown in Fig. 1, due to the existence of the shortcut of using the background for classification, the CLIP tends to predict the spider as a crab when the background is the beach.

To improve the zero-shot generalization ability of VLFMs, several methodologies have been proposed. These methods aim to enforce the model to employ invariant features during test phases, thereby alleviating the effect of decision shortcuts. Currently, two mainstream approaches worthy of consideration are: (1) Region-aware CLIP (Liang et al. 2023; Wei et al. 2023; Sun et al. 2023): This line introduces additional region information to the CLIP, encouraging it to avoid interference from spurious features. However, a significant limitation of these methods is the necessity to finetune the CLIP's weights or even alter its original architecture (Sun et al. 2023). Such alterations are costly and may also detrimentally affect the model's generalization abilities on in-distribution data due to changes in the model's architecture and weights. (2) Prompt Tuning (Shu et al. 2022): This strategy involves test-time optimization of task prompts using data from downstream tasks. The optimization goal is to force the model to learn invariant representations across various augmented versions of the original visual context (e.g., rotations & cropping), thereby mitigating the impact of visual shortcuts. Compared to the first approach, prompt tuning is advantageous as it does not require changing the original model's architecture or weights and can be applied out-of-the-box. However, these methods attempt to improve VLFMs' robustness by telling the models on which features to rely upon through various approaches, which heavily depends on the annotations of the invariant features. What's worse, forcing the model to

*These authors contributed equally.
†The project was conducted during the internship in AI Lab, Tencent with the mentor Bingzhe Wu (bingzhewu@tencent.com).
‡Corresponding to zhangchangqing@tju.edu.cn.

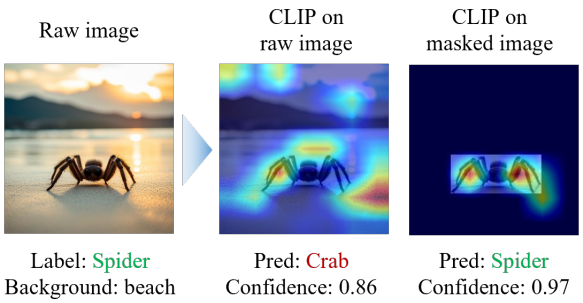

Figure 1: **Observation:** The attentive regions obtained by Grad-CAM. By eliminating beach background, CLIP changes its answer to be correct. This indicates the CLIP has learned a rich set of features (CLIP knows the object should be a spider), but the spurious features lead to decision shortcuts.

make classification based on certain areas introduces the risk of incorporating new decision shortcuts. In this paper, we present a method which could effectively overcome the potential decision shortcuts by erasing spurious features instead of forcing the model to make classifications based on certain features, without the necessity for altering the model's architecture and weights.

Our motivation stems from the following observation: *VLFMs have already learned a rich set of features, which include both the desired invariant causal features and the unwanted decision shortcuts*, and the reason why VLFMs fail on certain tasks is the inference process predominantly activates these decision shortcuts when applied to specific tasks. Specifically, we observe that a simple intervention by removing background information in the visual context can shift CLIP's focus towards invariant features (the object), thereby significantly improving model performance (As shown in Fig. 1, CLIP's zero-shot classification performance on foreground improved by 48% in terms of accuracy for the worst group). In other words, this implies that if we can teach VLFMs how to overcome the impact of spurious features, their performance could be significantly enhanced.

Based on this insight, we propose a test-time prompt tuning paradigm termed Spurious Feature Eraser (SEraser), encouraging VLFMs to leverage causal invariant features by disregarding any potential spurious features during inference. Unlike prior methods that directly inform VLFMs which part of the image contains invariant features, our method is more practical and flexible since it only needs to identify potential spurious features. Specifically, as shown in Fig. 2, we sample auxiliary images first then regularize VLFMs to predict a uniform distribution on these auxiliary images through test-time prompt tuning, thereby enforcing VLFMs to learn how to "erase" spurious features. With the help of optimized prompt, the model can recognize spurious features, thus overcoming potential decision shortcuts inherently. Extensive experiments confirm that our approach significantly enhances the model stability against decision shortcuts compared to existing state-of-the-art methods. The core contributions of this paper are summarized as follows:

- For improving vision-language foundation models' performance, we propose a novel test-time prompt tuning method to overcome potential decision shortcuts by flexibly "erasing" the spurious features, which inherently enables vision-language foundation models to recognize and disregard spurious features.
- We develop a new evaluation paradigm for vision-language foundation models, and utilize it as a supplementary benchmark to assess VLFMs' reliability. This evaluation framework effectively reflects the ability of overcoming decision shortcuts for VLFMs. And it can avoid data leakage, which may lead to unfair evaluation.
- Our method significantly improves the zero-shot classification performance of VLFMs, especially when they exhibit clear decision shortcuts on that task. As the empirical result in the benchmark dataset Waterbirds (Koh et al. 2021)), our method can even bring over 25% improvement on the worst group.

Method

In this section, we first introduce the problem setting and corresponding notations. Building upon the observation that "decision shortcuts" hinder the generalization capabilities of VLFMs, we introduce the Spurious Feature Eraser (SEraser) to improve the reliability of current VLFMs. Furthermore, we show the reasons why previous evaluation paradigms are not suitable for VLFMs, and consequently propose a new evaluation paradigm.

Notations

In the context of a vision-language foundation model $\mathcal{M}$, such as CLIP, we explore its application in downstream tasks, particularly in zero-shot scenarios. For a given test image sample x with a potential label set $\mathcal{Y} = \{1, 2, \cdots, K\}$, zero-shot classification requires using $\mathcal{M}$ to categorize x into its appropriate class, denoted as $\hat{y} \in \mathcal{Y}$. Specifically, each category-specific text input is prefixed with a hand-crafted prompt, for example, "`a photo of a`", leading to class descriptions like "`a photo of a waterbird`". These descriptions are then transformed into their respective embeddings $z^k_{\text{text}},\ k = 1, \cdots, K$, using the text encoder. Simultaneously, the image x is processed through the image encoder, resulting in its image embedding z_{img}. Subsequently, the normalized prediction distribution $\hat{P}(y|x) = p(\hat{y} = k|x)^K_i$ is computed using the softmax function, which is based on the cosine similarity between the image and text: $p(\hat{y} = k|x) = \frac{\exp(\cos(z^k_{\text{text}}, z_{\text{img}})/\tau)}{\sum_j^K \exp(\cos(z^j_{\text{text}}, z_{\text{img}})/\tau)}$, where $\cos(\cdot, \cdot)$ and τ indicate the cosine similarity between two embeddings and temperature, respectively.

Observation: CLIP Contains Rich Features

Based on the testing protocol outlined above, we craft a series of tasks to assess the reliability of the CLIP when confronted with decision shortcuts. Through comprehensive evaluation, we arrive at two core conclusions: (1) The CLIP, during its

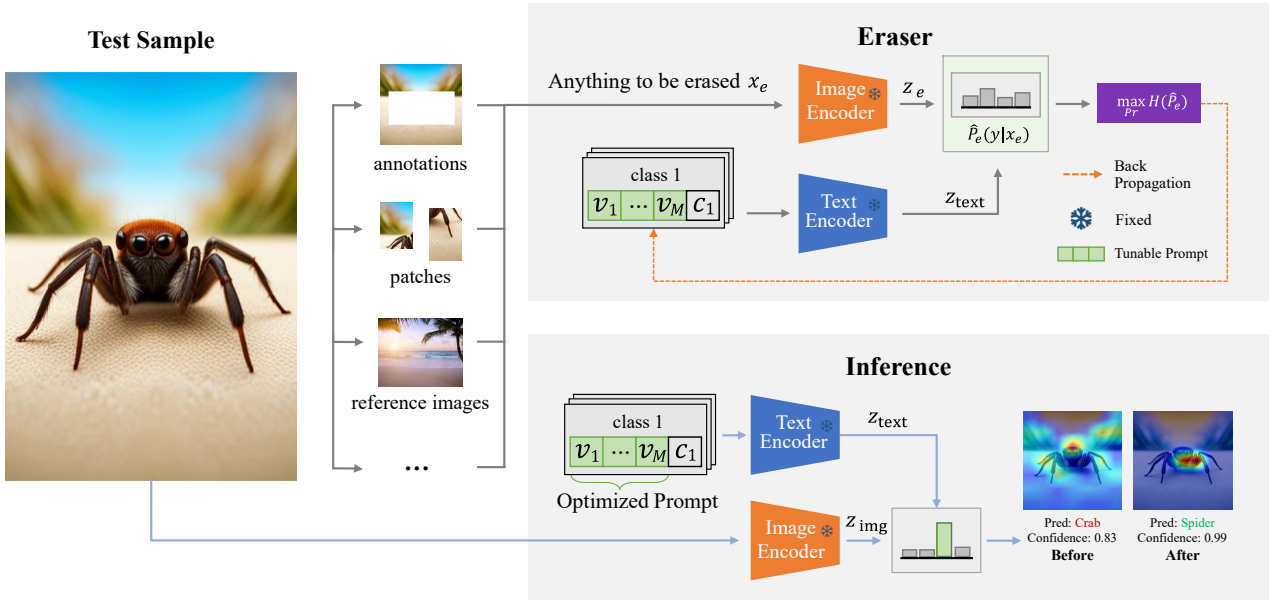

Figure 2: Framework of Spurious Feature Eraser (SEraser). Given a test sample, we first generate auxiliary images according to different strategies for test-time prompt tuning (termed as Eraser), then conduct zero-shot classification on the test sample with the optimized prompt from Eraser.

pre-training phase, has acquired a rich set of features. This set encompasses stable causal attributes that genuinely benefit downstream tasks, as well as spurious decision shortcuts that lead to erroneous associations. (2) In the unconstrained zero-shot inference stage, CLIP is highly susceptible to contextual influences, often resorting to the use of decision shortcuts for classification, so we can teach VLFMs how to overcome the impact of spurious features to improve their performance.

To illustrate this, we employ the SpiderCrab dataset as an example. In the classification task of differentiating crabs from spiders, our initial observations reveal markedly poor reliability of the CLIP in the most challenging test subgroup, which involves spiders against a beach background. The accuracy in this subgroup is strikingly low, hovering around only 42%. Utilizing visualization tools like Grad-CAM (Fig. 1), we further find that the model's focus was predominantly on background-related decision shortcuts, aligning with our second conclusion. However, a significant improvement is noted when we employ SAM model (Wang et al. 2023b) to extract the foreground context containing the core object as the model's input. This alteration leads to a substantial increase in accuracy, soaring to 90%, thereby supporting our first conclusion.

Spurious Feature Eraser

The experimental observations prompt us to contemplate how to compel foundational models to focus on and utilize causal features. In the following sections, we introduce a novel test-time adaptation paradigm that achieves this objective without modifying the original model's structure or weights. We begin by presenting the basic methodology of visual-language prompt tuning. Subsequently, we will introduce our proposed approach and the details of how to construct auxiliary images indicating spurious features.

Prompt tuning is an efficient and lightweight fine-tuning strategy specifically designed for foundation models (Liu et al. 2022). For VLFMs, prompt tuning operates by fine-tuning the prompt Pr of the sentences (Zhu et al. 2023). Specifically, the prompt is initialized by a hand-draft sentence, such as "`a photo of a [class`$_k$`]`", which will be tokenized as a vector $t_k = \{v_1, \cdots, v_M, c_k\}$, then we fix the class token c_k and fine-tune a set of M continuous context vectors $Pr := \{v_1, \cdots, v_M\}$:

$$Pr^* = \arg\min_{Pr} \mathcal{L}(\mathcal{M}, Pr, x), \tag{1}$$

where $\mathcal{L}$ is the optimization goal.

Prior work (Shu et al. 2022) has further leveraged the concept of minimizing entropy optimization to design an appropriate objective function, enhancing the model's reliability through test-time adaptation of prompts. Specifically, for a given test sample, the prompt is optimized by minimizing the entropy of averaged prediction distribution over augmented views and filters out the augmented views with low confidence to discard spurious features. However, these methods focus on instructing the models on which features to depend upon and heavily rely on the annotations of the invariant features. Consequently, they tend to fail when the annotations are not exactly precise.

The presence of spurious features in data often involves potential decision shortcuts, which can inadvertently influence models' performance. To mitigate this, our approach aims to minimize the impact of such spurious features. We achieve this by optimizing the test prompts through maximizing the entropy of the predictive distribution for these potential spurious features, effectively steering it towards a more uniform distribution. Specifically, suppose we want

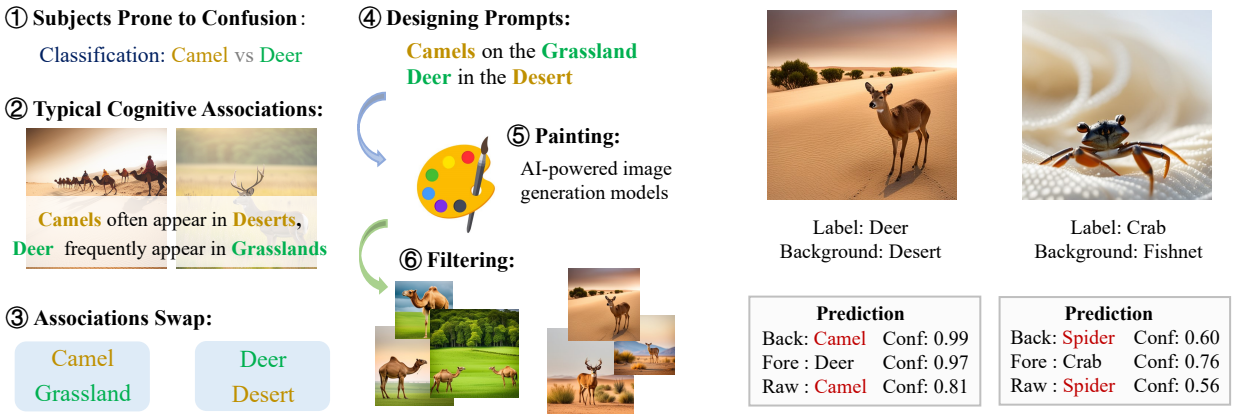

Figure 3: Pipeline of S2E and the predictions over classes of CLIP on different images. We can find that CLIP can correctly classify the objects on foreground, but its predictions on the images with background tend to be wrong when there are serious decision shortcuts on the background.

to erase the spurious features in an auxiliary image x_e, we minimize the Kullback-Leibler divergence between the prediction distribution on the auxiliary image $\hat{P}_e$ and the uniform distribution q:

$$\mathcal{L}_e = \mathrm{KL}(\hat{P}_e(y|x_e)\|q), \tag{2}$$

where $\mathrm{KL}(\cdot)$ indicates Kullback-Leibler divergence. The process is shown in Fig. 2, where the loss in Eq. 2 is used as the optimization objective to achieve the removal of decision shortcuts in VLFMs on the input image. Next, we will introduce in detail how to construct auxiliary images.

Construction of Auxiliary Images

The method proposed in this paper exhibits a high degree of flexibility, allowing for various choices in constructing auxiliary images. In this paper, we primarily introduce three simple yet effective strategies as examples (illustrated in Fig. 2). • **Annotations**: Assuming that we can obtain annotations of spurious features, such as expert annotations in medical imaging and background area annotations in target recognition, we can implement them as the features to be erased by SEraser, thereby achieving precise erasure of spurious features. In this paper, we simulate an expert's annotation by employing Segment Anything (Kirillov et al. 2023) to annotate the background. • **Patches**: In some tasks, obtaining the annotations of spurious features is difficult. Fortunately, we can still utilize alternative choices as auxiliary images. For instance, we can sample small patches from input images through multiple strategies as proxies, based on the assumption that the model should not make hasty judgments relying on partial information. • **Reference images**: Drawing inspiration from work utilizing reference data to enhance model robustness, we sample similar images from reference datasets as auxiliary images. For example, in a bird classification task, we can select images from an aircraft classification dataset as auxiliary images for SEraser.

Flexibility of Auxiliary Image Construction. If the auxiliary images are strictly prohibited from containing any invariant features, it would severely harm the flexibility of the proposed method. However, the approach remains effective even when these images do contain such invariant features, as it still improves the robustness of VLFMs. The method operates as a soft constraint, suggesting that decisions should not be solely based on certain features, as opposed to a rigid constraint that forces decisions to be based on specific features. *The flexibility enables the method to be more adaptable, and even when the auxiliary images contain partial invariant features, it is still effective since it prevents the model from making decisions based on partial features*. The similar strategy is also used in OOD detection (Bai et al. 2024b), which improves OOD detection performance using ID-like features. Therefore, the method can still ensure trustworthy decision even when auxiliary images contain invariant features. Moreover, we can design a more reasonable approximation of spurious features in auxiliary images using multiple strategies, thereby jointly improving the trustworthiness of decision. Ablation study demonstrates that even if partial invariant features in the image are erased, the proposed method can still enhance the robustness of VLFMs, resulting in a better performance.

In experiments, we find that current testing protocols and benchmarks, such as WILDS (Koh et al. 2021), are no longer suitable for VLFMs. These benchmarks are no longer as effective in evaluating VLFMs as they are for models such CNN models, and various issues arise. In the following section, we will discuss in detail why these benchmarks are no longer suitable and propose a new robustness evaluation paradigm for VLFMs to address these shortcomings.

S2E: Evaluation Protocol

In this section, we present a novel evaluation paradigm for assessing VLFMs. We find that current testing protocols and benchmarks, such as WILDS (Koh et al. 2021; Han et al. 2022), are no longer suitable for VLFMs, because they are built through large-scale self-supervised learning, and it will lead to *(1) data distribution shift in VLFMs paradigm no longer same as benchmarks* and *(2) testing data leakage*.

Method	Description	Augment	Tuning
Vanilla	basic comparative method, the original CLIP model is deployed for zero-shot classification directly.	None	✗
MASK	using SAM (Kirillov et al. 2023) to annotate the foreground and mask the background, then the masked image is classified in a zero-shot classification manner.	background annotation	✗
TPT	a ***prompt tuning*** method optimizes the prompt by minimizing the entropy of averaged prediction distribution over augmented views with confidence selection (Shu et al. 2022).	views augmentation	prompts
RoSHOT	a method (ROBOSHOT) which uses LLMs to obtain useful insights from task descriptions, then these insights are embedded and used to remove harmful and boost useful components in embeddings (Adila et al. 2023).	positive & negative insights	logits
α-CLIP	a ***region-aware*** version of CLIP with an auxiliary alpha channel to suggest attentive regions and fine-tuned with constructed millions of RGBA region-text pairs (Sun et al. 2023).	background annotation	architecture and weights
SEraser (Ours)	a test-time prompt tuning paradigm that optimizes a learnable prompt, helping the model to disregard decision shortcuts during the inference phase.	features to be erased (optional)	prompts

Table 1: Descriptions of different methods compared in experiments.

(1) In this context conventional robustness testing, most decision shortcuts are caused by carefully designed training datasets (i.e., distribution shifts in training data). A commonly employed benchmark is the WILDS dataset collection proposed by Stanford University (Koh et al. 2021). However, these data-centric robustness generalization testing protocols and benchmarks have become unsuitable for VLFMs developed through large-scale self-supervised learning. Specifically, decision shortcuts in traditional robustness testing benchmarks stem from carefully designed on limited-scale training data (e.g., decision shortcuts between image style and label designed in the PACS benchmark (Xu, Xiao, and López 2019)), whereas VLFMs are built using large-scale open-source data, rendering *these designed decision shortcuts often absent in VLFMs*.

(2) During the evaluation process of foundation models using benchmarks, a significant issue is *test data leakage*, and the same problem has been highlighted in LLM evaluation (Zhou et al. 2023a). These foundation models are trained on large-scale web data, and many samples in testset themselves or highly similar samples are present in the training data. For example, we find that alpha-CLIP performs exceptionally well on benchmarks such as Tiny-ImageNet. However, from the training details of alpha-CLIP, we find the released checkpoints are a retrained version of CLIP using ImageNet, implying its superior performance on the ImageNet relevant benchmarks. Similar issues of test data leakage in LLM evaluation are increasingly being acknowledged, and new simulated data are being generated to ensure the validity of model evaluation.

To bridge the existing gap in this field, this paper proposes a novel paradigm for evaluating the reliability of VLFMs, termed "shortcut-to-evaluate" (`S2E`). The overall pipeline is illustrated in the left portion of Fig. 3. Specifically, the process can be described as the following steps: (1) We identify a set of subjects that typically challenge the model's classification accuracy, such as differentiating between camels and deer. (2) We then examine their common cognitive associations - camels are typically associated with desert environments, while deer are often linked to grassland habitats. (3) Then, we interchange these environmental associations, positioning camels in grassland settings and deer in desert landscapes. This step is pivotal in creating decision-making shortcuts for the model to follow. (4) We refine these unusual combinations into specific instructions designed for advanced AI-powered image generation tools, such as DALL-E and Midjourney. (5) Using these customized instructions, we generate a series of test images, each crafted to evaluate the model's ability to handle these unconventional scenarios. (6) The final step involves a rigorous filtering process, discarding any images that do not meet our predefined criteria, such as those lacking the presence of the intended animals. In this paper, we generate two scenarios according to `S2E`.

Experiments

Setup

In the experiments, we evaluate different methods on different scenarios, including the real-word image data Tiny-ImageNet (Le and Yang 2015), CUB-200 (Wah et al. 2011) and the benchmark simulated data Waterbirds (Koh et al. 2021), and the datasets created by `S2E` CamelDeer and SpiderCrab. To intuitively demonstrate the decision shortcuts in VLFMs, we primarily focus on the samples of challenging classes in the dataset (for details please refer to Appendix). Except for Alpha-CLIP, which has its own separately designed visual encoder, all methods use the CLIP with a pre-

trained ViT-B-32 released by OpenAI (Radford et al. 2021) as the visual encoder, which is a representative architecture of vision encoder (Conde and Turgutlu 2021; Wu et al. 2023; Chen et al. 2023b).[1]

Main Results

Results on Real-world Scenarios We evaluate the performance of the proposed method on real-world datasets, and the results are presented in Table 2. "Patches" denotes the strategy of cropping patches of the test image to serve as auxiliary images, while "Images" refers to the approach of employing reference images as auxiliary images. "Both" indicates jointly employing both strategies in deploying SEraser. Specifically, **Patches** strategy divides the input image into 8x8 patches and utilizes the four corner patches as auxiliary images for SEraser implementation to improve the diversity of auxiliary images (adjacent patches often exhibit lower information complementarity (Wang et al. 2023a)). On the other hand, the "Images" involves selecting images from benchmarks that do not belong to any category in the classification task as OOD datasets and searching for the most similar images based on cosine similarity to serve as **Reference images**. "Both" represents the deployment of SEraser using auxiliary images obtained from the aforementioned strategies (for further details, please refer to Appendix). The results demonstrate that different strategies improve zero-shot classification performance, with the combined application of both strategies yielding higher performance than using either strategy independently. It is worth noting that Alpha-CLIP (α-CLIP) achieves outstanding performance on the open-source benchmark Tiny-ImageNet, which can be attributed to the checkpoint being retrained on ImageNet.

Results on Simulated Scenarios We conduct experiments using the Waterbirds benchmark as well as simulated tasks generated through the `S2E`. We adapt SAM (Kirillov et al. 2023) for image segmentation, which identifies the background as auxiliary images by creating bounding boxes (i.e., simulated an expert's **Annotations**). If an image contains multiple foreground boxes, we combine them as a single foreground. Furthermore, since only one test image is input in a zero-shot manner prompt during test-time adaptation, we apply regularization on the model to minimize the prediction distribution entropy on features that are not to be erased and employ RandAugment (Cubuk et al. 2020) on the images to prevent trivial solutions. Both Alpha-CLIP (region-aware) and our proposed method substantially enhance the performance of the CLIP, particularly in the worst group. Alpha-CLIP benefits from extensive data-driven region-aware finetuning and the annotation information of spurious features during the testing phase, resulting in a markedly improved performance compared to the original CLIP. On the other hand, our method achieves the best performance through test-time adaptation, without finetuning the model's weights. The performance of TPT is suboptimal due to its test-time adaptation objective, which minimizes the average entropy by randomly cropping the input image. TPT assumes that the views with higher confidence represent invariant features, while those with lower confidence are spurious features. However, this assumption is occasionally unsuitable. As illustrated in Fig. 3, higher confidence in the desert background than in the foreground leads the TPT models to make decisions based on the background. RoSHOT exhibits improvement in some tasks but a decline in others. We can observe that our method has achieved significant improvements across various datasets, particularly in the aspect of worst group performance.

[1]Codes: https://github.com/MaHuanAAA/SEraser

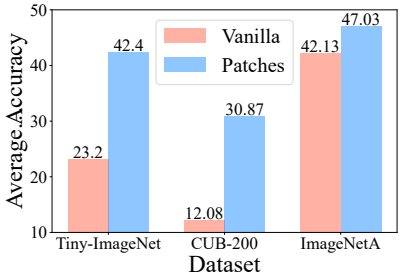

Figure 4: Performance under different patch sampling strategies.

Ablation Study

Effectiveness Across Different Model Architectures. The results presented before demonstrate that the proposed paradigm significantly enhances the zero-shot classification capability of CLIP ViT-B-32. To evaluate the method's flexibility across other models, we evaluate its performance on other widely-used architectures, including CLIP ViT-L-14 and BLIP-2. We conduct these evaluations using the Waterbirds benchmark, a widely employed benchmark dataset. As shown in Table 4, our method achieves promising performance on both models. Specifically, the CLIP with ViT-L14 image encoder is improved significantly after deploying SEraser, and BLIP-2 shows a significant improvement on the worst group, validating the effectiveness of SEraser across diverse architectures (for results on other datasets, please refer to Appendix).

Flexibility in Constructing Auxiliary Images. The flexibility of the proposed method in the presence of invariant information in auxiliary images is examined through the following ablation study. We modify the sampling strategy of **Patches** to investigate this aspect. Specifically, we replace the corner patches with a random patch sampling strategy. This approach implies that the patch sampling is entirely random, without any prior knowledge, and thus, it can sample patches that contain invariant features and are adjacent. As illustrated in Fig. 4, even when SEraser is implemented using these randomly sampled patches, the proposed method still enhances the performance of VLFMs. This improvement can be attributed to the soft constraint proposed in SEraser, which ensures that "VLFMs cannot make hasty decisions based on partial information", accordingly, assisting VLFMs in alleviating decision shortcuts on then patches even containing partial invariant features.

Dataset	Vanilla	TPT	RoSHOT	α-CLIP*	Patches	Images	Both
Tiny-ImageNet	23.2	29.6	**49.2**	76.0	42.4	41.2	$42.8_{(\blacktriangle 19.6)}$
CUB-200	12.1	8.7	25.5	44.3	26.2	24.2	$\mathbf{28.9}_{(\blacktriangle 16.8)}$
ImageNetA	42.1	**49.7**	38.9	51.5	47.4	45.4	$\mathbf{49.7}_{(\blacktriangle 7.6)}$
Average	25.8	29.3	37.9	57.3	38.7	36.9	$\mathbf{40.5}_{(\blacktriangle 14.7)}$

* *Clarification: α-CLIP was trained on multiple open-source datasets including ImageNet, which renders the above evaluation unsuitable for this model.*

Table 2: Zero-shot classification performance for different choices of how to construct auxiliary images on real-world scenarios, where "Average" indicates the averaged performance on all datasets and the ▲Green mark represents the improvement relative to Vanilla.

Dataset		Vanilla	MASK	TPT	RoSHOT	α-CLIP	Ours
Waterbirds	AVG.	67.67	71.97	66.88	68.86	67.59	$\mathbf{78.24}_{(\blacktriangle 10.57)}$
	W.G.	40.04	51.53	34.38	52.28	43.15	$\mathbf{65.25}_{(\blacktriangle 25.21)}$
CamelDeer*	AVG.	83.20	93.60	77.67	80.40	92.00	$\mathbf{95.67}_{(\blacktriangle 12.47)}$
	W.G.	66.40	87.20	55.33	60.80	84.40	$\mathbf{91.60}_{(\blacktriangle 25.2)}$
SpiderCrab*	AVG.	66.00	91.40	83.53	73.00	86.20	$\mathbf{95.33}_{(\blacktriangle 29.33)}$
	W.G.	42.00	90.40	72.53	50.40	86.00	$\mathbf{94.67}_{(\blacktriangle 52.67)}$
Average	AVG.	72.29	85.67	76.03	74.09	81.93	$\mathbf{89.75}_{(\blacktriangle 17.46)}$
	W.G.	49.48	76.38	54.08	54.49	71.18	$\mathbf{83.72}_{(\blacktriangle 34.24)}$

* We submit a subset of these datasets in supplementary materials limited by the maximum file size.

Table 3: Zero-shot classification performance on simulated scenarios, where "AVG." and "W.G." indicate the average performance on the entire test set and the worst-performing group (i.e., the group with the lowest accuracy), respectively (for error bars please refer to Appendix).

Models		Vanilla	MASK	Ours
CLIP-L14	AVG.	83.71	85.48	$\mathbf{87.78}_{(\blacktriangle 4.07)}$
	W.G.	32.87	40.81	$\mathbf{58.88}_{(\blacktriangle 26.01)}$
BLIP-2	AVG.	**57.65**	54.38	$55.56_{(\blacktriangledown 2.09)}$
	W.G.	28.19	35.05	$\mathbf{34.74}_{(\blacktriangle 6.55)}$
Average	AVG.	70.68	69.93	$\mathbf{71.67}_{(\blacktriangle 0.99)}$
	W.G.	30.53	37.93	$\mathbf{46.81}_{(\blacktriangle 16.28)}$

Table 4: Zero-shot classification performance of different VLFMs on Waterbirds.

Related Work

• **Region-aware CLIP**: One straightforward strategy is to mask the background (Liang et al. 2023), forcing the model to focus on the foreground. Some other approaches prompt the CLIP model by circling the foreground of the input image, guiding CLIP to focus on the area of interest. For example, Red-Circle (Shtedritski, Rupprecht, and Vedaldi 2023) and FGVP (Yang et al. 2023) utilize a circle or mask contour to indicate where CLIP should concentrate its attention. Alpha-CLIP (Sun et al. 2023) enhances CLIP by incorporating regions of interest through an additional alpha channel input. This supplementary channel is derived using SAM, which ranges from [0, 1], where 1 represents the foreground and 0 represents the background. • **Prompt Tuning**: Test-time Prompt Tuning (Shu et al. 2022) based on view-augmentation optimizes the prompt to encourage consistent predictions across augmented views by minimizing the marginal entropy and filtering out noisy augmentations with low confidence. However, it is important to note that this assumption may not always hold true. For instance, in an image containing a landbird and water background, spurious features (water background) can actually lead to high confidence rather than predicting a uniform distribution. ROBOSHOT (Adila et al. 2023) uses zero-shot to obtain useful insights from task descriptions. Instead of to find a optimized prompt, the authors utilize the prompts generated by LMs to push the model focus on invariant features. The insights generated from LMs are embedded and used to remove harmful and boost useful components in embeddings. PerceptionCLIP (An et al. 2023) finds that providing CLIP with contextual attributes improves zero-shot classification and reduces reliance on spurious features, and (Chuang et al. 2023) measure spurious correlations through biased prompts and then debias the model using orthogonal projection.

Conclusion

In this paper, we propose a method to address the challenges faced by vision-language foundation models in decision shortcuts. Starting from the observation that CLIP contains both desired invariant causal features and undesired decision shortcuts, we find that the latter is responsible for the underperformance in specific tasks. Then, a test-time prompt tuning paradigm is introduced to overcome the possible decision shortcuts, which optimizes a learnable prompt to encourage the model to focus on genuine causal invariant features while restraining decision shortcuts in inference. A comparative analysis is conducted to validate the effectiveness of the proposed method against existing approaches.

Acknowledgments

This work is supported by the National Natural Science Foundation of China Grant No.62376193, Grant No.61925602, and Grant No.U23B2049.

References

Adila, D.; Shin, C.; Cai, L.; and Sala, F. 2023. Zero-Shot Robustification of Zero-Shot Models With Foundation Models. *arXiv preprint arXiv:2309.04344*.

Agarwal, S.; Krueger, G.; Clark, J.; Radford, A.; Kim, J. W.; and Brundage, M. 2021. Evaluating clip: towards characterization of broader capabilities and downstream implications. *arXiv preprint arXiv:2108.02818*.

An, B.; Zhu, S.; Panaitescu-Liess, M.-A.; Mummadi, C. K.; and Huang, F. 2023. More Context, Less Distraction: Visual Classification by Inferring and Conditioning on Contextual Attributes. *arXiv preprint arXiv:2308.01313*.

Bai, Y.; Han, Z.; Cao, B.; Jiang, X.; Hu, Q.; and Zhang, C. 2024a. ID-like Prompt Learning for Few-Shot Out-of-Distribution Detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 17480–17489.

Bai, Y.; Han, Z.; Cao, B.; Jiang, X.; Hu, Q.; and Zhang, C. 2024b. ID-like Prompt Learning for Few-Shot Out-of-Distribution Detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

Cao, X.; Zhang, C.; Fu, H.; Liu, S.; and Zhang, H. 2015. Diversity-induced multi-view subspace clustering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 586–594.

Chen, K.; Jiang, X.; Hu, Y.; Tang, X.; Gao, Y.; Chen, J.; and Xie, W. 2023a. Ovarnet: Towards open-vocabulary object attribute recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 23518–23527.

Chen, Y.; Qi, X.; Wang, J.; and Zhang, L. 2023b. DisCoCLIP: A Distributed Contrastive Loss for Memory Efficient CLIP Training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 22648–22657.

Chuang, C.-Y.; Jampani, V.; Li, Y.; Torralba, A.; and Jegelka, S. 2023. Debiasing vision-language models via biased prompts. *arXiv preprint arXiv:2302.00070*.

Conde, M. V.; and Turgutlu, K. 2021. CLIP-Art: Contrastive pre-training for fine-grained art classification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 3956–3960.

Cubuk, E. D.; Zoph, B.; Shlens, J.; and Le, Q. V. 2020. Randaugment: Practical automated data augmentation with a reduced search space. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 3008–3017.

Fan, L.; Krishnan, D.; Isola, P.; Katabi, D.; and Tian, Y. 2023. Improving CLIP Training with Language Rewrites. *arXiv preprint arXiv:2305.20088*.

Gu, X.; Lin, T.-Y.; Kuo, W.; and Cui, Y. 2021. Open-vocabulary object detection via vision and language knowledge distillation. *arXiv preprint arXiv:2104.13921*.

Han, Z.; Zhang, C.; Fu, H.; and Zhou, J. T. 2022. Trusted multi-view classification with dynamic evidential fusion. *IEEE transactions on pattern analysis and machine intelligence*, 45(2): 2551–2566.

Kamath, A.; Singh, M.; LeCun, Y.; Synnaeve, G.; Misra, I.; and Carion, N. 2021. Mdetr-modulated detection for end-to-end multi-modal understanding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 1780–1790.

Kirillov, A.; Mintun, E.; Ravi, N.; Mao, H.; Rolland, C.; Gustafson, L.; Xiao, T.; Whitehead, S.; Berg, A. C.; Lo, W.-Y.; et al. 2023. Segment anything. *arXiv preprint arXiv:2304.02643*.

Koh, P. W.; Sagawa, S.; Marklund, H.; Xie, S. M.; Zhang, M.; Balsubramani, A.; Hu, W.; Yasunaga, M.; Phillips, R. L.; Gao, I.; et al. 2021. Wilds: A benchmark of in-the-wild distribution shifts. In *International Conference on Machine Learning*, 5637–5664. PMLR.

Le, Y.; and Yang, X. 2015. Tiny imagenet visual recognition challenge. *CS 231N*, 7(7): 3.

Li, L. H.; Zhang, P.; Zhang, H.; Yang, J.; Li, C.; Zhong, Y.; Wang, L.; Yuan, L.; Zhang, L.; Hwang, J.-N.; et al. 2022. Grounded language-image pre-training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 10965–10975.

Li, Z.; Evtimov, I.; Gordo, A.; Hazirbas, C.; Hassner, T.; Ferrer, C. C.; Xu, C.; and Ibrahim, M. 2023. A whac-a-mole dilemma: Shortcuts come in multiples where mitigating one amplifies others. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 20071–20082.

Liang, F.; Wu, B.; Dai, X.; Li, K.; Zhao, Y.; Zhang, H.; Zhang, P.; Vajda, P.; and Marculescu, D. 2023. Open-vocabulary semantic segmentation with mask-adapted clip. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 7061–7070.

Liu, X.; Ji, K.; Fu, Y.; Tam, W.; Du, Z.; Yang, Z.; and Tang, J. 2022. P-tuning: Prompt tuning can be comparable to fine-tuning across scales and tasks. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, 61–68.

Ma, H.; Zhang, C.; Bian, Y.; Liu, L.; Zhang, Z.; Zhao, P.; Zhang, S.; Fu, H.; Hu, Q.; and Wu, B. 2023a. Fairness-guided few-shot prompting for large language models. *Advances in Neural Information Processing Systems*, 36: 43136–43155.

Ma, H.; Zhang, Q.; Zhang, C.; Wu, B.; Fu, H.; Zhou, J. T.; and Hu, Q. 2023b. Calibrating multimodal learning. In *International Conference on Machine Learning*, 23429–23450. PMLR.

Patashnik, O.; Wu, Z.; Shechtman, E.; Cohen-Or, D.; and Lischinski, D. 2021. Styleclip: Text-driven manipulation of stylegan imagery. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2085–2094.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, 8748–8763. PMLR.

Shi, J.; Zheng, S.; Yin, X.; Lu, Y.; Xie, Y.; and Qu, Y. 2023. CLIP-guided Federated Learning on Heterogeneous and Long-Tailed Data. *arXiv preprint arXiv:2312.08648*.

Shtedritski, A.; Rupprecht, C.; and Vedaldi, A. 2023. What does clip know about a red circle? visual prompt engineering for vlms. *arXiv preprint arXiv:2304.06712*.

Shu, M.; Nie, W.; Huang, D.-A.; Yu, Z.; Goldstein, T.; Anandkumar, A.; and Xiao, C. 2022. Test-time prompt tuning for zero-shot generalization in vision-language models. *Advances in Neural Information Processing Systems*, 35: 14274–14289.

Sun, Z.; Fang, Y.; Wu, T.; Zhang, P.; Zang, Y.; Kong, S.; Xiong, Y.; Lin, D.; and Wang, J. 2023. Alpha-CLIP: A CLIP Model Focusing on Wherever You Want. *arXiv preprint arXiv:2312.03818*.

Wah, C.; Branson, S.; Welinder, P.; Perona, P.; and Belongie, S. J. 2011. The Caltech-UCSD Birds-200-2011 Dataset.

Wang, H.; Guo, S.; Ye, J.; Deng, Z.; Cheng, J.; Li, T.; Chen, J.; Su, Y.; Huang, Z.; Shen, Y.; et al. 2023a. Sam-med3d. *arXiv preprint arXiv:2310.15161*.

Wang, H.; Vasu, P. K. A.; Faghri, F.; Vemulapalli, R.; Farajtabar, M.; Mehta, S.; Rastegari, M.; Tuzel, O.; and Pouransari, H. 2023b. Sam-clip: Merging vision foundation models towards semantic and spatial understanding. *arXiv preprint arXiv:2310.15308*.

Wei, Y.; Zhang, Y.; Ji, Z.; Bai, J.; Zhang, L.; and Zuo, W. 2023. Elite: Encoding visual concepts into textual embeddings for customized text-to-image generation. *arXiv preprint arXiv:2302.13848*.

Wu, K.; Peng, H.; Zhou, Z.; Xiao, B.; Liu, M.; Yuan, L.; Xuan, H.; Valenzuela, M.; Chen, X. S.; Wang, X.; et al. 2023. Tinyclip: Clip distillation via affinity mimicking and weight inheritance. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 21970–21980.

Xu, J.; Liu, S.; Vahdat, A.; Byeon, W.; Wang, X.; and De Mello, S. 2023. Open-vocabulary panoptic segmentation with text-to-image diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2955–2966.

Xu, J.; Xiao, L.; and López, A. M. 2019. Self-supervised domain adaptation for computer vision tasks. *IEEE Access*, 7: 156694–156706.

Yang, G.; Qiao, Y.; Shit, J.; and Wang, Z. 2022. Long-Tailed Object Mining Based on CLIP Model for Autonomous Driving. In *2022 4th International Conference on Control and Robotics (ICCR)*, 348–352. IEEE.

Yang, L.; Wang, Y.; Li, X.; Wang, X.; and Yang, J. 2023. Fine-Grained Visual Prompting. *arXiv preprint arXiv:2306.04356*.

Zhou, K.; Zhu, Y.; Chen, Z.; Chen, W.; Zhao, W. X.; Chen, X.; Lin, Y.; Wen, J.-R.; and Han, J. 2023a. Don't Make Your LLM an Evaluation Benchmark Cheater. *arXiv preprint arXiv:2311.01964*.

Zhou, Z.; Hu, S.; Li, M.; Zhang, H.; Zhang, Y.; and Jin, H. 2023b. Advclip: Downstream-agnostic adversarial examples in multimodal contrastive learning. In *Proceedings of the 31st ACM International Conference on Multimedia*, 6311–6320.

Zhu, B.; Niu, Y.; Han, Y.; Wu, Y.; and Zhang, H. 2023. Prompt-aligned gradient for prompt tuning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 15659–15669.