Beyond Sole Strength: Customized Ensembles for Generalized Vision-Language Models

Zhihe Lu [1] **Jiawang Bai** [1 2] **Xin Li** [1 3] **Zeyu Xiao** [1 3] **Xinchao Wang** [1]

Abstract

Fine-tuning pre-trained vision-language models (VLMs), *e.g.*, CLIP, for the open-world generalization has gained increasing popularity due to its practical value. However, performance advancements are limited when relying solely on intricate algorithmic designs for a single model, even one exhibiting strong performance, *e.g.*, CLIP-ViT-B/16. This paper, for the first time, explores the collaborative potential of leveraging much weaker VLMs to enhance the generalization of a robust single model. The affirmative findings motivate us to address the generalization problem from a novel perspective, *i.e.*, ensemble of pre-trained VLMs. We introduce three customized ensemble strategies, each tailored to one specific scenario. Firstly, we introduce the zero-shot ensemble, automatically adjusting the logits of different models based on their confidence when only pre-trained VLMs are available. Furthermore, for scenarios with extra few-shot samples, we propose the training-free and tuning ensemble, offering flexibility based on the availability of computing resources. The code is available at https://github.com/zhiheLu/Ensemble_VLM.git.

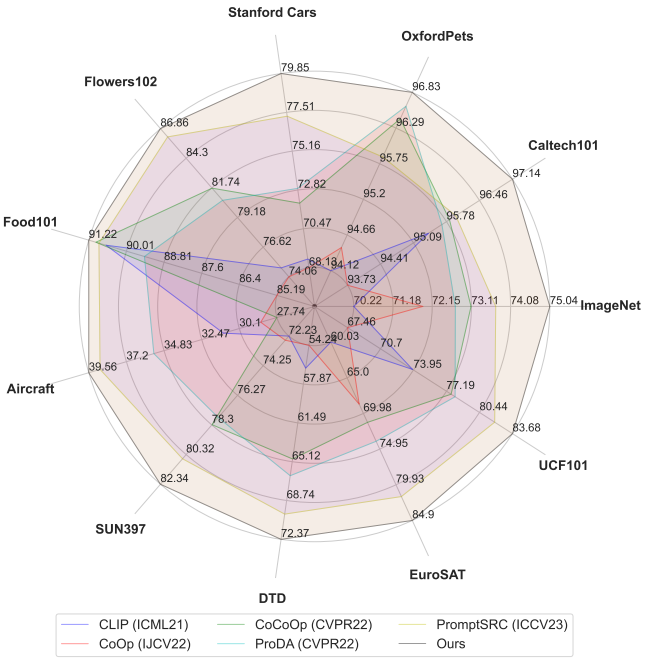


Figure 1. **Comparing existing methods on base-to-new generalization**. The results indicate that the proposed method outperforms existing arts on 11 diverse datasets, often by large margins.

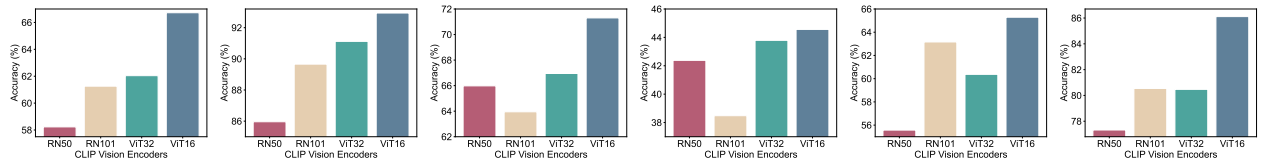
1. Introduction

Pre-trained large-scale vision-language models (VLMs), *e.g.*, CLIP (Radford et al., 2021) and ALIGN (Jia et al., 2021), have demonstrated remarkable recognition performance on open-vocabulary downstream tasks even in a zero-shot evaluation manner. This advancement overcomes the constraint observed in prior supervised deep models (He et al., 2016), which are limited to classify seen classes only, *i.e.*, test set should share the same label space with the training set. Encouraged by the impressive zero-shot generalization of VLMs, recent works (Yu et al., 2023; Li et al., 2023b; Zhou et al., 2022b;a) have explored how to adapt the VLMs for better performance on downstream tasks in an efficient tuning fashion, where only few-shot samples and limited learnable parameters are used.

For the efficient adaptation of VLMs, prompt learning (Zhou et al., 2022b) is a popular solution. The concept of prompt learning was first proposed in natural language processing (NLP) (Shin et al., 2020; Jiang et al., 2020; Zhong et al., 2021), which is to learn a prompt that dynamically suits the downstream tasks instead of prompt engineering. CoOp (Zhou et al., 2022b) was the first work to introduce prompt learning for adapting VLMs and it outperforms both hand-crafted prompts and the linear probe model. However, a critical problem of CoOp is that the learned prompt on

[1]National University of Singapore [2]Tsinghua University [3]University of Science and Technology of China. Correspondence to: Xinchao Wang <xinchao@nus.edu.sg>.

Proceedings of the 41st International Conference on Machine Learning, Vienna, Austria. PMLR 235, 2024. Copyright 2024 by the author(s).



(a) ImageNet (b) Caltech101 (c) Flowers102 (d) DTD (e) Stanford Cars (f) Food101

Figure 2. **Zero-shot evaluation of pre-trained CLIP vision encoders on varying datasets**. The bar charts show that the "weak" models may perform better than strong ones, *e.g.*, RN50 *vs.* RN101 in (c, d) and RN101 *vs.* ViT-B/32 in (e, f), encouraging us to leverage diverse models for enhanced ensemble.

base classes performs poorly when applied to unseen new classes even under the same data distribution. CoCoOp (Zhou et al., 2022a) hence has been proposed to improve the performance of this base-to-new set-up by conditioning the prompt learning with the image context. To further advance the base-to-new generalization, recent works have been proposed to leverage multi-modal prompt learning (MaPLe) (Khattak et al., 2023a), self-regulating constraints (PromptSRC) (Khattak et al., 2023b), or synthesized prompts (SHIP) (Wang et al., 2023). It is worth noting that the default CLIP model [1] used in these methods is ViT-B/16 (Dosovitskiy et al., 2021), one of the strongest CLIP models in Transformer structure (Vaswani et al., 2017). However, their performance on new classes is still not satisfactory, *e.g.*, 70.43% (CoCoOp, CVPR22) $\rightarrow$ 70.54% (MaPLe, CVPR23) $\rightarrow$ 70.26% (SHIP, ICCV23) $\rightarrow$ 70.73% (PromptSRC, ICCV23) on ImageNet. One possible reason is that with the advanced algorithms, the knowledge of a single model has already been fully explored for generalization, thereby leading to limited improvement. As such, one question naturally arises: *can we leverage the prior knowledge from multiple CLIP models for better generalization even though some of them are weaker performers?*

To this end, we first investigate the zero-shot performance of four widely used CLIP models, *i.e.*, RN50, RN101, ViT-B/32 and ViT-B/16, on 11 diverse datasets as shown in Table 1. Intuitively, a larger model should yield better performance, but we have observed that this is not always the case across all datasets. For instance, the "weak" models may perform better than strong ones, *e.g.*, RN50 *vs.* RN101 in Figure 2 (c, d) and RN101 *vs.* ViT-B/32 in Figure 2 (e, f). This indicates that models with distinct architectures are inclined to predict different samples correctly, as those models may develop biases even when trained on the same dataset. We posit that these biases can be leveraged for improved generalization in open-world scenarios. Indeed, zero-shot experiments (see Table 2) demonstrate enhanced generalization by combining any weaker CLIP model with CLIP-ViT-B/16 while using all models performs the best.

[1] CLIP model specifically refers to vision encoder in this paper.

Inspired by the favorable results of a simple ensemble, we further design three types of ensemble strategies, each tailored for a practical scenario. **First**, given solely CLIP models, we introduce the zero-shot ensemble (ZS_{En}), which employs confidence-aware weighting for the weaker models while preserving the dominance of the strongest model. This confidence-aware weighting dynamically adjusts the weights of the models' logits based on their prediction confidences of input samples. **Second**, in scenarios where additional few-shot samples (D_{train}) are available but without training resources, we propose the training-free ensemble (TF_{En}). This ensemble method leverages a greedy search to determine the "optimal" weights for model ensemble, identified when the accuracy of the ensemble is maximized on D_{train}. **Third**, when pre-trained models, D_{train} and training resources exist simultaneously, the tuning ensemble (T_{En}) is proposed by learning a sample-aware weight generator (SWIG). Specifically, the proposed SWIG takes the visual features of multiple models as input and automatically generates the ensemble weights through optimizing on the training data.

To assess the proposed methods, we employ ZS_{En} for zero-shot generalization, TF_{En} and T_{En} for base-to-new generalization on 11 diverse datasets. Additionally, we examine the effectiveness of T_{En} on cross-dataset evaluation. On all tasks, our methods always yield new state-of-the-art performance, often surpassing the second best by large margins (*e.g.*, see the results of base-to-new generalization in Figure 1). More importantly, this work offers a new possibility for the enhanced generalization of VLMs by overcoming the limitation of existing works that rely solely on designing intricate algorithms for a single VLM.

We summarize our contributions: **(i)** Having observed the saturated generalization performance achieved by fully exploring the knowledge from a single robust VLM, we, for the first time, propose to advance the generalization through ensemble learning. **(ii)** We propose three ensemble strategies, each tailored for a practical situation. *First*, in cases where only CLIP models are available, we introduce the confidence-aware weighting based zero-shot ensemble (ZS_{En}). *Second*, when additional few-shot sam-

ples (D_{train}) are accessible but without extra training resources, we present the training-free ensemble based on greedy search (TF$_{En}$). *Third*, in scenarios where pre-trained models, D_{train}, and training resources coexist, we propose the tuning ensemble (T$_{En}$) by incorporating a sample-aware weight generator (SWIG). **(iii)** We assess ZS$_{En}$ for zero-shot generalization, achieving an average accuracy gain of **2.61%** across 11 diverse datasets. Additionally, we evaluate T$_{En}$ for base-to-new and cross-dataset generalization, yielding new state-of-the-art performance. **(iv)** This work establishes a pathway for improved generalization in VLMs, surpassing the limitations of prior methods that depend solely on crafting intricate algorithms for a single VLM.

2. Related Work

2.1. Vision-Language Models

We mainly focus on language driven visual representation learning (LDVRL) (Socher et al., 2013; Frome et al., 2013; Elhoseiny et al., 2013; Lei Ba et al., 2015; Joulin et al., 2016; Gomez et al., 2017; Li et al., 2017; Anderson et al., 2018; Sariyildiz et al., 2020; Radford et al., 2021; Desai & Johnson, 2021). LDVRL based VLMs often adopt two encoders for language and vision, respectively, and are optimized by tailored matching constraints (Socher et al., 2013; Gomez et al., 2017). Early works employ varied approaches for language and vision modeling. For language, they utilize methods like unsupervised pre-trained models (Socher et al., 2013) or skip-gram text modeling (Frome et al., 2013; Mikolov et al., 2013a;b). Meanwhile, for vision, these works explore techniques like sparse coding and vector quantization (Coates & Ng, 2011; Socher et al., 2013) or utilize features such as Classeme (Torresani et al., 2010; Elhoseiny et al., 2013). In recent works (Radford et al., 2021; Jia et al., 2021; Li et al., 2021), a common trend is the utilization of two deep neural networks, *e.g.*, Transformers (Vaswani et al., 2017; Dosovitskiy et al., 2020; Touvron et al., 2021), to independently embed language and vision inputs. These works generally pre-train on million/billion-level image-text pairs from internet, *e.g.*, ∼400M for CLIP (Radford et al., 2021) and ∼1B for ALIGN (Jia et al., 2021), employing a contrastive loss. The resulting pre-trained VLMs showcase impressive zero-shot generalization across a range of downstream tasks. In this paper, we endeavor to improve their zero-shot generalization by effectively leveraging the knowledge embedded in multiple models.

2.2. Zero-Shot Generalization of VLMs

Zero-shot generalization/learning (ZSG) (Chao et al., 2016; Xian et al., 2017; Wang et al., 2019; Yi et al., 2022) aims to recognize novel classes by training only on base classes. Past ZSG works often learn the compositions of attributes (Huynh & Elhamifar, 2020) and word embeddings (Frome et al., 2013; Jiang et al., 2020) on base classes to tackle "seen-class bias" issue. The advent of large-scale VLMs introduces novel possibilities for addressing ZSG. These VLMs, trained to align image-and-text pairs, can be employed for ZSG on novel classes simply by adopting prompts of those novel classes for the language encoder.

To further advance the performance of ZSG on downstream tasks, prompt learning (Zhou et al., 2022b;a) has been proposed to learn a prompt on base classes that exhibits better generalization on novel classes within the same data distribution. MaPLe (Khattak et al., 2023a) explores multi-model prompt learning by jointly learning hierarchical prompts at two branches of CLIP, aiming for improved generalization performance. Recent works have introduced approaches leveraging self-regulating constraints (Khattak et al., 2023b) and synthesized prompts (Wang et al., 2023). Despite the intricate algorithms designed, the performance improvement on ZSG remains limited. One potential reason is that the knowledge stored in a single model has been thoroughly explored, making further improvement challenging. To that end, we take the initial step in exploring knowledge from multiple VLMs to enhance generalization.

2.3. Ensemble Learning

Ensemble learning (EL) (Wolpert, 1992; Breiman, 1996; Freund et al., 1996; LeBlanc & Tibshirani, 1996; Buja & Stuetzle, 2000; Breiman, 2001; Friedman, 2001; Ha et al., 2005; Deng et al., 2012; Cortes et al., 2014; Kang et al., 2020) is a machine learning paradigm where multiple models, often of diverse types, are combined to improve overall predictive performance and generalization. Technically, EL methods can be broadly categorized into three main types: Bagging (Bootstrap Aggregating) (Breiman, 1996; Buja & Stuetzle, 2000; Breiman, 2001; Ha et al., 2005), Boosting (Freund et al., 1996; Friedman, 2001; Cortes et al., 2014) and Stacking (Wolpert, 1992; LeBlanc & Tibshirani, 1996; Deng et al., 2012; Kang et al., 2020). Bagging was initially proposed in 1996 by Breiman (Breiman, 1996). The technique involves training multiple models of the same structure on diverse subsets obtained through bootstrap sampling. Boosting (Freund et al., 1996; Friedman, 2001; Cortes et al., 2014) is employed in ensemble models to transform several weak learners into one with improved generalization. In contrast, Stacking (Wolpert, 1992; LeBlanc & Tibshirani, 1996; Deng et al., 2012; Kang et al., 2020) is an integration technique where a meta-learning model is used to combine the outputs of base models. Recently, Cola (Chen et al., 2024) leverages large language models to coordinate multiple VLMs for visual reasoning. In this paper, without relying on other strong models, we resort to ensemble learning for better generalization of VLMs, with a specific emphasis on prediction fusion strategies in both training-free and tuning fashions.

3. Methodology

In this section, we first briefly introduce the VLM employed in our method, namely CLIP (Radford et al., 2021). Subsequently, we present three specific ensemble strategies designed to leverage VLMs. The VLMs employed in this work are four widely used CLIP models: CLIP-RN50, CLIP-RN101, CLIP-ViT-B/32, and CLIP-ViT-B/16. Note that the backbones here mainly refer to the vision encoder.

3.1. Preliminaries

The CLIP model is pre-trained on ∼400M image-text pairs with the objective of aligning two modalities within a unified embedding space using a contrastive learning loss. This pre-training enables CLIP to effectively capture broad visual concepts and learn general visual representations. During inference, a given image can be classified into a pre-defined category by computing the similarity between the image feature $\mathbf{f}$ extracted from the vision encoder and the textual embeddings $\{\mathbf{c}_i | i \in \{1, \ldots, K\}\}$ from the language encoder. The inputs to the language encoder are textual prompts for K categories, composed of a template, *e.g.*, "a photo of a $\{class\}$", and category names. This classification approach allows CLIP to be directly applied to a new recognition task with only the category names of interest. We formulate the inference process as follows:

$$\mathbf{Logit}_i = \frac{exp(cos(\mathbf{f}, \mathbf{c}_i)/\tau)}{\sum_{j=1}^{K} exp(cos(\mathbf{f}, \mathbf{c}_i)/\tau)} \quad (1)$$

, where $cos(\cdot, \cdot)$ is the cosine similarity function and τ is the learned temperature parameter.

3.2. Zero-shot Ensemble

The proposed zero-shot ensemble (ZS$_{En}$) aims to enhance the generalization capability of VLMs using solely pre-trained CLIP models as shown in Figure 3. Specifically, our ensemble strategy is designed with two key perspectives: (i) maintaining the dominance of the best-performing model, *i.e.*, CLIP-ViT-B/16; (ii) dynamically assigning weights to individual models based on their confidence for the given samples. The first consideration is grounded in the fact that CLIP-ViT-B/16, with inputting large patches and advanced Transformer architecture, can capture superior visual representations, leading to heightened transferability on downstream tasks. This is evidenced by the substantial zero-shot performance gaps observed between CLIP-ViT-B/16 and other variants. Consequently, preserving the dominance of CLIP-ViT-B/16 is deemed essential. To automate the weight generation for model ensemble, we propose to use confidence-aware weights by dynamically adjusting the contribution of individual models' logits based on their prediction confidences on input samples. The formulation of above process is defined as

$$\mathbf{p}_i = max(\mathbf{Logit}_i : \{p_1, p_2, \ldots, p_K\}) \quad (2)$$

$$\omega_i = \frac{exp(\mathbf{p}_i)}{\sum_{j=1}^{n-1} exp(\mathbf{p}_j)} \quad (3)$$

$$\mathbf{Logit}_e = \sum_{i=1}^{n-1} \omega_i \cdot \mathbf{Logit}_i + \mathbf{Logit}_n \quad (4)$$

, where $\mathbf{p}_i$ is the maximum prob of $\mathbf{Logit}_i$, $\mathbf{Logit}_i$ is the logit from the i-th VLM for an input image, $\mathbf{Logit}_n$ is the strongest VLM of n ensemble VLMs. It is worth noting that our ZS$_{En}$ enhances the overall generalization capability of VLMs without the need for additional training data or computing resources.

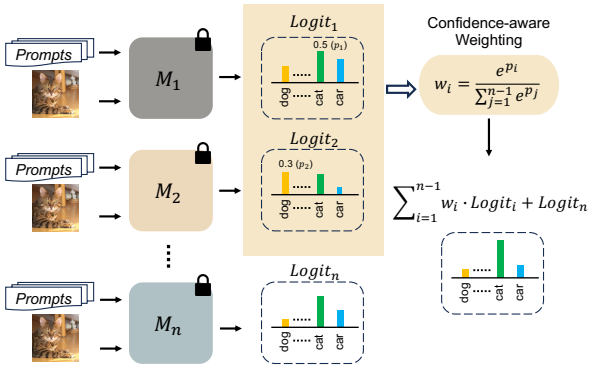


Figure 3. **Illustration of our zero-shot ensemble (ZS$_{En}$).** We assign a weight 1.0 to the best performing model, *i.e.*, CLIP-ViT-B/16, and use the confidence-aware weights for other VLMs.

3.3. Training-free Ensemble

We propose a training-free ensemble (see Figure 4) to further enhance the generalization of VLMs in scenarios where additional few-shot samples D_{train} are available but there are constraints on computing resources for training. Note that the classes in D_{train} are disjoint from those in D_{test} during inference, rendering this set-up both more challenging and practical. Specifically, we implement the training-free ensemble by employing greedy search to find the optimal accuracy on a training set. The desired weights $\{\omega_i\}$ are then determined when the optimal accuracy is achieved. The rationale behind this approach is that samples within the same data distribution can be utilized to search for optimal parameters for classifying new classes, aligning with methodologies for base-to-new generalization (Zhou et al., 2022a; Khattak et al., 2023a;b). To be specific, we formulate

the above process as follows:

$$\{\omega_i\} = \underset{\{\omega_i\}}{argmax}\left(\frac{1}{N}\sum_{j=1}^{N}\delta(y_i, \hat{y}_i)|D_{train}\right) \tag{5}$$

$$\hat{y}_i = argmax\left(\sum_{j=1}^{n-1}\omega_i \cdot \mathbf{Logit}_i + \mathbf{Logit}_n\right) \tag{6}$$

, where N is the total number of samples in D_{train}, y_i is the true label of the i-th sample, $\hat{y}_i$ is the predicted label of the i-th sample, $\delta(\cdot,\cdot)$ is the Kronecker delta function, which is equal to 1 if $y_i = \hat{y}_i$ and 0 otherwise.

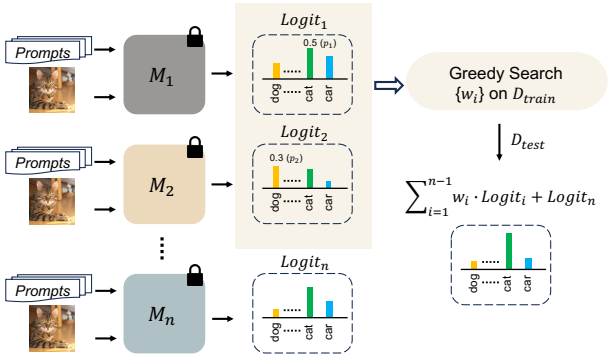


Figure 4. **Illustration of our training-free ensemble ($\mathbf{TF}_{En}$).** We assign a weight 1.0 to the best performing model, *i.e.*, CLIP-ViT-B/16, and determine the weights of other VLMs by greedy searching on a given "training" set without training.

3.4. Tuning Ensemble

Following the set-up of previous works (Zhou et al., 2022a; Khattak et al., 2023a;b), we introduce the tuning ensemble as illustrated in Figure 5. This ensemble method employs a meta-model, sample-aware weight generator (SWIG), to learn the generation of weights on the training set D_{train}. The advantage of the tuning ensemble, in comparison to the training-free version, lies in our SWIG's capability to dynamically generate weights for each sample based on the features of ensemble models. This facilitates test-sample-aware weight generation, with the corresponding features serving as inputs.

Concretely, to implement this tuning ensemble, we initially obtain the features $\{I_f^1, \ldots, I_f^n\}$ of the input image by passing it through the vision encoders of the employed VLMs. Subsequently, these features are concatenated and used as the input to our proposed SWIG, which generates weights $\{\omega_1, \ldots, \omega_n\}$ for the ensemble of VLMs. These dynamically generated weights serve as a form of sample-specific attention, allowing the ensemble to adapt to the characteristics of each input image. This personalized weight assignment enhances the tuning ensemble's ability to tailor its predictions based on the unique features of the test samples, thereby contributing to improved generalization performance. The aforementioned process can be formulated with the following equations:

$$I_f^1, I_f^2, \ldots, I_f^n = E_I^1(x), E_I^2(x), \ldots, E_I^n(x) \tag{7}$$

$$\{\omega_1, \ldots, \omega_n\} = \Theta_{SWIG}(concat(\{I_f^1, \ldots, I_f^n\})) \tag{8}$$

$$\mathbf{Logit}_e = \sum_{i=1}^{n}\omega_i \cdot \mathbf{Logit}_i \tag{9}$$

, where E_I^i is the vision encoder of the i-th VLM, x is the input image and Θ_{SWIG} is our sample-aware weight generator.

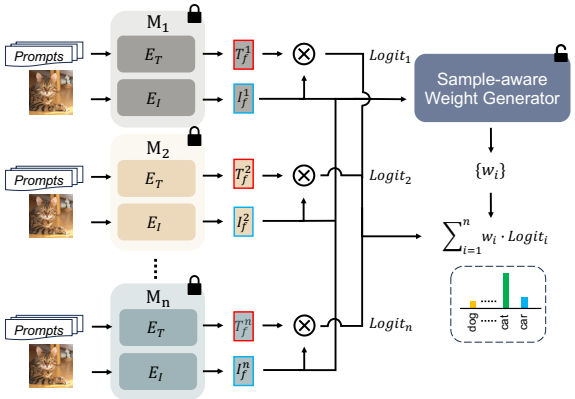


Figure 5. **Illustration of our tuning ensemble ($\mathbf{T}_{En}$).** The proposed sample-aware weight generator (SWIG) takes sample features as input to generate sample-aware weights, which are then used for weighted prediction.

Remark The choice of using image features, rather than logits, as the input to our SWIG is motivated by the unique prediction mechanism of VLMs. VLMs calculate the cosine similarity between the image feature and textual embeddings to generate logits. If logits were used as the input for SWIG, the generated weights might be biased towards the textual embeddings of base classes. This bias is undesirable for generalization to novel classes. By utilizing image features as the input for SWIG, we can effectively avoid this drawback. This approach ensures that SWIG remains sample-aware and generalizes well.

4. Experiments

In this section, we introduce the datasets and evaluation on zero-shot, base-to-new and cross-dataset generalization. All experiments are conducted on 11 diverse datasets, and the quantitative evaluation metric is classification accuracy. We also provide additional experimental results in the Appendix – A section.

Table 1. **Zero-shot generalization**. All compared methods are implemneted on the pre-trained CLIP (Radford et al., 2021) models without further training. Absolute gains over baseline are shown in blue.

Dataset	CLIP RN50	CLIP RN101	CLIP ViT-B/32	CLIP ViT-B/16	ZS_{En} Ours	Δ
Average	58.71	59.52	61.87	65.27	67.88	+2.61
ImageNet	58.23	61.26	62.04	66.72	70.66	+3.94
Caltech101	85.96	89.66	91.12	92.94	93.79	+0.85
OxfordPets	85.80	86.89	87.49	89.07	90.57	+1.50
Stanford Cars	55.57	63.16	60.37	65.29	70.76	+5.47
Flowers102	65.98	63.95	66.95	71.30	73.16	+1.86
Food101	77.32	80.54	80.47	86.11	86.78	+0.67
FGVC Aircraft	17.16	18.18	19.20	24.87	25.68	+0.81
SUN397	58.53	58.96	62.00	62.62	65.91	+3.29
DTD	42.38	38.48	43.79	44.56	49.35	+4.79
EuroSAT	37.42	32.62	45.12	47.69	50.20	+2.51
UCF101	61.43	61.04	62.07	66.77	69.84	+3.07

Table 2. **The ensemble of the CLIP-ViT-B/16 model with other weaker variants on ImageNet**.

CLIP ViT-B/16	CLIP RN50	CLIP RN101	CLIP ViT-B/32	Acc	Δ
✓				66.72	+0.00
✓	✓			67.75	+1.03
✓	✓	✓		68.34	+1.62
✓	✓	✓	✓	68.76	+2.04

4.1. Datasets

Following prior research (Zhou et al., 2022b;a; Khattak et al., 2023a;b), we employ a set of 11 diverse datasets, covering a large range of recognition tasks. Specifically, the benchmark comprises the following datasets: **(i)** ImageNet (Deng et al., 2009) and Caltech101 (Fei-Fei et al., 2004) for generic object classification; **(ii)** OxfordPets (Parkhi et al., 2012), StanfordCars (Krause et al., 2013), Flowers102 (Nilsback & Zisserman, 2008), Food101 (Bossard et al., 2014), and FGVCAircraft (Maji et al., 2013) for fine-grained classification; **(iii)** SUN397 (Xiao et al., 2010) for scene recognition; **(iv)** UCF101 (Soomro et al., 2012) for action recognition; **(v)** DTD (Cimpoi et al., 2014) for texture classification; **(vi)** EuroSAT (Helber et al., 2019) for satellite imagery recognition.

4.2. Implementation Details

In our experiments, we utilize four widely used CLIP models for ensemble learning: CLIP-RN50, CLIP-RN101, CLIP-ViT-B/32, and CLIP-ViT-B/16, with CLIP-ViT-B/16 being the best performer. In all three ensemble strategies, the **Logit** is computed by applying the softmax function to a model's prediction. For the training-free ensemble, the weight search space is fixed at $\{0.1, 0.2, \ldots, 1.0\}$ for each VLM, excluding CLIP-ViT-B/16. For the tuning ensemble, we set the initial learning rate to $5e-3$ and utilize the same adjusting scheduler as in (Zhou et al., 2022a; Khattak et al., 2023a;b). The sample-aware weight generator is a two-layer MLP ($f_{dim} \rightarrow f_{dim}/32$ and $f_{dim}/32 \rightarrow num_{weight}$), which is trained for 5 epochs with a batch size of 128. The tuning ensemble is applied to both pre-trained CLIP models and existing baseline methods.

Table 3. **Evaluating confidence-aware weighting (CAW) in different manners on ImageNet**. Baseline uses the mean of the logits from four models as the final prediction.

Method	Acc	Δ
CLIP ViT-B/16	66.72	+0.00
Baseline	68.76	+2.04
CAW of 3 models (w/o RN50)	70.01	+3.29
CAW of 4 models	70.19	+3.47
CLIP ViT-B/16 + CAW of 3 other models	70.66	+3.94

4.3. Evaluation Metrics

For zero-shot generalization, only the original test set of each dataset is used for computing accuracy since no training is needed. In the base-to-new generalization setup, we follow (Zhou et al., 2022b;a; Khattak et al., 2023a;b) to equally split the classes into two groups, *i.e.*, base and new classes, and then randomly sample a 16-shot training set from base classes. The learned model is then evaluated on both base and new classes from test set. We report three accuracy metrics: accuracy on base classes, accuracy on new classes and the harmonic mean (HM) of these two accuracies (Lu et al., 2023). In contrast, for cross-dataset generalization, a model is trained on ImageNet (Deng et al., 2009) and then conducted cross-validation on the remaining datasets. Note that the presented results are averaged over three runs, except for zero-shot generalization.

4.4. Zero-shot Generalization

We employ our zero-shot ensemble ZS_{En} for zero-shot generalization as neither additional training nor extra data is needed. The experimental results are shown in Table 1. Overall, the proposed ZS_{En} can always surpass the best performing single model, *i.e.*, CLIP-ViT-B/16, with the averaged performance gain 2.61% over 11 diverse datasets. In addition, two other interesting observations can be made.

First, the strategy of "weak helps strong" is employed successfully in our method, where models with weaker performance contribute to the ensemble, even when there are significant performance gaps between them, such as 58.71%

Table 4. **Comparison with state-of-the-art methods on base-to-new generalization**. All compared methods use the ViT-B/16 as the vision encoder. When employing our T_{En}, four of CoCoOp models, *i.e.*, RN50, RN101, ViT-B/32 and ViT-B/16, are utilized, while only two ViT-based models for PromptSRC as it is explicitly designed for ViT. †: reproduced by the official code. The best accuracy is in **bold**.

	Average			ImageNet			Caltech101			OxfordPets		
	Base	New	HM	Base	New	HM	Base	New	HM	Base	New	HM
CLIP (Radford et al., 2021)	69.34	74.22	71.70	72.43	68.14	70.22	96.84	94.00	95.40	91.17	97.26	94.12
CoOp (Zhou et al., 2022b)	82.69	63.22	71.66	76.47	67.88	71.92	98.00	89.81	93.73	93.67	95.29	94.47
CoCoOp (Zhou et al., 2022a)	80.47	71.69	75.83	75.98	70.43	73.10	97.96	93.81	95.84	95.20	97.69	96.43
CoCoOp† (Zhou et al., 2022a)	80.14	71.55	75.60	76.05	70.61	73.23	97.91	93.99	95.91	95.41	97.54	96.46
ProDA (Lu et al., 2022)	81.56	72.30	76.65	75.40	70.23	72.72	98.27	93.23	95.68	95.43	97.83	96.62
MaPLe (Khattak et al., 2023a)	82.28	75.14	78.55	76.66	70.54	73.47	97.74	94.36	96.02	95.43	97.76	96.58
Tip-Adapter + SHIP (Wang et al., 2023)	83.80	76.42	79.94	77.53	70.26	73.71	98.32	94.43	96.34	94.95	97.09	96.01
PromptSRC (Khattak et al., 2023b)	84.26	76.10	79.97	77.60	70.73	74.01	98.10	94.03	96.02	95.33	97.30	96.30
PromptSRC† (Khattak et al., 2023b)	83.81	75.69	79.54	77.57	70.22	73.71	98.06	93.96	95.97	94.59	97.15	95.85
TF_{En}	72.12	77.37	74.65	76.05	72.25	74.10	96.84	94.98	95.90	93.62	97.67	95.60
T_{En}	72.56	**77.72**	75.05	76.22	72.47	74.30	97.26	95.27	96.25	93.89	97.76	95.79
CoCoOp + T_{En}	83.56	75.27	79.20	78.39	**72.79**	**75.49**	98.32	**95.74**	97.01	95.69	**98.27**	**96.96**
PromptSRC + T_{En}	**85.48**	77.17	**81.11**	**78.74**	71.68	75.04	**98.58**	**95.74**	**97.14**	**95.96**	97.71	96.83

	StanfordCars			Flowers102			Food101			FGVCAircraft		
	Base	New	HM	Base	New	HM	Base	New	HM	Base	New	HM
CLIP (Radford et al., 2021)	63.37	74.89	68.65	72.08	77.80	74.83	90.10	91.22	90.66	27.19	36.29	31.09
CoOp (Zhou et al., 2022b)	78.12	60.40	68.13	97.60	59.67	74.06	88.33	82.26	85.19	40.44	22.30	28.75
CoCoOp (Zhou et al., 2022a)	70.49	73.59	72.01	94.87	71.75	81.71	90.70	91.29	90.99	33.41	23.71	27.74
CoCoOp† (Zhou et al., 2022a)	71.01	73.81	72.38	93.86	72.03	81.51	90.55	91.32	90.93	33.43	24.71	28.42
ProDA (Lu et al., 2022)	74.70	71.20	72.91	97.70	68.68	80.66	90.30	88.57	89.43	36.90	34.13	35.46
MaPLe (Khattak et al., 2023a)	72.94	74.00	73.47	95.92	72.46	82.56	90.71	92.05	91.38	37.44	35.61	36.50
Tip-Adapter + SHIP (Wang et al., 2023)	79.91	74.62	77.18	95.35	77.87	85.73	90.63	91.51	91.07	42.62	35.93	38.99
PromptSRC (Khattak et al., 2023b)	78.27	74.97	76.58	98.07	76.50	85.95	90.67	91.53	91.10	42.73	37.87	40.15
PromptSRC† (Khattak et al., 2023b)	79.19	75.45	77.27	97.63	77.07	86.14	90.35	91.46	90.90	42.73	35.63	38.86
TF_{En}	70.57	79.94	74.96	75.15	78.96	77.01	90.04	91.10	90.57	29.21	36.53	32.46
T_{En}	71.05	**80.21**	75.35	75.31	**79.15**	77.18	90.55	91.60	91.07	29.71	**37.25**	33.06
CoCoOp + T_{En}	78.69	78.73	78.71	97.82	77.23	86.31	**90.93**	**92.03**	**91.48**	36.51	30.17	33.04
PromptSRC + T_{En}	**81.26**	78.48	**79.85**	**98.77**	77.52	**86.86**	90.75	91.70	91.22	**43.22**	36.47	**39.56**

	SUN397			DTD			EuroSAT			UCF101		
	Base	New	HM	Base	New	HM	Base	New	HM	Base	New	HM
CLIP (Radford et al., 2021)	69.36	75.35	72.23	53.24	59.90	56.37	56.48	64.05	60.03	70.53	77.50	73.85
CoOp (Zhou et al., 2022b)	80.60	65.89	72.51	79.44	41.18	54.24	92.19	54.74	68.69	84.69	56.05	67.46
CoCoOp (Zhou et al., 2022a)	79.74	76.86	78.27	77.01	56.00	64.85	87.49	60.04	71.21	82.33	73.45	77.64
CoCoOp† (Zhou et al., 2022a)	79.46	77.24	78.33	76.01	54.99	63.81	86.10	56.49	68.22	81.73	74.27	77.82
ProDA (Lu et al., 2022)	78.67	76.93	77.79	80.67	56.48	66.44	83.90	66.00	73.88	85.23	71.97	78.04
MaPLe (Khattak et al., 2023a)	80.82	78.70	79.75	80.36	59.18	68.16	94.07	73.23	82.35	83.00	78.66	80.77
Tip-Adapter + SHIP (Wang et al., 2023)	81.32	77.64	79.43	81.83	61.47	70.21	93.38	**81.67**	**87.13**	85.99	78.10	81.85
PromptSRC (Khattak et al., 2023b)	82.67	78.47	80.52	83.37	62.97	71.75	92.90	73.90	82.32	87.10	78.80	82.74
PromptSRC† (Khattak et al., 2023b)	82.37	78.91	80.60	81.56	61.35	70.03	90.96	73.90	81.55	86.92	77.50	81.94
TF_{En}	72.90	77.08	74.93	58.72	64.82	61.62	56.60	77.85	65.55	73.63	79.86	76.62
T_{En}	73.18	77.49	75.27	59.26	**65.10**	62.04	57.69	78.60	66.54	74.03	**80.03**	76.91
CoCoOp + T_{En}	81.96	80.63	81.29	81.37	62.01	70.38	94.38	64.18	76.40	85.13	76.15	80.39
PromptSRC + T_{En}	**83.88**	**80.86**	**82.34**	**84.72**	63.16	**72.37**	**95.52**	76.41	84.90	**88.83**	79.10	**83.68**

(CLIP-RN50) *vs.* 65.27% (CLIP-ViT-B/16). This is a departure from conventional ensemble methods that typically involve models with similar performance. The effectiveness of weaker CLIP models contributing to the ensemble and aiding stronger ones can be attributed to the diverse structures and training on a large number of image-text pairs. While weaker models may have lower individual performance, they capture unique aspects of the data distribution and provide complementary information. The ensemble leverages this diversity, allowing the collective strength of the models to cover a broader range of patterns and features present in the data, thereby enhancing overall generalization performance.

Second, we found that the zero-shot performance of CLIP models with ResNet-based backbones does not consistently adhere to the expectation that larger models outperform smaller ones. On 5 out of 11 datasets, CLIP-RN101 performs either worse or comparably to CLIP-RN50. This trend contrasts with the behavior of Transformer-based models, which consistently improve with larger patch sizes. The discrepancy may be attributed to two main factors: (i) the ResNet-based vision encoder might not synergize effectively with the Transformer-based language encoder, hindering the capability of larger ResNet backbones. (ii) the ResNet-based structure might naturally exhibit weaker generalization capabilities compared to Transformer-based architectures.

Table 5. **Cross-dataset evaluation.** Our tuning ensemble T_{En} achieves the overall best performance.

	Source	Target										
	ImageNet	Caltech101	OxfordPets	StanfordCars	Flowers102	Food101	Aircraft	SUN397	DTD	EuroSAT	UCF101	*Average*
CLIP (Radford et al., 2021)	66.72	92.94	89.07	65.29	71.30	86.11	24.87	62.62	44.56	47.69	66.77	65.12
CoOp (Zhou et al., 2022b)	**71.51**	93.70	89.14	64.51	68.71	85.30	18.47	64.15	41.92	46.39	66.55	63.88
CoCoOp (Zhou et al., 2022a)	71.02	**94.43**	90.14	65.32	71.88	86.06	22.94	**67.36**	45.73	45.37	68.21	65.74
MaPLe (Khattak et al., 2023a)	70.72	93.53	90.49	65.57	72.23	86.20	24.74	67.01	46.49	48.06	68.69	66.30
PromptSRC (Khattak et al., 2023b)	71.27	93.60	90.25	65.70	70.25	86.15	23.90	67.10	46.87	45.50	68.75	65.81
T_{En}	70.88	93.91	**90.72**	**71.94**	**72.59**	**86.68**	**26.03**	66.07	**49.31**	**48.18**	**69.14**	**67.46**

Table 6. **Ablation Study of our SWIG** in terms of the number of output weights, input type and downsampling scales of middle layer on ImageNet.

Acc	# Output Weights		Input Type		Downsampling Scale				
	3	4	Logits	Features	4	8	16	32	64
Base	76.11	**76.22**	75.45	**76.22**	76.16	76.18	76.16	**76.22**	76.18
New	72.38	**72.47**	71.54	**72.47**	72.36	72.37	72.42	**72.47**	72.34
HM	74.20	**74.30**	73.44	**74.30**	74.21	74.23	74.24	**74.30**	74.21

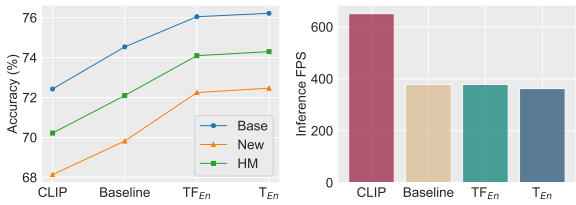
Ensemble of CLIP Models Table 2 presents the results of an ablation study on ensemble learning by incrementally adding a weak CLIP model to the strongest one. The adopted ensemble strategy involves the simple averaging of logits from the utilized models. The results reinforce the observation that weaker CLIP models contribute positively to the overall performance, supporting the "weak helps strong" phenomenon discussed earlier.

Effect of Confidence-aware Weighting The effectiveness of the proposed confidence-aware weighting (CAW) is evaluated in Table 3. We compare three different ways of using CAW for model ensemble: (i) applying CAW to the three strongest models, (ii) applying CAW to all four models, and (iii) applying CAW to the three weakest models combined with CLIP ViT-B/16 (ours). Our experiments demonstrate that all three approaches of using CAW can improve the performance of the ensemble, with our approach achieving the best results. This indicates that CAW is an effective technique for model ensemble, and that it is important to preserve the dominance of the best-performing model when large performance gaps exist among the ensemble models.

4.5. Base-to-New Generalization

We evaluate our training-free ensemble TF_{En} and tuning ensemble T_{En} on base-to-new generalization in Table 4. All baseline methods use ViT-B/16 as the default vision encoder. When employing T_{En} on baseline methods, we first train them with various backbones. Note that CoCoOp is trained with four backbones,*i.e.*, RN50, RN101, ViT-B/32 and ViT-B/16 while PromptSRC is trained with two ViT-based backbones as it is specifically designed for ViT. Given the baseline models with diverse backbones, we then only train our T_{En} for dynamic ensemble learning.

Overall, our ensemble strategies consistently demonstrate significant performance enhancements over baseline methods. For instance, TF_{En} and T_{En} improve zero-shot CLIP by 2.95% and 3.35%, respectively. CoCoOp + T_{En} exhibits a performance gain of 3.60%, while PromptSRC + T_{En} achieves a gain of 1.57%. Another interesting finding is that our T_{En} obtains the overall best performance on novel classes across 11 datasets. This can be attributed to two reasons: (i) the pre-trained CLIP models inherently capture generalized representations, and leveraging this knowledge effectively allows for superior performance on unseen classes; (ii) existing training methods often tend to overfit on base classes without effective constraints.



(a) Base-to-new performance (b) Inference FPS

Figure 6. **Comparing our ensemble strategies to baseline in terms of performance and inference time.** Baseline uses the mean of the logits from four models as the final prediction.

Further Analysis

Ensemble Strategies In Figure 6, we compare the performance and inference FPS with two baselines: zero-shot

Table 7. **Comparison with existing methods in domain generalization set-up**. Our CoCoOp + T_{En} shows the best performance on all target datasets.

	Source	Target				
	ImageNet	-V2	-S	-A	-R	Avg.
CLIP (Radford et al., 2021)	66.73	60.83	46.15	47.77	73.96	57.18
CoOp (Zhou et al., 2022b)	71.51	64.20	47.99	49.71	75.21	59.28
CoCoOp (Zhou et al., 2022a)	71.02	64.07	48.75	50.63	76.18	59.91
MaPLe (Khattak et al., 2023a)	70.72	64.07	49.15	50.90	76.98	60.28
PromptSRC (Khattak et al., 2023b)	71.27	64.35	49.55	50.90	77.80	60.65
CLIP + T_{En}	70.88	62.87	48.97	49.97	75.98	59.45
CoCoOp + T_{En}	**73.25**	**65.73**	**50.70**	**52.11**	**78.11**	**61.66**

CLIP-ViT-B/16 (**CLIP**) and mean of the logits from four CLIP models (**Baseline**). We found that both our TF_{En} and T_{En} can outperform **Baseline** with the comparable inference FPS, while achieving significant improvement over **CLIP** without sacrificing much inference time.

Ablation Study of SWIG We conduct comprehensive ablation studies of our SWIG in Table 6. First, we adjust the number of SWIG's outputs to three, representing the weights for three weak VLMs, while assigning a weight of 1.0 to CLIP-ViT-B/16. The performance was slightly inferior to using four weights. This suggests that tuning our SWIG on a given dataset can generate reasonable weights for the four models, with performance comparable to maintaining the dominance of CLIP-ViT-B/16. Second, we take the concatenated logits of four models as the input for our SWIG, but the performance is much worse as the SWIG is too biased on base classes as discussed in Sec 3.4. Finally, since SWIG consists of two linear layers: $f_{dim} \rightarrow f_{dim}/d_s$, $f_{dim}/d_s \rightarrow num_{weight}$, we ablate the downsampling scale d_s and found that our SWIG is not sensitive to d_s.

4.6. Cross-dataset Evaluation

Table 5 presents the results of cross-dataset evaluation for state-of-the-art methods. In general, our tuning ensemble demonstrates the best overall performance, surpassing other methods on 8 out of 10 datasets. Several noteworthy observations are discussed below. First, we observe that previous methods with intricate designs tend to overfit on the training dataset, leading to sub-optimal performance on other datasets. For instance, CoOp achieves the highest performance on ImageNet yet the lowest averaged accuracy over 10 datasets. Second, the zero-shot CLIP model, trained on a million-level paired data, already achieves impressive performance on diverse datasets, which is slightly worse than existing SOTA methods.

4.7. Domain Generalization

We provide the comparison results of existing methods on domain generalization in Table 7. Overall, the best performance is achieved by using our tuning ensemble T_{En} on CoCoOp, with the accuracy gain 1.75%. This further indicates the importance of ensemble learning on enhanced generalization of pre-trained vision-language models. Moreover, we found that CLIP + T_{En} shows much better performance than CLIP only, but is worse than other state-of-the-art generalization methods. This differs from the finding in the cross-dataset generalization, where CLIP + T_{En} surpasses other SOTA methods by large margins. The reason is that domain and cross-dataset generalization evaluate the model in distinct aspects. Concretely, cross-dataset generalization is more challenging as it requires the model trained on one dataset to perform well on other datasets with unseen classes and domains, while domain generalization considers domain transfer only. Another interesting observation is that our method can perform consistently better on both source and target domains. This stands in stark contrast to prior methods, which often excel in the source domain but exhibit a notable decline in performance when applied to the target domain. For instance, methods like CoOp may outperform others in the source domain but degrade when confronted with the target domain.

5. Conclusion

This paper represents the pioneering exploration of leveraging much weaker Vision-Language Models (VLMs) collaboratively to enhance the performance of a single robust one. Three tailored ensemble strategies are introduced, addressing distinct scenarios. Firstly, the proposed zero-shot ensemble dynamically adjusts logits based on model confidence when relying solely on pre-trained VLMs. Furthermore, for scenarios with additional few-shot samples, we propose both training-free and tuning ensembles, providing flexibility based on computing resource availability. Extensive experiments demonstrate the superiority of the proposed ensembles in diverse tasks including zero-shot evaluation and base-to-new generalization. Importantly, this work establishes a novel pathway for improved generalization in VLMs through ensemble learning, serving as a positive inspiration for future research endeavors.

Acknowledgments

This project is supported by the Ministry of Education, Singapore, under its Academic Research Fund Tier 2 (Award Number: MOE-T2EP20122-0006), and the National Research Foundation, Singapore, under its Medium Sized Center for Advanced Robotics Technology Innovation.

Impact Statement

This work pioneers the application of ensemble learning to enhance the generalization of vision-language models (VLMs) and offers three simple yet effective ensemble strategies, providing valuable insights to improve the generalization of pre-trained VLMs. More importantly, we made several interesting observations: (i) the "weak" VLMs, displaying substantial performance gaps compared to a strong individual VLM, can serve as valuable contributors in ensemble learning. (ii) contrary to the conventional wisdom that larger model sizes equate to better performance, we identified instances, particularly with ResNet-based (He et al., 2016) vision encoders for CLIP (Radford et al., 2021), where this expectation does not always hold. (iii) Zero-shot CLIP exhibits impressive generalization capabilities on unseen scenarios. Optimally harnessing this power can even surpass the performance of specifically designed generalization methods. In essence, our work provides fresh insights into the realm of ensemble learning for generalized VLMs and offers valuable findings that can contribute to advancements in the field.

References

Anderson, P., He, X., Buehler, C., Teney, D., Johnson, M., Gould, S., and Zhang, L. Bottom-up and top-down attention for image captioning and visual question answering. In *CVPR*, pp. 6077–6086, 2018.

Bossard, L., Guillaumin, M., and Gool, L. V. Food-101–mining discriminative components with random forests. In *ECCV*, pp. 446–461. Springer, 2014.

Breiman, L. Bagging predictors. *Machine learning*, 24: 123–140, 1996.

Breiman, L. Random forests. *Machine learning*, 45:5–32, 2001.

Buja, A. and Stuetzle, W. Smoothing effects of bagging. *Preprint. AT&T Labs-Research*, 2000.

Chao, W.-L., Changpinyo, S., Gong, B., and Sha, F. An empirical study and analysis of generalized zero-shot learning for object recognition in the wild. In *ECCV*, pp. 52–68. Springer, 2016.

Chen, L., Li, B., Shen, S., Yang, J., Li, C., Keutzer, K., Darrell, T., and Liu, Z. Large language models are visual reasoning coordinators. *NeurIPS*, 36, 2024.

Cimpoi, M., Maji, S., Kokkinos, I., Mohamed, S., and Vedaldi, A. Describing textures in the wild. In *CVPR*, pp. 3606–3613, 2014.

Coates, A. and Ng, A. Y. The importance of encoding versus training with sparse coding and vector quantization. In *ICML*, 2011.

Cortes, C., Mohri, M., and Syed, U. Deep boosting. In *ICML*, pp. 1179–1187. PMLR, 2014.

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., and Fei-Fei, L. Imagenet: A large-scale hierarchical image database. In *CVPR*, pp. 248–255. Ieee, 2009.

Deng, L., Yu, D., and Platt, J. Scalable stacking and learning for building deep architectures. In *ICASSP*, pp. 2133–2136. IEEE, 2012.

Desai, K. and Johnson, J. Virtex: Learning visual representations from textual annotations. In *CVPR*, pp. 11162–11173, 2021.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al. An image is worth 16x16 words: Transformers for image recognition at scale. In *ICLR*, 2020.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., and Houlsby, N. An image is worth 16x16 words: Transformers for image recognition at scale. *ICLR*, 2021.

Elhoseiny, M., Saleh, B., and Elgammal, A. Write a classifier: Zero-shot learning using purely textual descriptions. In *ICCV*, pp. 2584–2591, 2013.

Fei-Fei, L., Fergus, R., and Perona, P. Learning generative visual models from few training examples: An incremental bayesian approach tested on 101 object categories. In *CVPRW*, pp. 178–178. IEEE, 2004.

Freund, Y., Schapire, R. E., et al. Experiments with a new boosting algorithm. In *ICML*, volume 96, pp. 148–156. Citeseer, 1996.

Friedman, J. H. Greedy function approximation: a gradient boosting machine. *Annals of statistics*, pp. 1189–1232, 2001.

Frome, A., Corrado, G. S., Shlens, J., Bengio, S., Dean, J., Ranzato, M., and Mikolov, T. Devise: A deep visual-semantic embedding model. *NeurIPS*, 26, 2013.

Gomez, L., Patel, Y., Rusinol, M., Karatzas, D., and Jawahar, C. Self-supervised learning of visual features through embedding images into text topic spaces. In *CVPR*, pp. 4230–4239, 2017.

Ha, K., Cho, S., and MacLachlan, D. Response models based on bagging neural networks. *Journal of Interactive Marketing*, 19(1):17–30, 2005.

He, K., Zhang, X., Ren, S., and Sun, J. Deep residual learning for image recognition. In *CVPR*, pp. 770–778, 2016.

Helber, P., Bischke, B., Dengel, A., and Borth, D. 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.

Huynh, D. and Elhamifar, E. Fine-grained generalized zero-shot learning via dense attribute-based attention. In *CVPR*, pp. 4483–4493, 2020.

Jia, C., Yang, Y., Xia, Y., Chen, Y.-T., Parekh, Z., Pham, H., Le, Q., Sung, Y.-H., Li, Z., and Duerig, T. Scaling up visual and vision-language representation learning with noisy text supervision. In *ICML*, pp. 4904–4916. PMLR, 2021.

Jiang, Z., Xu, F. F., Araki, J., and Neubig, G. How can we know what language models know? *ACL*, 8:423–438, 2020.

Joulin, A., Maaten, L. v. d., Jabri, A., and Vasilache, N. Learning visual features from large weakly supervised data. In *ECCV*, pp. 67–84. Springer, 2016.

Kang, T., Chen, P., Quackenbush, J., and Ding, W. A novel deep learning model by stacking conditional restricted boltzmann machine and deep neural network. In *SIGKDD*, pp. 1316–1324, 2020.

Khattak, M. U., Rasheed, H., Maaz, M., Khan, S., and Khan, F. S. Maple: Multi-modal prompt learning. In *CVPR*, pp. 19113–19122, 2023a.

Khattak, M. U., Wasim, S. T., Naseer, M., Khan, S., Yang, M.-H., and Khan, F. S. Self-regulating prompts: Foundational model adaptation without forgetting. In *ICCV*, pp. 15190–15200, 2023b.

Krause, J., Stark, M., Deng, J., and Fei-Fei, L. 3d object representations for fine-grained categorization. In *ICCVW*, pp. 554–561, 2013.

LeBlanc, M. and Tibshirani, R. Combining estimates in regression and classification. *Journal of the American Statistical Association*, 91(436):1641–1650, 1996.

Lei Ba, J., Swersky, K., Fidler, S., et al. Predicting deep zero-shot convolutional neural networks using textual descriptions. In *ICCV*, pp. 4247–4255, 2015.

Li, A., Jabri, A., Joulin, A., and Van Der Maaten, L. Learning visual n-grams from web data. In *ICCV*, pp. 4183–4192, 2017.

Li, J., Li, D., Xiong, C., and Hoi, S. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *ICML*, pp. 12888–12900. PMLR, 2022.

Li, J., Li, D., Savarese, S., and Hoi, S. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *ICML*, pp. 19730–19742. PMLR, 2023a.

Li, X., Lian, D., Lu, Z., Bai, J., Chen, Z., and Wang, X. Graphadapter: Tuning vision-language models with dual knowledge graph. *NeurIPS*, 36, 2023b.

Li, Y., Liang, F., Zhao, L., Cui, Y., Ouyang, W., Shao, J., Yu, F., and Yan, J. Supervision exists everywhere: A data efficient contrastive language-image pre-training paradigm. *arXiv preprint arXiv:2110.05208*, 2021.

Lu, Y., Liu, J., Zhang, Y., Liu, Y., and Tian, X. Prompt distribution learning. In *CVPR*, pp. 5206–5215, 2022.

Lu, Z., He, S., Li, D., Song, Y.-Z., and Xiang, T. Prediction calibration for generalized few-shot semantic segmentation. *TIP*, 2023.

Maji, S., Rahtu, E., Kannala, J., Blaschko, M., and Vedaldi, A. Fine-grained visual classification of aircraft. *arXiv preprint arXiv:1306.5151*, 2013.

Mikolov, T., Chen, K., Corrado, G., and Dean, J. Efficient estimation of word representations in vector space. In *ICLR*, 2013a.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., and Dean, J. Distributed representations of words and phrases and their compositionality. *NeurIPS*, 26, 2013b.

Nilsback, M.-E. and Zisserman, A. Automated flower classification over a large number of classes. In *ICCVGIP*, pp. 722–729. IEEE, 2008.

Parkhi, O. M., Vedaldi, A., Zisserman, A., and Jawahar, C. Cats and dogs. In *CVPR*, pp. 3498–3505. IEEE, 2012.

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

Sariyildiz, M. B., Perez, J., and Larlus, D. Learning visual representations with caption annotations. In *ECCV*, pp. 153–170. Springer, 2020.

Shin, T., Razeghi, Y., Logan IV, R. L., Wallace, E., and Singh, S. Autoprompt: Eliciting knowledge from language models with automatically generated prompts. In *EMNLP*, 2020.

Socher, R., Ganjoo, M., Manning, C. D., and Ng, A. Zero-shot learning through cross-modal transfer. *NeurIPS*, 26, 2013.

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

Torresani, L., Szummer, M., and Fitzgibbon, A. Efficient object category recognition using classemes. In *ECCV*, pp. 776–789. Springer, 2010.

Touvron, H., Cord, M., Douze, M., Massa, F., Sablayrolles, A., and Jégou, H. Training data-efficient image transformers & distillation through attention. In *ICML*, pp. 10347–10357. PMLR, 2021.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. Attention is all you need. *NeurIPS*, 30, 2017.

Wang, W., Zheng, V. W., Yu, H., and Miao, C. A survey of zero-shot learning: Settings, methods, and applications. *TIST*, 10(2):1–37, 2019.

Wang, Z., Liang, J., He, R., Xu, N., Wang, Z., and Tan, T. Improving zero-shot generalization for clip with synthesized prompts. In *CVPR*, pp. 3032–3042, 2023.

Wolpert, D. H. Stacked generalization. *Neural networks*, 5 (2):241–259, 1992.

Xian, Y., Schiele, B., and Akata, Z. Zero-shot learning-the good, the bad and the ugly. In *CVPR*, pp. 4582–4591, 2017.

Xiao, J., Hays, J., Ehinger, K. A., Oliva, A., and Torralba, A. Sun database: Large-scale scene recognition from abbey to zoo. In *CVPR*, pp. 3485–3492. IEEE, 2010.

Yi, K., Shen, X., Gou, Y., and Elhoseiny, M. Exploring hierarchical graph representation for large-scale zero-shot image classification. In *ECCV*, pp. 116–132. Springer, 2022.

Yu, T., Lu, Z., Jin, X., Chen, Z., and Wang, X. Task residual for tuning vision-language models. In *CVPR*, pp. 10899–10909, 2023.

Zhong, Z., Friedman, D., and Chen, D. Factual probing is [mask]: Learning vs. learning to recall. In *NAACL*, 2021.

Zhou, K., Yang, J., Loy, C. C., and Liu, Z. Conditional prompt learning for vision-language models. In *CVPR*, pp. 16816–16825, 2022a.

Zhou, K., Yang, J., Loy, C. C., and Liu, Z. Learning to prompt for vision-language models. *IJCV*, 130(9):2337–2348, 2022b.

A. Appendix

We provide additional content to enhance the understanding of our method, which is listed as follows:

- Section A.1 offers more experimental results, including the effect of our tuning ensemble on state-of-the-art methods, evaluation on other VLMs, ensembel learning with random augmentations, parameter comparison and weight pattern analysis.
- Section A.2 shows the limitations of our method, inspiring the future works.

Table 8. **Effect of using tuning ensemble T_{En} on state-of-the-art methods**. When employing our T_{En}, four of CoCoOp models, *i.e.*, RN50, RN101, ViT-B/32 and ViT-B/16, are utilized, while only two ViT-based models for PromptSRC as it is explicitly designed for ViT. †: reproduced by the official code. $Mean_{En}$: averaging the predictions of individual models. The best accuracy is in **bold**.

	Average			ImageNet			Caltech101			OxfordPets		
	Base	New	HM	Base	New	HM	Base	New	HM	Base	New	HM
CoCoOp† (Zhou et al., 2022a)	80.14	71.55	75.60	76.05	70.61	73.23	97.91	93.99	95.91	95.41	97.54	96.46
CoCoOp + $Mean_{En}$	83.20	74.42	78.57	77.57	72.20	74.79	98.13	95.49	96.79	95.43	98.04	96.72
CoCoOp + T_{En}	83.56	75.27	79.20	78.39	**72.79**	**75.49**	98.32	**95.74**	97.01	95.69	**98.27**	**96.96**
PromptSRC† (Khattak et al., 2023b)	83.81	75.69	79.54	77.57	70.22	73.71	98.06	93.96	95.97	94.59	97.15	95.85
PromptSRC + $Mean_{En}$	85.12	76.63	80.65	78.14	71.42	74.63	98.45	95.25	96.82	95.55	97.62	96.57
PromptSRC + T_{En}	**85.48**	**77.17**	**81.11**	**78.74**	71.68	75.04	**98.58**	**95.74**	**97.14**	**95.96**	97.71	96.83

	StanfordCars			Flowers102			Food101			FGVCAircraft		
	Base	New	HM	Base	New	HM	Base	New	HM	Base	New	HM
CoCoOp† (Zhou et al., 2022a)	71.01	73.81	72.38	93.86	72.03	81.51	90.55	91.32	90.93	33.43	24.71	28.42
CoCoOp + $Mean_{En}$	78.23	78.46	78.34	97.69	75.93	85.45	90.85	91.89	91.37	36.39	27.05	31.03
CoCoOp + T_{En}	78.69	**78.73**	78.71	97.82	77.23	86.31	**90.93**	**92.03**	**91.48**	36.51	30.17	33.04
PromptSRC† (Khattak et al., 2023b)	79.19	75.45	77.27	97.63	77.07	86.14	90.35	91.46	90.90	42.73	35.63	38.86
PromptSRC + $Mean_{En}$	80.80	78.31	79.54	98.74	77.07	86.57	90.56	91.55	91.05	42.50	34.17	37.88
PromptSRC + T_{En}	**81.26**	78.48	**79.85**	**98.77**	**77.52**	**86.86**	90.75	91.70	91.22	**43.22**	**36.47**	**39.56**

	SUN397			DTD			EuroSAT			UCF101		
	Base	New	HM	Base	New	HM	Base	New	HM	Base	New	HM
CoCoOp† (Zhou et al., 2022a)	79.46	77.24	78.33	76.01	54.99	63.81	86.10	56.49	68.22	81.73	74.27	77.82
CoCoOp + $Mean_{En}$	81.61	80.41	81.01	81.02	61.15	69.70	93.77	62.15	74.75	84.54	75.88	79.98
CoCoOp + T_{En}	81.96	80.63	81.29	81.37	62.01	70.38	94.38	64.18	76.40	85.13	76.15	80.39
PromptSRC† (Khattak et al., 2023b)	82.37	78.91	80.60	81.56	61.35	70.03	90.96	73.90	81.55	86.92	77.50	81.94
PromptSRC + $Mean_{En}$	83.65	80.36	81.97	84.38	62.72	71.96	95.09	76.02	84.49	88.42	78.46	83.14
PromptSRC + T_{En}	**83.88**	**80.86**	**82.34**	**84.72**	**63.16**	**72.37**	**95.52**	**76.41**	**84.90**	**88.83**	**79.10**	**83.68**

Table 9. Zero-shot evaluation of Zero-Shot Ensemble (ZS_{En}) with BLIP, BLIP2 and CLIP-ViT-B/32 on ImageNet.

Method	Acc (%)
Zero-shot BLIP (Li et al., 2022)	49.14
Zero-shot BLIP2 (Li et al., 2023a)	36.41
Zero-shot CLIP-ViT-B/32 (Radford et al., 2021)	62.04
(BLIP, BLIP2) with Mean	51.39
(BLIP, BLIP2) with ZS_{En}	**56.38**
(BLIP, BLIP2, CLIP-ViT-B/32) with Mean	65.44
(BLIP, BLIP2, CLIP-ViT-B/32) with ZS_{En}	**67.88**

A.1. More Experiments

Effect of Tuning Ensemble Since we have analyzed the effect of using tuning ensemble T_{En} on CLIP in main text, we further show its effect on other state-of-the-art methods in Table 8, namely CoCoOp (CVPR22) (Zhou et al., 2022a) and PromptSRC (ICCV23) (Khattak et al., 2023b). In general, our T_{En} consistently demonstrates superior performance when compared to both the baseline model and its naïve ensemble counterpart. This trend underscores the efficacy of our method in effectively harnessing knowledge from multiple models. We also found that CoCoOp + T_{En} outperforms PromptSRC + T_{En} in some cases, *e.g.*, 75.49% *vs.* 75.04% on ImageNet. This is attributed to the advantage of employing more ensemble models, showcasing the effectiveness of our T_{En} in leveraging the collective power of multiple models.

Evaluation on Other VLMs We have expanded our experimental analysis to include CLIP-ViT-B/32, BLIP (Li et al., 2022) and BLIP2 (Li et al., 2023a) models, applying our Zero-Shot Ensemble (ZS_{En}) on ImageNet, as detailed in Table 9. It's important to clarify that this experiment does not presume prior knowledge of the best-performing model; instead, we implement confidence-aware weighting across both models indiscriminately. The results indicate a substantial enhancement in performance through our method, surpassing the results achieved by a simple mean ensemble.

Table 10. Base-to-new generalization of Training-free Ensemble (TF_{En}) and Tuning Ensemble (T_{En}) with BLIP, BLIP2 and CLIP-ViT-B/32 on ImageNet.

Method	Base Acc (%)	New Acc (%)	HM Acc (%)
Zero-shot BLIP (Li et al., 2022)	50.54	55.75	53.02
Zero-shot BLIP2 (Li et al., 2023a)	29.41	51.98	37.57
Zero-shot CLIP-ViT-B/32 (Radford et al., 2021)	67.45	64.02	65.69
(BLIP, BLIP2) with Mean	50.47	61.36	55.38
(BLIP, BLIP2) with TF_{En}	52.65	63.23	57.46
(BLIP, BLIP2) with T_{En}	**54.74**	**66.83**	**60.18**
(BLIP, BLIP2, CLIP-ViT-B/32) with Mean	68.86	70.00	69.43
(BLIP, BLIP2, CLIP-ViT-B/32) with TF_{En}	70.10	71.50	70.79
(BLIP, BLIP2, CLIP-ViT-B/32) with T_{En}	**71.04**	**73.92**	**72.45**

Table 11. Parameter comparison among various methods.

Method	CoOp	CoCoOp	MaPLe	PromptSRC	CoCoOp + T_{En}	PromptSRC + T_{En}
# Param	2048	34K	3.55M	46K	336K	292K
Acc (%)	71.66	75.60	78.55	79.54	**79.20**	**81.11**

Table 12. The weights searched by Training-Free Ensemble (TF_{En}) on diverse datasets.

Dataset	W_{RN50} (ZS Acc %)	W_{RN101} (ZS Acc %)	$W_{ViT-B/32}$ (ZS Acc %)
ImageNet	0.1 (58.23)	0.5 (61.26)	0.4 (62.04)
Caltech101	0.1 (85.96)	0.3 (89.66)	**0.4** (91.12)
OxfordPets	0.4 (85.80)	0.4 (86.89)	**0.6** (87.49)
StanfordCars	0.3 (55.57)	**0.8** (63.16)	0.4 (60.37)
Flowers102	0.3 (65.98)	0.8 (63.95)	**0.9** (66.95)
Food101	0.2 (77.32)	**0.3** (80.54)	0.2 (80.47)
FGVCAircraft	0.1(17.16)	0.7 (18.18)	0.2 (19.20)
SUN397	0.4 (58.53)	0.1 (58.96)	**0.7** (62.00)
DTD	0.6 (42.38)	0.6 (38.48)	**0.5** (43.79)
EuroSAT	0.8 (37.42)	0.5 (32.62)	0.1 (45.12)
UCF101	0.1 (61.43)	0.5 (61.04)	**0.8** (62.07)

Our TF_{En} and T_{En} methods were further assessed on CLIP-ViT-B/32, BLIP and BLIP2 models, with a focus on base-to-new generalization on ImageNet. The results, detailed in Table 10, underscore the efficacy of TF_{En} and T_{En} across various types of VLMs.

Ensemble Learning with Random Augmentations We have carried out experiments with the highest performing single model, ViT-B/16, applying four distinct random augmentations on the ImageNet. The augmentations are set to include "RandomResizedCrop(size=(224, 224), scale=(0.08, 1.0))" and "RandomHorizontalFlip(p=0.5)", applied four times to each image. These augmented images served as inputs for the model to do emsemble learning, resulting in an accuracy of 66.81%. This result slightly improves upon the baseline accuracy of 66.72% but falls significantly short of our ZS_{En} method's 70.66% accuracy. The limited performance enhancement using a single model, despite varied augmentations, suggests that a singular model may not adequately capture the multifaceted aspects of an image.

Parameter Comparison We list the parameters and accuracy of baseline methods in Table 11. From the numbers, we can see that our method achieves superior performance while preserving the nature of parameter efficiency.

Weight Pattern Analysis We have shown the searched weights of Training-Free Ensemble (TF_{En}) in Table 12. Note that the weight for ViT-B/16 is the constant 1.0. We found that the searching scheme tends to assign higher weights to the models with better zero-shot performance, evidenced on 8 out of 11 datasets. For models with similar performance, the weights assigned are also close.

A.2. Limitations

In this work, we take the initial stride towards incorporating ensemble learning into the generalization of VLMs. Our focus is on delineating potential avenues for leveraging ensemble strategies to enhance the performance of VLMs in generalized scenarios. However, it is essential to note that our exploration, at this juncture, only scratches the surface of the capabilities that ensemble learning might offer. Fully harnessing the power of ensemble learning in the context of VLM generalization represents a promising avenue for future investigation.