

ArGue: Attribute-Guided Prompt Tuning for Vision-Language Models

Xinyu Tian¹ Shu Zou¹ Zhaoyuan Yang² Jing Zhang¹
¹Australian National University ²GE Research

¹firstname.lastname@anu.edu.au, ²firstname.lastname@ge.com

Abstract

Although soft prompt tuning is effective in efficiently adapting Vision-Language (V&L) models for downstream tasks, it shows limitations in dealing with distribution shifts. We address this issue with Attribute-Guided Prompt Tuning (ArGue), making three key contributions. 1) In contrast to the conventional approach of directly appending soft prompts preceding class names, we align the model with primitive visual attributes generated by Large Language Models (LLMs). We posit that a model’s ability to express high confidence in these attributes signifies its capacity to discern the correct class rationales. 2) We introduce attribute sampling to eliminate disadvantageous attributes, thus only semantically meaningful attributes are preserved. 3) We propose negative prompting, explicitly enumerating class-agnostic attributes to activate spurious correlations and encourage the model to generate highly orthogonal probability distributions in relation to these negative features. In experiments, our method significantly outperforms current state-of-the-art prompt tuning methods on both novel class prediction and out-of-distribution generalization tasks. The code is available <https://github.com/Liam-Tian/ArGue>.

1. Introduction

Soft prompt tuning is increasingly favored in enabling Vision-Language (V&L) models [1, 16, 33] to be efficiently adapted to downstream tasks [20, 22, 24]. Models with a few soft tokens can achieve performance parity with, or even outperform, fully fine-tuned ones. Additionally, adapting to different downstream tasks typically necessitates prompt replacement rather than extensive model re-configuration [21, 38], further explaining the superiority of soft prompt tuning.

In typical classification tasks, prompt tuning often involves introducing a learnable context directly preceding the class name [20]. However, recent research in zero-shot recognition has emphasized the substantial benefits of incorporating visual attributes that describe the classes into

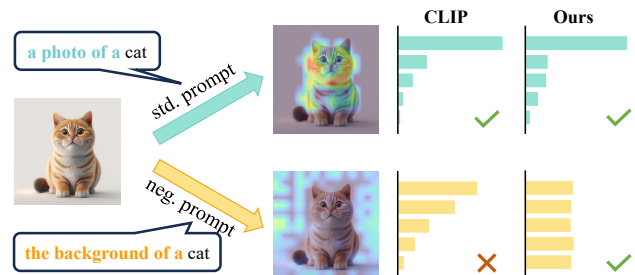


Figure 1. **The illustration of negative prompting.** Given an image of a cat (Left), we visualize the model rationale with Grad-CAM [36], which highlights the image pixels significantly determining the results (Middle). The standard prompt could be a photo of a cat, where vanilla models, e.g. CLIP [33], give high confidence on the ground truth class (the “CLIP” column). However, a negative prompt, e.g., the background of a cat, yields biased prediction since it activates the spurious correlation, i.e., background. In contrast, our attribute-guided model (the “Ours” column) disregards incorrect rationales and bases its predictions solely on class-specific semantics.

the input [28, 31, 35, 42, 44]. One observes that although class names, e.g., cat or bird, capture high-level semantics, during inference, primitive attributes, e.g., long tail or black paw, provide a more precise specification. This augmentation significantly enhances zero-shot classification accuracy, offering insights into transfer learning, particularly in few-shot scenarios.

In this paper, we investigate visual attributes for transfer learning by identifying the shortcuts existing in V&L models, which exhibit ease in adapting to new tasks but often provide incorrect rationales for their decisions [27]. For instance, a V&L model may correctly classify an object in the sky as a bird, not due to a comprehension of the semantic features, but because it detects spurious correlations between the bird and the sky. A model that predominantly highlights spurious correlations, e.g., the background, struggles to generalize effectively to out-of-distribution data.

To mitigate this challenge, we introduce *Attribute-Guided Prompt Tuning (ArGue)*. In contrast to the vanilla

prompt tuning methods that directly align image features with class names, ArGue encourages models to express high confidence in recognizing associated visual attributes generated by Large Language Models (LLMs) [3, 8, 32]. The underlying concept is that a model capable of identifying these primitive attributes captures the correct rationales for a class, rather than being influenced by spurious correlations. This approach offers two key advantages: firstly, attributes generated solely based on class names naturally circumvent shortcuts present in images, and secondly, these primitive attributes may be shared by other classes, enhancing models’ generalization capability.

Nevertheless, despite meticulous prompting, the inherent quality of attributes generated directly from LLMs remains uncertain. To address this, we present *Attribute Sampling* to select the most representative and non-redundant attributes that align well with the corresponding images. Particularly, the attribute pool is clustered, facilitating the selection of the most representative attributes per cluster while avoiding redundancy. Subsequently, within each cluster, we rank attributes based on their similarity to images in the feature space, opting for the most closely correlated attributes. This process enables the selection of the most semantically relevant visual attributes for the images. Empirically, we observe that reducing the number of attributes by 80% overall results in an accuracy improvement while conserving the computational resources.

Furthermore, rooted in attribute-guided prompt tuning, we introduce *Negative Prompting*, *i.e.*, ArGue-N. We contend that when presented with a negative attribute, one devoid of class-specific semantics and activating spurious correlations, the model should refrain from favoring any class. We provide a general negative prompt, *i.e.*, the background of a {class}, where the attribute, the background of a, activates the background of images which is semantically unrelated to classes. Upon employing a negative prompt, we enforce a uniform predictive probability distribution for the model (see Fig. 1 for an illustration of negative prompting). Despite the weak assumption of the general negative prompt, consistent performance enhancements are observed on out-of-distribution datasets.

In summary, our research focuses on leveraging visual attributes to encourage models to comprehend correct rationales, thereby improving robustness for transfer learning. The experiments reveal that our method outperforms existing state-of-the-art prompt tuning methods and, for the first time, surpasses pre-trained models on 10 out of 11 benchmark datasets in terms of novel class accuracy. Moreover, our method demonstrates consistent superior performance in out-of-distribution generalization against baselines. We aim for our work to serve as a foundational reference for the application of attributes in transfer learning, providing a strong baseline for the research community.

2. Related Work

Visual Attributes for Image Classification. Recent research emphasizes the use of visual attributes to enhance zero-shot recognition, moving beyond broad prompts like a photo of a {class} [28, 31, 35]. These attributes, *e.g.*, tail, paw, offer more distinguishing characteristics. Leveraging LLMs like GPT-3 [3], researchers can efficiently generate a wide array of class-specific attributes, surpassing manually crafted templates.

Despite the extensive research on zero-shot scenarios [18, 28, 31, 35, 42, 43], the role of attributes in transfer learning is under-explored. A pioneer study, Mao et al. [27], which is most related to ours, introduces an additional objective for V&L models to clarify their behaviors. However, they did not conduct an in-depth investigation into attributes, and manually curating attributes for datasets is quite costly. In contrast, we generate attribute pools through LLMs and efficiently select semantically related attributes via attribute sampling.

Prompt Engineering integrates foundational language models [3, 8, 32] into downstream tasks, allowing traditional tasks to be reframed as question-answering formats with carefully designed prompts [6, 11, 12, 17, 23, 39]. Manual prompt design is costly, driving the development of automated approaches like prompt tuning [20, 22, 24]. This technique optimizes soft tokens, reducing storage requirements and enhancing flexibility by enabling individual prompt replacement [21, 25, 38].

In the evolving field of V&L models [1, 16, 33], crafting text encoder prompts is pivotal for enhancing few-shot performance. CoOp [45] introduces soft prompts but at the expense of robustness. CoCoOp [46] tackles this by conditioning prompts on individual images, albeit with increased computational demand. LASP [4] proposes prompt regularization to align with pre-trained models’ generalization, yet overlooks their inherent biases. Our work extends LASP by utilizing attributes to guide models toward class-specific semantics and further correcting pre-trained model rationales through negative prompting.

3. Method

3.1. Preliminary

Prompt Engineering for Zero-shot Recognition. The Contrastive Language-Image Pre-training (CLIP) demonstrates the impressive understanding capability of V&L models for open-set concepts, showcasing competitive classification performance in zero-shot scenarios. Consider an image classification task where the dataset is defined as pairs $\mathcal{D} = \{(x, c)\}$, with x representing the image and $c \in \{1, \dots, C\}$ as its corresponding label. The classification problem is reformulated by calculating the similarity between visual and textual features within the CLIP space.

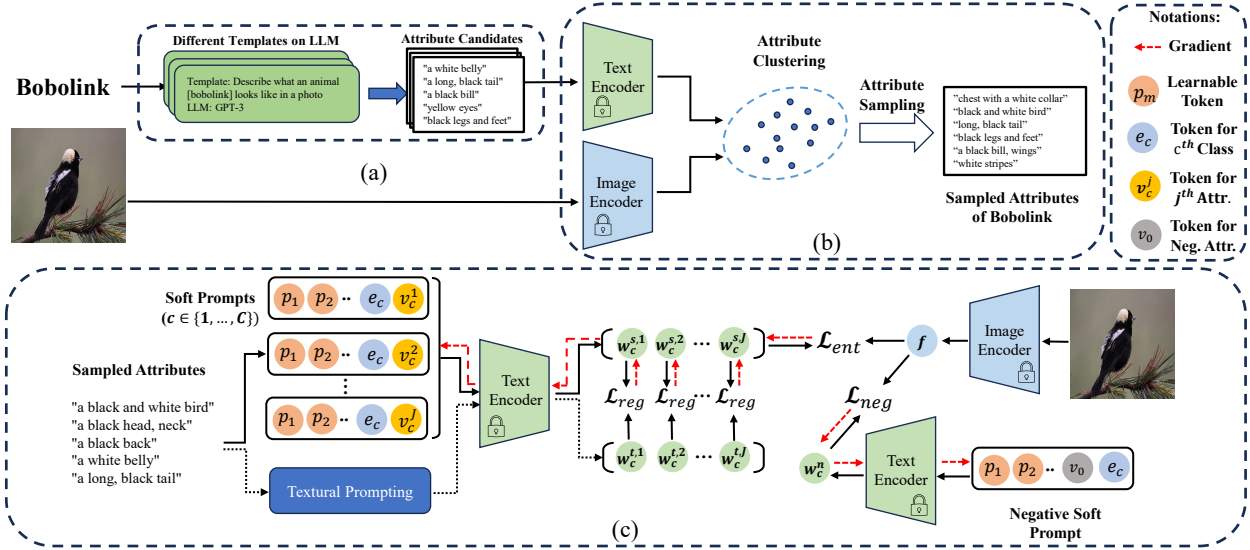


Figure 2. **The pipeline of ArGue.** In (a), we instruct the LLMs to generate attribute candidates using various LLM templates. In (b), we extract semantically relevant attributes through an assessment of their similarity to images, as described in Sec. 3.3. In (c), with guidance from the selected attributes and the application of negative prompting, we construct a set of soft tokens tailored to the task, which is detailed in Sec. 3.4 and Sec. 3.5.

Specifically, for each image x , it undergoes transformation via the vision encoder $h_I(\cdot)$ to compute a feature vector $\mathbf{f} = h_I(x)$. Simultaneously, a series of textual inputs $\{t_c\}_{c=1}^C$ are generated by appending a customized template to each class name, e.g., $t_c = \text{a photo of a } \{\text{class}_c\}$. These textual inputs are then processed through the text encoder $h_T(\cdot)$ to derive the textual features or known as weight vectors, denoted as $\{\mathbf{w}_c^t\}_{c=1}^C$, where $\mathbf{w}_c^t = h_T(t_c)$. The predictive probability for the image x classified to y is

$$P_t(y | x) = \frac{\exp(\cos(\mathbf{f}, \mathbf{w}_y^t) / \tau)}{\sum_{c=1}^C \exp(\cos(\mathbf{f}, \mathbf{w}_c^t) / \tau)}, \quad (1)$$

where $\cos(\cdot)$ computes the visual/text cosine similarity, and τ is a temperature scalar.

Prompt Tuning for Few-shot Learning. Prompt tuning aims to replace the manually designed discrete templates with a set of learnable continuous tokens $\{\mathbf{p}_m\}_{m=1}^M$ and optimize these tokens with a few labeled samples. Specifically, let $\mathbf{s}_c = \{\mathbf{p}_1, \mathbf{p}_2, \dots, \mathbf{p}_M, \mathbf{e}_c\}$ be the concatenation of the learnable tokens and the word embedding \mathbf{e}_c of a specific class c . With prompt tuning, the soft prompt \mathbf{s}_c is used instead of the discrete prompt t_c , leading to the learnable text embedding $\mathbf{w}_c^s = h_T(\mathbf{s}_c)$ with predictive distribution

$$P_s(y | x) = \frac{\exp(\cos(\mathbf{f}, \mathbf{w}_y^s) / \tau)}{\sum_{c=1}^C \exp(\cos(\mathbf{f}, \mathbf{w}_c^s) / \tau)}. \quad (2)$$

Finally, with the few labeled samples, a cross entropy loss is employed to align the logits with the ground truth to optimize the learnable tokens $\{\mathbf{p}_m\}_{m=1}^M$.

3.2. ArGue: Attribute-Guided Prompt Tuning

The pipeline of our method has been presented in Fig. 2. As discussed in Sec. 3.1, the word embedding of a specific class name is concatenated with the learnable tokens for conventional prompt tuning [20, 22, 45]. However, we contend that this practice represents a shortcut for CLIP to attain high accuracy without suitable rationales [27]. For instance, when presented with a class name of bird, CLIP may establish a semantic connection with the sky, introducing a dependence on the background rather than capturing the semantics of birds. This reliance on spurious correlations substantially undermines generalization capabilities.

To mitigate this challenge, instead of directly learning from class names, we advocate training a model that exhibits high confidence in the associated visual attributes, leading to the proposed *attribute-guided prompt tuning*. This approach is grounded in two fundamental intuitions. Firstly, in contrast to high-level class names, aligning explicitly with visual attributes encourages the model to prioritize inherent semantics of the class. Secondly, visual attributes representing low-level features may be shared with multiple classes, facilitating generalization to novel classes or out-of-distribution data.

A direct approach to obtain these visual attributes involves prompting LLMs with inquiries about the visual characteristics of specific classes. Notably, the LLM input exclusively consists of class names, thereby inherently circumventing shortcuts present in images. Formally, given any label c , we obtain a list of J attributes

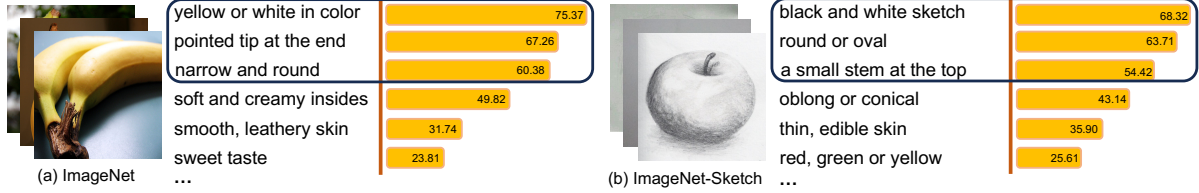


Figure 3. **Two example classes from (a) ImageNet and (b) ImageNet-Sketch for the attribute sampling procedure.** We demonstrate several attributes inside each class and the number within the yellow bar indicates its similarity to images in CLIP space. For each class, we designate 3 clusters, resulting in the selection of 3 attributes with the highest similarity score and they are framed with the black box.

$\text{attr}_c = \mathcal{U}(\text{class}_c)$, where \mathcal{U} is the language model. It’s worth noting that the templates for prompting LLMs have been pre-defined (see Supp. Mat. A). Now we let $\mathbf{s}_c^j = \{\mathbf{p}_1, \mathbf{p}_2, \dots, \mathbf{p}_M, \mathbf{e}_c, \mathbf{v}_c^j\}$, where $j \in [1, J]$, be the concatenation of the learnable tokens $\{\mathbf{p}_m\}_{m=1}^M$, the word embedding \mathbf{e}_c of class_c , and the word embedding \mathbf{v}_c^j of j^{th} attribute for class_c . We then define $\mathbf{w}_c^{s,j} = h_T(\mathbf{s}_c^j)$ as the attribute-guided soft embedding. Finally, for each sample (x, c) , we determine the probability distribution by averaging the logits over the attributes for each class, *i.e.*,

$$P_s(y | x) = \frac{\sum_{j=1}^J \exp(\cos(\mathbf{f}, \mathbf{w}_y^{s,j})/\tau)}{\sum_{c=1}^C \sum_{j=1}^J \exp(\cos(\mathbf{f}, \mathbf{w}_c^{s,j})/\tau)}. \quad (3)$$

The prompts are optimized with a typical cross entropy loss

$$\mathcal{L}_{ent} = -\sum_{c=1}^C y_c \log P_s(c | x). \quad (4)$$

Essentially, optimizing Eq. 4 implies our expectation for the model to exhibit high confidence in every attribute assigned to the ground truth class while minimizing its association with any other attributes.

3.3. Attribute Sampling

While LLMs can generate attributes associated with the class names, we find that some attributes exhibit a stronger semantic correlation with visual features than others. Our subsequent experiments further highlight that the removal of ineffective attributes not only reduces memory consumption but also improves the model’s accuracy. We thus work on selecting optimal attributes from an attribute pool. It is essential to note that while our primary task is few-shot adaptation, this method is equally applicable to attribute-based zero-shot recognition [28, 31, 35].

Our selection process revolves two main criteria: 1) the selected attributes should be both representative and non-redundant; 2) the selected attributes should be semantically related to the class-specific images. Consequently, our method involves two distinct steps. Firstly, given the attributes attr_c associated with class c from the attribute pool, we partition them into N clusters denoted as $\{\mathcal{A}_c^1, \mathcal{A}_c^2, \dots, \mathcal{A}_c^N\}$ based on their feature similarity in the CLIP space. This clustering strategy aims to ensure that

each cluster represents a distinct aspect, *e.g.*, color or shape, in the descriptions. Subsequently, within each cluster, we rank the attributes by assessing their similarity to visual features within the CLIP space, and select the one with the highest relevance. This approach filters out: 1) non-visual attributes, *e.g.*, sweet, edible, and 2) incorrect visual attributes that are semantically unrelated to the images.

An illustrative example could be found in ImageNet-Sketch [40], where the predominant content comprises sketches, devoid of the real colors of objects. Nevertheless, LLMs tend to generate class-specific colors despite careful prompting, *e.g.*, red for apple. In this situation, our attribute sampling approach initially groups attributes related to color into one cluster and subsequently identifies the most pertinent colors for sketches, *i.e.*, black and white. Fig. 3 offers concrete examples of this process.

3.4. Prompt Regularization

One issue of soft prompt learning within the few-shot setting is that the model may overfit training samples, leading to performance degradation for unseen data during testing [4]. Prompt regularization is a methodology that compels soft prompts to reside in proximity to natural texts in the feature space [4, 47], which is effective in dealing with the over-fitting issue. In this paper, we employ and interpret this technique through the lens of shortcut learning.

Empirically, the adaptation of pre-trained models often results in the acquisition of shortcuts, implying that spurious correlations, *e.g.*, background, may be given undue weight in the decision-making process. Therefore, prompt regularization is shown to be an effective approach for aligning semantic understanding with pre-trained models. Specifically, we define $t_c^j = \text{a photo of a } \{\text{class}_c\} \{\text{attr}_{c,j}\}$, which constitutes a textual prompt for the text encoder. Subsequently, we establish $\mathbf{w}_c^{t,j} = h_T(t_c^j)$. Recall that $\mathbf{w}_c^{s,j}$ represents the features for the attribute-guided soft prompts. The predictive distribution determining whether a soft prompt \mathbf{w}^s corresponds to its textual counterpart $\mathbf{w}_y^{t,k}$ is

$$P_{ts}(y, k | \mathbf{w}^s) = \frac{\exp(\cos(\mathbf{w}^s, \mathbf{w}_y^{t,k})/\tau)}{\sum_{c=1}^C \sum_{j=1}^J \exp(\cos(\mathbf{w}^s, \mathbf{w}_c^{t,j})/\tau)}. \quad (5)$$

The cross entropy loss is then used to optimize the prompts

$$\mathcal{L}_{reg} = -\sum_{c=1}^C \sum_{j=1}^J y_{cj} \log P_{ts}(c, j | \mathbf{w}^s). \quad (6)$$

That is, we establish a positive pair for each soft prompt in conjunction with its corresponding textual prompt, while any other textual prompt is designated as a negative pair. Consequently, the optimization of Eq. 6 is carried out in a contrastive manner.

In summary, we combine the loss terms as follows

$$\mathcal{L} = \mathcal{L}_{ent} + \beta \mathcal{L}_{reg}, \quad (7)$$

where β represents a predefined weight to balance the two components. We designate our method as Attribute-Guided Prompt Tuning (ArGue) for incorporating and sampling primitive visual attributes to bypass the incorrect rationales in the images.

3.5. Negative Prompting

In preceding sections, we explore the process of selecting attributes that maintain semantic and intrinsic relevance to our images. In this section, we further study the effects of attributes, but in the other way. We introduce the concept of negative prompting, where our objective is to explicitly enumerate attributes lacking class-specific information. We expect the model to display no preference for any class when presented with these negative attributes.

To illustrate, consider the cat image in Fig. 1, where CLIP is expected to confidently identify standard prompts like a photo of a cat. However, when introduced to a negative prompt, *e.g.*, the background of a cat, the model should provide a uniform prediction without a dominant class. In this context, the background of a exemplifies a typical negative attribute devoid of class-specific information while activating spurious correlations from the images. It serves as the general negative attribute in this paper. Although it is possible to provide more specific negative attributes, manually labeling them for each class is a labor-intensive task. Additionally, our experiments reveal that the general negative attribute, despite being a weak assumption, performs remarkably well across most datasets. A discussion on manually curating class-specific negative attributes is provided in Supp. Mat. E.

Moreover, it’s noteworthy that negative prompting follows a format akin to attribute-guided prompts, involving the integration of class names into the prompt structure. Empirical findings [35] suggest that when models overly lean on the class name, the impact of the attribute tends to be weakened. Considering that the negative prompt includes the class name, the model is designed to lessen the influence of negative attributes while concurrently diminishing the significance of class names. As a result, the model adeptly identifies and engages with areas indicated by class-specific

attributes, prioritizing them over class names for precise activation.

Formally, consider a negative attribute attr_0 , we define the embedding of the negative prompt as $\mathbf{n}_c = \{\mathbf{p}_1, \mathbf{p}_2, \dots, \mathbf{p}_M, \mathbf{v}_0, \mathbf{e}_c\}$, where \mathbf{v}_0 is the word embedding of the negative attribute. Then we let $\{\mathbf{w}_c^n\}_{c=1}^C$, where $\mathbf{w}_c^n = h_T(\mathbf{n}_c)$. The predictive probability that the negative prompt is classified to class y is

$$P_n(y | x) = \frac{\exp(\cos(\mathbf{f}, \mathbf{w}_y^n)/\tau)}{\sum_{c=1}^C \exp(\cos(\mathbf{f}, \mathbf{w}_c^n)/\tau)}. \quad (8)$$

To ensure that the model exhibits no preference for either class, we enforce the probability to be uniform. In other words, we aim to maximize the entropy of the distribution.

$$\mathcal{L}_{neg} = \sum_{c=1}^C \log P_n(c | x). \quad (9)$$

In summary, we aggregate all the introduced components

$$\mathcal{L} = \mathcal{L}_{ent} + \beta \mathcal{L}_{reg} + \gamma \mathcal{L}_{neg}, \quad (10)$$

where γ denotes the weight that accentuates the importance of negative prompting. We formally designate the comprehensive method as ArGue-N, signifying its inclusion of negative prompting within our attribute-guided prompt tuning framework.

4. Experiment

The evaluation primarily focuses on two tasks similar to [4, 46]: novel class prediction and out-of-distribution generalization. In the novel class prediction task, each dataset is equally partitioned into base and novel classes. The model undergoes training on the base classes, followed by the evaluation of test sets encompassing both base and novel classes. For the out-of-distribution generalization task, the model is transferred from an in-distribution dataset to several distinct yet related variants. Furthermore, we conduct a comprehensive analysis to validate and enhance our understanding of the proposed methodology.

Datasets. In the novel class prediction task, we employ 11 datasets, encompassing ImageNet [7], Caltech101 [10], OxfordPets [30], StanfordCars [19], Flowers102 [29], Food101 [2], FGVCAircraft [26], SUN397 [41], UCF101 [37], DTD [5] and EuroSAT [13]. For the out-of-distribution generalization task, we designate ImageNet [7] as the in-distribution or source set, and extend the model’s capabilities to four variants, including ImageNetV2 [34], ImageNet-Sketch [40], ImageNet-A [15] and ImageNet-R [14]. For a fair comparison, following [45, 46], we randomly sample 16 images, *i.e.*, 16 shots for each class, to form the training set. Each result represents an average over three runs with different initializations.

Dataset	CLIP [33]			CoOp [45]			CoCoOp [46]			LASP [4]			ArGue			ArGue-N			
	Base	New	H	Base	New	H	Base	New	H	Base	New	H	Base	New	H	Base	New	H	Δ
Average	69.34	74.22	71.70	82.69	63.22	71.66	80.47	71.69	75.83	83.18	76.11	79.48	83.69	78.07	80.78	83.77	78.74	81.18	+1.70
ImageNet	72.43	68.14	70.22	76.47	67.88	71.92	75.98	70.43	73.10	76.25	71.17	73.62	76.92	72.06	74.41	76.95	71.86	74.32	+0.70
Caltech101	96.84	94.00	95.40	98.00	89.91	93.73	97.96	93.81	95.84	98.17	94.33	96.21	98.43	95.20	96.79	98.63	94.70	96.63	+0.42
OxfordPets	91.17	97.26	94.12	93.67	95.29	94.47	95.20	97.69	96.43	95.73	97.87	96.79	95.36	97.95	96.64	96.23	98.59	97.40	+0.61
StanfordCars	63.37	74.89	68.85	78.12	60.40	68.13	70.49	73.59	72.01	75.23	71.77	73.46	75.64	73.38	74.49	75.06	74.18	74.62	+1.16
Flowers102	72.08	77.80	74.83	97.60	59.67	74.06	94.87	71.75	81.71	97.17	73.53	83.71	98.34	75.41	85.36	98.62	77.96	87.08	+3.37
Food101	90.10	91.22	90.66	88.33	82.26	85.19	90.70	91.29	90.99	91.20	91.90	91.54	92.33	91.96	92.14	91.42	92.40	91.91	+0.37
FGVCAircraft	27.19	36.29	31.09	40.44	22.30	28.75	33.41	23.71	27.74	38.05	33.20	35.46	40.46	38.03	39.21	41.29	38.80	40.01	+4.55
SUN397	69.36	75.35	72.23	80.60	65.89	72.51	79.74	76.86	78.27	80.70	79.30	80.00	81.52	80.74	81.13	81.89	80.48	81.18	+1.18
DTD	53.24	59.90	56.37	79.44	41.18	54.24	77.01	56.00	64.85	81.10	62.57	70.64	81.60	66.55	73.31	80.33	67.03	73.08	+2.44
EuroSAT	56.48	64.05	60.03	92.19	54.74	68.90	87.49	60.04	71.21	95.00	83.37	88.86	94.43	88.24	91.23	95.10	90.68	92.84	+3.98
UCF101	70.53	77.50	73.85	84.69	56.05	67.46	82.33	73.45	77.64	85.53	78.20	81.70	85.56	79.29	82.31	86.00	79.43	82.58	+0.88

Table 1. **The comparison with baselines on novel class prediction.** We report performance of both ArGue and its variant, ArGue-N. H is the harmonic mean of the test accuracy on base and new class. Δ is the absolute difference between ArGue-N and previous best results.

Dataset		CLIP	CoOp	CoCoOp	LASP	ArGue	ArGue-N
OOD	ImageNet	66.73	71.51	71.02	71.34	71.57	71.84
	ImageNetV2	60.83	64.20	64.07	64.04	64.57	65.02
	ImageNet-Sketch	46.15	47.99	48.75	47.93	48.92	49.25
	ImageNet-A	47.77	49.71	50.63	49.11	50.93	51.47
	ImageNet-R	73.96	75.21	76.18	75.36	76.56	76.96

Table 2. **The comparison against baselines for out-of-distribution generalization.** We employ ImageNet as our in-distribution set for adaptation and subsequently transfer our models to four related out-of-distribution variants.

Baselines. A primary point of reference is LASP [4], upon which we build our models. Additionally, we contrast our approach with CoCoOp [46], which conditions on images but significantly escalates computational requirements. Two baseline models, CLIP [33] and CoOp [45], are included, representing zero-shot performance and vanilla prompt tuning, respectively.

Implementation Details. By default, we employ a pre-trained CLIP model with a ViT-B/16 vision encoder backbone [9]. The soft token length M is configured to be 4 and is initialized with the word embedding of a photo of a. The choice of epoch numbers, learning rate, optimizer, and batch size aligns with the baselines [4, 45, 46] (SGD optimizer with a learning rate of 0.032 and a batch size of 32). Additionally, we set β to 20 following [4] and γ to 3 based on empirical observations (see Supp. Mat. G for parameter analysis of γ). For each class in datasets, we generate a total of $J = 15$ attributes with GPT-3 [3], while only sampling $N = 3$ representative attributes for training. We determine

N based on a 20% proportion relative to the total number of attributes. Insufficient attributes may not comprehensively elucidate the class, while an excessive N introduces redundancy, thereby amplifying computational burden (see Supp. Mat. H for further analysis).

4.1. Novel Class Prediction

The superiority of ArGue-N over state of the art. Table 1 provides a comparative analysis of our methods against baseline models for novel class prediction, showcasing ArGue-N’s consistent outperformance of LASP, the current state-of-the-art, by 1.70% on average across base and novel classes. Notably, it excels on more challenging benchmark datasets, demonstrating a remarkable 3.98% improvement on EuroSAT and an impressive 4.55% gain on FGVCAircraft. Additionally, CLIP serves as a robust baseline for novel class accuracy due to its large-scale pre-training. For the first time, ArGue-N outperforms CLIP on novel classes in 10 out of 11 datasets, marking a notable milestone.

The comparison between ArGue and ArGue-N. ArGue-N exhibits an overall advantage over ArGue, with an absolute improvement of 0.40% on average. It’s worth noting that this advantage is contingent upon dataset characteristics. When spurious correlations predominantly reside in the background of the dataset, *e.g.*, OxfordPets (+0.76%), Flowers102 (+1.72%), the efficacy of negative prompting becomes pronounced. Conversely, in specialized datasets, *e.g.*, DTD (-0.23%), ArGue-N tends to converge towards ArGue, as images cannot be distinguished between background and foreground, *e.g.*, textures. Nonetheless, the general negative prompt yields favorable results across the majority of datasets without any manual supervision.

Dataset	Baseline [45]			Attr.			+ Reg.			+ Samp. (ArGue)			+ Neg. (ArGue-N)		
	base	new	H	base	new	H	base	new	H	base	new	H	base	new	H
Average	82.69	63.22	71.66	83.51	75.50	79.30	83.54	77.57	80.44	83.69	78.07	80.78	83.77	78.74	81.18
ImageNet	76.47	67.88	71.92	76.77	71.43	74.00	76.67	71.90	74.21	76.92	72.06	74.41	76.95	71.86	74.32
Caltech101	98.00	89.91	93.73	98.54	93.16	95.77	98.36	94.75	96.52	98.43	95.20	96.79	98.63	94.70	96.63
OxfordPets	93.67	95.29	94.47	95.13	96.79	95.95	95.17	98.02	96.57	95.36	97.95	96.64	96.23	98.59	97.40
StanfordCars	78.12	60.40	68.13	77.52	70.38	73.78	75.98	72.29	74.09	75.64	73.38	74.49	75.06	74.18	74.62
Flowers102	97.60	59.67	74.06	98.56	72.45	83.51	98.17	75.01	85.04	98.34	75.41	85.36	98.62	77.96	87.08
Food101	88.33	82.26	85.19	92.19	89.47	90.81	92.14	91.97	92.05	92.33	91.96	92.14	91.42	92.40	91.91
FGVCAircraft	40.44	22.30	28.75	38.36	37.55	37.95	39.31	38.05	38.67	40.46	38.03	39.21	41.29	38.80	40.01
SUN397	80.60	65.89	72.51	81.14	78.82	79.96	81.07	80.06	80.56	81.52	80.74	81.13	81.89	80.48	81.18
DTD	79.44	41.18	54.24	81.27	65.92	72.79	81.62	65.98	72.97	81.60	66.55	73.31	80.33	67.03	73.08
EuroSAT	92.19	54.74	68.90	94.10	78.95	85.86	94.78	86.41	90.40	94.43	88.24	91.23	95.10	90.68	92.84
UCF101	84.69	56.05	67.46	84.98	75.55	79.99	85.62	78.80	82.07	85.56	79.29	82.31	86.00	79.43	82.58

Table 3. **Components analysis.** CoOp [45] is chosen as the baseline as it is the vanilla prompt tuning method without any modification.

4.2. Out-of-Distribution Generalization

ArGue outperforms baselines. Table 2 presents results by transferring from ImageNet to four variants. ArGue consistently exhibits strengths across all five datasets, with a notably substantial enhancement observed in OOD datasets. This observation is comprehensible as the distribution shift does not alternate class-specific semantics or introduce novel classes. ArGue empowers the model to comprehend the visual attributes associated with each existing class, reinforcing its robustness across different variants.

ArGue-N eliminates shortcuts. As shown in Table 2, ArGue-N consistently outperforms ArGue across four distinct variants. This observation suggests that ImageNet exhibits spurious correlations between background elements and class labels, and the utilization of negative prompting encourages the model to eliminate these shortcuts, refocusing its attention on the inherent semantics of the categories. The OOD datasets, in an adversarial manner, effectively eradicate these shortcuts. For instance, consider ImageNet-sketch, where objects are exclusively represented through sketches, completely devoid of any background context.

4.3. Attribute Sampling Analysis

We provide visual examples for a more comprehensive analysis of the influence of our attribute sampling procedure. In Fig. 3, we select one class from ImageNet and ImageNet-Sketch, respectively. Utilizing LLMs, we generate attributes for each class, thus creating an attribute pool. Subsequently, we apply attribute sampling to exclude ineffective attributes (see Sec. 3.3). The attributes that undergo filtering can be categorized into two primary types.

Non-visual attributes. Despite our meticulous guidance to LLMs to acquire visual attributes, it is possible for non-

visual attributes, *e.g.*, edible, sweet, to surface. Attribute sampling may place these attributes within any cluster, but their resemblance to the images is lower in comparison to other visual attributes, resulting in their exclusion from the selection process.

Semantically unrelated visual attributes refer to attributes that possess visual features but do not correspond to the image content. For instance, in scenarios like ImageNet-Sketch, where images only contain black sketches, the attribute pool may still include descriptions of object colors, *e.g.*, red for apples. In our clustering process, we tend to group attributes with similar semantics together, *e.g.*, {red, yellow, black}, {round, square, oblong}. Subsequently, color descriptions that do not align with the actual image content are regarded as dissimilar and are therefore excluded from the selection process.

4.4. Ablation Study

Simply introducing attributes improves the baseline by large margins. Table 3 presents the performance as we progressively include components. As evident from the table, the transition from the baseline to the vanilla solution guided solely by generated attributes without any additional components (the “Attr.” column), leads to a substantial 7.64% improvement on average. When juxtaposed with the observations in Table 1, it becomes clear that even without the inclusion of our proposed components, this level of performance outperforms CLIP and CoCoOp significantly and matches LASP, explaining the potential of attributes for novel class prediction.

Attribute sampling contributes more gains with less computation. During the sampling process, we select 20% attributes from the pool, resulting in an average perfor-

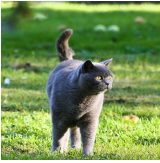
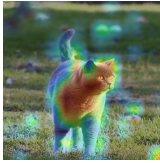



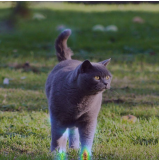
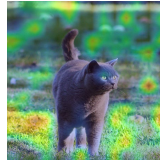

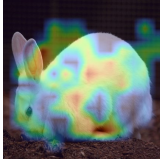




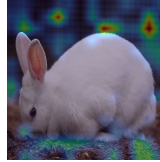
(A) Image	(B) Comp. with baseline methods			(C) Attr. visualization	(D) Neg. prompt	
Class Label: British Shorthair	Standard Prompt: A photo of a British shorthair			A photo of a cat which has a long tail	A photo of a cat which has a small paw	The background of a British shorthair
	CLIP	CoOp	ArGue-N			
						
Class Label: White Rabbit	Standard Prompt: A photo of a white rabbit			A photo of a rabbit which has big ears	A photo of a rabbit which has black eyes	The background of a white rabbit
	CLIP	CoOp	ArGue-N			
						

Figure 4. **The Grad-CAM visualization of our method and baselines.** (A) contains visual images. (B) features a comparison between our method and baselines using standard prompts, where CoOp and ArGue-N replace the template A photo of a with their respective soft tokens. (C) reveals the rationale of ArGue-N concerning various visual attributes. (D) showcases the negative prompt used during training.

mance improvement of 0.34% compared with the vanilla one (the “+Reg.” column). This indicates that with the judicious selection of attributes, significant enhancements can be achieved by merely introducing 1 to 2 additional prompts to the baseline.

4.5. Grad-CAM Visualization

To further enhance our comprehension of the learned rationales in ArGue-N, we employ Grad-CAM [36] to visualize the class activation map of the model in Fig. 4.

ArGue-N relies on correct rationales. In Fig. 4 (B), we conduct a comparative analysis to showcase the rationales learned by ArGue-N. We compare ArGue-N with baselines using the standard prompt that solely includes class names. It indicates that while CLIP broadly captures class-specific semantics, it also incorporates dependencies from the background. Moreover, CoOp exhibits a significant emphasis shift from the foreground to the background. Conversely, ArGue-N 1) more precisely captures the pixels determining intrinsic semantics and 2) nearly eliminates the background’s influence on the classification results.

ArGue-N comprehends primitive attributes. In Fig. 4 (C), we provide visualizations illustrating the rationales captured by ArGue-N using various primitive attributes in the prompts. These visual representations demonstrate ArGue-N’s proficiency in localizing the mentioned visual attributes while notably reducing the influence of the background. This observation supports our claim that when a

model exhibits high confidence in associated attributes, it accurately captures the correct rationales while mitigating the impact of spurious correlations.

Negative prompting diminishes reliance on class names.

The findings in Fig. 4 (C) reveal that ArGue-N precisely identifies the areas indicated by the attributes while disregarding the class names. For example, when prompted with a photo of a cat which has a long tail, the model accurately activates the tail rather than the entire cat. This phenomenon aligns with our assertion that incorporating class names within negative prompts contributes to reducing the model’s dependence on them.

5. Conclusion

We delve into an under-explored area, *i.e.*, leveraging visual attributes to guide the model toward correct rationales during adaptation. We propose ArGue, motivated by the intuition that a model exhibiting high confidence in associated visual attributes comprehends the class-specific semantics. We further introduce attribute sampling to enhance the quality of attributes while conserving computational resources by removing ineffective attributes. Finally, we present negative prompting, where, when provided with prompts that activate spurious correlations, the model is constrained with uniform predictive distribution. As attributes become increasingly prevalent in multi-modal zero-shot recognition, we aim for our work to initiate the incorporation of attributes into few-shot adaptation and serve as a strong baseline.

References

- [1] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, Roman Ring, Eliza Rutherford, Serkan Cabi, Tengda Han, Zhitao Gong, Sina Samangooei, Marianne Monteiro, Jacob L. Menick, Sebastian Borgeaud, Andy Brock, Aida Nematzadeh, Sahand Sharifzadeh, Mikolaj Binkowski, Ricardo Barreira, Oriol Vinyals, Andrew Zisserman, and Karén Simonyan. Flamingo: a visual language model for few-shot learning. In *NeurIPS*, 2022. 1, 2
- [2] Lukas Bossard, Matthieu Guillaumin, and Luc Van Gool. Food-101 - mining discriminative components with random forests. In *ECCV (6)*, pages 446–461. Springer, 2014. 5
- [3] Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners. In *NeurIPS*, 2020. 2, 6
- [4] Adrian Bulat and Georgios Tzimiropoulos. LASP: text-to-text optimization for language-aware soft prompting of vision & language models. In *CVPR*, pages 23232–23241. IEEE, 2023. 2, 4, 5, 6
- [5] Mircea Cimpoi, Subhansu Maji, Iasonas Kokkinos, Sammy Mohamed, and Andrea Vedaldi. Describing textures in the wild. In *CVPR*, pages 3606–3613. IEEE Computer Society, 2014. 5
- [6] Joe Davison, Joshua Feldman, and Alexander M. Rush. Commonsense knowledge mining from pretrained models. In *EMNLP/IJCNLP (1)*, pages 1173–1178. Association for Computational Linguistics, 2019. 2
- [7] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *CVPR*, pages 248–255. IEEE Computer Society, 2009. 5
- [8] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: pre-training of deep bidirectional transformers for language understanding. In *NAACL-HLT (1)*, pages 4171–4186. Association for Computational Linguistics, 2019. 2
- [9] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *ICLR*. OpenReview.net, 2021. 6
- [10] Li Fei-Fei, Rob Fergus, and Pietro Perona. Learning generative visual models from few training examples: An incremental bayesian approach tested on 101 object categories. In *CVPR Workshops*, page 178. IEEE Computer Society, 2004. 5
- [11] Tianyu Gao, Adam Fisch, and Danqi Chen. Making pre-trained language models better few-shot learners. In *ACL/IJCNLP (1)*, pages 3816–3830. Association for Computational Linguistics, 2021. 2
- [12] Adi Haviv, Jonathan Berant, and Amir Globerson. Bertese: Learning to speak to BERT. In *EACL*, pages 3618–3623. Association for Computational Linguistics, 2021. 2
- [13] Patrick Helber, Benjamin Bischke, Andreas Dengel, and Damian Borth. Eurosat: A novel dataset and deep learning benchmark for land use and land cover classification. *IEEE J. Sel. Top. Appl. Earth Obs. Remote. Sens.*, 12(7):2217–2226, 2019. 5
- [14] Dan Hendrycks, Steven Basart, Norman Mu, Saurav Kadavath, Frank Wang, Evan Dorundo, Rahul Desai, Tyler Zhu, Samyak Parajuli, Mike Guo, Dawn Song, Jacob Steinhardt, and Justin Gilmer. The many faces of robustness: A critical analysis of out-of-distribution generalization. In *ICCV*, pages 8320–8329. IEEE, 2021. 5
- [15] Dan Hendrycks, Kevin Zhao, Steven Basart, Jacob Steinhardt, and Dawn Song. Natural adversarial examples. In *CVPR*, pages 15262–15271. Computer Vision Foundation / IEEE, 2021. 5
- [16] Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc V. Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In *ICML*, pages 4904–4916. PMLR, 2021. 1, 2
- [17] Zhengbao Jiang, Frank F. Xu, Jun Araki, and Graham Neubig. How can we know what language models know. *Trans. Assoc. Comput. Linguistics*, 8:423–438, 2020. 2
- [18] Jae-Myung Kim, A. Sophia Koepke, Cordelia Schmid, and Zeynep Akata. Exposing and mitigating spurious correlations for cross-modal retrieval. In *CVPR Workshops*, pages 2585–2595. IEEE, 2023. 2
- [19] Jonathan Krause, Michael Stark, Jia Deng, and Li Fei-Fei. 3d object representations for fine-grained categorization. In *ICCV Workshops*, pages 554–561. IEEE Computer Society, 2013. 5
- [20] Brian Lester, Rami Al-Rfou, and Noah Constant. The power of scale for parameter-efficient prompt tuning. In *EMNLP (1)*, pages 3045–3059. Association for Computational Linguistics, 2021. 1, 2, 3
- [21] Brian Lester, Joshua Yurtsever, Siamak Shakeri, and Noah Constant. Reducing retraining by recycling parameter-efficient prompts. *CoRR*, abs/2208.05577, 2022. 1, 2
- [22] Xiang Lisa Li and Percy Liang. Prefix-tuning: Optimizing continuous prompts for generation. In *ACL/IJCNLP (1)*, pages 4582–4597. Association for Computational Linguistics, 2021. 1, 2, 3
- [23] Pengfei Liu, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hiroaki Hayashi, and Graham Neubig. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *ACM Comput. Surv.*, 55(9):195:1–195:35, 2023. 2
- [24] Xiao Liu, Kaixuan Ji, Yicheng Fu, Weng Tam, Zhengxiao Du, Zhilin Yang, and Jie Tang. P-tuning: Prompt tuning can

- be comparable to fine-tuning across scales and tasks. In *ACL* (2), pages 61–68. Association for Computational Linguistics, 2022. [1](#), [2](#)
- [25] Yajing Liu, Yuning Lu, Hao Liu, Yaozu An, Zhuoran Xu, Zhuokun Yao, Baofeng Zhang, Zhiwei Xiong, and Chengguang Gui. Hierarchical prompt learning for multi-task learning. In *CVPR*, pages 10888–10898. IEEE, 2023. [2](#)
- [26] Subhransu Maji, Esa Rahtu, Juho Kannala, Matthew B. Blaschko, and Andrea Vedaldi. Fine-grained visual classification of aircraft. *CoRR*, abs/1306.5151, 2013. [5](#)
- [27] Chengzhi Mao, Revant Teotia, Amrutha Sundar, Sachit Menon, Junfeng Yang, Xin Wang, and Carl Vondrick. Doubly right object recognition: A why prompt for visual rationales. In *CVPR*, pages 2722–2732. IEEE, 2023. [1](#), [2](#), [3](#)
- [28] Sachit Menon and Carl Vondrick. Visual classification via description from large language models. In *ICLR*. OpenReview.net, 2023. [1](#), [2](#), [4](#)
- [29] Maria-Elena Nilsback and Andrew Zisserman. Automated flower classification over a large number of classes. In *ICVGIP*, pages 722–729. IEEE Computer Society, 2008. [5](#)
- [30] Omkar M. Parkhi, Andrea Vedaldi, Andrew Zisserman, and C. V. Jawahar. Cats and dogs. In *CVPR*, pages 3498–3505. IEEE Computer Society, 2012. [5](#)
- [31] Sarah Pratt, Ian Covert, Rosanne Liu, and Ali Farhadi. What does a platypus look like? generating customized prompts for zero-shot image classification. In *ICCV*, pages 15691–15701, 2023. [1](#), [2](#), [4](#)
- [32] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019. [2](#)
- [33] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *ICML*, pages 8748–8763. PMLR, 2021. [1](#), [2](#), [6](#)
- [34] Benjamin Recht, Rebecca Roelofs, Ludwig Schmidt, and Vaishal Shankar. Do imagenet classifiers generalize to imagenet? In *ICML*, pages 5389–5400. PMLR, 2019. [5](#)
- [35] Karsten Roth, Jae Myung Kim, A. Sophia Koepke, Oriol Vinyals, Cordelia Schmid, and Zeynep Akata. Waffling around for performance: Visual classification with random words and broad concepts. In *ICCV*, pages 15746–15757, 2023. [1](#), [2](#), [4](#), [5](#)
- [36] Ramprasaath R. Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *ICCV*, pages 618–626. IEEE Computer Society, 2017. [1](#), [8](#)
- [37] Khurram Soomro, Amir Roshan Zamir, and Mubarak Shah. A dataset of 101 human action classes from videos in the wild. *Center for Research in Computer Vision*, 2(11), 2012. [5](#)
- [38] Tu Vu, Brian Lester, Noah Constant, Rami Al-Rfou’, and Daniel Cer. Spot: Better frozen model adaptation through soft prompt transfer. In *ACL (1)*, pages 5039–5059. Association for Computational Linguistics, 2022. [1](#), [2](#)
- [39] Eric Wallace, Shi Feng, Nikhil Kandpal, Matt Gardner, and Sameer Singh. Universal adversarial triggers for attacking and analyzing NLP. In *EMNLP/IJCNLP (1)*, pages 2153–2162. Association for Computational Linguistics, 2019. [2](#)
- [40] Haohan Wang, Songwei Ge, Zachary C. Lipton, and Eric P. Xing. Learning robust global representations by penalizing local predictive power. In *NeurIPS*, pages 10506–10518, 2019. [4](#), [5](#)
- [41] Jianxiong Xiao, James Hays, Krista A. Ehinger, Aude Oliva, and Antonio Torralba. SUN database: Large-scale scene recognition from abbey to zoo. In *CVPR*, pages 3485–3492. IEEE Computer Society, 2010. [5](#)
- [42] An Yan, Yu Wang, Yiwu Zhong, Chengyu Dong, Zexue He, Yujie Lu, William Yang Wang, Jingbo Shang, and Julian McAuley. Learning concise and descriptive attributes for visual recognition. In *ICCV*, pages 3090–3100, 2023. [1](#), [2](#)
- [43] Yue Yang, Artemis Panagopoulou, Shenghao Zhou, Daniel Jin, Chris Callison-Burch, and Mark Yatskar. Language in a bottle: Language model guided concept bottlenecks for interpretable image classification. In *CVPR*, pages 19187–19197. IEEE, 2023. [2](#)
- [44] Yue Yao, Xinyu Tian, Zheng Tang, Sujit Biswas, Huan Lei, Tom Gedeon, and Liang Zheng. Training with product digital twins for autoretail checkout. *arXiv preprint arXiv:2308.09708*, 2023. [1](#)
- [45] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Learning to prompt for vision-language models. *Int. J. Comput. Vis.*, 130(9):2337–2348, 2022. [2](#), [3](#), [5](#), [6](#), [7](#)
- [46] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Conditional prompt learning for vision-language models. In *CVPR*, pages 16795–16804. IEEE, 2022. [2](#), [5](#), [6](#)
- [47] Beier Zhu, Yulei Niu, Saeil Lee, Minhoe Hur, and Hanwang Zhang. Debaised fine-tuning for vision-language models by prompt regularization. In *AAAI*, pages 3834–3842. AAAI Press, 2023. [4](#)