Gradient-Regulated Meta-Prompt Learning for Generalizable Vision-Language Models

Juncheng Li [1, 2*†] Minghe Gao [1*] Longhui Wei [2, 3] Siliang Tang [1] Wenqiao Zhang [4]
Mengze Li [1] Wei Ji [4] Qi Tian [2] Tat-Seng Chua [4]
Yueting Zhuang [1‡]

[1] Zhejiang University, [2] Huawei Cloud, [3] University of Science and Technology of China [4] National university of Singapore
{junchengli, 22221320, siliang, mengzeli, yzhuang}@zju.edu.cn
{weilonghui1, tian.qi1}@huawei.com, {wenqiao, jiwei, dcscts}@nus.edu.sg

Abstract

Prompt tuning, a recently emerging paradigm, enables the powerful vision-language pre-training models to adapt to downstream tasks in a parameter- and data- efficient way, by learning the "soft prompts" to condition frozen pre-training models. Though effective, it is particularly problematic in the few-shot scenario, where prompt tuning performance is sensitive to the initialization and requires a time-consuming process to find a good initialization, thus restricting the fast adaptation ability of the pre-training models. In addition, prompt tuning could undermine the generalizability of the pre-training models, because the learnable prompt tokens are easy to overfit to the limited training samples. To address these issues, we introduce a novel Gradient-RegulAted Meta-prompt learning (GRAM) framework that jointly meta-learns an efficient soft prompt initialization for better adaptation and a lightweight gradient regulating function for strong cross-domain generalizability in a meta-learning paradigm using only the unlabeled image-text pre-training data. Rather than designing a specific prompt tuning method, our GRAM can be easily incorporated into various prompt tuning methods in a model-agnostic way, and comprehensive experiments show that GRAM brings about consistent improvement for them in several settings (i.e., *few-shot learning, cross-domain generalization, cross-dataset generalization,* etc.*) over 11 datasets. Further, experiments show that GRAM enables the orthogonal methods of textual and visual prompt tuning to work in a mutually-enhanced way, offering better generalizability beyond the uni-modal prompt tuning methods.*

*Equal Contribution.
†Work done when interning at Huawei Cloud.
‡Corresponding Author.

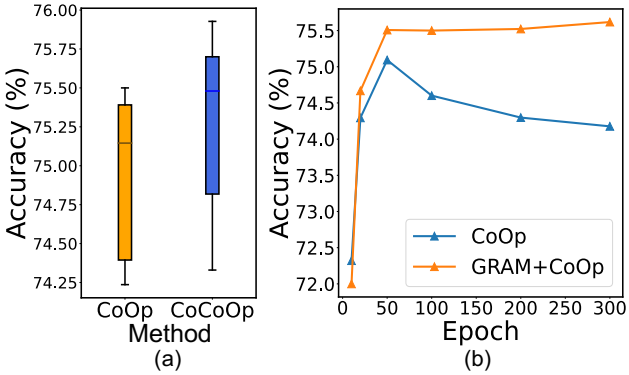

Figure 1: (a) Prompt tuning accuracy varies significantly with different initialization. (b) As the training continues, CoOp's performance drops severely while our GRAM prevents CoOp from overfitting to spurious correlations.

1. Introduction

Pre-trained on vast image-text pairs that cover almost an infinite range of concepts in the real-world, recent vision-language pre-training models [44, 16, 22] have exhibited impressive generalizability on a wide variety of downstream tasks [1, 32, 29, 28, 59]. By simply infilling a hand-crafted prompt template (*e.g.*, "`a photo of a [CLASS]`") with real class names as input to the text encoder, the pre-training models can achieve zero-shot image classification. While effective, a slight word change in prompt templates could lead to a huge difference in performance [62]. Thus, identifying suitable prompts for different tasks requires time-consuming attempts by experts on an extra large validation set. Instead of manually designing hard prompts (discrete language words), some recent prompt tuning methods [62, 61, 63, 17, 20] are proposed to learn a set of soft prompts (continuous embeddings) using a few labeled data.

Despite clear improvements on the downstream tasks,

prompt tuning for few-shot generalization still has two limitations: (1) **Initialization-sensitive issue:** performance is particularly sensitive to the initialization of soft prompts. Figure 1(a) shows that the average few-shot performance varies significantly due to different initialization. Every time we encounter a new task, we need to carefully tune different initialization, which restricts the pre-training models from fast adapting to new tasks. (2) **Generalizability degradation:** since all the prompt tokens are fine-tuned on limited training samples, it can easily overfit to some spurious correlations or in-distribution patterns, damaging the generalizability of the pre-training models. As shown in Figure 1(b), CoOp achieves the best results at the early stage. However, as the training continues, its generalizability decreases significantly.

In this paper, we propose a novel Gradient-RegulAted Meta-prompt learning (GRAM) framework to jointly meta-learn **an efficient soft prompt initialization** that learns to better adapt to new prompting tasks and **a lightweight gradient regulating function** that learns to transform the raw fine-tuning gradient into a consistent direction across domains to prevent prompt tuning from damaging the generalizability of the pre-training models.

Meta-learning [7], also known as learning to learn, optimizes the ability to quickly learn new tasks with only a few samples by transferring the knowledge from learning across a set of meta-training tasks. Typical meta-learning algorithms usually assume access to a distribution of well-annotated meta-training tasks. Differently, we resort to large-scale image-text pairs on the Internet, which is easily available and contains a broader set of visual concepts.

Specifically, we first design a Cross-Modal Hierarchical Clustering algorithm to organize the large-scale image-text data into a hierarchical structure, where the image-text data is first grouped into different semantic topics according to the text descriptions, and each topic of data is further grouped into multiple domains according to the image contents. Then, a diverse set of meta-training classification tasks can be derived by subsampling from the set of semantic topics. For each meta-training task, we simulate domain shift between support set and query set by sampling examples from different domains. The meta-optimization objective is then defined as: after fine-tuning the prompt initialization by one or a few steps using the regulated gradient over a few support set samples, the newly prompted pre-training model should directly perform well on the query set domain. The soft prompt initialization and the gradient regulating function are jointly updated according to the meta gradient directions over the query set samples, thus explicitly learning to better adapt to the new tasks and to avoid overfitting to specific in-domain biases.

Moreover, we provide analysis to show that the proposed gradient regulating function is learned to regulate the gradient into a consistent direction across domains, thus avoiding overfitting to some spurious correlations of a single domain. Note that, our method is model-agnostic. Comprehensive experiments show that GRAM is generalizable to different prompt tuning methods, significantly boosting all models' performance and generalizability. Further, GRAM enables the harmonious and efficient integration of two orthogonal methods - textual prompt tuning (*i.e.*, CoOp) and visual prompt tuning (*i.e.*, VPT). By jointly meta-learning an efficient initialization for both textual and visual prompts, GRAM ensures that both the textual and visual prompts are optimized for better adaptation to new tasks in a complementary way. The resulting UNIversal Gradient-RegulAted Meta-prompt (UNIGRAM) leverages this seamless integration to unlock the greater potential of both methods and achieves superior few-shot generalization performance. Our contributions are mainly three-fold:

- We propose an innovative Gradient-RegulAted Meta-prompt learning (GRAM) framework that explicitly optimizes the adaptation capability to new prompting tasks and the generalization capability to novel domains in a bi-level meta-learning paradigm using only unlabeled image-text pre-training data.
- GRAM can be easily incorporated into different prompt tuning methods in a plug-and-play fashion, and the extensive experiments over 11 datasets illustrate the superior generalizability of our GRAM on base-to-new, cross-domain, and cross-dataset generalization.
- In addition, GRAM enables the orthogonal methods of textual and visual prompt tuning to work in a mutually-enhanced manner, offering stronger generalizability.

2. Related Work

Prompt Tuning. Prompt tuning is first introduced in the NLP area [45] to close the gap between pre-training and downstream tasks. Petroni *et al.* [43] manually create cloze-style prompts to elicit knowledge from pre-trained language models in a "fill-in-the-blank" way. Further, prompt tuning is introduced in vision-language understanding [62], which can enhance the generalizability of large vision-language models [44, 16, 23] on a wide range of vision-language understanding tasks [60, 15, 27, 24, 25, 57, 14, 58, 30]. As manually designing suitable prompts for different tasks is time-consuming and usually sub-optimal, recent works [62, 17, 20] propose to optimize a set of continuous learnable prompt embeddings. Concretely, CoOp [62] optimizes continuous prompt embeddings to improve the few-shot generalizability of CLIP. CoCoOp [61] proposes to learn image-conditioned prompts to further improve the generalizability of CoOp. ProDA [33] learns a distribution of diverse prompts via Gaussian distribution to handle the varying visual representations. To further enhance

CLIP's adaption capability, Tip-Adapter builds a key-value cache model from the few-shot training samples to perform feature retrieval. Instead of designing a specific prompt tuning method, we propose a model-agnostic meta-prompt learning framework to improve the adaptation ability and cross-domain generalizability of the prompt tuning methods, which can be incorporated into existing methods in a plug-and-play fashion.

Meta-Learning. Meta-learning aims to enable efficient adaptation ability of models by leveraging the experience from learning across a set of tasks. Meta-learning approaches can be categorized as: *metric-based* [51, 53, 54], *memory-based* [37, 38, 41, 49], and *optimization-based* [39, 7, 13, 46, 8, 26]. Our framework is based on the optimization-based method (*i.e.*, MAML [7]). Rather than relying on human-annotated meta-training tasks, our method can automatically generate a diverse set of meta-training tasks by cross-modal hierarchical clustering. Li *et al.* [21] propose to synthesize domain shift during meta-training to learn a domain-generalizable initialization. Differently, we present a novel gradient regulating function that actively transforms the updated gradient into a domain-generalizable direction.

3. Method

In this section, we first introduce the preliminaries in Section 3.1. Next, we present the Cross-Modal Hierarchical Clustering to automatically construct a diverse set of meta-training tasks in Section 3.2. Then, in Section 3.3, we elaborate on how GRAM jointly meta-learns an efficient soft prompt initialization and a lightweight gradient regulating function from these meta-training tasks. Finally, we provide theoretical analysis to better understand how our GRAM can improve generalizability in Section 3.4.

3.1. Preliminaries

Contrastive Language-Image Pre-Training. CLIP [44] aims to learn an image encoder f_I and a text encoder f_T by contrastive language-image pre-training paradigm on tremendous image-text pairs, where the matched image-text pairs are optimized to get closer in the joint semantic space. After pre-training, CLIP can generalize to zero-shot visual recognition by reformulating classification as an image-text matching problem. Concretely, the "`[CLASS]`" name can be extended to an input sentence to the text encoder f_T by filling a prompt template like "`a photo of a [CLASS]`". Let $f_T(\mathbf{T}_i)$ denotes the class-extended text feature for the $i-$th class, and then the probability for the $i-$th class is defined as:

$$p(y=i|\mathbf{I}) = \frac{\exp(\text{sim}(f_T(\mathbf{T}_i), f_I(\mathbf{I}))/\tau)}{\sum_{j=1}^{J}\exp(\text{sim}(f_T(\mathbf{T}_j), f_I(\mathbf{I}))/\tau)} \quad (1)$$

where $\text{sim}(\cdot,\cdot)$ denotes cosine similarity, J is the number of classes, and τ is a learned temperature parameter.

Textual Prompt Tuning. To avoid the time-consuming process of identifying customized prompts for different tasks, Context Optimization (CoOp) [62] proposes to learn a set of M continuous prompt vectors $\{\mathbf{t}_1, \mathbf{t}_2, ..., \mathbf{t}_M\}$ to replace the hand-craft prompt template. The prompt sentence for the i-th class is constructed by concatenating the prompt vectors with the word embedding of the class name c_i:

$$\hat{\mathbf{T}}_i = [\mathbf{t}_1, \mathbf{t}_2, ..., \mathbf{t}_M, \mathbf{c}_i] \quad (2)$$

In downstream tasks, since the pre-training model is frozen, the learnable prompt vectors can be efficiently optimized by minimizing the cross-entropy loss using only a few samples.

Visual Prompt Tuning. Recently, similar prompt tuning ideas [2, 17] have been proposed for the vision Transformer encoder, where a set of learnable prompt vectors are concatenated with the input image patch tokens, with the goal of extracting more transferable visual features.

3.2. Cross-Modal Hierarchical Clustering

As large-scale image-text pairs are easily available on the Internet and cover more comprehensive semantic concepts than any existing human-annotated dataset, we present a Cross-Modal Hierarchical Clustering (CHC) algorithm to construct a diverse and structured set of meta-training tasks from the image-text pairs. As shown in Figure 2, CHC organizes the image-text pairs into a hierarchical structure through two steps: **semantic topic clustering** and **visual domain clustering**. The semantic topic clustering first groups image-text pairs into different clusters according to their text descriptions, where each cluster corresponds to a semantic topic. Next, the visual domain clustering further partitions the data in each of the semantic topics into consistent and distinct subsets based on their image features.

Semantic Topic Clustering. To group image-text data based on their underlying semantics, we employ BERTopic [9] to cluster the text descriptions. Specifically, for each image-text pair, we first use Sentence-BERT [48] to encode the text sentence into a dense sentence embedding. To avoid the semantic space collapse problem where the spatial locality becomes ill-defined and distance measures differ little in high dimensional space, we adopt UMAP [36] to reduce the dimensionality of sentence embeddings while preserving the local and global features of high-dimensional data. Then, we cluster the reduced embeddings by the standard clustering algorithm HDBSCAN [35].

After obtaining L clusters of image-text data $\mathcal{P} = \{\mathcal{C}^l\}_{l=1}^{L}$, we extract the semantic topic word for each cluster $\mathcal{C}^l$ according to a cluster-wise TF-IDF [18], which measures the importance of a word to a cluster. Specifically, we treat all text sentences in a cluster as a single document by

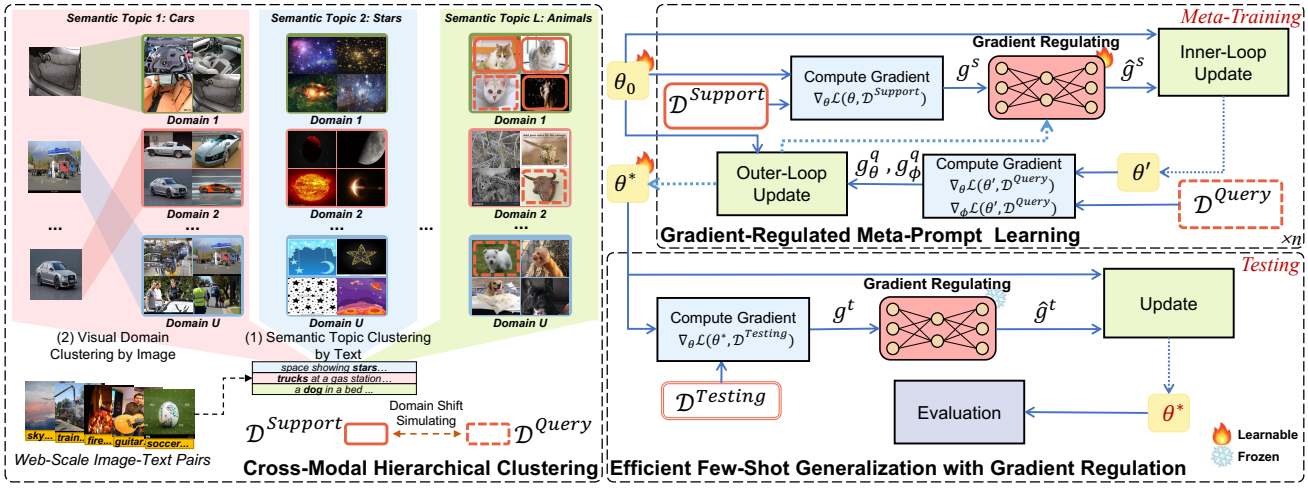

Figure 2: Overview of the proposed GRAM framework.

concatenating the sentences. Then, the cluster-wise TF-IDF score for word w in the cluster $\mathcal{C}^l$ is defined as:

$$\text{TF}-\text{IDF}_{w,l} = \frac{N_{w,l}}{N_l} \cdot \log\left(\frac{L}{L_w + 1}\right) \tag{3}$$

where $N_{w,l}$ is the number of times that word w occurs in cluster $\mathcal{C}^l$, N_l is the total number of words in cluster $\mathcal{C}^l$, L is the number of clusters, and L_w is the number of clusters that contain word w. The first term models the frequency of word w in cluster $\mathcal{C}^l$, and the second term measures how much information word w provides.

Thus, the cluster-wise TF-IDF allows us to extract the most representative word as the semantic topic for each cluster by choosing the word with the highest TF-IDF score.

Visual Domain Clustering. After grouping image-text pairs into multiple semantic topics, we perform visual domain clustering to partition each semantic topic of data into several consistent and distinct domains based on the image features. Specifically, we extract the image features using the pre-trained vision encoder. Next, we run k-means clustering to further group each topic of image-text pairs $\mathcal{C}^l \in \mathcal{P}$ into several domains: $\mathcal{C}^l = \{\mathcal{H}_u^l\}_{u=1}^U$, where $\mathcal{H}_u^l = \{(\mathbf{I}_i, Y^l)\}_{i=1}^{N_u^l}$. Here we omit the paired text of image $\mathbf{I}_i$, and Y^l is the selected semantic topic word for cluster $\mathcal{C}^l$. Thus, we obtain a hierarchical structure of image-text pairs, facilitating to construct diverse meta-training tasks and simulate domain shift during meta-training process.

3.3. Gradient-Regulated Meta-Prompt Learning

As shown in Figure 2, GRAM is a bi-level meta-learning paradigm that jointly meta-learns an efficient soft prompt initialization θ for better adaptation and a lightweight gradient regulating function R^ϕ to prevent prompt tuning from damaging the generalizability of the pre-training models. As GRAM is a model-agnostic method, θ can represent the parameters of any type of prompt tuning method.

Automatic Meta-Training Task Generation. To construct a K_t-way image classification task τ_t, we first sample K_t clusters from $\mathcal{P} = \{\mathcal{C}^l\}_{l=1}^L$. Each sampled cluster corresponds to a category of images with the class label Y^l. Then, we sample a few image instances from the selected clusters to construct the support set $\mathcal{D}_t^{\text{support}}$ and query set $\mathcal{D}_t^{\text{query}}$ for the meta-training task τ_t. Note that, for the support set, we restrict the images to be only sampled from a single domain of each selected cluster. As for the query set, we uniformly sample more images from all domains of each selected cluster. In this way, we **simulate the train/test domain shift** during meta-training. The support set samples are **domain-specific** and the query set samples are **more representative**. Following the above procedure, we construct a diverse set of meta-training tasks $\mathcal{T} = \{\tau_t\}_{t=1}^T$.

Meta-Training Overview. Our bi-level meta-learning paradigm mainly consists of two optimization steps. In the *inner-loop*, the initialization θ is adapted to each meta-training task τ_t according to the gradient regulated by R^ϕ over a few support set samples $\mathcal{D}_t^{\text{support}}$, and then, in the *outer-loop*, a meta-learning objective evaluates the adaptation and generalization capabilities of the adapted model on a distinguished query set $\mathcal{D}_t^{\text{query}}$. The initialization θ and gradient regulating function R^ϕ are jointly optimized according to the performance of the adapted model on the query set across a wide range of meta-training tasks.

Inner-Loop. Formally, we consider adapting θ to a new task τ_t. In the inner-loop, the prompt parameters are first updated via gradient descent over support set $\mathcal{D}_t^{\text{support}}$:

$$\theta_t' \longleftarrow \theta - \alpha \nabla_\theta \mathcal{L}(\theta, \mathcal{D}_t^{\text{support}}) \tag{4}$$

where $\mathcal{L}$ and α denote the loss function and the inner-loop learning rate, respectively.

While straightforward, the above update step on limited samples might overfit to some domain-specific patterns, undermining the generalizability of the pre-training models on

other domains. Thus, we propose to meta-learn an effective and efficient **gradient regulating function**, which can transform the raw gradient into a more consistent direction across domains while ignoring spuriously correlations.

Concretely, gradient regulating function R^ϕ parameterized by ϕ performs affine transformation to modulate the raw gradient for generalizable fine-tuning. Given $\mathbf{g}_t = \nabla_\theta \mathcal{L}(\theta, \mathcal{D}_t^{\text{support}}) \in \mathbb{R}^{d\times M}$ as input, R^ϕ first generates two modulation vectors $\gamma_t \in \mathbb{R}^{d\times M}$ and $\beta_t \in \mathbb{R}^{d\times M}$ as follows:

$$\gamma_t = \tanh(\mathbf{W}^\gamma \mathbf{g}_t + \mathbf{b}^\gamma), \quad \beta_t = \tanh(\mathbf{W}^\beta \mathbf{g}_t + \mathbf{b}^\beta) \quad (5)$$

where $\mathbf{W}^\gamma, \mathbf{b}^\gamma, \mathbf{W}^\beta$ and $\mathbf{b}^\beta$ are learnable parameters. Then, the raw gradient $\mathbf{g}_t$ is regulated as:

$$\hat{\mathbf{g}}_t = \gamma \odot \mathbf{g}_t + \beta_t \quad (6)$$

Consequently, Equation 4 can be transformed as:

$$\theta_t' \longleftarrow \theta - \alpha R^\phi(\nabla_\theta \mathcal{L}(\theta, \mathcal{D}_t^{\text{support}})) \quad (7)$$

Outer-Loop. After adapting the soft prompt parameters to the task τ_t according to the support set $\mathcal{D}_t^{\text{support}}$, the initialization parameters θ and the parameters of the gradient regulating function ϕ are optimized for the performance of the adapted parameters θ' on the query set $\mathcal{D}_t^{\text{query}}$:

$$\theta \longleftarrow \theta - \lambda_1 \nabla_\theta \mathcal{L}(\theta_t', \mathcal{D}_t^{\text{query}}) \quad (8)$$

$$\phi \longleftarrow \phi - \lambda_2 \nabla_\phi \mathcal{L}(\theta_t', \mathcal{D}_t^{\text{query}}) \quad (9)$$

where λ denotes the outer-loop learning rate. Overall, the meta-optimization objective can be formulated as:

$$\min_{\theta,\phi} \sum_{\tau_t \in \mathcal{T}} \mathcal{L}(\theta - \alpha R^\phi(\nabla_\theta \mathcal{L}(\theta, \mathcal{D}_t^{\text{support}})), \mathcal{D}_t^{\text{query}}) \quad (10)$$

The meta-optimization is performed across a wide range of meta-training tasks. Thus, the initialization θ is explicitly optimized to better adapt to new tasks and R^ϕ is optimized to transform the raw gradient so that the model can generalize to the unseen domains of the query sets. The overall methodological flow is summarized in Algorithm 1.

Testing. At test-time, the optimized θ^* is deployed as the soft prompt initialization and adapted to testing tasks using the regulated gradient over few-shot samples as Equation 7.

Universal Gradient-Regulated Meta-Prompt. As a model-agnostic approach, our later experimental results show that GRAM can significantly boost both textual and visual prompt tuning, respectively. Based on this observation, we further present UNIversal Gradient-RegulAted Meta-prompt (UNIGRAM) to explore whether our GRAM enables the visual and textual prompts to cooperate in a complementary way. Without any architecture modification, we meta-learn an effective initialization $\theta = [\theta_T, \theta_V]$ and a corresponding gradient regulating function, where θ_T and θ_V represent the parameters of textual prompt vectors and visual prompt vectors, respectively.

Algorithm 1 Gradient-Regulated Meta-Prompt Learning

```
Randomly initialize θ, φ
while not converged do
    Sample a batch of tasks {τ_t}_{t=1}^B from T
    for all τ_t = {D_t^support, D_t^query} do
        Evaluate ∇_θ L(θ, D_t^support) on D_t^support
        Regulate ∇_θ L(θ, D_t^support) via R^φ
        θ'_t ⟵ θ − αR^φ(∇_θ L(θ, D_t^support))
    end for
    θ ⟵ θ − λ_1 ∇_θ Σ_{τ_t} L(θ'_t, D_t^query)
    φ ⟵ φ − λ_2 ∇_φ Σ_{τ_t} L(θ'_t, D_t^query)
end while
```

3.4. How GRAM Improves Generalizability

In this subsection, we analyze formally how our GRAM can improve generalizability. Let us consider the first order Taylor expansion of the meta-optimization objective at a point $\mathbf{x}_0$ (we omit the subscript t for clarity):

$$\mathcal{L}(\mathbf{x}, \mathcal{D}^{\text{query}}) = \mathcal{L}(\mathbf{x}_0, \mathcal{D}^{\text{query}}) + \nabla_{\mathbf{x}_0} \mathcal{L}(\mathbf{x}_0, \mathcal{D}^{\text{query}}) \cdot (\mathbf{x} - \mathbf{x}_0) \quad (11)$$

Assume we have $\mathbf{x} = \theta - \alpha R^\phi(\nabla_\theta \mathcal{L}(\theta, \mathcal{D}_t^{\text{support}}))$ and $\mathbf{x}_0 = \theta$. Then, Equation 10 can be reformulated as:

$$\begin{aligned} \min_{\theta,\phi} \quad & \mathcal{L}(\mathbf{x}, \mathcal{D}^{\text{query}}) = \min_{\theta,\phi} \quad \mathcal{L}(\theta, \mathcal{D}^{\text{query}}) \\ & - \alpha R^\phi(\nabla_\theta \mathcal{L}(\theta, \mathcal{D}^{\text{support}})) \cdot \nabla_\theta \mathcal{L}(\theta, \mathcal{D}^{\text{query}}) \end{aligned} \quad (12)$$

where the first term represents the loss on the query set and the second term represents the inner product between the regulated gradient over the support set and the gradient over the query set. We are therefore jointly learning to minimize the loss on the query set and maximize the similarity between the gradients. A high similarity means a "similar gradient direction" between the support set domain and the query set domain, which indicates that tuning on the support domain will improve the performance on the query domain.

Recall that we simulate domain shift between the support set and the query set, where the support set samples are domain-specific while the query set samples are more representative across domains. Thus, the gradient over the query set samples represents a more general direction, which is consistent across domains. To improve the gradient alignment, the parameters of the gradient regulating function ϕ are meta-optimized to regulate the raw gradient over the support set into a more generalizable direction, thus avoiding overfitting to some domain-specific correlations.

4. Experiments

In this section, we evaluate our approach on three settings: (1) generalization from base to new classes within a dataset (Section 4.2); (2) cross-domain generalization (Section 4.3); (3) cross-dataset generalization (Section 4.4).

Table 1: Accuracy (%) of base-to-new generalization evaluation over 11 datasets. Prompts are learned from the base classes (16 shots). H: Harmonic mean, which is used to highlight the generalization trade-off. The best results are highlighted in red.

(a) **Average over 11 datasets**.

	Base	New	H
CLIP	69.34	74.22	71.70
CoCoOp	80.47	71.69	75.83
CoOp	77.58	73.11	75.28
+ GRAM	**78.74**	**74.93**	**76.79**
VPT	72.53	72.34	72.43
+ GRAM	**74.04**	**74.21**	**74.12**
UNIGRAM	80.34	75.92	78.07

(b) ImageNet.

	Base	New	H
CLIP	72.43	68.14	70.22
CoCoOp	75.98	70.43	73.10
CoOp	76.21	69.98	72.96
+ GRAM	**76.42**	**70.17**	**73.16**
VPT	74.45	69.22	71.74
+ GRAM	**74.76**	**69.54**	**72.06**
UNIGRAM	76.60	70.69	73.53

(c) Caltech101.

	Base	New	H
CLIP	96.84	94.00	95.40
CoCoOp	97.96	93.81	95.84
CoOp	97.49	94.67	96.06
+ GRAM	**98.07**	**95.00**	**96.51**
VPT	96.92	93.44	95.15
+ GRAM	**97.33**	**94.11**	**95.69**
UNIGRAM	98.07	95.11	96.57

(d) OxfordPets.

	Base	New	H
CLIP	91.17	97.26	94.12
CoCoOp	95.20	97.69	96.43
CoOp	94.48	95.88	95.17
+ GRAM	**94.71**	**97.52**	**96.09**
VPT	92.63	94.96	93.78
+ GRAM	**93.50**	**97.06**	**95.25**
UNIGRAM	94.94	97.94	96.42

(e) StanfordCars.

	Base	New	H
CLIP	63.37	74.89	68.65
CoCoOp	70.49	73.59	72.01
CoOp	71.26	73.92	72.57
+ GRAM	**72.08**	**74.80**	**73.41**
VPT	65.06	74.68	69.54
+ GRAM	**65.65**	**75.10**	**70.06**
UNIGRAM	73.50	75.38	74.43

(f) Flowers102.

	Base	New	H
CLIP	72.08	77.80	74.83
CoCoOp	94.87	71.75	81.71
CoOp	87.87	74.11	80.41
+ GRAM	**92.60**	**75.64**	**83.27**
VPT	76.23	71.55	73.82
+ GRAM	**77.10**	**74.64**	**75.85**
UNIGRAM	95.20	76.21	84.65

(g) Food101.

	Base	New	H
CLIP	90.10	91.22	90.66
CoCoOp	90.70	91.29	90.99
CoOp	90.63	91.17	90.90
+ GRAM	**90.68**	**91.91**	**91.29**
VPT	89.27	90.50	89.88
+ GRAM	**89.86**	**91.32**	**90.58**
UNIGRAM	90.84	92.12	91.48

(h) FGVCAircraft.

	Base	New	H
CLIP	27.19	36.29	31.09
CoCoOp	33.41	23.71	27.74
CoOp	30.66	35.73	33.00
+ GRAM	**31.19**	**36.50**	**33.64**
VPT	28.23	32.21	30.09
+ GRAM	**28.81**	**34.50**	**31.40**
UNIGRAM	32.25	38.00	34.89

(i) SUN397.

	Base	New	H
CLIP	69.36	75.35	72.23
CoCoOp	79.74	76.86	78.27
CoOp	79.78	76.04	77.87
+ GRAM	**80.09**	**76.97**	**78.50**
VPT	75.14	76.89	76.00
+ GRAM	**75.74**	**77.64**	**76.68**
UNIGRAM	80.43	77.91	79.15

(j) DTD.

	Base	New	H
CLIP	53.24	59.90	56.37
CoCoOp	77.01	56.00	64.85
CoOp	69.46	55.67	61.81
+ GRAM	**72.87**	**59.49**	**65.50**
VPT	56.71	57.25	56.98
+ GRAM	**58.25**	**58.00**	**58.12**
UNIGRAM	73.62	62.38	67.56

(k) EuroSAT.

	Base	New	H
CLIP	56.48	64.05	60.03
CoCoOp	87.49	60.04	71.21
CoOp	74.79	61.50	67.50
+ GRAM	**76.00**	**69.92**	**72.83**
VPT	67.57	59.69	63.39
+ GRAM	**77.26**	**68.26**	**72.48**
UNIGRAM	86.26	71.38	78.12

(l) UCF101.

	Base	New	H
CLIP	70.53	77.50	73.85
CoCoOp	82.33	73.45	77.64
CoOp	80.72	75.55	78.05
+ GRAM	**81.47**	**76.33**	**78.82**
VPT	75.65	75.31	75.48
+ GRAM	**76.21**	**76.17**	**76.19**
UNIGRAM	82.00	78.06	79.98

4.1. Experimental Setup

Datasets. For base-to-new generalization and cross-dataset generalization, we use 11 image recognition datasets, which cover a diverse set of recognition tasks: ImageNet [5] and Caltech101 [6] for generic object recognition; OxfordPets [42], StanfordCars [19], Flowers102 [40], Food101 [3] and FGVCAircraft [34] for fine-grained classification; UCF101 [52] for action recognition; SUN397 [56] for scene recognition; DTD [4] for texture classification; and EuroSAT [10] for satellite imagery classification. For cross-domain generalization, we train our model on ImageNet and evaluate the domain generalizability on four variants of ImageNet: ImageNetV2 [47], ImageNetSketch [55], ImageNet-A [12], and ImageNet-R [11].

Baselines. We use the following baselines: (1) Handcrafted prompt method: Zero-Shot CLIP [44]; (2) Textual prompt tuning methods: CoOp [62], CoCoOp [61]; (3) Visual prompt tuning method: VPT [17].

Training Details. For a fair comparison, all methods use CLIP-ViT-B/16 as the pre-training model, and the number of prompt tokens is set to 4, which has been suggested by [61] that a shorter context length can lead to better performance. For our UNIGRAM, we use 2 textual and visual prompt tokens, respectively. In all three settings, we evaluate the 16-shot performance, and all methods follow the same training epochs (*i.e.*, 10 epochs), training schedule,

Table 2: Accuracy (%) of cross-domain generalization evaluation. Prompts are learned from the source dataset (16 shots).

	Source	Target				
	ImageNet	ImageNetV2	ImageNet-Sketch	ImageNet-A	ImageNet-R	Average
CLIP	66.73	60.83	46.15	47.77	73.96	57.18
CoCoOp	71.02	64.07	48.75	50.63	76.18	59.91
CoOp	71.35	64.28	48.67	50.65	76.50	60.03
+ GRAM	**71.62**	**64.66**	**49.06**	**51.12**	**76.76**	**60.40**
VPT	68.92	61.84	47.64	46.50	75.86	57.96
+ GRAM	**69.09**	**62.33**	**47.92**	**47.13**	**76.26**	**58.41**
UNIGRAM	71.65	64.81	49.54	51.51	77.34	60.80

Table 3: Accuracy (%) of cross-dataset generalization evaluation. Prompts are learned from the source dataset (16 shots).

	Source	Target										
	ImageNet	Caltech101	OxfordPets	StanfordCars	Flowers102	Food101	Aircraft	SUN397	DTD	EuroSAT	UCF101	Average
CoCoOp	71.02	94.43	90.14	65.32	71.88	86.06	22.94	67.36	45.73	45.37	68.21	65.74
CoOp	71.35	93.60	89.84	64.74	70.83	85.97	23.03	66.16	44.21	45.95	68.65	65.30
+ GRAM	**71.62**	**94.28**	**90.17**	**65.76**	**71.92**	**86.33**	**23.76**	**66.75**	**45.75**	**51.15**	**69.16**	**66.50**
VPT	68.92	93.07	89.44	64.77	67.79	84.91	23.72	66.16	45.02	37.74	67.00	63.96
+ GRAM	**69.09**	**93.62**	**90.03**	**65.56**	**68.83**	**85.32**	**24.88**	**66.77**	**45.69**	**42.01**	**67.65**	**65.04**
UNIGRAM	71.65	94.67	90.83	66.78	73.12	86.69	25.27	67.97	48.06	52.63	71.03	67.71

and data augmentation settings in CoCoOp. Considering the results of CoOp reported in [61] is obtained by training 200 epochs, which extremely undermines the generalizability of CoOp, we re-train CoOp for 10 epochs using the officially released code and find that fewer training epochs significantly improve the generalizability of CoOp. For CHC, we use CC3M [50], which consists of 3.1 million image-text pairs. Before clustering, we adopt a filtering model to filter out mismatched noisy image-text pairs.

4.2. Generalization From Base to New Classes

Prompt tuning with a few training samples (16 shots) of the base classes, we evaluate the adaptation ability of models on the remaining testing samples of the base classes and the generalization ability of models on the unseen classes. Table 1 summarizes the results. **(1)** Overall, the proposed GRAM method is capable of generalizing to different baseline models, ranging from textual prompt tuning to visual prompt tuning. Our GRAM can not only boost their few-shot adaptation ability on the base classes but also significantly improves their generalizability on the unseen classes. **(2)** As for the average accuracy of the base classes over 11 datasets, our GRAM largely improves CoOp and VPT by 1.16% and 1.51%, respectively, indicating that our unsupervised meta-learning empowers the adaptation ability of existing methods. **(3)** As for the average accuracy of the new classes over 11 datasets, our GRAM brings about 1.82% and 1.87% improvements on CoOp and VPT, respectively, which demonstrates that the proposed gradient regulating function can effectively mitigate the overfitting problem. **(4)** Moreover, our GRAM enables the visual and textual prompt tuning to work in a mutually-enhanced way. Our UNIGRAM achieves stronger few-shot generalization performance beyond its uni-modal components, improving the average accuracy of unseen classes from 73.11% (CoOp) to 75.92%. Note that, UNIGRAM largely outperforms CoCoOp by 14.29%, 6.38%, and 11.34% on FGVCAircraft, DTD, and EuroSAT datasets, respectively.

4.3. Cross-Domain Generalization

Contrastively pre-trained vision-language models have demonstrated strong generalizability, but prompt tuning on limited data from a specific dataset might undermine the generalizability of the pre-training models. In this section, we evaluate the out-of-distribution generalization performance of prompt tuning methods. Following [61], we evaluate the cross-domain generalization performance by transferring the prompts learned from ImageNet to four other variants of ImageNet with domains shift.

As shown in Table 2, our GRAM consistently improves the domain generalizability of CoOp and VPT on all target datasets while at the same time maintaining a higher performance on the source dataset. This indicates that meta-learning to regulate the gradient can effectively prevent the

Table 4: Accuracy (%) of cross-domain generalization evaluation. Prompts are learned from the source dataset (4 shots).

	Source	Target				
	ImageNet	ImageNetV2	ImageNet-Sketch	ImageNet-A	ImageNet-R	Average
CLIP	66.73	60.83	46.15	47.77	73.96	57.18
CoCoOp	70.13	63.05	46.48	49.36	73.80	58.17
CoOp	69.86	62.83	46.90	48.98	74.55	58.32
+ GRAM	**70.49**	**63.72**	**48.42**	**51.13**	**76.39**	**59.92**
VPT	70.11	62.66	46.57	47.99	74.26	57.87
+ GRAM	**70.46**	**63.93**	**48.32**	**49.93**	**76.37**	**59.64**
UNIGRAM	70.84	64.01	48.29	51.20	76.76	60.07

Table 5: Accuracy (%) of cross-dataset generalization evaluation. Prompts are learned from the source dataset (4 shots).

	Source	Target										
	ImageNet	Caltech101	OxfordPets	StanfordCars	Flowers102	Food101	Aircraft	SUN397	DTD	EuroSAT	UCF101	Average
CoCoOp	70.13	93.33	88.76	64.49	69.00	85.48	19.09	64.03	42.58	45.61	66.43	63.88
CoOp	69.86	93.70	89.14	64.51	68.71	85.30	18.47	64.15	41.92	45.39	66.55	63.78
+ GRAM	**70.49**	**94.98**	**90.94**	**64.80**	**69.08**	**85.62**	**19.60**	**64.17**	**41.33**	**46.56**	**66.66**	**64.37**
VPT	70.11	93.30	87.76	62.64	67.91	83.64	20.62	64.51	41.21	40.85	63.37	62.58
+ GRAM	**70.46**	**93.94**	**88.04**	**63.42**	**67.86**	**83.81**	**21.27**	**64.47**	**41.67**	**41.51**	**63.54**	**62.95**
UNIGRAM	70.84	93.69	91.37	64.84	69.54	85.99	21.56	64.69	42.60	46.61	66.68	64.76

models from overfitting to some spurious correlations of a single domain. In addition, GRAM brings about clear improvement by harmonically combining visual and textual prompts. On the ImageNet-R, UNIGRAM significantly surpasses CoCoOp by 1.16%.

4.4. Cross-Dataset Generalization

We further consider a more challenging setting, that is, generalizing across different datasets. The models are only prompt tuned on the source dataset and required to transfer to other 10 datasets in a zero-shot manner. As illustrated in Table 3, equipped with our GRAM, the average transfer performance of CoOp and VPT increases 1.20% and 1.08%, respectively. This validates that GRAM can also improve the cross-dataset generalizability of different methods. Further, our UNIGRAM not only achieves the highest performance on the source dataset but also demonstrates stronger cross-dataset generalization performance over existing methods, outperforming CoCoOp by 1.97 points.

4.5. Extremely Few-Shot Generalization

We further consider extremely few-shot scenarios to better evaluate the adaptation and generalization abilities of our approach. We measure the 4-shot performance instead of the 16-shot performance, where we keep the same training details and evaluation metrics as the 16-shot setting.

Table 6: Ablation results (%) over 11 datasets.

		Base	New	H
0	**UNIGRAM**	80.34	75.92	78.07
1	-domain shift simulating	79.68	75.56	77.57
2	-gradient regulating	77.90	74.79	76.31
3	-meta-learning = prompt pre-training	75.90	74.32	75.10
4	-CHC = supervised meta-learning	78.68	74.47	76.52
5	**joint textual&visual prompt tuning**	77.41	72.85	75.06

We report the cross-domain and cross-dataset generalization performance in Table 4 and Table 5. When the training samples are extremely limited, we find that our GRAM also demonstrates a strong ability to boost the cross-domain and cross-dataset generalizability of CoOp and VPT while maintaining a higher performance on the source dataset. Furthermore, our UNIGRAM exhibits superior generalizability over existing methods by harmonically combining the visual and textual prompt tuning.

4.6. In-Depth Analysis

Effectiveness of Individual Components. In Table 6, we train the following ablation models: (1) w/o domain shift simulating: support set samples are uniformly sampled from all domains, without simulating domain shift. (2) w/o gradient regulating: we remove the gradient regulating function and update the model using raw gradient. (3) w/o meta-learning: we remove the bi-level meta-

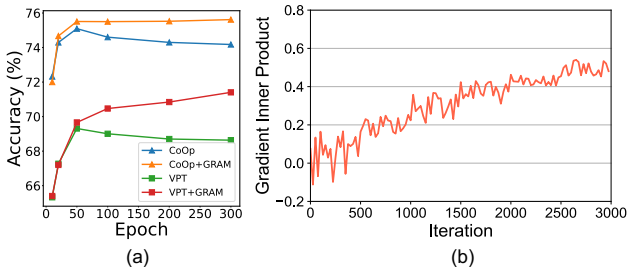

Figure 3: (a) Test accuracy during training. (b) Normalized gradient inner product during training.

learning paradigm and directly use image classification over clustering data as a "pre-training" task to learn a better soft prompt initialization. (4) w/o Cross-Modal Hierarchical Clustering (CHC): instead of using unlabeled data, we directly use an annotated image classification dataset (*i.e.*, WebVision [31]) for meta-training. We construct meta-training tasks by sub-sampling from the set of classes. (5) joint textual&visual prompt tuning: we further consider a straightforward approach that tunes the visual and textual prompts jointly.

The results of Row 1 indicate that using our proposed cross-modal hierarchical clustering to simulate domain shift is crucial for our gradient regulating function to learn to avoid overfitting. Without domain shift simulating, the generalization performance on the new classes is degraded. Also, Row 2 validates the superiority of the proposed gradient regulating function on preventing the model from being misled by some domain-specific correlations. Our gradient regulating function takes up 15% of the relative gain on accuracy (new classes). Further, the results of Row 3 show that the main performance gain does not directly come from the pre-training image-text pairs. Instead, our GRAM provides a novel way to utilize the unlabeled data to address the limitations of prompt tuning. Then, according to Row 4, we notice that the large-scale unlabeled data is a better choice for meta-learning, which covers a wider range of semantics and domains than existing supervised datasets. Finally, from the results of Row 5, we observe that direct joint tuning of the textual and visual prompts performs slightly worse than CoOp. In contrast, Row 0 demonstrates that our approach enables the visual and textual prompt tuning to work in a mutually-enhanced way.

Analysis on Adaptation and Generalization. We report the averaged few-shot performance per epoch. As shown in Figure 3(a), CoOp+GRAM and VPT+GRAM can better adapt to testing datasets, demonstrating the stronger adaptation ability brought by GRAM. Besides, as the training continues, the performance of CoOp and VPT on the new classes drops seriously. In contrast, the proposed gradient regulating function effectively prevents CoOp+GRAM and VPT+GRAM from overfitting to training data.

Table 7: Ablation results (%) with respect to different prompt token numbers over 11 datasets.

Prompt Token Number	Base	New	H
2	77.49	76.09	76.78
4	80.34	75.92	78.07
6	80.09	75.63	77.80
8	80.31	74.79	77.45

Visualization of Gradient Regulating. To verify whether our gradient regulating function can regulate the gradient conflict between support set and query set, we report the normalized gradient inner product between support set and query set during training. As shown in Figure 3(b), we clearly observe that the normalized gradient inner product gradually increases during training. This indicates that our gradient regulating function is learned to regulate the gradient over the support set into a more generalizable direction.

Analysis on the Number of Prompt Tokens. We explore the impact of different numbers of the learnable prompt tokens by varing the number of prompt tokens from 2 to 8. We report the average accuracy over 11 datasets in Table 7. By increasing the number of prompt tokens from 2 to 4, the performance on the base classes is clearly improved while the performance on the new classes drops slightly. Then, continuing increasing the token numbers will damage the generalization performance on the new classes.

5. Conclusions

In this paper, we point out the initialization-sensitive issue and the generalizability degradation issue of current prompt tuning methods for few-shot generalization. We introduce a model-agnostic meta-prompting method GRAM, which jointly learns an efficient soft prompt initialization for better adaptation and a lightweight gradient regulating function for strong cross-domain generalizability using only unlabeled image-text pairs. Extensive experiments on several settings (*e.g.*, cross-domain generalization, cross-dataset generalization) over 11 datasets demonstrate that GRAM can boost existing methods in a plug-and-play fashion. Further experiments show that our GRAM enables both visual and textual prompts to work in a complementary way, exhibiting stronger few-shot generalization ability.

Acknowledgment

This work has been supported in part by the National Key R&D Program of China (2022ZD0160101), Zhejiang NSF (LR21F020004), the NSFC (No. 62272411), Alibaba-Zhejiang University Joint Research Institute of Frontier Technologies, and Ant Group. We thank all the reviewers for their valuable suggestions.

References

[1] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. Vqa: Visual question answering. In *Proceedings of the IEEE international conference on computer vision*, pages 2425–2433, 2015. 1

[2] Hyojin Bahng, Ali Jahanian, Swami Sankaranarayanan, and Phillip Isola. Visual prompting: Modifying pixel space to adapt pre-trained models. *arXiv preprint arXiv:2203.17274*, 2022. 3

[3] Lukas Bossard, Matthieu Guillaumin, and Luc Van Gool. Food-101–mining discriminative components with random forests. In *European conference on computer vision*, pages 446–461. Springer, 2014. 6

[4] Mircea Cimpoi, Subhransu Maji, Iasonas Kokkinos, Sammy Mohamed, and Andrea Vedaldi. Describing textures in the wild. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3606–3613, 2014. 6

[5] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009. 6

[6] Li Fei-Fei, Rob Fergus, and Pietro Perona. Learning generative visual models from few training examples: An incremental bayesian approach tested on 101 object categories. In *2004 conference on computer vision and pattern recognition workshop*, pages 178–178. IEEE, 2004. 6

[7] Chelsea Finn, Pieter Abbeel, and Sergey Levine. Model-agnostic meta-learning for fast adaptation of deep networks. In *International conference on machine learning*, pages 1126–1135. PMLR, 2017. 2, 3

[8] Kevin Frans, Jonathan Ho, Xi Chen, Pieter Abbeel, and John Schulman. Meta learning shared hierarchies. *arXiv preprint arXiv:1710.09767*, 2017. 3

[9] Maarten Grootendorst. Bertopic: Neural topic modeling with a class-based tf-idf procedure. *arXiv preprint arXiv:2203.05794*, 2022. 3

[10] Patrick Helber, Benjamin Bischke, Andreas Dengel, and Damian Borth. Eurosat: A novel dataset and deep learning benchmark for land use and land cover classification. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 12(7):2217–2226, 2019. 6

[11] Dan Hendrycks, Steven Basart, Norman Mu, Saurav Kadavath, Frank Wang, Evan Dorundo, Rahul Desai, Tyler Zhu, Samyak Parajuli, Mike Guo, et al. The many faces of robustness: A critical analysis of out-of-distribution generalization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8340–8349, 2021. 6

[12] Dan Hendrycks, Kevin Zhao, Steven Basart, Jacob Steinhardt, and Dawn Song. Natural adversarial examples. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15262–15271, 2021. 6

[13] Sepp Hochreiter, A Steven Younger, and Peter R Conwell. Learning to learn using gradient descent. In *International Conference on Artificial Neural Networks*, pages 87–94. Springer, 2001. 3

[14] Wei Ji, Renjie Liang, Lizi Liao, Hao Fei, and Fuli Feng. Partial annotation-based video moment retrieval via iterative learning. In *Proceedings of the 31th ACM international conference on Multimedia*, 2023. 2

[15] Wei Ji, Renjie Liang, Zhedong Zheng, Wenqiao Zhang, Shengyu Zhang, Juncheng Li, Mengze Li, and Tat-seng Chua. Are binary annotations sufficient? video moment retrieval via hierarchical uncertainty-based active learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 23013–23022, 2023. 2

[16] Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In *International Conference on Machine Learning*, pages 4904–4916. PMLR, 2021. 1, 2

[17] Menglin Jia, Luming Tang, Bor-Chun Chen, Claire Cardie, Serge Belongie, Bharath Hariharan, and Ser-Nam Lim. Visual prompt tuning. *arXiv preprint arXiv:2203.12119*, 2022. 1, 2, 3, 6

[18] Thorsten Joachims. A probabilistic analysis of the rocchio algorithm with tfidf for text categorization. Technical report, Carnegie-mellon univ pittsburgh pa dept of computer science, 1996. 3

[19] Jonathan Krause, Michael Stark, Jia Deng, and Li Fei-Fei. 3d object representations for fine-grained categorization. In *Proceedings of the IEEE international conference on computer vision workshops*, pages 554–561, 2013. 6

[20] Brian Lester, Rami Al-Rfou, and Noah Constant. The power of scale for parameter-efficient prompt tuning. *arXiv preprint arXiv:2104.08691*, 2021. 1, 2

[21] Da Li, Yongxin Yang, Yi-Zhe Song, and Timothy Hospedales. Learning to generalize: Meta-learning for domain generalization. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32, 2018. 3

[22] Juncheng Li, Xin He, Longhui Wei, Long Qian, Linchao Zhu, Lingxi Xie, Yueting Zhuang, Qi Tian, and Siliang Tang. Fine-grained semantically aligned vision-language pre-training. *Advances in neural information processing systems*, 35:7290–7303, 2022. 1

[23] Juncheng Li, Kaihang Pan, Zhiqi Ge, Minghe Gao, Hanwang Zhang, Wei Ji, Wenqiao Zhang, Tat-Seng Chua, Siliang Tang, and Yueting Zhuang. Empowering vision-language models to follow interleaved vision-language instructions, 2023. 2

[24] Jiacheng Li, Siliang Tang, Juncheng Li, Jun Xiao, Fei Wu, Shiliang Pu, and Yueting Zhuang. Topic adaptation and prototype encoding for few-shot visual storytelling. In *Proceedings of the 28th ACM International Conference on Multimedia*, pages 4208–4216, 2020. 2

[25] Juncheng Li, Siliang Tang, Linchao Zhu, Haochen Shi, Xuanwen Huang, Fei Wu, Yi Yang, and Yueting Zhuang. Adaptive hierarchical graph reasoning with semantic coherence for video-and-language inference. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1867–1877, 2021. 2

[26] Juncheng Li, Xin Wang, Siliang Tang, Haizhou Shi, Fei Wu, Yueting Zhuang, and William Yang Wang. Unsupervised re-

inforcement learning of transferable meta-skills for embodied navigation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12123–12132, 2020. 3

[27] Juncheng Li, Junlin Xie, Long Qian, Linchao Zhu, Siliang Tang, Fei Wu, Yi Yang, Yueting Zhuang, and Xin Eric Wang. Compositional temporal grounding with structured variational cross-graph correspondence learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3032–3041, 2022. 2

[28] Mengze Li, Han Wang, Wenqiao Zhang, Jiaxu Miao, Zhou Zhao, Shengyu Zhang, Wei Ji, and Fei Wu. Winner: Weakly-supervised hierarchical decomposition and alignment for spatio-temporal video grounding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 23090–23099, 2023. 1

[29] Mengze Li, Tianbao Wang, Haoyu Zhang, Shengyu Zhang, Zhou Zhao, Jiaxu Miao, Wenqiao Zhang, Wenming Tan, Jin Wang, Peng Wang, et al. End-to-end modeling via information tree for one-shot natural language spatial video grounding. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8707–8717, 2022. 1

[30] Mengze Li, Tianbao Wang, Haoyu Zhang, Shengyu Zhang, Zhou Zhao, Wenqiao Zhang, Jiaxu Miao, Shiliang Pu, and Fei Wu. Hero: Hierarchical spatio-temporal reasoning with contrastive action correspondence for end-to-end video object grounding. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 3801–3810, 2022. 2

[31] Wen Li, Limin Wang, Wei Li, Eirikur Agustsson, and Luc Van Gool. Webvision database: Visual learning and understanding from web data. *arXiv preprint arXiv:1708.02862*, 2017. 9

[32] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer, 2014. 1

[33] Yuning Lu, Jianzhuang Liu, Yonggang Zhang, Yajing Liu, and Xinmei Tian. Prompt distribution learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5206–5215, 2022. 2

[34] Subhransu Maji, Esa Rahtu, Juho Kannala, Matthew Blaschko, and Andrea Vedaldi. Fine-grained visual classification of aircraft. *arXiv preprint arXiv:1306.5151*, 2013. 6

[35] Leland McInnes, John Healy, and Steve Astels. hdbscan: Hierarchical density based clustering. *J. Open Source Softw.*, 2(11):205, 2017. 3

[36] Leland McInnes, John Healy, and James Melville. Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426*, 2018. 3

[37] Nikhil Mishra, Mostafa Rohaninejad, Xi Chen, and Pieter Abbeel. A simple neural attentive meta-learner. *arXiv preprint arXiv:1707.03141*, 2017. 3

[38] Tsendsuren Munkhdalai and Hong Yu. Meta networks. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 2554–2563. JMLR. org, 2017. 3

[39] Alex Nichol, Joshua Achiam, and John Schulman. On first-order meta-learning algorithms. *arXiv preprint arXiv:1803.02999*, 2018. 3

[40] Maria-Elena Nilsback and Andrew Zisserman. Automated flower classification over a large number of classes. In *2008 Sixth Indian Conference on Computer Vision, Graphics & Image Processing*, pages 722–729. IEEE, 2008. 6

[41] Boris Oreshkin, Pau Rodríguez López, and Alexandre Lacoste. Tadam: Task dependent adaptive metric for improved few-shot learning. In *Advances in Neural Information Processing Systems*, pages 721–731, 2018. 3

[42] Omkar M Parkhi, Andrea Vedaldi, Andrew Zisserman, and CV Jawahar. Cats and dogs. In *2012 IEEE conference on computer vision and pattern recognition*, pages 3498–3505. IEEE, 2012. 6

[43] Fabio Petroni, Tim Rocktäschel, Patrick Lewis, Anton Bakhtin, Yuxiang Wu, Alexander H Miller, and Sebastian Riedel. Language models as knowledge bases? *arXiv preprint arXiv:1909.01066*, 2019. 2

[44] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, pages 8748–8763. PMLR, 2021. 1, 2, 3, 6

[45] 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

[46] Sachin Ravi and Hugo Larochelle. Optimization as a model for few-shot learning. 2016. 3

[47] Benjamin Recht, Rebecca Roelofs, Ludwig Schmidt, and Vaishaal Shankar. Do imagenet classifiers generalize to imagenet? In *International Conference on Machine Learning*, pages 5389–5400. PMLR, 2019. 6

[48] Nils Reimers and Iryna Gurevych. Sentence-bert: Sentence embeddings using siamese bert-networks. *arXiv preprint arXiv:1908.10084*, 2019. 3

[49] Adam Santoro, Sergey Bartunov, Matthew Botvinick, Daan Wierstra, and Timothy Lillicrap. Meta-learning with memory-augmented neural networks. In *International conference on machine learning*, pages 1842–1850, 2016. 3

[50] Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2556–2565, 2018. 7

[51] Jake Snell, Kevin Swersky, and Richard Zemel. Prototypical networks for few-shot learning. In *Advances in Neural Information Processing Systems*, pages 4077–4087, 2017. 3

[52] Khurram Soomro, Amir Roshan Zamir, and Mubarak Shah. Ucf101: A dataset of 101 human actions classes from videos in the wild. *arXiv preprint arXiv:1212.0402*, 2012. 6

[53] Flood Sung, Yongxin Yang, Li Zhang, Tao Xiang, Philip HS Torr, and Timothy M Hospedales. Learning to compare: Relation network for few-shot learning. In *Proceedings of the*

IEEE Conference on Computer Vision and Pattern Recognition, pages 1199–1208, 2018. 3

[54] Oriol Vinyals, Charles Blundell, Timothy Lillicrap, Daan Wierstra, et al. Matching networks for one shot learning. In *Advances in neural information processing systems*, pages 3630–3638, 2016. 3

[55] Haohan Wang, Songwei Ge, Zachary Lipton, and Eric P Xing. Learning robust global representations by penalizing local predictive power. *Advances in Neural Information Processing Systems*, 32, 2019. 6

[56] Jianxiong Xiao, James Hays, Krista A Ehinger, Aude Oliva, and Antonio Torralba. Sun database: Large-scale scene recognition from abbey to zoo. In *2010 IEEE computer society conference on computer vision and pattern recognition*, pages 3485–3492. IEEE, 2010. 6

[57] Qifan Yu, Juncheng Li, Yu Wu, Siliang Tang, Wei Ji, and Yueting Zhuang. Visually-prompted language model for fine-grained scene graph generation in an open world. *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023. 2

[58] Wenqiao Zhang, Jiannan Guo, Mengze Li, Haochen Shi, Shengyu Zhang, Juncheng Li, Siliang Tang, and Yueting Zhuang. Boss: Bottom-up cross-modal semantic composition with hybrid counterfactual training for robust content-based image retrieval. *arXiv preprint arXiv:2207.04211*, 2022. 2

[59] Wenqiao Zhang, Haochen Shi, Jiannan Guo, Shengyu Zhang, Qingpeng Cai, Juncheng Li, Sihui Luo, and Yueting Zhuang. Magic: Multimodal relational graph adversarial inference for diverse and unpaired text-based image captioning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 3335–3343, 2022. 1

[60] Wenqiao Zhang, Siliang Tang, Yanpeng Cao, Shiliang Pu, Fei Wu, and Yueting Zhuang. Frame augmented alternating attention network for video question answering. *IEEE Transactions on Multimedia*, 22(4):1032–1041, 2019. 2

[61] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Conditional prompt learning for vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16816–16825, 2022. 1, 2, 6, 7

[62] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Learning to prompt for vision-language models. *International Journal of Computer Vision*, 130(9):2337–2348, 2022. 1, 2, 3, 6

[63] Beier Zhu, Yulei Niu, Yucheng Han, Yue Wu, and Hanwang Zhang. Prompt-aligned gradient for prompt tuning. *arXiv preprint arXiv:2205.14865*, 2022. 1