

DiMPLe - Disentangled Multi-Modal Prompt Learning: Enhancing Out-Of-Distribution Alignment with Invariant and Spurious Feature Separation

Umaina Rahman¹, Mohammad Yaqub¹, Dwarikanath Mahapatra²

¹Mohamed Bin Zayed University of Artificial Intelligence, ² Khalifa University

Abstract

We introduce **DiMPLe** (**Disentangled Multi-Modal Prompt Learning**), a novel approach to disentangle invariant and spurious features across vision and language modalities in multi-modal learning. Spurious correlations in visual data often hinder out-of-distribution (OOD) performance. Unlike prior methods focusing solely on image features, DiMPLe **disentangles** features **within and across modalities** while maintaining consistent alignment, enabling better generalization to **novel classes** and robustness to **distribution shifts**. Our method combines three key objectives: (1) mutual information minimization between invariant and spurious features, (2) spurious feature regularization, and (3) contrastive learning on invariant features. Extensive experiments demonstrate DiMPLe demonstrates superior performance compared to CoOp-OOD, when averaged across 11 diverse datasets, and achieves absolute gains of 15.27 in base class accuracy and 44.31 in novel class accuracy. The code is available at github.com/rumaima/DiMPLe.

1. Introduction

Contrastive Language–Image Pretraining (CLIP) [21] revolutionized multi-modal learning by linking textual and visual data [11], [24], [5], [7], [17]. The semantic associations present in the shared image-text embedding space allowed it to make predictions on novel and unseen categories, inspiring further research in the domain of multi-modal understanding. Prompt tuning methods like CoOp [31] refined CLIP’s performance through context optimization of learnable prompt vectors, and demonstrate improved text-visual alignment and model adaptability. By employing different prompt-initialization strategies, CoOp successfully showed how prompt engineering could facilitate model adaptability and enable improved alignment between text and visual representations. However, despite CoOp’s effectiveness, it exhibited limited generalizability to novel classes. This limitation stemmed from spurious correlations within the visual data, which often misled the model’s predictions when applied to out-of-distribution (OOD) samples.

CoOp-OOD [29], an enhancement to the original CoOp, addressed this issue by decoupling invariant and spurious features in the images. By maximizing the similarity between invariant visual features and the corresponding textual representations, it emphasized on robust features, while regulating the predictions based on spurious ones. Additionally, minimizing the conditional mutual information between invariant and spurious image features ensured a clearer separation within the feature space.

Nonetheless, a crucial consideration in this approach remains unaddressed: spurious features within an image should logically correspond to spurious elements within textual representations, while invariant features should be aligned exclusively with invariant text components. Without explicit mapping, the disentanglement process may lead to ambiguous associations between spurious and invariant features, making it difficult for the model to distinguish relevant signals from noise during inference as seen in Fig. 1 (a) and (c). This ambiguity can propagate through the network, potentially amplifying the influence of spurious correlations under domain shifts or when encountering novel categories. By explicitly aligning spurious and invariant features within and across modalities, the model ensures consistent feature associations, thereby enhancing its interpretability and resilience in applications where distribution shifts are inevitable.

In light of these limitations, we propose **DiMPLe** - a **Disentangled Multi-Modal Prompt Learning** a principled approach aimed at separating invariant and spurious features across both vision and language modalities. DiMPLe achieves this by minimizing the conditional mutual information between the invariant and spurious features within each modality. Our contribution can be summarised as:

1. DiMPLe is a coherent approach to feature disentanglement across vision-language modalities avoiding ambiguities across modality mappings (Fig. 1 (b) and (d)).
2. DiMPLe goes beyond spurious correlation mitigation by employing class-conditioned mutual information minimization that maintains separation between invariant and

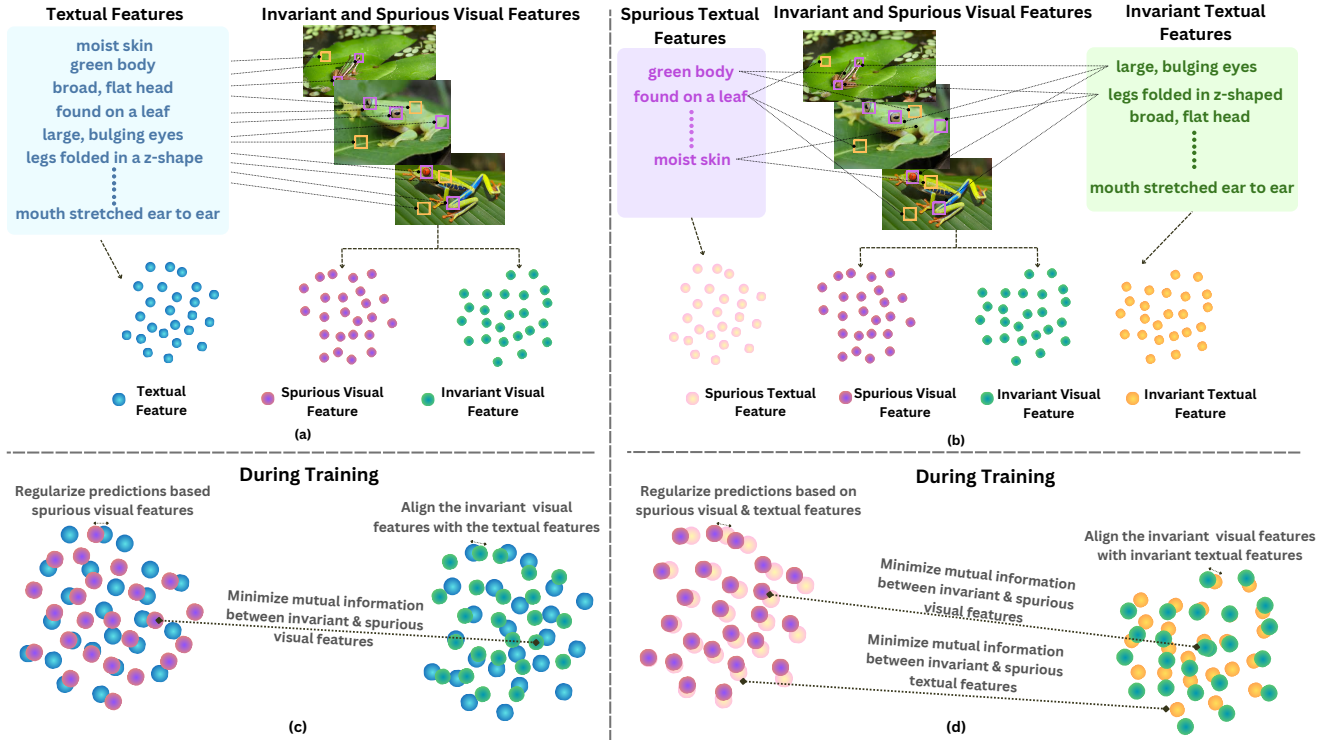


Figure 1. (a) Without the explicit mapping in previous methods, spurious elements in the text embeddings may inadvertently align with the visual model’s invariant features, weakening the model’s robustness and reducing its effectiveness on novel classes and out-of-distribution images. (b) Importance of a principled method (**DiMPLe**) for disentangling multi-modal representations, to enhance generalization for novel classes as well as to improve the robustness of multi-modal models against distribution shifts for the same class by explicitly mapping the invariant and spurious features for each modality. (c) In previous methods, during training, the general textual features are used to align with the invariant image features as well as spurious image features. (d) **DiMPLe** removes the ambiguous mapping by aligning the invariant textual and image features as well as associating spurious textual features with their spurious image counterparts.

spurious components across modalities and ensures effective separation of unwanted correlations and associating spurious textual features with its image counterpart.

3. **DiMPLe** achieves significant improvements in both base and novel class performance. Ensuring that invariant features remain isolated and consistent, it transfers better across diverse datasets against distribution shifts for the same class, compared to existing methods.

2. Related Works

Vision-Language Models (VLMs) like CLIP have revolutionized the paradigm of multi-modal training by leveraging the semantic alignment between visual and textual features. The effectiveness of these models heavily depends on how the text prompts are constructed and optimized. Early approaches [11], [1], [17], relied on hand-crafted prompts, but recent work such as [31], [12], [32], [13], [28] have shown that learnable prompts can significantly improve performance. CoOp [31] introduced a paradigm shift by replacing hand-crafted prompts with learnable context vec-

tors while keeping the class names fixed. This approach demonstrated superior performance compared to manual prompt engineering, particularly in base category classification. However, CoOp’s performance degraded significantly when applied to novel classes, indicating limitations in its generalization capabilities. MaPLe [15] advanced the field by introducing a meta-learning framework for prompt optimization. The key innovation was the separation of prompts into two components: a shared meta-prompt that captures domain-general knowledge, and class-specific prompts that encode category-specific features. This decomposition allowed for better generalization to unseen classes while maintaining strong performance on base categories. MaPLe’s meta-learning strategy enables the model to learn how to adapt prompts quickly for new categories, addressing a key limitation of previous approaches.

Out-of-distribution (OOD) generalization has emerged as a critical challenge for Vision-Language Models, particularly in real-world applications where test distributions may differ significantly from training data. Co-CoOp [30] specifically addressed the OOD challenge by introducing

This decomposition enables us to isolate the stable, class-relevant features from those influenced by spurious correlations, thus achieving unbiased alignment in OOD scenarios.

Conditional Independence Regularization: We enforce a conditional independence constraint between the invariant representations $\mathbf{z}_{i,u}$ and its spurious counterpart $\mathbf{z}_{i,s}$ given the label Y for the vision (v) and text (t) modalities, using a mutual information regularization term $\mathcal{L}_{cmi}$ to minimize the conditional mutual information:

$$\begin{aligned}\mathcal{L}_{cmi.v} &= I(\mathbf{z}_{v,i,u}; \mathbf{z}_{v,i,s} | Y) \text{ for image features,} \\ \mathcal{L}_{cmi.t} &= I(\mathbf{z}_{t,i,u}; \mathbf{z}_{t,i,s} | Y) \text{ for text features,} \\ \mathcal{L}_{cmi} &= \text{avg}(\mathcal{L}_{cmi.v}, \mathcal{L}_{cmi.t})\end{aligned}\quad (3)$$

Practically, we estimate $\mathcal{L}_{cmi}$ based on the HSIC metric proposed by [14] which minimizes the conditional mutual information by promoting the independence of variables. It achieves this by computing the Conditional Hilbert-Schmidt Independence Criterion (HSIC) between two sets of features: invariant (x) and spurious (y) conditioned on the class label (z), capturing distinct and relevant signals.

Deep Multi-modal Prompt Tuning: We incorporate multi-modal deep prompting by introducing learnable prompt tokens in both vision and language branches up to the J -th transformer layer, where $J < K$, with K being the total number of layers in the CLIP model. In the language branch, we introduce b learnable prompt tokens $\{P_i \in \mathbb{R}^{d_l}\}_{i=1}^b$, and the prompts are progressively updated within each transformer layer:

$$[P_i, W_i] = L_i([P_{i-1}, W_{i-1}]) \quad (4)$$

In the vision branch, we add learnable prompts $\{\tilde{P}_i \in \mathbb{R}^{d_v}\}_{i=1}^b$ that are incorporated at each transformer block, updating the feature representations as:

$$[c_i, E_i, \tilde{P}_i] = V_i([c_{i-1}, E_{i-1}, \tilde{P}_{i-1}]) \quad (5)$$

Vision-Language Prompt Coupling via Linear Projection: We incorporate vision-language coupling by conditioning vision prompts $\tilde{P}_i$ on the language prompts P_i through a linear projection F_i that maps text embeddings P_i into the vision space. This coupling is represented as:

$$\tilde{P}_i = F_i(P_i), \quad \tilde{P}_i \in \mathbb{R}^{d_v} \quad (6)$$

This coupling function bridges the vision and language modalities by enabling prompt learning in a shared embedding space, enhancing consistency and complementary representation across both branches.

Disentangled Cross-modal Alignment: We use a two-fold contrastive learning loss to align the invariant features with class-relevant prompts and spurious ones with unrelated prompts in order to avoid spurious interference:

(1) We minimize the cross-entropy loss $\mathcal{L}_{ce}^u$ between the invariant image $\mathbf{z}_{v,i,u}$ and text $\mathbf{z}_{t,i,u}$ features:

$$\mathcal{L}_{ce}^u = - \sum_{i=1}^N y_i \log p_u(y_i | x_i) \quad (7)$$

$$p_u(y_i | x_i) = \frac{\exp(\text{sim}(\mathbf{z}_{v,i,u}, \mathbf{z}_{t,i,u})/\tau)}{\sum_{j=1}^C \exp(\text{sim}(\mathbf{z}_{v,j,u}, \mathbf{z}_{t,j,u})/\tau)}, \quad (8)$$

(2) For the spurious features, we use a uniformity constraint to prevent classification alignment on these features for both modalities, using the Kullback-Leibler (KL) divergence loss $\mathcal{L}_r^{sp}$, which helps prevent shortcut learning:

$$\mathcal{L}_r^{sp} = \sum_{i=1}^N \ell_{KL}(p_s(y_i | x_i) | p_0) \quad (9)$$

The combined objective function integrates all components:

$$\mathcal{L} = \mathcal{L}_{ce}^u + \alpha \mathcal{L}_r^{sp} + \beta \mathcal{L}_{cmi} \quad (10)$$

where α and β balance the effects of spurious, invariant, and conditional independence terms. Here, p_u is the predicted probability distribution over classes using invariant features, p_s is the predicted probability distribution using spurious features and p_0 is the uniform distribution.

Method	Base Acc.	Novel Acc.	HM
Independent Vision-Language Prompting			
1: CoOP-ODD*	25.99	29.98	27.6
2: DiMPLLe*	76.73	72.36	74.48
Conditioned Vision-Language Prompting			
1: CoOP-ODD†	62.72	65.6	64.13
2: DiMPLLe	76.09	73.35	74.70

Table 1. Comparison of DiMPLLe with different prompting designs in base-to-novel generalization. Results are averaged over 11 datasets. IVL-P is independent vision-language prompting, CVL-P is conditioned vision-language prompting where the vision prompts are conditioned on the textual prompts. CoOP-ODD* is CoOP-ODD w/ IVL-P, CoOP-ODD† is CoOP-ODD w/ CVL-P and DiMPLLe* is DiMPLLe using IVL-P. HM is the harmonic mean.

4. Experiments and Results

Datasets: For evaluating base-to-novel class generalization and cross-dataset performance, we assess our model on 11 image classification datasets covering a broad range of tasks. These include two general-object datasets, ImageNet [4] and Caltech101 [6]; five fine-grained datasets, OxfordPets [20], StanfordCars [16], Flowers102 [19], Food101 [2], and FGVCAircraft [18]; a scene recognition dataset, SUN397 [27]; an action recognition dataset, UCF101 [25]; a texture dataset, DTD [3]; and a satellite-image dataset,

EuroSAT [8]. For domain generalization, we use ImageNet as the source dataset and test it on four of its variants, including ImageNetV2 [22], ImageNetSketch [26], ImageNet-A [10], and ImageNet-R [9].

Baselines: In addition to evaluating DiMPLe against existing CoOp-OOD [29] as discussed in the preliminary section, we introduce two more baselines by incorporating vision prompts into the CoOp-OOD pipeline. As discussed in Table 1, we specifically create two variants: (1) CoOp-OOD*, which extends CoOp-OOD with vision prompts that operate independently from the language prompts, and (2) CoOp-OOD†, which incorporates vision prompts conditioned on the learnable textual prompt vectors. Both CoOp-OOD* and CoOp-OOD† are trained using the same loss functions that guide CoOp-OOD, ensuring consistency in optimization objectives across the baselines.

Implementation Details: To implement our approach, we used few-shot learning, selecting 16 random samples per class and fine-tuning a pre-trained ViT-B/16 CLIP model. The model architecture utilized specific dimensions ($d_l = 512$, $d_v = 768$, and $d_{vl} = 512$) and employed prompt tuning with a depth of 9 layers and prompt lengths of 2 for both language and vision components. We use $512 \rightarrow 512$ simple linear projections to map the original text and vision embeddings (both of shape $[\ast, 512]$) into disentangled invariant and spurious subspaces, each also of dimension 512. The training was conducted on a single NVIDIA A100 GPU over 5 epochs, using an SGD optimizer with a batch size of 4 and a learning rate of 0.0035, with results averaged across 3 runs. We initialized the first layer’s prompt vector using CLIP’s pre-trained word embeddings for ‘a photo of a [category]’. Subsequent layers were randomly initialized using a normal distribution. In cross-dataset evaluation scenarios, we trained on all 1000 ImageNet classes using a reduced prompt depth of 3 layers, training for 2 epochs with a learning rate of 0.0026, measuring performance through base and novel class accuracies and their harmonic mean.

4.1. Base-to-Novel Class Generalization

We evaluate the generalizability of DiMPLe in a zero-shot setup by assessing its performance on both base and novel classes, highlighting its ability to handle out-of-distribution (OOD) data. Each dataset is divided into base classes, used for few-shot prompt tuning, and novel classes, which test generalization to unseen categories without retraining. Table 3 presents a detailed comparison of DiMPLe with baselines such as CLIP [21] and CoOp-OOD [29]. We also include CoOp-OOD†, an enhanced version that incorporates conditioned vision prompts to improve adaptation to distribution shifts. Results are reported across 11 datasets as well as their average, using three key metrics: base accuracy, novel accuracy, and harmonic mean (HM), which captures the balance between in-distribution and OOD performance.

DiMPLe consistently outperforms both zero-shot and fine-tuned baselines, particularly on novel classes. For example, it achieves absolute gains of 15.27 on base accuracy, 44.31 on novel accuracy, and 35.39 on HM when averaged across datasets. These improvements are largely attributed to DiMPLe’s disentanglement strategy, which separates invariant features from spurious ones, improving adaptability to unseen domains. Notably, performance gains are most evident on fine-grained datasets such as StanfordCars and FGVC-Aircraft, where spurious cues are more likely to mislead standard models.

4.2. Domain Generalization

We examine the robustness of our method by testing it on out-of-distribution datasets. Like in cross-dataset evaluation, we directly evaluate our ImageNet-trained model on four distinct ImageNet variants- ImageNet-A, ImageNet-R, ImageNet-S, and ImageNetV2. These variants represent different types of domain shifts, testing the model’s ability to generalize beyond the original ImageNet training data. The results of DiMPLe are compared against CLIP and CoOp-OOD. CLIP represents a general-purpose vision-language model baseline, while CoOp-OOD, is an approach specifically designed for handling OOD shifts, which we anticipate will perform differently from models not specifically optimized for OOD. Table 2 showcases that across all target datasets, DiMPLe’s balanced scores suggest that it offers consistent performance across different types of domain shifts without compromising in-distribution accuracy.

	Source	Target			
	ImageNet	ImageNetV2	ImageNet-S	ImageNet-A	ImageNet-R
CLIP	66.73	60.83	46.15	47.77	73.96
CoOp-OOD	38.7	34.53	24.13	25.5	27.77
DiMPLe	69.73	61.2	45.67	44.07	73.87

Table 2. Evaluation under Domain Generalization setting: DiMPLe shows superior performance with respect to CoOp-OOD.

For the in-distribution ImageNet dataset, CLIP achieves a base accuracy of 66.73%, while DiMPLe attains a close score of 69.73%, indicating competitive performance in the source domain. In contrast, CoOp-OOD significantly underperforms with an accuracy of 38.7%, suggesting that CoOp-OOD sacrifices in-distribution performance to handle OOD scenarios. For Out-of-Distribution evaluation, DiMPLe performs similarly to CLIP, while CoOp-OOD struggles significantly across all benchmarks. On ImageNetV2, which introduces minor shifts, DiMPLe (61.2%) slightly improves over CLIP (60.83%), while CoOp-OOD struggles at 34.53%. On ImageNet-S, which assesses robustness to synthetic distortions, CLIP (46.15%) and DiMPLe (45.67%) perform similarly, but CoOp-OOD lags sig-

nificantly at 24.13%. For ImageNet-A, which contains adversarially challenging images, CLIP (47.77%) outperforms DiMPLe (44.07%), while CoOp-OOD remains the weakest (25.5%). Finally, on ImageNet-R, which tests adaptation to visual shifts, DiMPLe (73.87%) remains close to CLIP (73.96%), whereas CoOp-OOD struggles at 27.77%. These results highlight that DiMPLe consistently outperforms CoOp-OOD and remains competitive with CLIP, demonstrating stronger domain generalization.

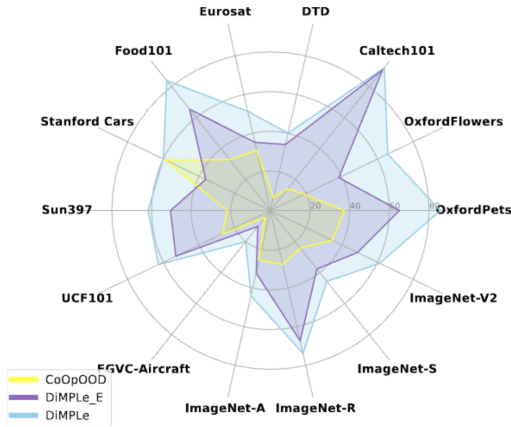


Figure 3. A radar plot showing the results using CoOp-OOD, DiMPLe, DiMPLe-E for Cross-Dataset (11 datasets) and Domain-Generalization (4 ImageNet variants) evaluation. For training, all 1000 classes of ImageNet were used and the respective models were evaluated using the above datasets in a zero-shot setting.

4.3. Cross-Dataset Evaluation

To test the potential for cross-dataset transfer, we apply our ImageNet-trained model to other datasets. Our model undergoes few-shot training on all 1000 ImageNet classes. When comparing CoOp-OOD and DiMPLe across the cross-dataset evaluation, DiMPLe consistently outperforms CoOp-OOD by substantial margins, indicating DiMPLe’s robustness in cross-domain generalization. For example, we can see in Fig. 3 on OxfordPets, DiMPLe achieves 85.67% accuracy, far surpassing CoOp-OOD’s 37.33%. This pattern holds across most datasets, with DiMPLe scoring 65.7% on OxfordFlowers and 92.1% on Caltech101, whereas CoOp-OOD scores only 18.93% and 14.07%, respectively. Notably, DiMPLe also excels on challenging datasets such as ImageNet-A and ImageNet-R, achieving 44.07% and 73.87% accuracy, respectively, compared to CoOp-OOD’s 25.5% and 27.77%.

4.4. Results Discussion

Dataset Size and Sample Diversity Trade-off: The impact of dataset size and sample diversity on performance can be critically analyzed by considering how both factors influence model generalization and adaptability across different domains. DiMPLe’s performance across varied datasets,

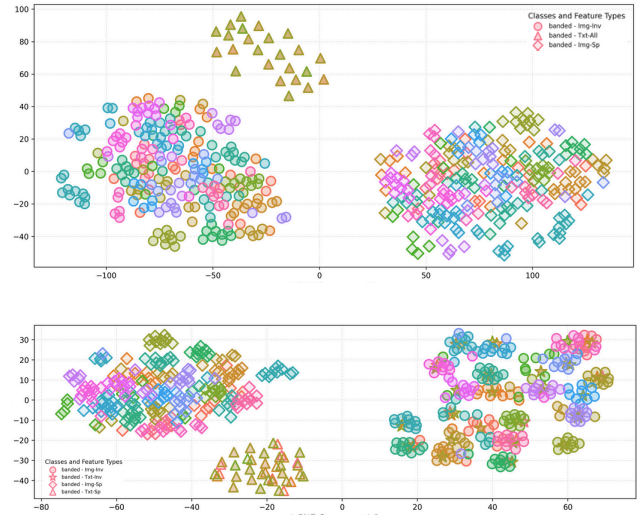


Figure 4. tSNE visualizations showing class discrimination ability in CoOp-OOD (top) compared to DiMPLe (bottom).

ranging from large, well-established datasets like ImageNet to smaller, more diverse datasets like OxfordPets and Stanford Cars, highlights key trends and challenges in model robustness. As datasets grow in size and complexity, there is an increased risk of overfitting to certain biases in the data (e.g., class imbalance, spurious correlations). DiMPLe mitigates this risk with its disentangling mechanism, which ensures that it learns more invariant and less dataset-specific features, allowing it to generalize better on both large-scale and small, diverse datasets. It benefits from large datasets for in-distribution performance but does not sacrifice its adaptability on smaller, more diverse datasets.

Class Discrimination Ability-CoOp-OOD vs DiMPLe:

The ability to distinctly represent different classes in the feature space, is a crucial aspect of robust image classification, especially in scenarios involving OOD data. We use t-SNE plots to visualize embedding distribution, providing insights into class separation and cluster compactness. The t-SNE visualizations in Fig. 4 compare the the class discrimination capabilities of CoOp-OOD and DiMPLe, revealing notable differences between the two approaches as evident. Firstly, the compactness of clusters in DiMPLe embeddings 4 (bottom) indicates lower intra-class variance, meaning that samples from the same class are more tightly grouped. CoOp-OOD embeddings 4 (top), however, show looser clusters. Secondly, these visualizations confirm our hypothesis of the importance of segregating the invariant and spurious features for both modalities. Fig. 4 (top) presents a tSNE visualization using the CoOp-OOD method. It shows how the invariant image features (circle) are separated from spurious image features (diamond), and all the textual features (triangle) are closer to invariant image features in the embedding space compared to the spurious ones. In contrast to this 4 (bottom) presents a tSNE

(a) Average over 11 datasets				(b) ImageNet.				(c) Caltech101			
	Base	Novel	HM		Base	Novel	HM		Base	Novel	HM
CLIP	69.34	74.22	71.70	CLIP	72.43	68.14	70.22	CLIP	96.84	94.00	95.40
ProDA	81.56	72.30	76.65	ProDA	75.40	70.23	72.72	ProDA	98.27	93.23	95.68
MaPLe	82.28	75.14	78.55	MaPLe	76.66	70.54	73.47	MaPLe	97.74	94.36	96.02
CoOp-OOD	60.82	29.04	39.31	CoOp-OOD	49.13	17.03	25.29	CoOp-OOD	26.10	19.20	22.12
CoOp-OOD [†]	62.72	65.60	64.13	CoOp-OOD [†]	62.20	55.90	59.00	CoOp-OOD [†]	91.20	92.57	91.88
DiMPLe	76.09	73.35	74.70	DiMPLe	71.87	62.67	66.96	DiMPLe	97.43	93.53	95.44
(d) OxfordPets				(e) StanfordCars				(f) Flowers102			
	Base	Novel	HM		Base	Novel	HM		Base	Novel	HM
CLIP	91.17	97.26	94.12	CLIP	63.37	74.89	68.65	CLIP	72.08	77.80	74.83
ProDA	95.43	97.83	96.62	ProDA	74.70	71.20	72.91	ProDA	97.70	68.68	80.66
MaPLe	95.43	97.76	96.58	MaPLe	72.94	74.00	73.47	MaPLe	95.92	72.46	82.56
CoOp-OOD	72.97	34.73	47.06	CoOp-OOD	76.67	69.77	73.06	CoOp-OOD	79.13	16.17	26.85
CoOp-OOD [†]	85.2	96.03	90.29	CoOp-OOD [†]	55.83	62.23	58.86	CoOp-OOD [†]	58.00	61.93	59.9
DiMPLe	91.57	97.73	94.55	DiMPLe	67.97	74.07	70.89	DiMPLe	84.33	75.00	79.39
(g) Food101				(h) FGVC Aircraft				(i) SUN397			
	Base	Novel	HM		Base	Novel	HM		Base	Novel	HM
CLIP	90.10	91.22	90.66	CLIP	27.19	36.29	31.09	CLIP	69.36	75.35	72.23
ProDA	90.30	88.57	89.43	ProDA	36.90	34.13	35.46	ProDA	78.67	76.93	77.79
MaPLe	90.71	92.05	91.38	MaPLe	37.44	35.61	36.50	MaPLe	80.82	78.70	79.75
CoOp-OOD	74.40	34.27	46.93	CoOp-OOD	33.27	4.83	8.44	CoOp-OOD	45.23	30.03	36.1
CoOp-OOD [†]	84.57	86.27	85.41	CoOp-OOD [†]	16.40	23.23	19.23	CoOp-OOD [†]	71.00	66.60	68.73
DiMPLe	89.80	91.23	90.51	DiMPLe	30.57	36.80	33.40	DiMPLe	75.43	74.27	74.85
(j) DTD				(k) EuroSAT				(l) UCF101			
	Base	Novel	HM		Base	Novel	HM		Base	Novel	HM
CLIP	53.24	59.90	56.37	CLIP	56.48	64.05	60.03	CLIP	70.53	77.50	73.85
ProDA	80.67	56.48	66.44	ProDA	83.90	66.00	73.88	ProDA	85.23	71.97	78.04
MaPLe	80.36	59.18	68.16	MaPLe	94.07	73.23	82.35	MaPLe	83.00	78.66	80.77
CoOp-OOD	57.20	9.53	16.34	CoOp-OOD	85.70	38.63	53.25	CoOp-OOD	69.17	45.3	54.75
CoOp-OOD [†]	41.70	44.23	42.93	CoOp-OOD [†]	55.30	63.5	59.12	CoOp-OOD [†]	68.23	69.1	68.66
DiMPLe	69.80	62.30	65.84	DiMPLe	80.10	63.93	71.11	DiMPLe	78.10	75.33	76.69

Table 3. Comparison of DiMPLe with CLIP, ProDA, MaPLe, CoOp-OOD and CoOp-OOD[†] for base-to-novel generalization.

visualization using our approach i.e., DiMPLe. This empirically shows how the textual invariant features (star) are aligned with the image invariant features (circle) whereas the textual spurious features (triangle) which are closer to the image spurious features are well separated from the invariant ones. Thereby involving the multi-modal invariant features to effectively contribute towards the accurate prediction of a class and minimizing the influence of the spurious ones. Thus we can conclude that by explicitly disentangling invariant and spurious features across the vision-language modalities, DiMPLe ensures that invariant embeddings primarily focus on class-relevant information.

4.5. Ablation Experiments

Optimal points and strategy of disentanglement: In order to maximize effectiveness, we must identify the optimal points and strategies for disentangling invariant and spurious features across the two modalities. This leads us to the following question: *how and at what stage should we disentangle the invariant and spurious features across the two modalities?* To answer this question, we analyze an early disentanglement strategy (please see supplementary B) within the DiMPLe framework which disentangles the features during prompt initialization stage. In this approach, two learnable prompts are designed for each modality and are initialized with templates targeting invariant ('a photo

Effect of Grouping		
Model (w/o Group Info)	CelebA	
	Avg	Worst
CoOp-OOD	78.10	31.11
DiMPLe	87.36	70.0

Table 4. Using group information during training. CoOp-OOD vs DiMPLe.

capturing core and invariant features of [category]’ and spurious features (‘a photo with features that keep changing in [category]’), which are then projected into the visual space using learnable linear layers (μ_u and η_s). This design ensures the parallel decomposition of features in both modalities, maintaining alignment through the shared visual-language space. As evident in Fig. 3 in cross-dataset evaluation, DiMPLe consistently outperforms DiMPLe-E across a variety of datasets, showing better generalization.

Ablation on Loss components: Fig. 5 demonstrates the impact of different loss components on model performance. The baseline model using only the unsupervised cross-entropy loss $\mathcal{L}_{ce}^u$ performs well, but adding the class-wise mutual information loss $\mathcal{L}_{cmi}$ consistently enhances accuracy across datasets. In contrast, adding only the spurious feature regularization loss $\mathcal{L}_r^{sp}$ leads to suboptimal results, particularly in Food101 and Flowers102, highlighting that suppressing spurious features without reinforcing task-relevant ones is insufficient. While combining $\mathcal{L}_{ce}^u$ and $\mathcal{L}_{cmi}$ improves robustness, the full model—integrating all three losses—achieves the best overall performance, with $\mathcal{L}_r^{sp}$ encouraging generalization by reducing reliance on spurious correlations. We use weighting coefficients α and β for $\mathcal{L}_r^{sp}$ and $\mathcal{L}_{cmi}$, respectively, and find that $\alpha = 1.0$, $\beta = 2.0$ yields the best trade-off; both lower and higher values degrade performance. As a guideline, we recommend fixing $\alpha = 1.0$ and tuning β in [1, 2].

Effect of Number of Tokens and Prompt Depth Tables 5 and 6 examine the impact of varying the number of tokens and prompt depth on model performance respectively, for base-t-novel generalization. For the number of tokens, increasing from $n = 1$ to $n = 2$ shows minor variations across datasets, with performance remaining stable. However, further increasing to $n = 4$ leads to a slight decline in accuracy, particularly for the novel classes in all datasets, suggesting while additional tokens may enhance feature expressiveness, excessive tokenization could introduce redundancy, leading to diminishing returns. In case of prompt depth, performance trends vary depending on the dataset. Increasing the depth from $J = 3$ to $J = 6$ results in an improvement for most datasets, with notable gains in Flowers102. For $J = 9$ and $J = 12$, performance either plateaus or slightly declines indicating deeper prompts can enhance representation learning up to a certain extent, beyond which they may lead to overfitting or reduced generalization.

Effect of Number of Tokens						
# Tokens	Caltech101		Food101		Flowers102	
	Base	Novel	Base	Novel	Base	Novel
$n = 1$	97.4	94.4	89.9	91.2	82.7	75.5
$n = 2$	97.4	93.5	89.8	91.2	84.3	75.0
$n = 4$	97.4	92.8	88.6	90.3	88.0	74.6

Table 5. Varying the number of tokens.

Effect of Prompt Depth						
Prompt Depth J	Caltech101		Food101		Flowers102	
	Base	Novel	Base	Novel	Base	Novel
$J = 3$	97.7	95.1	90.4	91.8	79.3	73.3
$J = 6$	97.9	92.7	90.6	91.7	83.6	75.8
$J = 9$	97.4	93.5	89.8	91.2	84.3	75.0
$J = 12$	97.6	92.5	89.2	91.3	81.4	75.2

Table 6. Varying the prompt depth J .

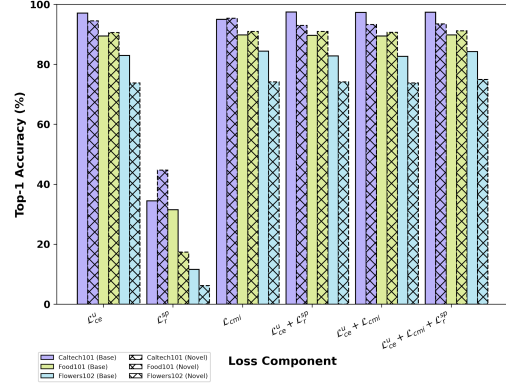


Figure 5. Impact of loss components on model performance.

Using group information during training: Table 4 evaluates model performance on **CelebA**, a hair color prediction dataset with four groups: non-blond females (G1, 3.9%), non-blond males (G2, 73.9%), blond females (G3, 21.1%), and blond males (G4, 1.1%). The minority group G4 is severely underrepresented, and gender acts as a spurious feature. CoOp-OOD, which optimizes for OOD generalization, achieves an average accuracy of 78.10%, but its worst-group accuracy drops significantly to 31.11%, showing poor generalization to the minority group. DiMPLe, without explicitly incorporating group information, achieves an improved average accuracy of 87.36% and significantly enhances the worst-group accuracy to 70.0%. This suggests that DiMPLe generalizes better across groups, mitigating the reliance on spurious correlations and achieving robust performance for underrepresented groups.

5. Conclusion

DiMPLe, enhances adaptation to OOD data by isolating invariant and spurious features at textual and visual levels. A central novelty of our work lies in eliminating spurious correlations within the textual modality using disentanglement, an overlooked yet critical limitation in prior vision-language models. More importantly, we uncover a key insight: invariant image features can mis-align with spurious textual prompts, reducing generalization. By explicitly addressing this through disentangled prompting, our method not only improves generalization, but also establishes a new paradigm for building reliable, prompt-tuned VLMs.

References

- [1] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. Flamingo: a visual language model for few-shot learning. *Advances in neural information processing systems*, 35:23716–23736, 2022. 2
- [2] Lukas Bossard, Matthieu Guillaumin, and Luc Van Gool. Food-101—mining discriminative components with random forests. In *Computer vision—ECCV 2014: 13th European conference, zurich, Switzerland, September 6–12, 2014, proceedings, part VI 13*, pages 446–461. Springer, 2014. 4
- [3] 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. 4
- [4] 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. 4
- [5] Yu Du, Fangyun Wei, Ziheng Zhang, Miaoqing Shi, Yue Gao, and Guoqi Li. Learning to prompt for open-vocabulary object detection with vision-language model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14084–14093, 2022. 1
- [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. 4
- [7] Xiuye Gu, Tsung-Yi Lin, Weicheng Kuo, and Yin Cui. Open-vocabulary object detection via vision and language knowledge distillation. *arXiv preprint arXiv:2104.13921*, 2021. 1
- [8] 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. 5
- [9] 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. 5
- [10] 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. 5
- [11] 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
- [12] Menglin Jia, Luming Tang, Bor-Chun Chen, Claire Cardie, Serge Belongie, Bharath Hariharan, and Ser-Nam Lim. Visual prompt tuning. In *European Conference on Computer Vision*, pages 709–727. Springer, 2022. 2
- [13] Woojeong Jin, Yu Cheng, Yelong Shen, Weizhu Chen, and Xiang Ren. A good prompt is worth millions of parameters: Low-resource prompt-based learning for vision-language models. *arXiv preprint arXiv:2110.08484*, 2021. 2
- [14] Florian Kalinke and Zoltán Szabó. Nyström m -hilbert-schmidt independence criterion, 2023. 4
- [15] Muhammad Uzair Khattak, Hanoona Rasheed, Muhammad Maaz, Salman Khan, and Fahad Shahbaz Khan. Maple: Multi-modal prompt learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19113–19122, 2023. 2, 3
- [16] 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. 4
- [17] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, pages 19730–19742. PMLR, 2023. 1, 2
- [18] Subhransu Maji, Esa Rahtu, Juho Kannala, Matthew Blaschko, and Andrea Vedaldi. Fine-grained visual classification of aircraft. *arXiv preprint arXiv:1306.5151*, 2013. 4
- [19] 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. 4
- [20] 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. 4
- [21] 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, 5
- [22] 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. 5
- [23] Yang Shu, Xingzhuo Guo, Jialong Wu, Ximei Wang, Jianmin Wang, and Mingsheng Long. Clipood: Generalizing clip to out-of-distributions. In *International Conference on Machine Learning*, pages 31716–31731. PMLR, 2023. 3
- [24] Amanpreet Singh, Ronghang Hu, Vedanuj Goswami, Guillaume Couairon, Wojciech Galuba, Marcus Rohrbach, and Douwe Kiela. Flava: A foundational language and vision alignment model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15638–15650, 2022. 1
- [25] K Soomro. Ucf101: A dataset of 101 human actions classes from videos in the wild. *arXiv preprint arXiv:1212.0402*, 2012. 4

- [26] 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. [5](#)
- [27] 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. [4](#)
- [28] Yuan Yao, Ao Zhang, Zhengyan Zhang, Zhiyuan Liu, Tat-Seng Chua, and Maosong Sun. Cpt: Colorful prompt tuning for pre-trained vision-language models. *AI Open*, 5:30–38, 2024. [2](#)
- [29] Jie Zhang, Xiaosong Ma, Song Guo, Peng Li, Wenchao Xu, Xueyang Tang, and Zicong Hong. Amend to alignment: Decoupled prompt tuning for mitigating spurious correlation in vision-language models. In *Forty-first International Conference on Machine Learning*. [1](#), [3](#), [5](#)
- [30] 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. [2](#)
- [31] 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](#)
- [32] Beier Zhu, Yulei Niu, Yucheng Han, Yue Wu, and Hanwang Zhang. Prompt-aligned gradient for prompt tuning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15659–15669, 2023. [2](#)