Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

DG-CLIP: A Vision-Language Model with Dynamic Fusion and Gradient-Aware Injection

Xinyu Yan

2420812030l3@lut.edu.cn

Lanzhou University of Technology

Guofang Zhang

Lanzhou University of Technology

Channi Li

Lanzhou University of Technology

Hao Wang

Lanzhou University of Technology

Yi Zhang

Lanzhou University of Technology

Gang Pan

Lanzhou University of Technology

Research Article

Keywords:

Posted Date: November 25th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11622282/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

DG-CLIP: A Vision-Language Model with Dynamic Fusion and Gradient-Aware Injection

Xinyu Yan[1], Guofang Zhang[1*], Channi Li[1], Hao Wang[1], Yi Zhang[2], Gang Pan[1]

[1] School of Computer Science and Artificial Intelligence, Lanzhou University of Technology, Lanzhou, China .

[2] Beijing University of Posts and Telecommunications, Beijing, China .

*Corresponding author(s). E-mail(s): 242081203013@lut.edu.cn;

Abstract

Vision-language models (VLMs) excel at bridging vision and text, yet they face two key limitations: the static nature of feature fusion and the instability of fine-tuning large pre-trained models. Static fusion struggles to capture deep, evolving cross-modal interactions, while full fine-tuning is computationally expensive and prone to catastrophic forgetting. To tackle these issues, we propose DG-CLIP, a novel fine-tuning strategy that integrates two core innovations: a Dynamic Context Fusion Network (DCFN) and a Gradient-Scaled Residual Injection (GSRI) module. DCFN enables adaptive, multi-scale fusion via cross-modal attention and dynamic weighting, producing context-rich representations. GSRI introduces a gradient scaling layer that precisely controls gradient flow during backpropagation, allowing efficient task adaptation without compromising general knowledge. Integrated into CLIP, DG-CLIP achieves state-of-the-art performance on 11 benchmarks, with notable gains in fine-grained classification and texture recognition. Our work advances multimodal fusion and provides a robust solution for efficient model adaptation in few-shot learning and base-to-novel generalization.

Keywords: CLIP, Dynamic Context Fusion, Parameter-Efficient, Transfer Learning

1 Introduction

Vision-Language Models (VLMs) have become a fundamental framework for multimodal representation learning by aligning visual and textual information in a shared embedding space [1–3]. Among them, CLIP has shown strong zero-shot transfer ability through large-scale contrastive pre-training, and has been widely adopted in image classification, retrieval, and other downstream vision-language tasks [4–7].

Despite these advances, adapting CLIP-like VLMs to downstream tasks still faces two major challenges. First, many existing methods rely on static or weakly adaptive fusion strategies, such as direct concatenation or linear projection, which are insufficient for modeling task-dependent and fine-grained cross-modal interactions [8–11]. Second, fully fine-tuning large pre-trained VLMs is computationally expensive and may disturb the general knowledge learned during pre-training, while existing parameter-efficient methods still need more stable mechanisms to inject task-specific knowledge without causing negative interference [12–16].

To address these challenges, we propose **Dynamic Fusion and Gradient-Aware Injection for CLIP** (**DG-CLIP**), a parameter-efficient fine-tuning framework built upon CLIP. DG-CLIP introduces two complementary modules:

a **Dynamic Contextual Fusion Network** (DCFN) for adaptive multimodal feature fusion, and a **Gradient-Scaled Residual Injection** (GSRI) module for stable and efficient task adaptation.

Specifically, DCFN is designed to overcome the limitations of static fusion. It extracts multi-scale visual and textual representations, models cross-modal interactions through attention, and dynamically assigns fusion weights according to the input context. In this way, the model can adaptively emphasize different levels of visual and linguistic cues, leading to more context-aware joint representations.

GSRI focuses on parameter-efficient and stable fine-tuning. Instead of updating the entire pre-trained backbone, it injects lightweight residual modules into the CLIP encoders. A Gradient Scale Layer is further introduced to regulate the gradient flow of the injected modules during backpropagation, allowing the model to learn task-specific knowledge while reducing interference with the original pre-trained representations.

By combining dynamic cross-modal fusion with gradient-aware residual injection, DG-CLIP enables CLIP to better adapt to downstream recognition tasks while preserving its generalization ability. We evaluate DG-CLIP on 11 widely used benchmarks, demonstrating its effectiveness in both base-to-novel generalization and few-shot learning settings. In summary, the main contributions of this paper are as follows:

- We propose a Dynamic Contextual Fusion Network (DCFN), which performs adaptive multimodal fusion through multi-scale feature extraction, cross-modal attention, and dynamic weighting.
- We introduce a Gradient-Scaled Residual Injection (GSRI) module, which enables parameter-efficient fine-tuning by controlling the gradient flow of lightweight injected modules, improving training stability while preserving pre-trained knowledge.
- We integrate DCFN and GSRI into the CLIP framework and conduct comprehensive experiments on 11 open-source datasets, validating the effectiveness of DG-CLIP in base-to-novel generalization and few-shot learning.

2 Related Work

2.1 Dynamic Fusion in Vision-Language Models

Recent work explores dynamic fusion to enhance cross-modal interactions in VLMs, with methods such as degradation-aware fusion [17], environment conditioned fusion [18], text-guided multi-view fusion [19], instruction-aligned 3D reconstruction [20], bidirectional progressive fusion for medical imaging [21], and multi-granularity knowledge integration [22]. Despite these advances, existing dynamic fusion approaches often focus on task-specific adaptations or operate at a single level of feature granularity, lacking a comprehensive mechanism to simultaneously capture multi-scale representations and adaptively weight them based on cross-modal context. Many methods also rely on static post-fusion strategies that do not fully leverage the dynamic interplay between modalities throughout the fusion process. Our proposed Dynamic Context Fusion Network (DCFN) addresses these limitations by introducing multi-scale feature extraction with a dynamic weight generator that adaptively fuses features based on input complexity, combined with cross-modal attention and an adaptive gating mechanism to enable deeper inter-modal interactions while preserving original information—capabilities absent in prior static or simplistic fusion designs.

2.2 Parameter-Efficient Fine-Tuning for VLMs

Parameter-efficient fine-tuning (PEFT) for VLMs has gained significant traction [23], with approaches including parallel decoding and action chunking [24], counterfactual reasoning [25], language-aligned action representation [26], MLLM-guided joint inference [27], similarity coverage analysis [28], customized adapter-based planning [29], and skip-based adaptation without additional modules [30]. While these methods reduce parameter counts and mitigate catastrophic forgetting, they often struggle to balance task-specific adaptation with the preservation of general knowledge, and many suffer from gradient instability during fine-tuning due to simplistic adaptation mechanisms or external modules that

disrupt the original feature space. Our Gradient-Scaled Residual Injection (GSRI) module directly tackles these challenges through a gradient-aware mechanism: a Gradient Scale Layer asymmetrically scales gradients during backpropagation while maintaining identity mapping in the forward pass, ensuring that injected modules learn task-specific knowledge without interfering with the pre-trained backbone. The bottleneck architecture with optional parameter sharing further enhances parameter efficiency, while the dual GSL design maintains consistent residual magnitudes, providing more stable and effective fine-tuning compared to existing approaches.

3 Methods

3.1 DG-CLIP

The DG-CLIP model inherits CLIP's contrastive learning paradigm, comprising a visual encoder E_V and a text encoder E_T. These two encoders respectively map images and text into a shared embedding space. Unlike the original CLIP, DG-CLIP integrates a gradient injection mechanism within its encoders for efficient finetuning, and introduces a dynamic fusion mechanism after the encoder outputs for adaptive multimodal feature fusion. The overall architecture of DG-CLIP is illustrated in Fig. 1.

Given a batch of N image-text pairs (I, T), DG-CLIP's processing workflow is as follows: First, the image I and text T are fed separately through the visual encoder E_V and text encoder E_T, both integrated with the gradient injection mechanism, to generate the initial visual features F_V and text features F_T, respectively:

$$\mathrm{F_V} = \varepsilon_{\mathrm{V}}(\mathrm{I}) \tag{1}$$

$$\mathrm{F_T} = \varepsilon_{\mathrm{T}}(\mathrm{T}) \tag{2}$$

Were, I denotes the input image, and T denotes the input text sequence. $\mathcal{E}_{\mathrm{V}}$ is the visual encoder, which incorporates the gradient injection mechanism, and $\mathcal{E}_{\mathrm{T}}$ is the text encoder, similarly integrated with the gradient injection mechanism.$\mathcal{F}_V \in \mathbb{R}^{B\times D}$ represents the raw visual features output by the encoder, where B is the batch size and D is the feature dimension. $\mathcal{F}_T \in \mathbb{R}^{B\times D}$ represents the raw text features output by the encoder. Subsequently, these raw features, O and Z, are fed into the dynamic fusion network for further dynamic fusion and enhancement, ultimately yielding the final image features v and text features t:

$$(\mathrm{v}, \mathrm{t}) = \prod(\mathrm{F_V}, \mathrm{F_T}). \tag{3}$$

Where $\prod$ denotes the Dynamic Cross-modal Fusion Network (DCFN), whose detailed structure will be elaborated in Section 2.2. $\mathrm{v}, \mathrm{t} \in \mathrm{R}^{B\times D}$ represents the final image and text features output by $\prod$. Finally, these fused features, v and t, are then L2-normalized and used to compute the contrastive learning loss. The core objective of DG-CLIP is to maximize the similarity of matched image-text pairs through contrastive learning while simultaneously minimizing the similarity of mismatched pairs. Its loss function is defined as:

$$L = -\frac{1}{N}\sum_{i=1}^{N}\left[\log\frac{\exp(\tau\cdot\mathrm{sim}(v_i,t_i))}{\sum_{j=1}^{N}\exp(\tau\cdot\mathrm{sim}(v_j,t_j))} + \log\frac{\exp(\tau\cdot\mathrm{sim}(v_i,t_i))}{\sum_{j=1}^{N}\exp(\tau\cdot\mathrm{sim}(v_j,t_j))}\right] \tag{4}$$

Where N denotes the batch size. $\mathrm{v_i}, \mathrm{t_i} \in \mathrm{R}^{\mathrm{B}\times\mathrm{D}}$ are the D-dimensional feature vectors for the i-th image and the i-th text, respectively. The similarity metric employed is cosine similarity sim. $\mathcal{T}$ is a learnable temperature parameter, which controls the sharpness of the similarity distribution.

3.2 DCFN

We propose a Dynamic Contextual Fusion Network (DCFN) module, whose architecture is depicted in Figure 2. This module sequentially integrates multi-scale feature extraction, dynamic weight allocation, cross-modal attention, and an adaptive fusion gating mechanism. This comprehensive approach enables the refined, adaptive fusion and enhancement of visual and textual features, specifically designed to overcome the inherent limitations of traditional static fusion mechanisms. This network corresponds to $\prod$ in Eq. (3).

The DCFN processes image and text features independently at first, with cross-modal interaction occurring in a later stage. The module's core components and operational flow are outlined below: Initially, for the input modal features, denoted as $F_m \in \mathbb{R}^{B\times D}$, where $\mathrm{m} \in \{\mathrm{V}, \mathrm{T}\}$ represent visual and textual modalities, respectively, the DCFN employs a series of feature projectors P_k to extract multi-scale feature representations. Each projector

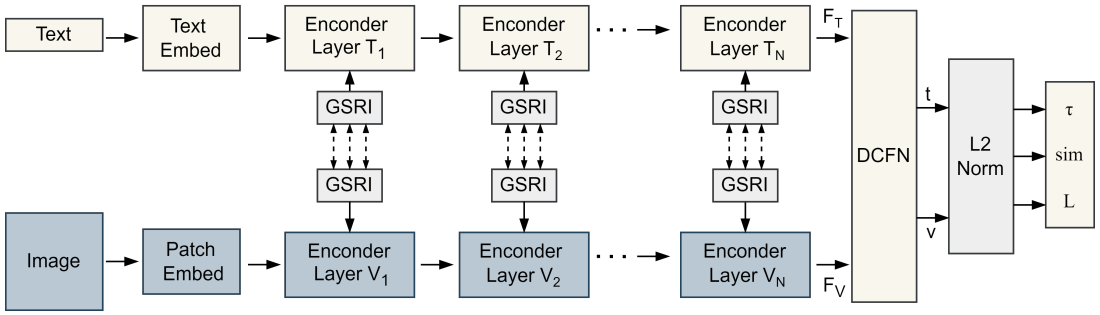


Fig. 1 Overall Architecture Schematic of DG-CLIP.

P_k comprises two linear transformations l followed by an activation function $\boldsymbol{\rho}$.

$$F_{m,k} = P_k(F_m) = l_k, 2(\rho(l_k, 1(F_m))). \quad (5)$$

Here,$F_{m,k} \in \mathbb{R}^{B\times D}$ represents the feature extracted for modality m at scale k, where $\urcorner \in \{1, \cdots, \mathbb{K}\}$ denotes the different scale indices l_k. 1/2 are the first and second linear transformations, respectively, projecting D-dimensional features to $\frac{D}{sk}$. s,k dimensions, where s and k are scale factors. $\boldsymbol{\rho}$ denotes the Rectified Linear Unit (ReLU) activation function. To adaptively fuse these multi-scale features, the DCFN introduces a dynamic weight generator. This generator takes the original modal features F_m as input and, through a compact network, predicts the fusion weights for each scale feature.

$$W_m = S(l_w, 2(\rho(l_w, 1(F_m)))). \quad (6)$$

Here, $W_m \in \mathbb{R}^{B\times K}$ represents the dynamic fusion weight vector for modality m, with its elements $W_{m,k}$ denoting the weights for scale k. The Softmax function S is applied to ensure that all scale weights sum to 1. Having obtained the multi-scale features $F_{m,k}$ and dynamic weights $W_{m,k}$, the DCFN fuses them through a weighted summation to yield the preliminary fused feature $\hat{F}_m$. Subsequently, this preliminary fused feature then undergoes further enhancement via a Multi-Layer Perceptron (MLP):

$$\hat{F}_m = \sum_{k=1}^{K} W_{m,k} \odot F_{m,k}. \quad (7)$$

$$F'_m = M(\hat{F}_m). \quad (8)$$

Specifically: $\hat{F}_m \in \mathbb{R}^{B\times D}$ represents the feature for modality m resulting from its weighted multi-scale fusion. The symbol $\odot$ denotes element-wise multiplication. $F'_m \in \mathbb{R}^{B\times D}$ is the feature for modality m after undergoing feature enhancement. M is a feature enhancement network comprising a linear transformation, $\boldsymbol{\rho}$ activation, and Dropout. Subsequently, to effectively capture deep inter-modal interactions, the DCFN introduces a cross-modal attention mechanism. This mechanism allows the features of one modality to serve as the Query, attending to the features of another modality as the Key and Value. This process, in turn, generates features that are inherently enhanced by the information from the other modality:

$$F_{V\to T} = \mathrm{N}\left(A\left(F_V, F_T, F_T\right) + F_V\right). \quad (9)$$

$$F_{T\to V} = \mathrm{N}\left(A\left(F_T, F_V, F_V\right) + F_T\right). \quad (10)$$

Here, $F_V, F_T \in R^{B\times D}$ refers to the image and text features after they have been processed internally within their respective modalities. The notation A(Query, Key, Value) signifies the multi-head attention mechanism. $F_{V\to T} \in R^{B\times D}$ denotes the image features, enhanced by the text features, with the image features serving as the Query. Conversely, $F_{V\to T} \in R^{B\times D}$ denotes the text features, enhanced by the image features, with the text features serving as the Query. N stands for Layer Normalization.

Finally, an adaptive fusion gating mechanism is employed to merge the original features with those enhanced by the cross-modal attention. This gating mechanism dynamically determines how much of the original information to retain and how much of the enhanced information to incorporate:

$$G_m = \sigma(l([F_m; F_{m\to other}])) \quad (11)$$

$$F_{final} = G_m \odot F_m + (1 - G_m) \odot F_{m\to other} \quad (12)$$

Here, F_m represents the original feature for modality m. $F_{m\to other}$ is the feature for modality m after being enhanced by cross-modal attention.

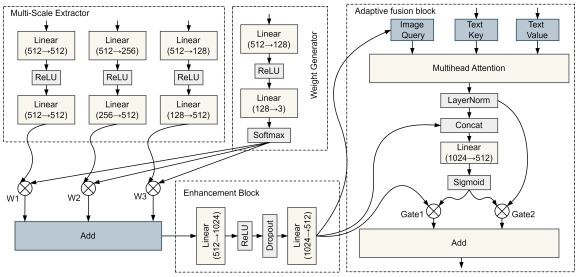


Fig. 2 Architecture of the Dynamic Fusion Network.

L is a linear transformation that projects the concatenated features into a D-dimensional space. σ is a Sigmoid function, which outputs a gating value $G_m \in \mathbb{R}^{B\times D}$ ranging from 0 to 1. The symbol $\odot$ signifies element-wise multiplication. F_{final} represents the final fused feature for modality m, which will then be utilized for subsequent contrastive learning.

Through this multi-stage dynamic fusion process, the DCFN is able to generate more context-aware and adaptive multi-modal feature representations. This, in turn, significantly enhances DG-CLIP's understanding capabilities in complex scenarios.

3.3 GSRI

GSRI is crucial for achieving parameter-efficient fine-tuning within DG-CLIP, with its architecture depicted in Figure 3. It operates by injecting small residual networks directly into the Transformer blocks of the pre-trained model. Crucially, it then leverages a Gradient Scale Layer (GSL) to precisely control the gradient flow within these injected modules. This allows it to efficiently guide the model towards learning task-specific knowledge, all without interfering with the general knowledge embedded in the original pre-trained model. The GSL serves as the core component of this gradient injection mechanism. Its unique characteristic lies in the asymmetrical behavior between its forward and backward passes. During the forward pass, the GSL functions as an identity mapping; however, during the backward pass, it scales the gradients by a predetermined scaling factor, s. We denote the GSL operation using G.

$$Forward: y = G(x, s) = x. \tag{13}$$

$$Backward: \frac{\partial L}{\partial x} = s \cdot \frac{\partial L}{\partial y}. \tag{14}$$

Here, x denotes the input feature tensor, y is the output feature tensor, s is the gradient scaling factor, and L represents the final loss function. This mechanism ensures that the injected module has minimal impact on the backbone network during forward inference, thereby enabling more stable parameter-efficient fine-tuning.

The injected module itself is a lightweight residual network, typically comprising a downsampling linear transformation D, an optional shared module S, and an upsampling linear transformation U, thus forming a bottleneck structure. It is injected into the residual path of the Transformer block, operating in parallel with the original self-attention and Multi-Layer Perceptron branches. For the i-th Transformer block, let $X_i \in \mathbb{R}^{L\times B\times D}$ denote its input feature tensor. Here, L, B, and D denote the sequence length, batch size, and feature dimension, respectively. The standard Transformer block updates can then be expressed as:

$$X_i' = X_i + SelfAttention(N(X_i)). \tag{15}$$

$$X_{i+1} = X_i + M(N(X_i')). \tag{16}$$

Where M denotes the Multi-Layer Perceptron (MLP). The injected module augments the Transformer block by adding an additional residual term, Δ $\mathrm{X_i}$, after the M branch. The injection process proceeds as follows: First, the current layer's input feature, $\boldsymbol{X}_\mathrm{i}$, is processed by the GSL to control its gradient:

$$\tilde{X}_i = G(X_i, s_i) \tag{17}$$

Then, the injected module processes (A) to generate the residual term Δ $\mathrm{X_i}$

$$\Delta X_i = \mathrm{G}(U_i(\mathrm{S}(D_i(\mathrm{N}(\tilde{X}_i)))) \cdot s_i, 1/s_i) \tag{18}$$

Here, S_i represents the gradient scaling factor for the current layer i. N denotes Layer Normalization. D_i is the injected module's downsampling linear transformation, which reduces the feature dimension from D to D/r, where r is the reduction ratio. S is an optional shared injection module, allowing parameters to be shared across multiple layers. U_i is the injected module's upsampling linear transformation, restoring the feature dimension from D/r back to D. The second $G(\cdot, 1/s_i)$ is designed to counteract the multiplicative effect S_i introduced by the first GSL during forward propagation. This ensures the magnitude of the residual term is consistent with the main backbone

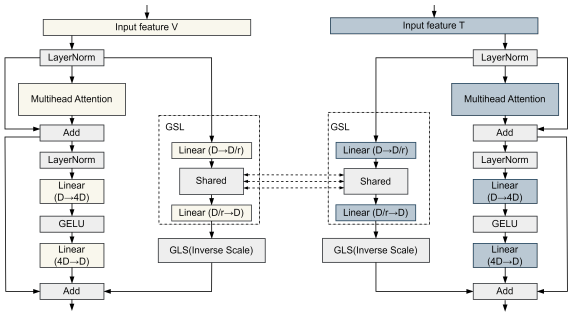


Fig. 3 Architecture of the GSRI.

network. Simultaneously, during backward propagation, it ensures the gradients of the injected module are scaled by S_i. Finally, the output X_{i+1} of the Transformer block is updated to:

$$X_{i+1} = X_i^{'} + M(N(X_i^{'})) + \Delta X_i. \quad (19)$$

Through this gradient-aware injection mechanism, DG-CLIP can effectively guide the model to learn task-specific knowledge during fine-tuning with a minimal number of parameters. Crucially, it also prevents catastrophic forgetting of the pre-trained backbone and significantly enhances training stability and efficiency.

4 Experiments

4.1 Experimental Setup

Tasks. We evaluate DG-CLIP under two standard transfer settings. In *base-to-novel generalization*, the classes of each dataset are split into base and novel categories following the protocol of CoOp [31]. The model is trained on the base classes with 16 shots per class and evaluated on both base and novel classes. We report the accuracy on base classes, the accuracy on novel classes, and their harmonic mean (HM), which reflects the balance between adaptation and generalization.

In *few-shot learning*, the model is trained with 1, 2, 4, 8, and 16 labeled samples per class, and then evaluated on the test set containing all classes. This setting is used to examine whether DG-CLIP can adapt effectively when only limited labeled data are available.

Datasets. We consider a suite of 11 datasets used in CoOp[31]. Specifically, these include ImageNet[32], Caltech101 (CAL)[33], Oxford Pets (PETS)[34], Stanford Cars (CARS)[35], Oxford Flowers 102 (FLWR)[36], Food-101 (FOOD)[37], FGVC Aircraft (AIR)[38], SUN39[39], Describable Textures (DTD)[40], EuroSAT (ESAT)[41], and UCF-101 (UCF)[42], and we use the splits of [31].

4.2 Comparison with State-of-the-Art Methods

To comprehensively evaluate the effectiveness of the DG-CLIP model, we conducted generalization experiments on 11 widely-used image classification datasets. We compared DG-CLIP with various state-of-the-art (SOTA) methods proposed in recent years[12, 31, 43–49] , with the results presented in Table 1. The experimental results indicate that DG-CLIP significantly outperforms existing competitors in terms of overall performance. Specifically, DG-CLIP achieved an average HM score of 80.19%, surpassing the previously best-performing QMaPLe and MaPLe, and substantially outperforming CoOp and Co-CoOp.

Notably, DG-CLIP demonstrated exceptional performance in addressing the trade-off between 'catastrophic forgetting' and 'generalization capability.' For Base classes, DG-CLIP achieved an average accuracy of 77.80%, surpassing most fine-tuning methods and demonstrating that the GSRI module effectively preserves the pre-trained model's general knowledge through gradient scaling. For Novel classes, DG-CLIP achieved an average accuracy of 71.48%, significantly higher than CoOp and MaPLe. This result strongly validates that the DCFN module, through dynamic context fusion, successfully extracts more robust cross-modal features, thereby greatly enhancing the model's zero-shot inference capability on unseen categories.

On ImageNet, a large-scale general-purpose dataset, DG-CLIP exhibited the best performance, with its HM reaching 74.51%, outperforming all compared methods. DG-CLIP's advantages were particularly evident in more challenging fine-grained classification tasks. For instance, on the FGVCAircraft dataset, where existing methods generally perform poorly, MaPLe's HM was only 36.50%, whereas DG-CLIP boosted it to 39.48%. Furthermore, its accuracy on Novel classes reached 37.61%, significantly higher than MaPLe's 35.61%. Similarly, on the StanfordCars and Flowers102 datasets, DG-CLIP's HM reached 77.49% and 86.41% respectively, both setting new state-of-the-art records. This indicates that DG-CLIP's

Table 1 Accuracy comparison on Base-to-novel generalization of DG-CLIP with previous methods.

Method		Avg	ImageNet	Caltech	Pets	Cars	Flowers	Food101	Aircraft	SUN397	DTD	EuroSAT	UCF101.
CLIP[7]	B	69.34	72.43	96.84	91.17	63.37	72.08	90.10	27.19	69.36	53.24	56.48	70.53
	N	74.22	68.14	94.00	97.26	74.89	**77.80**	91.22	36.29	75.35	59.90	64.05	77.50
	HM	71.70	70.22	95.40	94.12	68.65	74.83	90.66	31.09	72.23	56.37	60.03	73.85
CoOp[31]	B	82.69	76.47	98.00	93.67	78.12	97.60	88.33	40.44	80.60	79.44	92.19	84.69
	N	63.22	67.88	89.81	95.29	60.40	59.67	82.26	22.30	65.89	41.18	54.74	56.05
	HM	71.66	71.92	93.73	94.47	68.13	74.06	85.19	28.75	72.51	54.24	68.69	67.46
Co-CoOp[43]	B	80.47	75.98	97.96	95.20	70.49	94.87	90.70	33.41	79.74	77.01	87.49	82.33
	N	71.69	70.43	93.81	97.69	73.59	71.75	91.29	23.71	76.86	56.00	60.04	73.45
	HM	75.83	73.10	95.84	96.43	72.01	81.71	90.99	27.74	78.27	64.85	71.21	77.64
ProGrad[12]	B	82.48	77.02	98.02	95.07	77.68	95.54	90.37	40.54	81.26	77.35	90.11	84.33
	N	70.75	66.66	93.89	97.63	68.63	71.87	89.59	27.57	74.17	52.35	60.89	74.94
	HM	76.16	71.46	95.91	96.33	72.88	82.03	89.98	32.82	77.55	62.45	72.67	79.35
KgCoOp[44]	B	80.73	75.83	97.72	94.65	71.76	95.00	90.50	36.21	80.29	77.55	85.64	82.89
	N	73.60	69.96	94.39	97.76	75.04	74.73	91.70	33.55	76.53	54.99	64.34	76.67
	HM	77.00	72.78	96.03	96.18	73.36	83.65	91.09	34.83	78.36	64.35	73.48	79.65
MaPLe[45]	B	82.28	76.66	97.74	95.43	72.94	95.92	**90.71**	37.44	80.82	80.36	94.07	83.00
	N	75.14	70.54	94.36	97.76	74.00	72.46	92.05	35.61	**78.70**	59.18	73.23	78.66
	HM	78.55	73.47	96.02	96.58	73.47	82.56	91.38	36.50	79.75	68.16	82.35	80.77
ProDA[46]	B	81.56	75.4	98.27	95.43	74.70	97.70	90.30	36.90	78.67	80.67	83.90	85.23
	N	72.30	70.23	93.23	97.83	71.20	68.68	88.57	34.13	76.93	56.48	66.00	71.97
	HM	76.65	72.72	95.68	96.62	72.91	80.66	89.43	35.46	77.79	66.44	73.88	78.04
PLOT[47]	B	83.98	77.30	**98.53**	94.50	79.07	97.93	89.80	**42.13**	82.20	81.97	93.70	86.60
	N	71.72	69.87	92.80	96.83	74.80	73.53	91.37	33.73	73.63	43.80	62.67	75.90
	HM	77.37	73.40	95.58	95.65	76.88	83.99	90.58	37.46	77.68	57.09	75.11	80.90
TCP[48]	B	**84.13**	77.27	98.23	94.67	**80.80**	97.73	90.57	41.97	**82.63**	82.77	91.63	**87.13**
	N	75.36	69.87	94.67	97.20	74.13	75.57	91.37	34.43	78.20	58.07	74.73	80.77
	HM	79.51	73.38	96.42	95.92	77.32	85.23	90.97	37.83	**80.35**	68.25	82.32	83.83
QMaPLe[49]	B	83.02	76.93	97.97	95.67	75.00	96.43	90.63	39.10	81.33	80.77	**94.30**	85.10
	N	75.57	70.73	95.00	97.63	73.67	74.33	**92.10**	34.90	78.27	57.63	79.47	77.50
	HM	79.12	73.70	96.46	96.64	74.33	83.95	**91.36**	36.88	79.77	67.27	**86.25**	81.12
DG-CLIP (Ours)	B	83.59	**77.8**	98.13	**96.17**	79.21	**98.86**	90.11	41.54	81.91	**83.45**	85.26	87.07
	N	**77.4**	**71.48**	94.88	**97.87**	**75.84**	76.74	91.39	**37.61**	78.34	**62.80**	83.00	**81.40**
	HM	**80.19**	**74.51**	**96.48**	**97.01**	**77.49**	**86.41**	90.75	**39.48**	80.09	**71.67**	84.11	**84.14**

dynamic fusion mechanism can keenly capture subtle visual differences, demonstrating a significant advantage when handling challenging fine-grained tasks.

On the DTD dataset, the complexity of texture features places extremely high demands on a model's generalization capability. Experimental data shows that DG-CLIP's HM on DTD reached 71.67%, significantly surpassing CoOp's 54.24% and MaPLe's 68.16%. Specifically for Novel class recognition, DG-CLIP achieved 62.80%, 3.62 percentage points higher than MaPLe. In the EuroSAT satellite image classification task, DG-CLIP also performed exceptionally well, with an HM of 84.11%. These results demonstrate that DG-CLIP is not only suitable for natural object recognition but also maintains exceptionally high generalization stability through adaptive feature injection when dealing with data exhibiting large cross-domain distribution shifts, such as textures and aerial imagery.

4.2.1 Few-Shot Learning

Table 2 summarizes the few-shot results with the ViT-B/16 backbone. DG-CLIP shows competitive performance under extremely limited supervision and becomes stronger when more training samples are available. In the 8-shot setting, DG-CLIP achieves 71.4% on ImageNet, 74.9% on SUN397, and 68.1% on DTD. In the 16-shot setting, it further reaches 73.2% on ImageNet, 76.9% on SUN397, and 74.8% on DTD, outperforming or matching strong baselines on several datasets.

These results indicate that DG-CLIP benefits from a small but sufficient amount of supervision. Compared with static adaptation methods, the proposed dynamic fusion module can better exploit task-specific visual-textual cues, while GSRI reduces the risk of unstable updates to the pre-trained backbone. Although the 1-shot setting remains challenging and the performance varies

Table 2 Performance comparison under few-shot setting with ViT-B/16 backbone.

Shots	Method	ImageNet	Caltech	Pets	Food	SUN397	DTD	UCF101
1-shot	CoOp[31]	65.7	92.5	90.2	84.3	67.0	50.1	71.2
	MaPLe[45]	69.7	93.6	91.4	85.4	**69.3**	50.0	71.1
	PLOT [47]	66.5	**94.3**	**91.9**	86.2	66.8	**54.6**	**74.3**
	DG-CLIP *[Ours]*	**68.9**	94.1	88.2	**86.3**	69.1	44.6	71.1
2-shot	CoOp	67.0	93.1	89.9	84.4	67.0	54.1	74.1
	MaPLe	**70.0**	**94.9**	91.8	**86.5**	70.7	50.6	74.0
	PLOT	68.3	94.7	**92.3**	86.3	68.1	**56.7**	**76.8**
	DG-CLIP *[Ours]*	69.9	94.4	89.2	86.3	**71.1**	50.0	74.6
4-shot	CoOp	68.8	94.5	92.5	84.5	69.7	59.5	77.6
	MaPLe	70.6	95.0	**93.3**	**86.7**	71.4	59.0	77.1
	PLOT	70.4	**95.1**	92.6	86.5	71.7	**62.4**	**79.8**
	DG-CLIP *[Ours]*	**71.0**	94.0	93.1	85.7	**73.2**	60.2	79.4
8-shot	CoOp	70.6	94.5	91.3	82.7	71.9	64.8	80.0
	MaPLe	71.3	95.1	**93.1**	**87.2**	73.2	63.0	79.5
	PLOT	71.3	**95.5**	93.0	86.6	73.9	66.5	**82.8**
	DG-CLIP *[Ours]*	**71.4**	**95.5**	92.7	86.8	**74.9**	**68.1**	82.3
16-shot	CoOp	71.9	95.8	92.0	84.2	74.9	69.7	83.1
	MaPLe	71.9	95.4	93.2	**87.4**	74.5	68.4	81.4
	PLOT	72.6	**96.0**	**93.6**	87.1	76.0	71.4	**85.3**
	DG-CLIP *[Ours]*	**73.2**	95.9	93.5	87.3	**76.9**	**74.8**	84.5

across datasets, the overall trend shows that DG-CLIP provides stronger adaptation as the number of labeled samples increases.

4.3 Ablation Experiments

To validate the effectiveness of each core component within the DG-CLIP framework, we conduct comprehensive ablation studies across 11 datasets by isolating individual architectural contributions. Specifically, we evaluate four distinct configurations: the standard CLIP model employing traditional static fusion serving as the baseline; the baseline augmented solely with the Dynamic Context Fusion Network (DCFN); the baseline incorporating only the Gradient-Scaled Residual Injection (GSRI) module; and our complete framework integrating both modules. The experimental results across the Base, Novel, and HM metrics are summarized in Table 3.

As shown in Table 3, independently introducing DCFN consistently outperforms the baseline, especially in fine-grained and texture-heavy scenarios. Specifically, DCFN boosts the HM from 56.09% to 70.47% on DTD, and from 58.10% to 76.87% on EuroSAT. This confirms that its multi-scale extraction and cross-modal attention yield more context-aware joint representations than static fusion. Meanwhile, GSRI mitigates catastrophic forgetting efficiently, raising the ImageNet HM from 70.21% to 73.35% and the FGVCAircraft HM from 30.09% to 38.79%. The pronounced gains on novel classes verify that the gradient-scaling layers effectively balance pre-trained knowledge retention and task-specific adaptation.

Crucially, the full DG-CLIP framework achieves the best performance across all datasets, peaking at an average HM of 80.19% (vs. the baseline's 70.69%). This demonstrates a clear synergy: DCFN enhances multi-modal fusion at the feature level, while GSRI stabilizes fine-tuning at the parameter level. Their combination simultaneously drives base-class precision and novel-class generalization, thoroughly validating our architectural design.

5 Discussion

The DG-CLIP model proposed in this paper aims to address the dual challenges of static feature fusion and inefficient fine-tuning prevalent in existing vision-language models. Through extensive experimentation across 11 benchmark datasets, we have validated the effectiveness of its two core

Table 3 Ablation experiment table.

Method		Avg	Cars	Pets	ImageNet	Flowers	Caltech	Food101	Aircraft	SUN397	DTD	EuroSAT	UCF101.
Baseline	B	68.07	63.32	89.31	89.31	68.85	96.32	89.39	27.49	69.38	53.7	50.26	68.36
	N	73.85	74.94	96.87	68.11	**77.23**	94.1	90.95	33.23	75.43	58.7	68.85	73.99
	HM	70.69	68.64	92.94	70.21	72.8	95.2	90.16	30.09	72.28	56.09	58.1	71.06
DCFN	B	82.62	78.34	95.64	95.64	97.34	97.74	89.62	41.48	**82.34**	81.48	81.64	85.63
	N	75.53	74.7	97.87	69.83	74.96	93.12	**91.6**	35.99	78.23	62.08	72.62	79.88
	HM	78.74	76.48	96.74	73.51	84.7	95.37	90.6	38.54	80.23	70.47	76.87	82.66
GSRI	B	83.36	**79.36**	95.27	95.27	98.2	97.61	**90.15**	41.12	82.1	**83.8**	**85.6**	85.99
	N	75.89	73.85	97.82	69.43	75.53	92.58	91.23	36.71	**78.58**	61.35	80.49	77.18
	HM	79.25	76.51	96.53	73.35	85.39	95.03	90.69	38.79	**80.3**	70.84	82.97	81.35
DG-CLIP	B	**83.59**	79.21	**96.17**	**96.17**	**98.86**	**98.13**	90.11	**41.54**	81.91	83.45	85.26	**87.07**
	N	**77.4**	**75.84**	**97.87**	**71.48**	76.74	**94.88**	91.39	**37.61**	78.34	**62.8**	**83**	**81.4**
	HM	**80.19**	**77.49**	**97.01**	**74.51**	**86.41**	**96.48**	**90.75**	**39.48**	80.09	**71.67**	**84.11**	**84.14**

components: the Dynamic Context Fusion Network (DCFN) and the Gradient-Scaled Residual Injection (GSRI).

The experimental results highlight the significant advantage of the DCFN module in handling complex visual tasks. Unlike traditional static concatenation or linear projection, DCFN leverages multi-scale feature extraction and an adaptive gating mechanism to dynamically adjust its fusion strategy based on the semantic complexity of the input content. This capability was particularly evident in fine-grained classification and texture recognition tasks, where DG-CLIP's performance on these datasets substantially outperformed the baseline model. This suggests that explicit cross-modal attention and dynamic weight allocation effectively capture subtle visual cues, leading to the generation of more discriminative joint feature representations.

The GSRI module was conceived to enable efficient task adaptation without compromising pre-trained knowledge. Comparative experiments revealed that DG-CLIP maintained high accuracy on Base categories while simultaneously achieving substantial improvements on Novel categories. This confirms the pivotal role of the Gradient Scaling Layer (GSL) in striking a balance between adapting to novel tasks and preserving existing knowledge. Compared to full parameter fine-tuning, GSRI requires updating only a minimal number of parameters. This not only reduces computational overhead but also effectively mitigates catastrophic forgetting, enabling the model to maintain strong generalization capabilities in few-shot or zero-shot scenarios.

While DG-CLIP demonstrates strong performance, it is not without its limitations. Firstly, the DCFN module introduces additional computational latency during inference due to its cross-modal attention computations. Although this enhances accuracy, it could pose a challenge for edge devices with stringent real-time constraints compared to a simpler CLIP model.

To address the computational overhead associated with DCFN, we will explore using knowledge distillation or more efficient attention mechanisms to optimize the model structure, aiming to achieve a more favorable trade-off between inference speed and accuracy. Furthermore, we plan to extend the DG-CLIP architecture to more modalities and more complex downstream tasks, such as open-vocabulary object detection and multimodal dialogue systems, to validate its versatility. Presently, the scaling factors within GSRI are either predefined or learned globally. Future work could investigate adaptive gradient scaling strategies based on layer-wise or sample difficulty, thereby further enhancing fine-tuning granularity and efficiency.

6 Conclusion

In this paper, we introduced DG-CLIP, a novel vision-language model that integrates dynamic context fusion with gradient-aware injection. By incorporating the DCFN module, we achieved adaptive deep fusion of multi-scale features, significantly enhancing the model's ability to understand complex contexts. Furthermore, through the design of the GSRI module, we proposed a parameter-efficient fine-tuning paradigm that effectively addresses the problem of catastrophic forgetting and improves training stability. Experimental results on 11 public datasets demonstrate that DG-CLIP outperforms existing state-of-the-art (SOTA)

methods in both zero-shot classification and generalization performance. DG-CLIP not only offers a novel perspective on multimodal feature fusion but also provides a practical and viable solution for the efficient deployment of large pre-trained models in real-world applications.

Statements and Declarations

Funding

No funding was received for this study.

Competing interests

The authors declare no competing financial or non-financial interests relevant to this work.

Data availability

This study uses eleven existing benchmark datasets: ImageNet [32], Caltech101 [33], Oxford Pets [34], Stanford Cars [35], Oxford Flowers 102 [36], Food-101 [37], FGVC Aircraft [38], SUN397 [39], Describable Textures Dataset [40], EuroSAT [41], and UCF101 [42]. These datasets are available through their original sources, subject to the access conditions of the respective providers. Summary experimental results are reported in this article. The authors plan to release detailed experimental records following acceptance of the manuscript.

Code availability

The source code has not yet been publicly released. The authors plan to make it publicly available following acceptance of the manuscript.

Author contributions

Xinyu Yan contributed to the conceptualization and methodology, implemented the software, and conducted the experiments. Guofang Zhang analyzed and interpreted the results and prepared the original draft. Channi Li, Hao Wang, Yi Zhang, and Gang Pan contributed to manuscript revision and provided research guidance.

References

[1] Liu, Y., Hong, Q., Huang, L., Gomez-Villa, A., Goswami, D., Liu, X., others, Tian, Y.: Continual learning for VLMs: A survey and taxonomy beyond forgetting. arXiv preprint arXiv:2508.04227 (2025)

[2] Gavit, D., Mazumder, D., Das, S., Patro, J.: On VLMs for Diverse Tasks in Multimodal Meme Classification. arXiv preprint arXiv:2505.20937 (2025)

[3] Li, Z., Wu, X., Du, H., Nghiem, H., Shi, G.: Benchmark evaluations, applications, and challenges of large vision language models: A survey. arXiv preprint arXiv:2501.02189 (2025)

[4] Cooper, A., Kato, K., Shih, C.H., Yamane, H., Vinken, K., Takemoto, K., others, Boix, X.: Rethinking vlms and llms for image classification. Scientific Reports **15**(1), 19692 (2025)

[5] Duan, Y., Huang, X., Chen, D.: Enhancing Subsequent Video Retrieval via Vision-Language Models (VLMs). arXiv preprint arXiv:2503.17415 (2025)

[6] Liu, X., Tai, Y.W., Tang, C.K.: Agentic 3D Scene Generation with Spatially Contextualized VLMs. arXiv preprint arXiv:2505.20129 (2025)

[7] Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., others, Sutskever, I.: Learning transferable visual models from natural language supervision. In: ICML, pp. 8748–8763 (2021)

[8] Liu, W., Wu, G., Wang, H., Ren, F.: Cross-modal data fusion via vision-language model for crop disease recognition. Sensors **25**(13), 4096 (2025)

[9] Kuo, W., Cui, Y., Gu, X., Piergiovanni, A.J., Angelova, A.: F-vlm: Open-vocabulary object detection upon frozen vision and language models. arXiv preprint arXiv:2209.15639 (2022)

[10] Huang, J., Li, Y., Feng, J., Wu, X., Sun, X., Ji, R.: Clover: Towards a unified video-language alignment and fusion model. In: CVPR, pp. 14856–14866 (2023)

[11] Wang, W., Lv, Q., Yu, W., Hong, W., Qi, J., Wang, Y., others, Ding, M.: Cogvlm: Visual expert for large language models (2023)

[12] Zhu, B., Niu, Y., Han, Y., Wu, Y., Zhang, H.: Prompt-aligned gradient for prompt tuning. In: ICCV, pp. 15659–15669 (2023)

[13] Kim, K., Oh, Y., Ye, J.C.: ZegOT: Zero-shot Segmentation Through Optimal Transport of Pixels to Text Prompts (2025)

[14] Das, D., Talon, D., Mancini, M., Wang, Y., Ricci, E.: One vlm to keep it learning: Generation and balancing for data-free continual visual question answering. In: 2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), pp. 5635–5645. IEEE (2025)

[15] Zhou, D.W., Zhang, Y., Wang, Y., Ning, J., Ye, H.J., Zhan, D.C., Liu, Z.: Learning without forgetting for vision-language models. IEEE TPAMI (2025)

[16] Jin, W., Wang, Z., Zhang, Y.: Category-instance distillation based on visual-language models for rehearsal-free class incremental learning. IET Computer Vision **19**(1), 12327 (2025)

[17] Zhang, T., Zhao, J., Zhu, Y., Cui, G.: Coupled Degradation Modeling and Fusion: A VLM-Guided Degradation-Coupled Network for Degradation-Aware Infrared and Visible Image Fusion. arXiv preprint arXiv:2510.11456 (2025)

[18] Taparia, A., Ngu, N., Leiva, M., Kricheli, J.S., Corcoran, J., Bastian, N.D., others, Senanayake, R.: VLC Fusion: Vision-Language Conditioned Sensor Fusion for Robust Object Detection. arXiv preprint arXiv:2505.12718 (2025)

[19] Sun, P., Song, Y., Zhu, X., Liu, X., Wang, Q., Liu, Y., others, Chu, X.: City-vlm: Towards multidomain perception scene understanding via multimodal incomplete learning. In: ACM MM, pp. 3448–3457 (2025)

[20] Fan, Z., Zhang, J., Li, R., Zhang, J., Chen, R., Hu, H., others, Ranjan, R.: VLM-3R: Vision-Language Models Augmented with Instruction-Aligned 3D Reconstruction. arXiv preprint arXiv:2505.20279 (2025)

[21] Sultan, R.I., Zhu, H., Li, C., Zhu, D.: BiPVL-Seg: Bidirectional Progressive Vision-Language Fusion with Global-Local Alignment for Medical Image Segmentation. arXiv preprint arXiv:2503.23534 (2025)

[22] Chen, K., Li, Y., Zhu, X., Zhang, W., Hu, B.: A vision-language model with multi-granular knowledge fusion in medical imaging. World Wide Web **28**(1), 5 (2025)

[23] Khattak, M.U., Wasim, S.T., Naseer, M., Khan, S., Yang, M.-H., Khan, F.S.: Self-regulating prompts: Foundational model adaptation without forgetting. In: ICCV, pp. 15190–15200 (2023)

[24] Kim, M.J., Finn, C., Liang, P.: Fine-tuning vision-language-action models: Optimizing speed and success. arXiv preprint arXiv:2502.19645 (2025)

[25] Zhang, J., Cai, K., Fan, Y., Wang, J., Wang, K.: CF-VLM: CounterFactual Vision-Language Fine-tuning. arXiv preprint arXiv:2506.17267 (2025)

[26] Hancock, A.J., Wu, X., Zha, L., Russakovsky, O., Majumdar, A.: Actions as language: Fine-tuning vlms into vlas without catastrophic forgetting. arXiv preprint arXiv:2509.22195 (2025)

[27] Tu, R.C., Jin, Z., Liao, J., Luo, X., Wang, Y., Shen, L., Tao, D.: MLLM-Guided VLM Fine-Tuning with Joint Inference for Zero-Shot Composed Image Retrieval. arXiv preprint arXiv:2505.19707 (2025)

[28] He, J., Tang, S., Liu, A., Cheng, L., Wu, J., Wei, Y.: Efficient vision language model fine-tuning for text-based person anomaly search. In: Companion Proceedings of the ACM on Web Conference 2025, pp. 1568–1572 (2025)

[29] Chen, P., Wu, Z., Sun, J., Wang, D., Zhou, P., Cao, N., others, Li, X.: Alignbot: Aligning vlm-powered customized task planning with user reminders through fine-tuning for household robots. In: 2025 IEEE International Conference on Robotics and Automation (ICRA), pp. 12549–12556. IEEE (2025)

[30] Wu, S., Zhang, J., Zeng, P., Gao, L., Song, J., Shen, H.T.: Skip tuning: Pre-trained vision-language models are effective and efficient adapters themselves. In: CVPR, pp. 14723–14732 (2025)

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

[32] Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: Imagenet: A large-scale hierarchical image database. In: CVPR, pp. 248–255. IEEE (2009)

[33] Li, F.-F., Andreetto, M., Ranzato, M., Perona, P.: Caltech 101 [Dataset]. CaltechDATA (2022). https://doi.org/10.22002/D1.20086 . https://doi.org/10.22002/D1.20086

[34] Parkhi, O.M., Vedaldi, A., Zisserman, A., Jawahar, C.V.: Cats and dogs. In: CVPR (2012)

[35] Krause, J., Stark, M., Deng, J., Fei-Fei, L.: 3d object representations for fine-grained categorization. In: Proceedings of the IEEE International Conference on Computer Vision Workshops, pp. 554–561 (2013)

[36] Nilsback, M.E., Zisserman, A.: Automated flower classification over a large number of classes. In: 2008 Sixth Indian Conference on Computer Vision, Graphics & Image Processing, pp. 722–729. IEEE (2008)

[37] Bossard, L., Guillaumin, M., Gool, L.V.: Food-101–mining discriminative components with random forests. In: ECCV, pp. 446–461. Springer (2014)

[38] Maji, S., Rahtu, E., Kannala, J., Blaschko, M., Vedaldi, A.: Fine-grained visual classification of aircraft. arXiv preprint arXiv:1306.5151 (2013)

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

[40] Cimpoi, M., Maji, S., Kokkinos, I., Mohamed, S., Vedaldi, A.: Describing textures in the wild. In: CVPR, pp. 3606–3613 (2014)

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

[42] Soomro, K., Zamir, A.R., Shah, M.: Ucf101: A dataset of 101 human actions classes from videos in the wild. arXiv preprint arXiv:1212.0402 (2012)

[43] Zhou, K., Yang, J., Loy, C.C., Liu, Z.: Conditional prompt learning for vision-language models. In: CVPR, pp. 16816–16825 (2022)

[44] Yao, H., Zhang, R., Xu, C.: Visual-language prompt tuning with knowledge-guided context optimization. In: CVPR, pp. 6757–6767 (2023)

[45] Khattak, M.U., Rasheed, H., Maaz, M., Khan, S., Khan, F.S.: Maple: Multi-modal prompt learning. In: CVPR, pp. 19113–19122 (2023)

[46] Lu, Y., Liu, J., Zhang, Y., Liu, Y., Tian, X.: Prompt distribution learning. In: CVPR, pp. 5206–5215 (2022)

[47] Chen, G., Yao, W., Song, X., Li, X., Rao, Y., Zhang, K.: Plot: Prompt learning with optimal transport for vision-language models. arXiv preprint arXiv:2210.01253 (2022)

[48] Yao, H., Zhang, R., Xu, C.: Tcp: Textual-based class-aware prompt tuning for visual-language model. In: CVPR, pp. 23438–23448 (2024)

[49] Hao, T., Ding, X., Feng, J., Yang, Y., Chen, H., Ding, G.: Quantized prompt for efficient generalization of vision-language models. In: ECCV, pp. 54–73. Springer (2024)