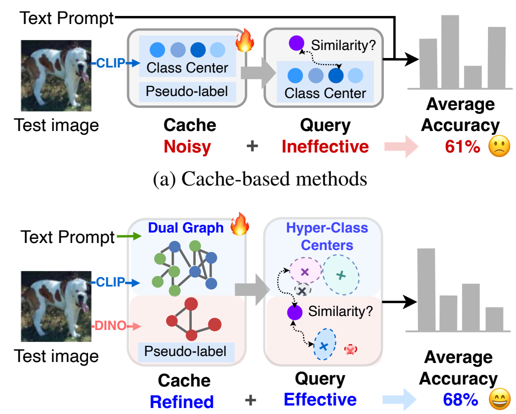
COSMIC: Clique-Oriented Semantic Multi-space Integration for Robust CLIP Test-Time Adaptation

Fanding Huang[1,*], Jingyan Jiang[2,*], Qinting Jiang[1], Hebei Li[3], Faisal Nadeem Khan[1,†], Zhi Wang[1,†]

[1]Shenzhen International Graduate School, Tsinghua University
[2]Shenzhen Technology University [3]University of Science and Technology of China

Abstract

*Recent vision-language models (VLMs) face significant challenges in test-time adaptation to novel domains. While cache-based methods show promise by leveraging historical information, they struggle with both caching unreliable feature-label pairs and indiscriminately using single-class information during querying, significantly compromising adaptation accuracy. To address these limitations, we propose **COSMIC** (Clique-Oriented Semantic Multi-space Integration for CLIP), a robust test-time adaptation framework that enhances adaptability through multi-granular, cross-modal semantic caching and graph-based querying mechanisms. Our framework introduces two key innovations: Dual Semantics Graph (DSG) and Clique Guided Hyper-class (CGH). The Dual Semantics Graph constructs complementary semantic spaces by incorporating textual features, coarse-grained CLIP features, and fine-grained DINOv2 features to capture rich semantic relationships. Building upon these dual graphs, the Clique Guided Hyper-class component leverages structured class relationships to enhance prediction robustness through correlated class selection. Extensive experiments demonstrate COSMIC's superior performance across multiple benchmarks, achieving significant improvements over state-of-the-art methods: 15.81% gain on out-of-distribution tasks and 5.33% on cross-domain generation with CLIP RN-50. Code is available at github.com/hf618/COSMIC.*



(a) Cache-based methods

(b) Clique-Oriented Semantic Multi-space Integration for CLIP (Ours)

Figure 1. (a) In the conventional cache-based method, the cache has only dull information with coarse-grained clip visual features and simple query way via similarity between samples and cached visual class centers. (b) In our COSMIC, the cache has diverse structural information via extra fine-grained DINOv2 visual features and effective query way via similarity between samples and meticulously designed hyper-class centers.

1. Introduction

Vision-language models (VLMs), such as CLIP [28] and ALIGN [17], have demonstrated remarkable performance across various downstream tasks, including semantic segmentation [21, 36] and video understanding [32, 35]. This success can be attributed primarily to the alignment of visual and textual features on large-scale datasets, enabling these models to exhibit robust image understanding capabilities in open-world scenarios. However, when deployed in real-world applications, these models often encounter test samples that significantly deviate from the training dataset, resulting in performance degradation [42, 43].

Recently, researchers [30] have explored the test-time adaptation (TTA) scenario for CLIP inference. TTA focuses on adapting models solely during testing by fine-tuning a small subset of the model's parameters, such as prompts (i.e., prompt learning), or even employing training-free cache-based methods to enhance the zero-shot image classification of CLIP. Prompt learning methods [1, 8, 30] optimize textual learnable prompts by minimizing entropy with confidence selection, ensuring consistent predictions across various augmented views of each test image. Using

[*]These authors contributed equally to this work.
[†]Co-corresponding authors.

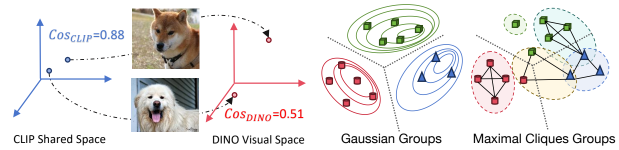
the updated prompt, it generates adapted textual class features, making predictions based on their similarity to sample features. While effective, these approaches suffer from computational inefficiency due to numerous visual augmentations and iterative backpropagation steps.

In contrast, as shown in Fig. 1 (a), cache-based methods [18, 41] enhance model performance by utilizing historical information. Specifically, they propose an extra cache to store visual features with pseudo-labels to generate visual *class centers* which are the average of previous visual features of each class. When querying the cache, features of the new test image are compared to find similar class centers. Then, labels corresponding to those centers are chosen to provide more information for the final prediction. Due to its training-free design and ability to leverage global historical information, the cache-based approach surpasses prompt learning in both effectiveness and efficiency.

However, two issues remain unaddressed in previous research: (1) Noisy pseudo-labels during cache construction contaminate the cache with unreliable feature-label pairs; (2) During querying, each class center encapsulates only single-class information, leading to blind propagation of a single pseudo-label regardless of its reliability. These dual flaws in cache construction and query mechanisms critically undermine adaptation accuracy.

To address the aforementioned issues, we introduce **COSMIC** (Clique-Oriented Semantic Multi-space Integration for CLIP), a robust CLIP test-time adaptation framework. Our key idea is to leverage multi-granularity and cross-modal semantic information to enrich the semantic content in the cache with limited samples while utilizing graph structures to organize and query the cache robustly. Specifically, as shown in Fig. 1 (b), we design two core components: (1) *Dual Semantics Graph (DSG)* can enhance the semantic diversity of pseudo-labels by incorporating fine-grained DINOv2 visual features and cross-modal text features. Additionally, it bridges the gap between textual semantics and more granular visual semantics, thereby reducing the proportion of noisy feature and pseudo-label pairs in the cache. (2) *Clique Guided Hyper-class (CGH)* connects class centers and merges pseudo-labels to include information from multiple categories, thus improving the probability of containing correct label information in retrieved pseudo-labels. This, in turn, allows for selective use of this information in subsequent prediction phases, improving the model's robustness and accuracy.

For the Dual Semantics Graph, we initially construct complementary feature spaces with varying semantic granularity: (1) The *CLIP Shared Semantic space* unifies text embeddings and visual class centers calculated by historical test features. (2) The *Auxiliary Fine-grained Visual space* incorporates fine-grained visual class centers from self-supervised DINOv2 [26]. As shown in Fig. 2 (a),



(a) Different Visual Affinity. (b) Different Feature Cluster.

Figure 2. (a) For various types of dogs, like "shiba inu" and "great Pyrenees" DINOv2 features offer more refined similarity perception than CLIP. (b) Samples in a single Gaussian cluster come from the same class, while the maximal clique cluster can explicitly represent cross-class correlations.

CLIP's visual discrimination is less refined than DINOv2's for subtly varied pet images. To capture intricate feature dynamics, we transform dual Euclidean spaces into dual graph spaces, modeling nonlinear interactions between classes.

For the Clique Guided Hyper-class, we design hyper-classes within each graph to represent latent semantics from various classes. Unlike previous approaches using Gaussian distributions [9], as shown in Fig. 2 (b), our graph-based feature modeling facilitates robust relationships with fewer samples and effectively clusters samples from diverse classes. Therefore, we can efficiently search for maximal cliques to form the Clique Guided Hyper-class, which represents the centroids of diverse, highly-affiliated class centers in the dual graph. Then test features query the hyper-class centers by similarity to select highly correlated classes, referred to as inlier classes. Based on the logits of these inlier classes, we generate adapted predictions for test-time adaptation, ensuring a more robust and accurate model performance.

In summary, our contributions are as follows:

- We introduce *COSMIC*, a training-free test-time adaptation method for CLIP that explicitly captures complementary semantic relationships.
- To refine cache construction, we design a *Dual Semantics Graph* that integrates intra-modal, cross-modal, coarse-grained, and fine-grained semantics.
- To query cache adaptively, we design *Clique Guided Hyper-Class* to represent class clusters with high affinity, enabling more robust querying of test samples.

2. Related Work

2.1. Vision-Language Models

Pre-trained on large and high-quality datasets, vision-language models like CLIP [28] achieve strong generalization by aligning textual and visual features via contrastive learning. For few-shot image classification, prompt learning improves adaptability by optimizing learnable tokens on the textual [42, 43] or visual [19] encoders.

To avoid backpropagation costs, methods like Tip-Adapter [38], CaFo [39] and GraphAdapter [22] use vision adapters for VLM generalization. Although these methods enrich feature semantics in few-shot settings, we focus on

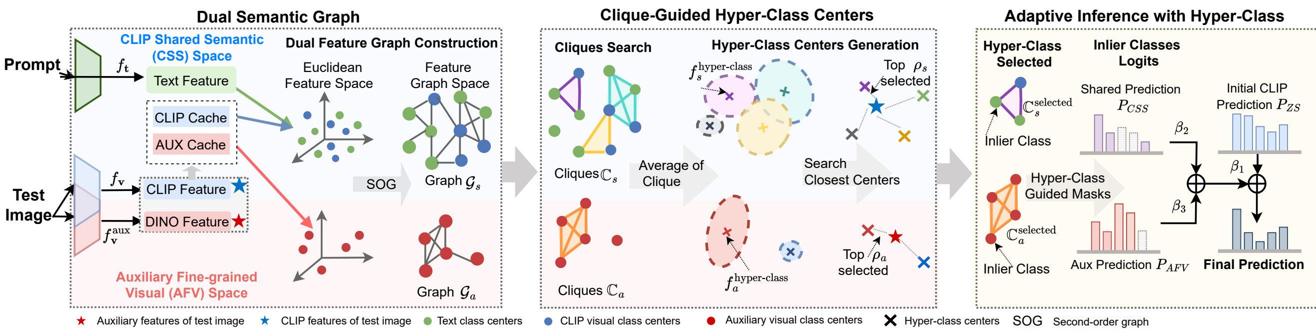


Figure 3. **Overview of COSMIC.** To refine cache with cross-modal, multi-granular class features, we construct *Dual Semantics Graph* with complementary semantics, incorporating both joint modalities and fine-grained visual information. To efficiently query the compatibility of diverse semantics, we propose novel *Clique Guided Hyper-class* to model different communities in the cache as the test domain evolves, enabling adaptive querying of test samples.

zero-shot test-time adaptation, rigorously exploring cross-modal and multi-granular class centers to address the non-linear relationships in new domains, better aligning with real-world deployment scenarios.

2.2. Test Time Adaptation

Test-time adaptation addresses distribution shifts between training and test data. Recent approaches for VLMs include: TPT [30] leveraging prompt tuning for view consistency; SwapPrompt [23] employing self-contrastive learning for rapid adaptation; and PromptAlign [1] aligning cross-domain statistical features.

In contrast, cache-based methods like TDA [18] and DMN [41] construct dynamic caches to retrieve adapted visual class centers during testing. While this method enables CLIP to generalize via historical samples, such caches rely solely on CLIP's coarse text-image aligned features—lacking vision-specific contrastive knowledge critical for similarity awareness, as demonstrated by DINOv2 [26]. Our work bridges this gap by enhancing cached features with textual and fine-grained visual semantics.

2.3. Pre-training of Visual Encoders

Modern vision systems predominantly rely on a two-stage paradigm: large-scale pretraining followed by task-specific fine-tuning. Despite its effectiveness, this approach poses limitations due to label dependency. Self-supervised learning mitigates this bottleneck by learning annotation-free representations, achieving state-of-the-art performance in semantic segmentation [14, 15] and depth estimation [16].

Transitioning from multimodal to view-consistent learning overcomes critical visual granularity limitations. While CLIP [28] excels in cross-modal alignment through vision-language pretraining, its text-centric objective compromises spatial details vital for dense prediction. DINOv2 [26] resolves this by learning augmentation-invariant features through self-distillation, preserving pixel-level nuances for annotation-free localization without textual guidance.

3. Method

3.1. Preliminaries

CLIP model consists of two parallel encoders: a text encoder F_t and a visual encoder F_v, which map features from different modalities into a shared embedding space. In a zero-shot K class classification task, given a test image $\boldsymbol{x}_{\text{test}}$, the visual feature $\boldsymbol{w}_{\mathbf{v}} = F_v(\boldsymbol{x}_{\text{test}})$ serves as a query to the text features of all class descriptions. The probability that $\boldsymbol{x}_{\text{test}}$ belongs to the class k can then be expressed as

$$p_k^{\text{zs}}(\boldsymbol{x}_{\text{test}}) = \frac{\exp(\cos(\boldsymbol{w}_{\mathbf{v}}, f_{\mathbf{t}_k})/\tau)}{\sum_{i=1}^{K}\exp(\cos(\boldsymbol{w}_{\mathbf{v}}, f_{\mathbf{t}_i})/\tau)}, \tag{1}$$

where $f_{\mathbf{t}_i}$ represents the textual class feature for the i^{th} class, and τ denotes the temperature parameter in the softmax function.

Inspired by few-shot adaptation of CLIP method Tip-Adapter [38], **TDA** [18] is a typical cache-based method for TTA that utilizes a cache to store knowledge from high-confidence historical test samples. This cache operates as a queue-style dictionary, where the keys represent the pseudo-labels of the samples, and the values store the corresponding visual features. The decision to add a sample to the cache depends on its prediction entropy. Specifically, for a test sample $\boldsymbol{x}_{\text{test}}$, after undergoing random $\mathcal{N}$ augmentations $\mathcal{A}$ (including the original image), the marginal entropy of the sample predictions is calculated as:

$$\begin{aligned} H(\boldsymbol{x}_{\text{test}}) &= -\sum_{i=1}^{K}\tilde{p}_i(\boldsymbol{x}_{\text{test}})\log\tilde{p}_i(\boldsymbol{x}_{\text{test}}), \\ \tilde{p}_i(\boldsymbol{x}_{\text{test}}) &= \frac{1}{\mathcal{N}_{\max}}\sum_{j=1}^{\mathcal{N}_{\max}} p_i^{\text{zs}}(\mathcal{A}_j(\boldsymbol{x}_{\text{test}})), \end{aligned} \tag{2}$$

where $p_i^{\text{zs}}(\mathcal{A}_j(\boldsymbol{x}_{\text{test}}))$ represents probability of class i generated by the model for the j^{th} augmented view of the test image. N_{max} is the number of first $R \times \mathcal{N}$ high-confidence view predictions, where R is the selection ratio. Based on this, the logic for storing $\boldsymbol{w}_{\mathbf{v}}$ predicted as class i into cache set $\mathbb{M}_i$ is as follows:

$$\mathbb{M}_i = \begin{cases} \mathbb{M}_i \cup \boldsymbol{w}_{\mathbf{v}}, & \text{if } |\mathbb{M}_i| < L, \\ \mathbb{M}_i \setminus \{f_{\max}\} \cup \boldsymbol{w}_{\mathbf{v}}, & \text{if } |\mathbb{M}_i| = L \text{ and} \\ & H(\boldsymbol{x}_{\text{test}}) < H_{\max}(\mathbb{M}_i), \\ \mathbb{M}_i, & \text{otherwise.} \end{cases} \tag{3}$$

Here, $f_{\max}$ denotes the feature in cached features set $\mathbb{M}_i$ with the highest entropy $H_{\max}(\mathbb{M}_i)$, and L denotes the maximum capacity of the cache. Using this cache, the adapted predictions are computed as follows:

$$\mathbf{P}_{\text{adapted}}(\boldsymbol{x}_{\text{test}}) = \varphi(\boldsymbol{w}_{\mathbf{v}}\mathbf{M}^\top)\mathbf{L}_{\text{pseudo}}, \tag{4}$$

where $\varphi(x) = \exp(-\alpha(1-x))$ is an adaptation function, $\mathbf{M} \in \mathbb{R}^{KL\times d}$ is the cached features matrix, and $\mathbf{L}_{\text{pseudo}} \in \mathbb{R}^{KL\times K}$ is the one-hot pseudo-label matrix.

3.2. Dual Semantics Graph (DSG)

3.2.1. CLIP Shared Semantic (CSS) Space

CLIP aims to align text and visual features in a shared feature space. Leveraging this, we combine text features $\mathbf{f}_{\mathbf{t}} = [f_{\mathbf{t}_1} f_{\mathbf{t}_2} \cdots f_{\mathbf{t}_K}]^\top \in \mathbb{R}^{K\times d_1}$ with cached visual class centers $\mathbf{f}_{\mathbf{v}} = [f_{\mathbf{v}_1} f_{\mathbf{v}_2} \cdots f_{\mathbf{v}_K}]^\top \in \mathbb{R}^{K\times d_1}$ in the same Euclidean space, forming a unified feature set $\mathbf{f}_{\mathbf{s}} = \mathbf{f}_{\mathbf{t}} \cup \mathbf{f}_{\mathbf{v}} = [f_{\mathbf{s}_1} f_{\mathbf{s}_2} \cdots f_{\mathbf{s}_{2K}}]^\top \in \mathbb{R}^{2K\times d_1}$, termed the CLIP Shared Semantic space with d_1 dimension. The text feature $f_{\mathbf{t}_i}$ are CLIP embeddings of class i description, defined as:

$$f_{\mathbf{t}_i} = F_t(\text{Text}_i), \tag{5}$$

where Text_i is the input text of class i. Inspired by DMN [41], we construct the visual center $f_{\mathbf{v}_i}$ for class i using a weighted combination of visual features, determined by the cosine similarity between test and cached visual features, illustrated as:

$$f_{\mathbf{v}_i} = \varphi\left(\boldsymbol{w}_{\mathbf{v}}\mathbf{M}_i^\top\right)\mathbf{M}_i, \tag{6}$$

where $\mathbf{M}_i \in \mathbb{R}^{l_1\times d_1}$ represents the cached visual feature matrix from the i^{th} category with a capacity l_1 of cache.

3.2.2. Auxiliary Fine-grained Visual (AFV) Space

While the CLIP Shared Semantic space leverages both text and visual features for alignment, the cache model eliminates the need for text features during inference. However, relying solely on the CLIP visual encoder under noisy pseudo-labels can lead to inaccurate similarity perception, especially for easily confused images. Therefore, we introduce an auxiliary visual feature branch using self-supervised encoders like DINOv2 [26] to generate more finer-grained visual features of $\boldsymbol{x}_{\text{test}}$.

Analogously to the CLIP visual feature cache, we establish a DINOv2 visual feature cache using the same criteria as in Eq. 3. To average historical test information with fine-grained semantics, we compute the auxiliary visual center $f^{\text{aux}}_{\mathbf{v}_i}$ via the centroid of cached DINOv2 features for class i:

$$\begin{aligned} \mathbf{f}^{\text{aux}}_{\mathbf{v}} &= \left[f^{\text{aux}}_{\mathbf{v}_1} \quad f^{\text{aux}}_{\mathbf{v}_2} \quad \ldots \quad f^{\text{aux}}_{\mathbf{v}_C}\right]^\top \in \mathbb{R}^{K\times d_2}, \\ f^{\text{aux}}_{\mathbf{v}_i} &= \frac{1}{l_2}\sum_{j=1}^{l_2}\mathbf{M}^{\text{aux}}_i[j,:], \end{aligned} \tag{7}$$

where l_2 and d_2 are the capacity of the auxiliary cache and dimension of AFV space, respectively. $\mathbf{M}^{\text{aux}}_i \in \mathbb{R}^{l_2\times d_2}$ is the auxiliary cached visual feature matrix for the i^{th} class.

3.2.3. Second-Order Graph Construction

In contrast to Euclidean space, graph space provides a more precise representation of complex affinity relationships [37, 40] between intra- and inter-modal features. Specifically, we construct two types of graphs: the First-Order Graph (FOG), which captures direct pairwise similarities, and the Second-Order Graph (SOG), which extends this by modeling higher-order relationships through the product of adjacency matrices, enabling the discovery of more complex and hidden patterns. In each FOG, nodes represent class center features, and edges capture bidirectional compatibility. Let $\mathbf{F} \in \mathbb{R}^{N\times D}$ represent the feature matrix, where $N = 2K, D = d_1$ for CSS space and $N = K, D = d_2$ for AFV space. The adjacency matrix of First-Order Graph is defined as:

$$\mathcal{W}_{FOG} = [w_{ij}]_{N\times N},\ w_{ij} = \begin{cases} 1 & \text{if } \frac{\mathbf{F}_i\cdot\mathbf{F}_j}{\|\mathbf{F}_i\|\|\mathbf{F}_j\|} > t^{\text{aff}} \\ 0 & \text{otherwise} \end{cases}, \tag{8}$$

where t^{aff} is a threshold changed by test iterations. Overly concentrated class centers in specific domains can lead to excessive redundancy in graph information. To address this, the Second-Order Graph (SOG) captures higher-order dependencies through the adjacency matrix:

$$\mathcal{W}_{SOG} = \mathcal{W}_{FOG} \odot (\mathcal{W}_{FOG} \times \mathcal{W}_{FOG}). \tag{9}$$

Unlike FOG, SOG models indirect relationships, such as transitive or multi-hop connections, providing a more comprehensive view of node interactions. Additionally, the SOG is sparser than the FOG, reducing noise and improving computational efficiency. As testing progresses, the CSS and AFV graphs accumulate higher-quality features from new domains. To adapt to this dynamic, we adopt an increasing rule for the threshold $t^{\text{aff}}_i = \min\left(1, t^{\text{aff}}_0 + g\cdot i\right)$ where t^{aff}_0 is the initial value, i is the current number of samples tested, and g is a constant growth rate.

3.3. Clique Guided Hyper-class (CGH)

Maximal Cliques Search. Given an undirected graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, clique $\mathcal{C} = (\mathcal{V}', \mathcal{E}'), \mathcal{V}' \subseteq \mathcal{V}, \mathcal{E}' \subseteq \mathcal{E}$ is a complete subgraph of $\mathcal{G}$. A maximal clique is a clique that cannot be extended by including any additional node. We employ the modified Bron-Kerbosch algorithm [6] for maximal clique search, which is a highly efficient algorithm with a worst-case time complexity of $\mathcal{O}(b(n-b)3^{(b/3)})$, where b denotes the graph's degeneracy. When the t^{aff} is high, the graph becomes sparser, leading to a lower degeneracy, which further accelerates the search process. For CSS and AFV spaces, we define their graphs:

$$\mathcal{G}_{\text{CSS}} = (\mathcal{V}_{\text{CSS}}, \mathcal{E}_{\text{CSS}}), \mathcal{G}_{\text{AFV}} = (\mathcal{V}_{\text{AFV}}, \mathcal{E}_{\text{AFV}}), \tag{10}$$

where $\mathcal{G}_{\text{CSS}}$ and $\mathcal{G}_{\text{AFV}}$ represent the graphs in the CSS and AFV spaces, respectively. For each graph $j \in \{\text{CSS}, \text{AFV}\}$, we apply the modified Bron-Kerbosch algorithm to both graphs:

$$\mathbb{C}_j = \text{BronKerbosch}(\mathcal{G}_j), \tag{11}$$

where $\mathbb{C}_j = \{\mathcal{C}_{j1}, \mathcal{C}_{j2}, ..., \mathcal{C}_{jm_j}\}$ is the set of maximal cliques in graph $\mathcal{G}_j$. These two sets of maximal cliques, $\mathbb{C}_{\text{CSS}}$ and $\mathbb{C}_{\text{AFV}}$, enable us to capture a more comprehensive understanding of feature affinities across diverse levels of semantics and granularities.

Hyper-classes Generation. Building on the maximal cliques identified in both $\mathcal{G}_{\text{CSS}}$ and $\mathcal{G}_{\text{AFV}}$, we further reveal latent patterns by searching for *Clique-Guided Hyper-classes*. Given the dense connectivity within maximal cliques, we define hyper-class centers as the centroids of node classes within clique i.

$$f_{ji}^{\text{hyper-class}} = \frac{1}{|\mathcal{C}_{ji}|} \sum_{f \in \mathcal{C}_{ji}} f, \tag{12}$$

where f represents class nodes in the respective graph. This approach models community structures between class centers through graph-based contextual reasoning, overcoming the locality bias of similarity-query paradigms. The affinity between test image feature $\boldsymbol{w_v}$ and clique i in $\mathcal{G}_j$ is:

$$\rho_j(\boldsymbol{w}_{\mathbf{v}}, \mathcal{C}_{ji}) = \cos\left(\boldsymbol{w}_{\mathbf{v}}, f_{ji}^{\text{hyper-class}}\right). \tag{13}$$

To identify the hyper-classes most proximal to the test sample, we sort the cliques by affinity: $\rho_j(\boldsymbol{w}_{\mathbf{v}}, \mathcal{C}_{j(1)}) \geq \rho_j(\boldsymbol{w}_{\mathbf{v}}, \mathcal{C}_{j(2)}) \geq \cdots \geq \rho_j(\boldsymbol{w}_{\mathbf{v}}, \mathcal{C}_{j(m_j)})$. Then we select the top r proportion of the closest hyper-classes in each graph:

$$\mathbb{C}_j^{\text{selected}} = \{\mathcal{C}_{j(1)}, \mathcal{C}_{j(2)}, \ldots, \mathcal{C}_{j(k_j)}\}, \tag{14}$$

where $k_j = \lceil r \cdot m_j \rceil$. This procedure generates two hyper-class guided masks $\mathcal{M}_j$, effectively delineating the selected classes as inliers while designating the remainder as outliers (classes with low relevance to the test sample).

$$\begin{aligned} \mathcal{M}_j &= \begin{bmatrix} m_{j1} & m_{j2} & \ldots & m_{jN_j} \end{bmatrix} \in \mathbb{R}^{1\times N_j}, \\ m_{ji} &= \begin{cases} 1, & \text{if } i \in \bigcup_{\mathcal{C}\in\mathbb{C}_j^{\text{selected}}} \mathcal{C} \\ 0, & \text{otherwise} \end{cases}, \end{aligned} \tag{15}$$

where $N_{\text{CSS}} = 2K$ for the CLIP Shared Semantic space and $N_{\text{AFV}} = K$ for the Auxiliary Fine-grained Visual space.

3.4. Adaptive Inference with Hyper-class

The $\mathcal{M}_{\text{CSS}}$ is then applied to the initial logits of all classes in the CSS space:

$$\mathbf{P}_{\text{CSS}}^{\text{initial}} = \text{Softmax}(\boldsymbol{w}_{\mathbf{v}}\mathbf{f_s}^{\top}) \odot \mathcal{M}_{\text{CSS}} \in \mathbb{R}^{1\times 2K}. \tag{16}$$

The probability p_i for the i^{th} class in the CSS space is calculated as the average of the text and image class predictions:

$$p_i = (p_i^{\text{initial}} + p_{i+K}^{\text{initial}})/2, \quad \text{for} \quad i = 1, 2, \ldots, K. \tag{17}$$

Let p_i^{initial} denote the prediction value in the i^{th} column of $\mathbf{P}_{\text{CSS}}^{\text{initial}}$. The classification probability distribution in the CSS space is then given by:

$$\mathbf{P}_{\text{CSS}} = \begin{bmatrix} p_1 & p_2 & \ldots & p_K \end{bmatrix} \in \mathbb{R}^{1\times K}, \tag{18}$$

where K is the number of classes. Following a similar process, we obtain the class prediction probability distribution within the AFV space:

$$\mathbf{P}_{\text{AFV}} = \text{Softmax}(\boldsymbol{w}_{\mathbf{v}}^{\text{aux}}\mathbf{f}_{\mathbf{v}}^{\text{aux}\top}) \odot \mathcal{M}_{\text{AFV}} \in \mathbb{R}^{1\times K}, \tag{19}$$

where $\boldsymbol{w}_{\mathbf{v}}^{\text{aux}}$ is the auxiliary image feature of test image $\boldsymbol{x}_{\text{test}}$. The calculation of the $\mathcal{M}_{\text{AFV}}$ is similar to $\mathcal{M}_{\text{CSS}}$, but they mainly filter out outliers from various feature clusters, thereby preventing prediction bias caused by the accumulation of moderate logits values. We combine initial CLIP prediction with adaptive prediction into the final prediction:

$$\begin{aligned} \mathbf{P}_{\text{Final}} &= \beta_1\mathbf{P}_{\text{ZS}} + \beta_2\mathbf{P}_{\text{CSS}} + \beta_3\mathbf{P}_{\text{AFV}}, \\ \mathbf{P}_{\text{ZS}} &= \text{Softmax}(\boldsymbol{w}_{\mathbf{v}}\mathbf{f_t})^{\top} \in \mathbb{R}^{1\times K}, \end{aligned} \tag{20}$$

where $\beta_1, \beta_2, \beta_3$ represent the weights, and their values will be discussed in detail in Sec. 4.3.3.

4. Experiments

4.1. Experimental Settings

Datasets. In the VLM test-time adaptation setting, two main benchmark types are typically used. The first evaluates the model's robustness under out-of-distribution (OOD) shifts, while the second examines cross-domain (CD) generalization capabilities.

- For OOD shifts, we employ the ImageNet validation [4] along with four variants: ImageNet-A [13], ImageNet-V2 [29], ImageNet-R [12], and ImageNet-Sketch [33].
- For CD tasks, we utilize ten diverse sub-datasets, each representing a distinct domain: Aircraft [24], Caltech101 [7], Cars [20], DTD [3], EuroSAT [11], Flower102 [25], Food101 [2], Pets [27], SUN397 [34], and UCF101 [31].

Table 1. Top-1 accuracy (%) comparison on ImageNet and its OOD variants using CLIP with ResNet-50 and ViT-B/16 backbones. Our results are reported as mean±std over 3 random seeds. **Bold** indicates the highest performance.

Method	Adaptation Settings	ImageNet	ImageNet-A	ImageNet-V2	ImageNet-R	ImageNet-S	Average	OOD Average
CLIP-RN-50	-	58.16	21.83	51.41	56.15	33.37	44.18	40.69
CoOp [43] (IJCV'22)	Training Few-shot	63.33	23.06	55.40	56.60	34.67	46.61	42.43
CoCoOp [42] (CVPR'22)	Training Few-shot	62.81	23.32	55.72	57.74	34.48	46.81	42.82
TPT [30] (NeurIPS'22)	Training Zero-shot	60.74	26.67	54.70	59.11	35.09	47.26	43.89
DiffTPT [8] (ICCV'23)	Training Zero-shot	60.80	31.06	55.80	58.80	37.10	48.71	45.69
TDA [18] (CVPR'24)	Training-free Zero-shot	61.35	30.29	55.54	62.58	38.12	49.58	46.63
DMN [41] (CVPR'24)	Training-free Zero-shot	63.87	28.57	56.12	61.44	39.84	49.97	46.49
COSMIC (Ours)	Training-free Zero-shot	**75.19**±0.89	**49.07**±0.06	**63.81**±0.39	**79.87**±0.08	**56.44**±0.34	**64.88**±0.22	**62.30**±0.05
Method	Adaptation Settings	ImageNet	ImageNet-A	ImageNet-V2	ImageNet-R	ImageNet-S	Average	OOD Average
CLIP-ViT-B/16	-	66.73	47.87	60.86	73.98	46.09	59.11	57.20
CoOp [43] (IJCV'22)	Training Few-shot	71.51	49.71	64.20	75.21	47.99	61.72	59.28
CoCoOp [42] (CVPR'22)	Training Few-shot	71.02	50.63	64.07	76.18	48.75	62.13	59.91
TPT [30] (NeurIPS'22)	Training Zero-shot	68.98	54.77	63.45	77.06	47.94	62.44	60.81
DiffTPT [8] (ICCV'23)	Training Zero-shot	70.30	55.68	65.10	75.00	46.80	62.58	60.65
TDA [18] (CVPR'24)	Training-free Zero-shot	69.51	60.11	64.67	80.24	50.54	65.01	63.89
DMN [41] (CVPR'24)	Training-free Zero-shot	72.25	58.28	65.17	78.55	53.20	65.49	63.80
COSMIC (Ours)	Training-free Zero-shot	**78.19**±0.56	**73.32**±0.32	**69.62**±0.19	**85.60**±0.12	**62.79**±0.10	**73.90**±0.05	**72.83**±0.10

Implement Settings. We employed CLIP's officially pre-trained vision encoders, including ResNet50 [10] and ViT-B/16 [5]. Inspired by previous works [41], we utilized hand-crafted textual prompts and multi-view augmentations of the original images to calculate prediction confidence. For all experiments, we generated 16 views per image and fixed hyperparameters across each sub-dataset. To combine logits, we used an adaptive step search to determine optimal weights while also demonstrating the robustness of fixed weights. Unless otherwise specified, we employ DINOv2 ViT-L/14 as the auxiliary visual encoder. All experiments were conducted on a single Tesla V100S-PCIE-32GB GPU, using top-1 accuracy for classification performance.

Comparison Methods. We initially evaluate few-shot methods such as CoOp [43] and CoCoOp [42], using 16-shot annotated samples per class. Next, we assess training-based zero-shot methods. TPT [30] straightforwardly optimizes the text prompts by minimizing the multi-view marginal entropy. DiffTPT [8] is an advanced iteration of TPT, employing diffusion-based augmentations to refine prompts. Furthermore, we evaluate training-free zero-shot methods. TDA [18] is an adapter-based approach that builds positive and negative caches at test time without training. DMN [41] leverages a dynamic memory to compile information from past test data, bypassing backpropagation.

4.2. Comparison with SOTA

4.2.1. Results on the OOD Benchmark

In Tab. 1, we compare model performance on the in-domain ImageNet validation set and four OOD variants. Due to domain shifts, the zero-shot generalization ability of CLIP is limited in OOD scenarios. Notably, while CoOp [43] and CoCoOp [42] enhance CLIP's transferability by fine-tuning prompts on few-shot samples, this incurs additional training costs, making it impractical for real-world deployment. In contrast, our proposed method is training-free, allowing for immediate adaptation to unseen domains. Furthermore, our method achieves substantial improvements over CoCoOp, with gains of 19.48% and 12.92% across different CLIP visual encoders, demonstrating its superior performance in OOD generalization tasks.

Additionally, our method outperforms previous test-time adaptation techniques. For prompt-learning-based methods, our approach surpasses TPT [30] and DiffTPT [8], with improvements of 18.41%, 12.02%, and 16.61%, 12.18% on two different backbones with OOD shifts, respectively, highlighting the effectiveness of our training-free approach. In comparison to cache model-based methods, our approach achieves superior results over the strongest techniques, TDA [18] and DMN [41], with gains of 15.67%, 15.81%, and 8.94%, 9.03% on two backbones with OOD shifts, respectively, validating the effectiveness of our Dual Semantics Graph and Clique Guided Hyper-class to improve the refinement and querying in cache.

4.2.2. Results on the CD Benchmark

We conducted a comparison on a diverse cross-domain dataset against a variety of contemporary methods, as shown in Tab. 2. Notably, both few-shot and zero-shot prompt learning methods show limited improvements in generalization performance. This limitation arises because, when there is a significant shift both in textual and visual modality, it becomes challenging to identify an optimal prompt within a constrained parameter space. By circumventing complex prompt space searches, our method adaptively bridges domain gaps through feature affinity mod-

Table 2. Top-1 accuracy (%) comparison on 10 diverse cross-domain datasets using CLIP with ResNet-50 and ViT-B/16 backbones. Our results are reported as mean±std over 3 random seeds. **Bold** indicates the highest performance.

Method	Adaptation Settings	Aircraft	Caltech101	Cars	DTD	EuroSAT	Flower102	Food101	Pets	SUN397	UCF101	Average
CLIP-RN-50	-	15.66	85.88	55.70	40.37	23.69	61.75	73.97	83.57	58.80	58.84	55.82
CoOp [43] (IJCV'22)	Training Few-shot	15.12	86.53	55.32	37.29	26.20	61.55	75.59	87.00	58.15	59.05	56.18
CoCoOp [42] (CVPR'22)	Training Few-shot	14.61	87.38	56.22	38.53	28.73	65.57	76.20	88.39	59.61	57.10	57.23
TPT [30] (NeurIPS'22)	Training Zero-shot	17.58	87.02	58.46	40.84	28.33	62.69	74.88	84.49	61.46	60.82	57.66
DiffTPT [8] (ICCV'23)	Training Zero-shot	17.60	86.89	60.71	40.72	41.04	63.53	79.21	83.40	62.72	62.67	59.85
TDA [18] (CVPR'24)	Training-free Zero-shot	17.61	89.70	57.78	43.74	42.11	68.74	77.75	86.18	62.53	64.18	61.03
DMN [41] (CVPR'24)	Training-free Zero-shot	22.77	90.14	60.02	50.41	48.72	67.93	76.70	86.78	64.39	65.34	63.32
COSMIC (Ours)	Training-free Zero-shot	**25.49**±0.57	**94.77**±0.35	**66.69**±0.51	**55.44**±0.68	**48.97**±2.41	**77.78**±1.22	**83.53**±0.11	**92.63**±0.29	**69.73**±0.08	**71.51**±0.78	**68.65**±0.32

Method	Adaptation Settings	Aircraft	Caltech101	Cars	DTD	EuroSAT	Flower102	Food101	Pets	SUN397	UCF101	Average
CLIP-ViT-B/16	-	23.22	93.55	66.11	45.04	50.42	66.99	82.86	86.92	65.63	65.16	64.59
CoOp [43] (IJCV'22)	Training Few-shot	18.47	93.70	64.51	41.92	46.39	68.71	85.30	89.14	64.15	66.55	63.88
CoCoOp [42] (CVPR'22)	Training Few-shot	22.29	93.79	64.90	45.45	39.23	70.85	83.97	90.46	66.89	68.44	64.63
TPT [30] (NeurIPS'22)	Training Zero-shot	24.78	94.16	66.87	47.75	42.44	68.98	84.67	87.79	65.50	68.04	65.10
DiffTPT [8] (ICCV'23)	Training Zero-shot	25.60	92.49	67.01	47.00	43.13	70.10	87.23	88.22	65.74	62.67	64.92
TDA [18] (CVPR'24)	Training-free Zero-shot	23.91	94.24	67.28	47.40	58.00	71.42	86.14	88.63	67.62	70.66	67.53
DMN [41] (CVPR'24)	Training-free Zero-shot	30.03	95.38	67.96	55.85	**59.43**	74.49	85.08	92.04	70.18	72.51	70.30
COSMIC (Ours)	Training-free Zero-shot	**31.44**±0.56	**96.80**±0.42	**71.31**±0.46	**58.23**±1.40	58.82±0.40	**82.14**±0.49	**86.60**±0.09	**94.19**±0.09	**72.33**±0.06	**76.20**±0.55	**72.81**±0.09

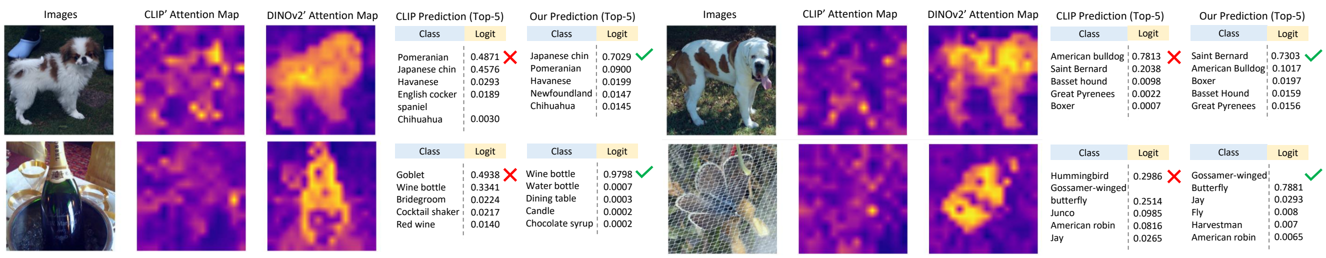


Figure 4. Visualization of the attention maps from CLIP ViT-B/16 and DINOv2 ViT-B/14, along with their top-5 prediction results.

Table 3. Efficiency and effectiveness comparison of TTA methods on Flower102 [25] dataset using CLIP ViT-B/16.

Method	Views	Time Per Image (s)	Top-1 Accuracy (%)	Gain (%)
CLIP [28]	1	0.098	66.99	-
TPT [30]	64	0.198	68.98	1.99
TDA [18]	64	**0.119**	71.42	4.43
DMN [41]	128	0.483	74.49	7.50
COSMIC	**16**	0.368	**82.46**	**15.47**

eling. For example, our approach achieves gains over TPT [30] of 10.99% and 7.71% across two CLIP visual backbones, respectively, validating its robustness.

Similarly, cache model-based methods typically adapt solely to the image distribution of the new domain. However, our method unifies query processing for both textual and cached image features, enabling it to adapt to cross-modal information distributions simultaneously. Our approach achieves improvements of 7.62% and 5.28% over TDA [18] across two backbones, further supporting the efficacy of our unified strategy.

4.2.3. Computation Efficiency

As shown in Tab. 3, we evaluated the efficiency of the Flowers dataset [25] on a single Tesla V100S-PCIE-32GB GPU. Note that the number of augmentation views is based on the specifications stated in the original papers for each method. By using a minimal number of image augmentations, we effectively avoided excessive overhead while achieving a 15.47% improvement over the original CLIP. Additionally, our supplementary material includes analysis and experiments on accelerating inference, highlighting further optimization potential.

Table 4. Graph ablation with CLIP ViT-B/16 and DINOv2 ViT-S/14. CSS denotes the prediction from the CLIP Shared Semantic Graph, while AFV represents the prediction from the Auxiliary Fine-grained Visual Graph.

CLIP	CSS	AFV	ImageNet-Val [4]		Flower102 [25]	
			Top-1 Accuracy (%)	Gain (%)	Top-1 Accuracy (%)	Gain (%)
✓			69.91	-	72.64	-
	✓		71.88	1.97	74.46	1.82
		✓	67.05	-2.86	77.71	5.07
✓	✓		72.06	2.15	75.07	2.43
✓		✓	73.16	3.25	79.74	7.10
✓	✓	✓	**73.79**	**3.88**	**81.61**	**8.97**

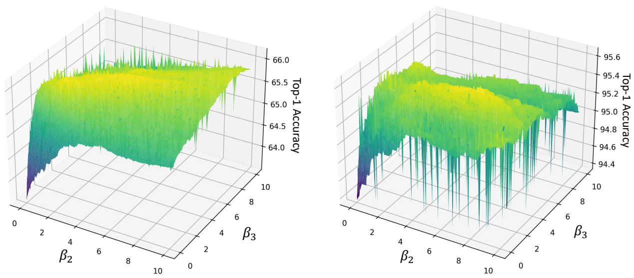
4.3. Ablation Study

4.3.1. Ablation of Graph Type

We evaluated the performance of different graphs on ImageNet Val [4] and Flowers102 [25]. As shown in the Tab. 4, when CSS and AFV prediction are fused with the original CLIP predictions, significant improvements are observed, with gains of 2.15%, 3.25%, and 2.43%, 7.10%, respectively. This indicates that each graph contributes to enhancing the generalization ability of CLIP. Moreover, the performance is optimized when CSS and AFV are used in

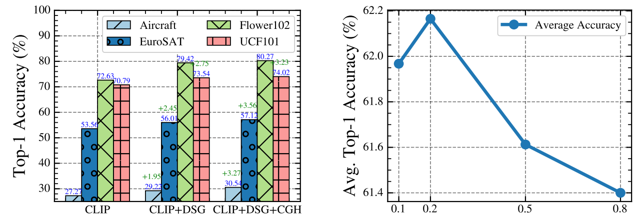
Table 5. Ablation of various auxiliary visual encoder in ImageNet Val [4] with CLIP ViT-B/16.

DINOv2 Backbone	Extra Parameters	Top-1 Accuracy (%)	Gain (%)
-	-	69.91	-
ViT-S/14	21 M	73.79	3.88
ViT-B/14	86 M	76.73	6.82
ViT-L/14	300 M	**77.54**	**7.63**



(a) ImageNet-V2 [29] (b) Caltech101 [7]

Figure 5. Ablation of β_2, β_3 with CLIP ViT-B/16 and DINOv2 ViT-S/14.



(a) Ablation of key modules. (b) Ablation of r.

Figure 6. Ablation of components with CLIP-ViT-B16.

conjunction, achieving improvements of 3.88% and 8.97%, which validates that the different target feature spaces enhance the perception of shared modal features and fine-grained visual capabilities. However, we note a -2.86% degradation when using AFV alone on ImageNet-Val, as relying solely on cached DINOv2 visual features is insufficient to accurately describe new images with domain gaps.

4.3.2. Ablation of Auxiliary Visual Encoder

Introducing an auxiliary visual encoder with various sizes consistently improves model accuracy, as shown in Tab. 5. Specifically, each encoder boosts accuracy over the baseline (69.91%), with gains of 3.67% (ViT-S/14), 6.82% (ViT-B/14), and 7.63% (ViT-L/14). Furthermore, the Tab. 5 highlights a clear trade-off between performance gains and additional parameters. The smallest encoder, ViT-S/14, with only 21M extra parameters, shows modest improvement, while the largest, ViT-L/14, requires 300M additional parameters for the highest gain.

4.3.3. Ablation of Logits Weights

To verify the robustness of different predictive logits on the final prediction in Eq. 20, we fixed the value of β_1 at 1 and set the step size to 0.05, varying β_2 and β_3 from 0 to 10 to observe their impact on model prediction accuracy. As shown in Fig. 5, we visualized this on two representative subsets ImageNet-V2 [29] and Caltech101 [7]. As shown, increasing either β_2 or β_3 significantly enhances accuracy, indicating that both semantic granularities independently improve generalization. These parameters exhibit strong synergistic properties, mutually promoting the fusion of complementary semantics and thereby achieving superior performance. The model demonstrates robustness, maintaining high accuracy even with perturbations around the optimal β values, suggesting that performance is resilient to variations in these hyperparameters.

4.3.4. Ablation of Components

In Fig. 6 (a), CLIP+DSG is equivalent to directly computing the similarity between test samples and class nodes within the DSG. Although DSG independently enhances model generalization, CGH highlights the importance of selecting an appropriate r to enhance the model's robustness by effectively filtering out noise and outliers, thereby improving prediction accuracy. As shown in Fig. 6 (b), we varied the selection ratio r of nearest anchors across multiple datasets to evaluate the impact of different components on model performance. The results indicate that the model achieves optimal performance at $r = 0.2$ with both DSG and CGH. This suggests that the similarity between test samples and diverse hyper-classes is more robust than that between samples and normal classes.

4.3.5. Visualization

As shown in Fig. 4, DINOv2's vision-contrastive encoder outperforms CLIP in capturing key objects. Leveraging finer-grained features, our method excels in recognizing "hard" classes, accurately distinguishing similar pets (e.g., Pomeranian vs. Japanese Chin) with higher confidence.

5. Conclusion

In CLIP test-time adaptation, we focus on exploring the potential of the cache-based method. As for how to refine cache, we introduced a *Dual Semantics Graph* to explore inter- and cross-modal affinities with various semantics. As for how to query cache, we introduce *Clique Guided Hyper-classes* within dual graphs to enhance the selection of correlated classes for robust predictions. Our method outperforms SOTA on multiple benchmarks, demonstrating strong zero-shot capabilities, while revealing the significant potential of structured semantic integration for robust visual understanding. **Limitation:** The current clique search methodology incurs non-trivial time complexity, imposing an additional computational burden on COSMIC. To address this, we aim to explore dual graph sparsification techniques for accelerated search in future iterations.

Acknowledgment

This work was supported by the following funding sources: the National Key Research and Development Project of China (Grant No. 2023YFF0905502), the National Natural Science Foundation of China (Grant Nos. 92467204 and 62472249), the Shenzhen Science and Technology Program (Grant Nos. JCYJ20220818101014030 and KJZD20240903102300001), the Scientific Research Startup Fund of TSIGS (Project No. QD2022004C), the National Natural Science Foundation of China (Grant No. W2432041), and the Natural Science Foundation of Top Talent of SZTU (Grant No. GDRC202413).

References

[1] Jameel Abdul Samadh, Mohammad Hanan Gani, Noor Hussein, Muhammad Uzair Khattak, Muhammad Muzammal Naseer, Fahad Shahbaz Khan, and Salman H Khan. Align your prompts: Test-time prompting with distribution alignment for zero-shot generalization. *Advances in Neural Information Processing Systems*, 36, 2024. 1, 3

[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. 5

[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. 5

[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. 5, 7, 8

[5] Alexey Dosovitskiy. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 6

[6] David Eppstein, Maarten Löffler, and Darren Strash. Listing all maximal cliques in sparse graphs in near-optimal time. In *Algorithms and Computation: 21st International Symposium, ISAAC 2010, Jeju Island, Korea, December 15-17, 2010, Proceedings, Part I 21*, pages 403–414. Springer, 2010. 5

[7] 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. 5, 8

[8] Chun-Mei Feng, Kai Yu, Yong Liu, Salman Khan, and Wangmeng Zuo. Diverse data augmentation with diffusions for effective test-time prompt tuning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2704–2714, 2023. 1, 6, 7

[9] Zongbo Han, Jialong Yang, Junfan Li, Qinghua Hu, Qianli Xu, Mike Zheng Shou, and Changqing Zhang. Dota: Distributional test-time adaptation of vision-language models. *arXiv preprint arXiv:2409.19375*, 2024. 2

[10] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 6

[11] 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

[12] 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

[13] 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

[14] Lukas Hoyer, Dengxin Dai, and Luc Van Gool. Daformer: Improving network architectures and training strategies for domain-adaptive semantic segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9924–9935, 2022. 3

[15] Fanding Huang, Zihao Yao, and Wenhui Zhou. Dtbs: Dual-teacher bi-directional self-training for domain adaptation in nighttime semantic segmentation. In *ECAI*, pages 1084–1091, 2023. 3

[16] Tak-Wai Hui. Rm-depth: Unsupervised learning of recurrent monocular depth in dynamic scenes. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1675–1684, 2022. 3

[17] 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

[18] Adilbek Karmanov, Dayan Guan, Shijian Lu, Abdulmotaleb El Saddik, and Eric Xing. Efficient test-time adaptation of vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14162–14171, 2024. 2, 3, 6, 7

[19] 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

[20] 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. 5

[21] Boyi Li, Kilian Q Weinberger, Serge Belongie, Vladlen Koltun, and René Ranftl. Language-driven semantic segmentation. *arXiv preprint arXiv:2201.03546*, 2022. 1

[22] Xin Li, Dongze Lian, Zhihe Lu, Jiawang Bai, Zhibo Chen, and Xinchao Wang. Graphadapter: Tuning vision-language

models with dual knowledge graph. *Advances in Neural Information Processing Systems*, 36, 2024. 2

[23] Xiaosong Ma, Jie Zhang, Song Guo, and Wenchao Xu. Swapprompt: Test-time prompt adaptation for vision-language models. *Advances in Neural Information Processing Systems*, 36, 2024. 3

[24] Subhransu Maji, Esa Rahtu, Juho Kannala, Matthew Blaschko, and Andrea Vedaldi. Fine-grained visual classification of aircraft. *arXiv preprint arXiv:1306.5151*, 2013. 5

[25] 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. 5, 7

[26] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *Transactions on Machine Learning Research Journal*, pages 1–31, 2024. 2, 3, 4

[27] 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. 5

[28] 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, 7

[29] 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, 8

[30] Manli Shu, Weili Nie, De-An Huang, Zhiding Yu, Tom Goldstein, Anima Anandkumar, and Chaowei Xiao. Test-time prompt tuning for zero-shot generalization in vision-language models. *Advances in Neural Information Processing Systems*, 35:14274–14289, 2022. 1, 3, 6, 7

[31] Khurram Soomro, Amir Roshan Zamir, and Mubarak Shah. Ucf101: A dataset of 101 human actions classes from videos in the wild. 2012. 5

[32] Mingkang Tang, Zhanyu Wang, Zhenhua Liu, Fengyun Rao, Dian Li, and Xiu Li. Clip4caption: Clip for video caption. In *Proceedings of the 29th ACM International Conference on Multimedia*, pages 4858–4862, 2021. 1

[33] 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

[34] 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. 5

[35] Hu Xu, Gargi Ghosh, Po-Yao Huang, Dmytro Okhonko, Armen Aghajanyan, Florian Metze, Luke Zettlemoyer, and Christoph Feichtenhofer. Videoclip: Contrastive pre-training for zero-shot video-text understanding. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6787–6800, 2021. 1

[36] Jiarui Xu, Shalini De Mello, Sifei Liu, Wonmin Byeon, Thomas Breuel, Jan Kautz, and Xiaolong Wang. Groupvit: Semantic segmentation emerges from text supervision. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18134–18144, 2022. 1

[37] Zihao Yao, Fanding Huang, Yannan Li, Wei Duan, Peng Qian, Nan Yang, and Willy Susilo. Mecon: A gnn-based graph classification framework for mev activity detection. *Expert Systems with Applications*, 269:126486, 2025. 4

[38] Renrui Zhang, Rongyao Fang, Wei Zhang, Peng Gao, Kunchang Li, Jifeng Dai, Yu Qiao, and Hongsheng Li. Tip-adapter: Training-free clip-adapter for better vision-language modeling. *arXiv preprint arXiv:2111.03930*, 2021. 2, 3

[39] Renrui Zhang, Xiangfei Hu, Bohao Li, Siyuan Huang, Hanqiu Deng, Yu Qiao, Peng Gao, and Hongsheng Li. Prompt, generate, then cache: Cascade of foundation models makes strong few-shot learners. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15211–15222, 2023. 2

[40] Xiyu Zhang, Jiaqi Yang, Shikun Zhang, and Yanning Zhang. 3d registration with maximal cliques. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17745–17754, 2023. 4

[41] Yabin Zhang, Wenjie Zhu, Hui Tang, Zhiyuan Ma, Kaiyang Zhou, and Lei Zhang. Dual memory networks: A versatile adaptation approach for vision-language models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 28718–28728, 2024. 2, 3, 4, 6, 7

[42] 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

[43] 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, 6, 7