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.

Attention-guided local dynamic alignment: Boosting zero-shot classification of vision-language models

Lei Chen
Naval University of Engineering
Li Duan
Naval University of Engineering
Rui Fang
Naval University of Engineering
Hongping Zhang
d25181002@nue.edu.cn

Naval University of Engineering

Research Article

Keywords: Vision-language models, Zero-shot learning, Local dynamic alignment, Collaborative iterative optimization

Posted Date: July 6th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-9752955/v1

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

Additional Declarations: No competing interests reported.

Attention-guided local dynamic alignment: Boosting zero-shot classification of vision-language models

Lei Chen, Li Duan, Rui Fang, Hongping Zhang*

Abstract

Vision-language models possess cross-modal generalization capabilities in zero-shot image classification. Existing methods improve classification accuracy by generating fine-grained text descriptions via large language models and aligning them with local image region features. However, these approaches suffer from two limitations: (1) random or semi-random cropping tends to introduce background noise and lose discriminative details, while failing to adaptively adjust the observation scale based on region-wise semantic importance; (2) there exists a many-to-many matching relationship between local regions and text descriptions, yet current methods adopt one-way independent weight computation and do not explicitly model the bidirectional correlation between regions and descriptions. To address these issues, this paper proposes attention-guided local dynamic alignment (ALDA), which achieves cross-modal fine-grained semantic alignment without training or fine-tuning. We reveal and model two core dynamic properties: spatial sampling dynamics, where the location and scale of discriminative regions vary with image content, and semantic matching dynamics, where region weights and description weights mutually define and co-evolve. ALDA employs attention-driven adaptive multi-scale sampling to locate discriminative regions and allocate observation scales, reducing background noise and focusing on details. It further performs bidirectional iterative propagation on a region-description bipartite graph to achieve mutual reinforcement of weights and matching convergence. Experimental results on six standard datasets and four out-of-distribution generalization benchmarks show that ALDA achieves an average accuracy of 69.17% on the ViT-B/16 backbone, with consistent performance improvements in fine-grained classification and natural adversarial scenarios, while maintaining inference efficiency and interpretability.

Keywords: Vision-language models; Zero-shot learning; Local dynamic alignment; Collaborative iterative optimization

1 Introduction

Vision-language models [1-6,40] achieve cross-modal semantic alignment through contrastive

learning on large-scale image-text pairs. CLIP [1], as a representative work, can be transferred to new categories without downstream training samples, possessing zero-shot classification capability. However, the zero-shot performance of CLIP heavily depends on the quality of the input prompt template. Although manually designed prompt templates (e.g., a photo of {class}") are simple, they struggle to capture discriminative visual features of categories. Designing more complex prompts manually requires domain expertise and is time-consuming, making large-scale application difficult [41-45].

To overcome this limitation, recent research has progressed in two directions. The first leverages LLMs to automatically generate fine-grained descriptions of categories [7-11], replacing category names with natural language descriptions such as "the beak is yellow" or "the wings have black stripes" to enrich the semantic expression of the text modality. The second introduces local region cropping strategies on the vision side, segmenting an image into multiple local regions to achieve precise matching between local visual features and fine-grained text descriptions [12-15]. Fig. 1 compares three cross-modal alignment strategies: global image-category label alignment, global image-fine-grained description alignment, and local region-fine-grained description alignment. Such local alignment methods show advantages in fine-grained classification and complex scenarios, yet an unsolved problem remains: under the zero-shot constraint of no training or fine-tuning, how to achieve accurate and robust cross-modal semantic alignment between local image regions and fine-grained text descriptions.

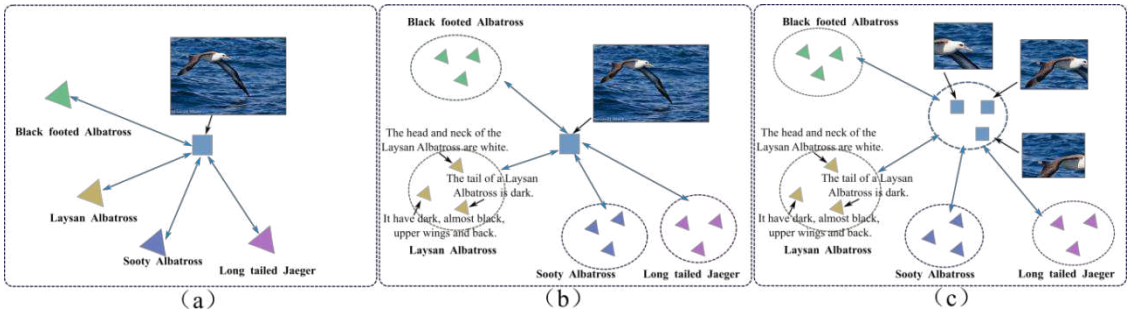


Fig. 1 Comparative evolution of visual-language cross-modal alignment strategies: (a) CLIP: Direct alignment between global images and category labels, ignoring local discriminative features; (b) DCLIP [8]/CuPL [11]: Alignment between global images and LLM fine-grained descriptions, exhibiting global-local semantic mismatches; (c) WCA [12]/ALDA: Precise alignment between local image regions and LLM descriptions, achieving fine-grained cross-modal matching.

Zero-shot classification aims to recognize categories never seen during pre-training [38,39].

Descriptions generated by LLMs (e.g., "the head and neck of the black-footed albatross are white") typically characterize local discriminative parts of an object rather than its overall appearance. Therefore, directly computing similarity between the whole image and these descriptions inherently suffers from modality mismatch. From an information-theoretic perspective, when an image can be decomposed into discriminative regions $\boldsymbol{x}_1$ and non-discriminative regions $\boldsymbol{x}_2$, the similarity between global features and fine-grained descriptions may be lower than that between discriminative local features and descriptions due to interference from background and non-discriminative regions. Hence, adopting a local alignment strategy helps improve matching accuracy.

Existing local alignment methods have design limitations. WCA [12] adopts uniform random cropping to obtain local regions, computes weights independently based on region-to-image similarity and description-to-label similarity, and then multiplies them for aggregation. This design lacks content awareness and the weight computation is one-way and independent. ABS [14] uses DINO [16] attention maps to localize object regions, constraining cropping centers to high-attention areas to reduce background cropping; the crop scale is randomly selected within a preset interval. However, the scale selection in ABS is independent of region discriminability, and its soft matching weight is still a one-way similarity from region to description, without modeling bidirectional information feedback between regions and descriptions. This paper attributes the above issues to two dynamic challenges:

Spatial sampling dynamics---where to look and at what scale. The location and optimal observation scale of discriminative regions vary with image content. As shown in Fig. 2, the semantic importance of a cat and background branches differs, but neither random cropping nor attention-location-based random scale cropping (e.g., ABS) can adaptively adjust the observation range according to region discriminability: highly discriminative parts require small scales to focus on details, while low-discriminability backgrounds require large scales to retain context. The scale should be conditioned on attention.

Semantic matching dynamics---which region to trust and which description to trust. Among the set of LLM-generated descriptions, different descriptions exhibit varying matching quality to different regions. For example, for an image of a bird, the description "yellow beak" is highly relevant to the beak region, while "black wings" matches the wing region. The importance of regions and descriptions are interrelated: a region that matches multiple high-confidence descriptions should have its importance increased; conversely, a description that matches multiple highly discriminative regions should also

have its importance enhanced. WCA [12] computes region weights and text weights independently and then aggregates them directly, failing to capture this bidirectional dependency. The soft matching in ABS [14] computes a weight for each region-description pair, but this weight is based only on the one-way similarity from region to description, without receiving feedback from the description side. Neither explicitly models the bidirectional coupling and co-evolution between regions and descriptions.

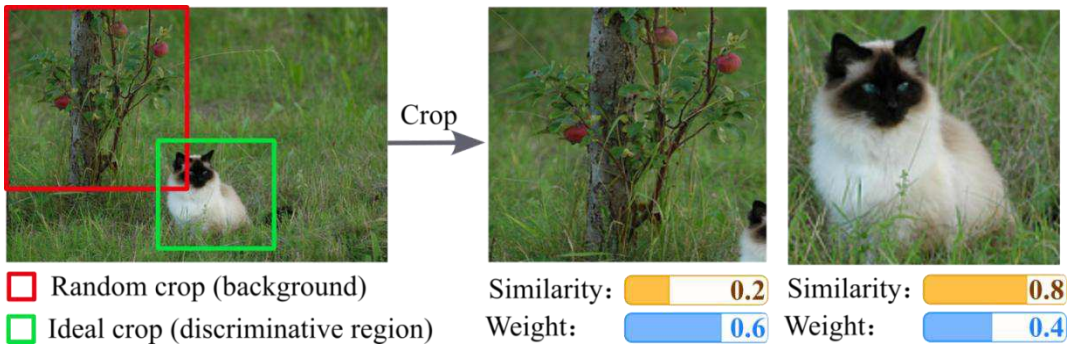


Fig. 2 Schematic diagram illustrating background noise and weight misallocation introduced by cropping. Red box: Background region with low similarity to category description (0.2) but assigned excessively high weight (0.6); Green box: Discriminative target region with high similarity (0.8) but suppressed weight (0.4).

This paper proposes attention-guided local dynamic alignment (ALDA) to tackle these challenges, with the following main contributions:

- For spatial sampling dynamics, we introduce an attention-driven adaptive sampling strategy. The DINO self-supervised attention map is used to modulate both the location and scale of sampling, achieving discriminability-driven multi-scale observation.
- For semantic matching dynamics, we present an iterative propagation co-alignment mechanism. The relationship between regions and descriptions is modeled as bidirectional message passing on a bipartite graph, where region weights and description weights are mutually reinforced and co-evolved through iterative updates, transforming the matching process from one-way computation to bidirectional cross-enhancement.
- Achieving a balance between performance and efficiency under strict zero-sample constraints. With N=40 sampling regions, the inference time per image is 0.1091 seconds (RTX 4090), reducing the time ratio from 15.6 to 8.3 compared to the ABS method with comparable accuracy. The method requires no training or fine-tuning, enabling plug-and-play operation.
- Comprehensive experimental validation and in-depth analysis. On six standard datasets and four

out-of-distribution generalization benchmarks, ALDA achieves an average accuracy of 69.17% on the ViT-B/16 backbone, outperforming mainstream baselines. Ablation studies quantify the contribution of each module.

2 Related work

2.1 Development lineage and representative works

Fig. 3 illustrates the timeline of zero-shot image classification methods for vision-language models. This field has undergone three stages of evolution.

Manual prompt engineering. CLIP [1] uses "a photo of {class}" as the default prompt. CLIP-E [1] integrates multiple manually designed templates and improves robustness via average pooling. However, manual prompt design relies on domain expertise and struggles to transfer to new categories or specific tasks.

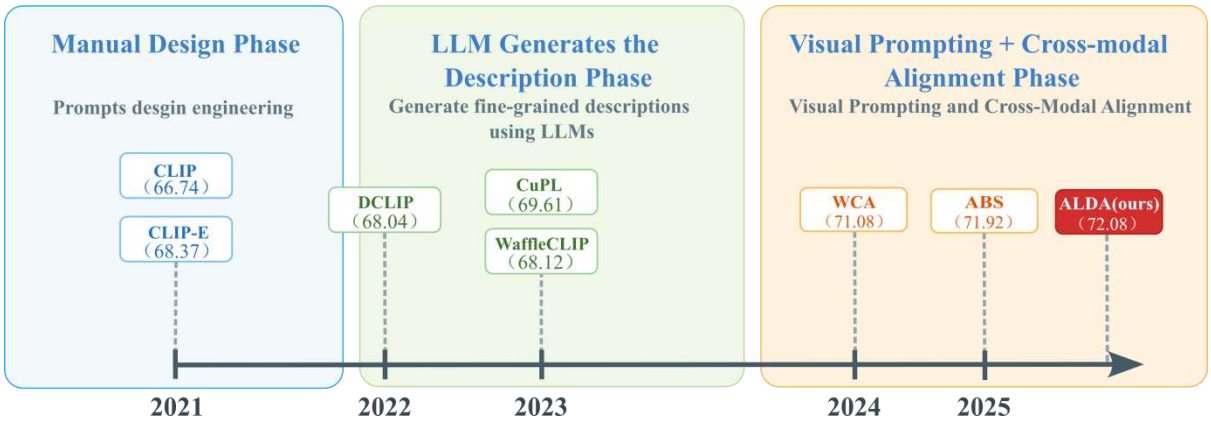


Fig. 3 Evolution and performance comparison (%) of CLIP-based zero-sample image classification methods. The approach consists of three stages: manual prompt engineering, LLM-generated descriptions, and visual prompts combined with cross-modal alignment. ALDA achieves 72.08% accuracy on the ViT-B/16 backbone, delivering optimal performance under the local dynamic alignment paradigm.

LLM-generated fine-grained descriptions. DCLIP [8] first employs GPT-3 to generate multiple visual descriptions per category (e.g., "a seabird", "a long, hooked bill", "long, narrow wings"), enriching the semantic expression of the text modality. CuPL [11] optimizes the generation template, guiding the LLM to produce more discriminative descriptions and improving description quality. WaffleCLIP [17] discovers that random strings can also boost performance, thereby reducing the cost of LLM calls.

Visual prompting and cross-modal alignment. WCA [12] proposes random cropping with

dynamic weighting, aggregating multiple local regions with learned weights to improve matching accuracy. However, its cropping lacks content awareness and weight computation is one-way and independent. ABS [14] introduces DINO [16] self-supervised attention maps to guide cropping locations, suppressing background noise, but the cropping scale is still randomly selected within a fixed interval, independent of region content; its soft matching is also one-way. The ALDA in this paper, building upon the attention-guided location of ABS, modulates the scale by attention and upgrades the matching mechanism to bidirectional iterative co-alignment.

Different from the above methods that model dynamics on the vision side or vision-text matching side, another line of research focuses on automatically searching for optimal text prompts using a few labeled samples. ProAPO [32] proposes a progressively automatic prompt optimization method based on evolutionary algorithms, editing and evolving from task-level templates to category-level descriptions, and introduces an entropy-constrained fitness score to mitigate overfitting. This method achieves performance improvement under the one-shot supervised setting. However, ProAPO's optimization process relies on labeled samples to evaluate prompt quality, belonging to supervised hyperparameter search in the natural language space. In contrast, our ALDA adopts a zero-shot setting, requiring no labeled samples, and achieves online alignment with LLM descriptions by modeling the intrinsic spatial and semantic dynamics within images. The two approaches are complementary in methodology.

2.2 Limitations of existing methods and targeted improvements of ALDA

Background noise and scale blindness. WCA's [12] uniform random cropping easily captures background. ABS [14] mitigates this by constraining cropping centers to object regions via attention maps, but the cropping scale is uniformly randomly selected within a preset interval ([0.6, 0.9]), independent of the discriminability of the sampling point. This means that a high-attention region containing key details and a medium-attention region containing only part of an edge may be assigned the same cropping scale, leading to insufficient observation of key details or excessive magnification of background regions. ALDA's improvement: It retains attention-guided positioning and makes the cropping scale a continuous function of attention value: high-attention points use small scales to focus on details, while low-attention points use large scales to retain context.

Unidirectional independent matching weights. WCA [12] computes region weights and text

weights are computed independently and then multiplied, with no interaction between them. ABS's [14] soft matching computes a softmax weight for each region against all descriptions, but this weight is based only on the one-way region-to-description similarity, without using description-to-region feedback to adjust region importance. Such one-way computation cannot capture the bidirectional dependency between regions and descriptions. ALDA's improvement: It performs iterative propagation on a region-description bipartite graph, where region weights are updated according to the quality of matched descriptions, and description weights are updated according to the discriminability of matched regions, enabling co-evolution.

DINO's self-supervised Vision Transformer [18-21] can focus its attention maps on object parts without annotations. ABS [14] first introduces DINO attention maps into visual prompting for CLIP, but only for cropping location guidance, without leveraging them for adaptive scale modulation or weight co-optimization. Building upon this, the present paper further uses attention for scale modulation and integrates an iterative propagation mechanism, extending from "attention for position" to "attention for position + scale + weight".

3 Methodology

The two core modules of ALDA are complementary in design: the attention-driven joint dynamic sampling is responsible for "where to look and at what scale", providing high-quality multi-scale candidate regions for subsequent matching; he iterative propagation co-alignment mechanism is responsible for "which region to attend to and which description to trust", achieving weight convergence through bidirectional cross-enhancement. Fig. 4 shows the overall framework.

3.1 Overall framework

Given a query image $\boldsymbol{x}$ and the set of LLM-generated descriptions $\{D_c\}$ for each category, ALDA operates as follows. First, the DINO model is used to extract a multi-layer fused attention map $\boldsymbol{A}_{\text{enhanced}}$ as a visual saliency prior. Guided by this prior, adaptive spatial dynamic sampling is performed: the sampling ratios for foreground/midground/background are determined based on attention concentration, positions are sampled from the attention probability distribution, and the cropping scale for each sampling point is computed according to its attention value, yielding a high-quality multi-scale region set $\{\boldsymbol{x}_i\}_{i=1}^{N}$. Subsequently, the CLIP visual encoder $f(.)$ extracts global and region features, and the text encoder $g(.)$ extracts category label features and description features.

After quickly filtering top-k candidate categories C_{cand} via global similarity, for each candidate category, the iterative propagation co-alignment is performed on the bipartite graph formed by its descriptions and sampled regions, where region weights and description weights are bidirectionally updated until convergence, producing an alignment score and the final prediction. The entire pipeline requires no training or fine-tuning, preserving CLIP's zero-shot capability.

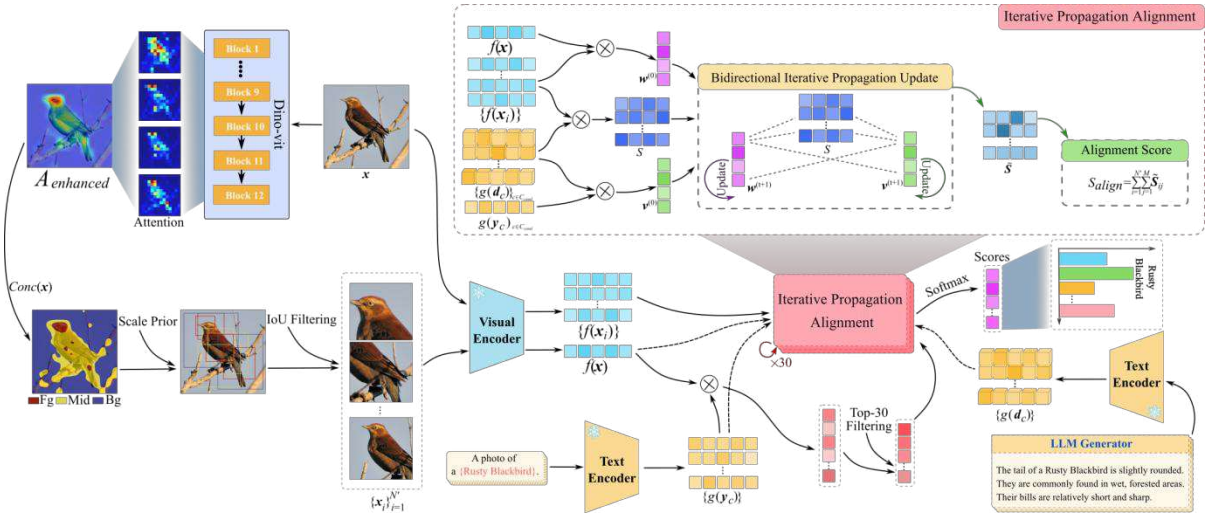


Fig. 4 Overall framework of ALDA. It comprises four modules: 1) DINO self-supervised attention extraction; 2) attention-guided adaptive sampling; 3) fast top-k candidate screening with global features; 4) region-description bidirectional iterative propagation co-alignment. The whole pipeline requires no training or fine-tuning, preserving the native zero-shot capability of CLIP.

3.2 Attention-guided adaptive sampling

This module addresses the dynamics of spatial sampling: The location and scale of discriminative regions vary with image content, necessitating joint adaptive sampling where "attention determines location and attention determines scale".

Multi-layer attention fusion. For an input image $\boldsymbol{x}$, DINO is used to extract the average attention between the [CLS] token and image patches from the last L Transformer layers, followed by weighted fusion (higher weights for deeper layers) to emphasize high-level semantics. After contrast enhancement and upsampling, the spatial saliency attention $\boldsymbol{A}_{\text{enhanced}}$ is obtained. This attention map highlights the foreground regions of the image without requiring annotations.

Attention concentration and dynamic allocation of sampling ratios. To quantify the spatial aggregation of salient regions, we define the attention concentration $Conc(\boldsymbol{x})$ as follow:

$$Conc(\boldsymbol{x}) = \alpha_1 \cdot \frac{top_{90\%} - \mu}{\sigma} + \alpha_2 \cdot (1 - \frac{\sum(A_{enhanced} > \mu + \sigma)}{P^2}) , \quad (1)$$

where α_1=0.5, α_2=0.5,$\alpha_1+\alpha_2$=1 , the two terms characterize the peak intensity of saliency and the area proportion of highly salient regions, respectively, using normalized weights. $Conc(\boldsymbol{x})\in[0,1]$, μ and σ are the mean and standard deviation of $\boldsymbol{A}_{enhanced}$, $\text{top}_{90\%}$ is the 90th percentile of the attention map $\boldsymbol{A}_{\text{enhanced}}$. A high $Conc(x)$ indicates that the foreground subject is prominent, calling for a focused sampling strategy; a low value indicates scattered semantics, necessitating diversity in spatial exploration. Based on $Conc(x)$, the sampling ratios for foreground and midground are dynamically determined: ρ_{fg} increases linearly with concentration, ρ_{mid} decreases correspondingly, and the background is uniformly sampled with a minimum coverage guarantee. Foreground and midground points are drawn from their respective attention probability distributions, while background points are uniformly sampled from the entire image region.

Attention-aware adaptive scale adjustment. For each sampling point (u,v), let its normalized attention value be $\hat{\mathbf{a}}_{uv}\in[0,1]$. We establish a continuous mapping from attention value to cropping scale:

$$scale_{(u,v)} = scale_{base} + (0.8 - scale_{base}) \cdot (1 - \hat{\boldsymbol{a}}_{uv}) , \quad (2)$$

where $scale_{base}\in[0.45,0.65]$ is determined by the global attention strength $\boldsymbol{A}_{enhanced}$, allowing the base scale to shift adaptively with the overall image saliency level. For high-attention points $\hat{\mathbf{a}}_{uv}\to1$, the cropping scale approaches $scale_{base}$ for fine-grained capture; for low-attention points $\hat{\mathbf{a}}_{uv}\to0$, the scale approaches 0.8, enlarging the receptive field. Finally, a crop box centered at (u, v) with side length $scale_{(u,v)} \cdot \min(H,W)$ is generated, and after IoU-based non-maximum suppression, the final N' image regions are obtained.

3.3 Global feature candidate class filtering

Vision-language models exhibit a characteristic of strong recall but weak fine-grained discrimination: they can quickly identify salient features but struggle to distinguish subtle differences. As shown in Fig. 5, the top-1 accuracy is not high, but the top-30 recall is extremely high. Leveraging the strong recall capability of CLIP and to improve computational efficiency, we compute the cosine similarity between the CLIP global features and all class labels, and select the top-k (k=30) candidate classes C_{cand}. Only for these candidate classes do we perform the subsequent fine-grained alignment, reducing computational overhead.

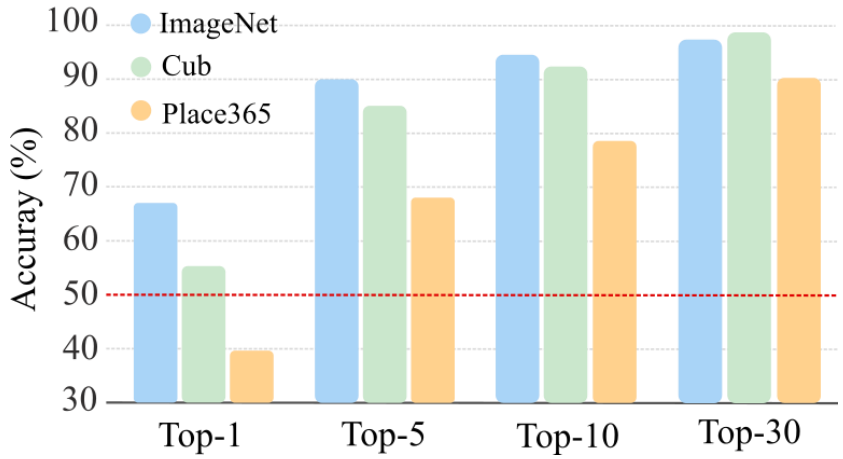


Fig. 5 Top-k accuracy of CLIP vision-language model on ImageNet. CLIP demonstrates strong recall but weak fine-grained discrimination: top-30 recall is nearly saturated, but top-1 accuracy is restricted, providing the basis for ALDA to first screen top-k candidates and then perform fine-grained alignment.

3.4 Bidirectional iterative propagation co-alignment

This module addresses the dynamics of semantic matching: the importance of regions and that of descriptions mutually define each other in a complex many-to-many relationship, requiring explicit modeling of bidirectional coupling and synergistic reinforcement.

Similarity matrix construction. For a candidate class c, we use CLIP to extract features of its M descriptions $\{g(\boldsymbol{d}_c^j)\}$ and compute the cosine similarity matrix $\boldsymbol{S}\in\mathbb{R}^{N'\times M}$ with the N' region features $\{f(\boldsymbol{x}_i)\}$:

$$\boldsymbol{S}_{ij}=\frac{f(\boldsymbol{x}_i)^{\top}g(\boldsymbol{d}_c^j)}{\|f(\boldsymbol{x}_i)\|\,\|g(\boldsymbol{d}_c^j)\|}\ , \tag{3}$$

Each $\boldsymbol{S}_{ij}$ in this matrix represents the degree of semantic match between region i and description j.

Adaptive temperature adjustment. The temperature is adaptively determined based on the matching entropy of the similarity matrix: $\tau=\tau_0+0.02\cdot H(\boldsymbol{S})$. Under high confidence (low entropy), τ is reduced to make weights sharper; under low confidence (high entropy), τ is increased to maintain exploration.

Weight initialization. The initial region weights $\boldsymbol{w}^{(0)}$ are derived from the similarity between the global image feature and regional features, and the initial description weights $\boldsymbol{v}^{(0)}$ from the similarity between the class label feature $g(\boldsymbol{y}_c)$ and the description features:

$$\boldsymbol{w}_i^{(0)}=\frac{\exp(f(\boldsymbol{x})^{\top}f(\boldsymbol{x}_i)/\tau)}{\sum_{k=1}^{N'}\exp(f(\boldsymbol{x})^{\top}f(\boldsymbol{x}_k)/\tau)}\ , \tag{4}$$

$$\boldsymbol{v}_j^{(0)}=\frac{\exp(g(\boldsymbol{y}_c)^{\top} g(\boldsymbol{d}_c^j)/\tau)}{\sum_{k=1}^{M}\exp(g(\boldsymbol{y}_c)^{\top} g(\boldsymbol{d}_c^k)/\tau)} \tag{5}$$

Bidirectional iterative propagation and cooperative update. This module remedies the unidirectional weight computation of WCA [12] and ABS [14]. ALDA places the region weights $\boldsymbol{w}\in\mathbb{R}^{N}$ and description weights $\boldsymbol{v}\in\mathbb{R}^{M}$ on the two sides of a bipartite graph and performs cooperative updates. With momentum coefficient α and T iterations:

$$\boldsymbol{w}^{(t+1)}=\alpha\cdot Softmax(\frac{\boldsymbol{S}\boldsymbol{v}^{(t)}}{\tau})+(1-\alpha)\cdot\boldsymbol{w}^{(t)}\ , \tag{6}$$

$$\boldsymbol{v}^{(t+1)}=\alpha\cdot Softmax(\frac{\boldsymbol{S}^{\top}\boldsymbol{w}^{(t+1)}}{\tau})+(1-\alpha)\cdot\boldsymbol{v}^{(t)} \tag{7}$$

This iterative process has a clear geometric interpretation: $\boldsymbol{S}\boldsymbol{v}^{(t)}$ aggregates regions weighted by the current description weights, enhancing those regions highly similar to currently high-weight descriptions; subsequently, $\boldsymbol{S}^{\top}\boldsymbol{w}^{(t+1)}$ uses the updated region weights to inversely enhance the matching descriptions. Through multiple steps of cross-reinforcement, the region and description weights rapidly concentrate on the main cluster of the bipartite graph, automatically discovering the most discriminative matching pairs while suppressing generic descriptions and background noise.

Final alignment score. The final alignment score between class c and image $\boldsymbol{x}$ is a bilinear matching score:

$$S_{align}(\boldsymbol{x},c)=\boldsymbol{w}^{\top}\boldsymbol{S}\boldsymbol{v} \tag{8}$$

This score integrates the fine-grained matching strengths of all region-description pairs, with natural interpretability: the product terms of high-weight regions and high-weight descriptions dominate the alignment score, allowing direct tracing back to key evidence for the prediction.

To intuitively reveal the dynamic adjustment of region–description matching during iterative propagation, we define the "weighted similarity matrix" after the t-th iteration, $\tilde{\boldsymbol{S}}^{(t)}\in\mathbb{R}^{N\times M}$:

$$\tilde{\boldsymbol{S}}_{ij}^{(t)}=\boldsymbol{w}_i^{(t)}\cdot\boldsymbol{S}_{ij}\cdot\boldsymbol{v}_j^{(t)} \tag{9}$$

where $\boldsymbol{w}_i^{(t)}$ and $\boldsymbol{v}_j^{(t)}$ are the region and description weights at iteration t, respectively. This matrix preserves the semantic associations of the original similarity $\boldsymbol{S}_{ij}$ while being bidirectionally modulated by the weight pair $(\boldsymbol{w}_i^{(t)},\boldsymbol{v}_j^{(t)})$. Fig. 6 visualizes the dynamic evolution of region weights and description weights within the iterative propagation co-alignment mechanism. Initially, the distributions are relatively uniform and no clear matching focus has emerged. After iterative propagation, the weights of fine-grained descriptions that highly match discriminative regions rise, while those of irrelevant

background regions are suppressed.

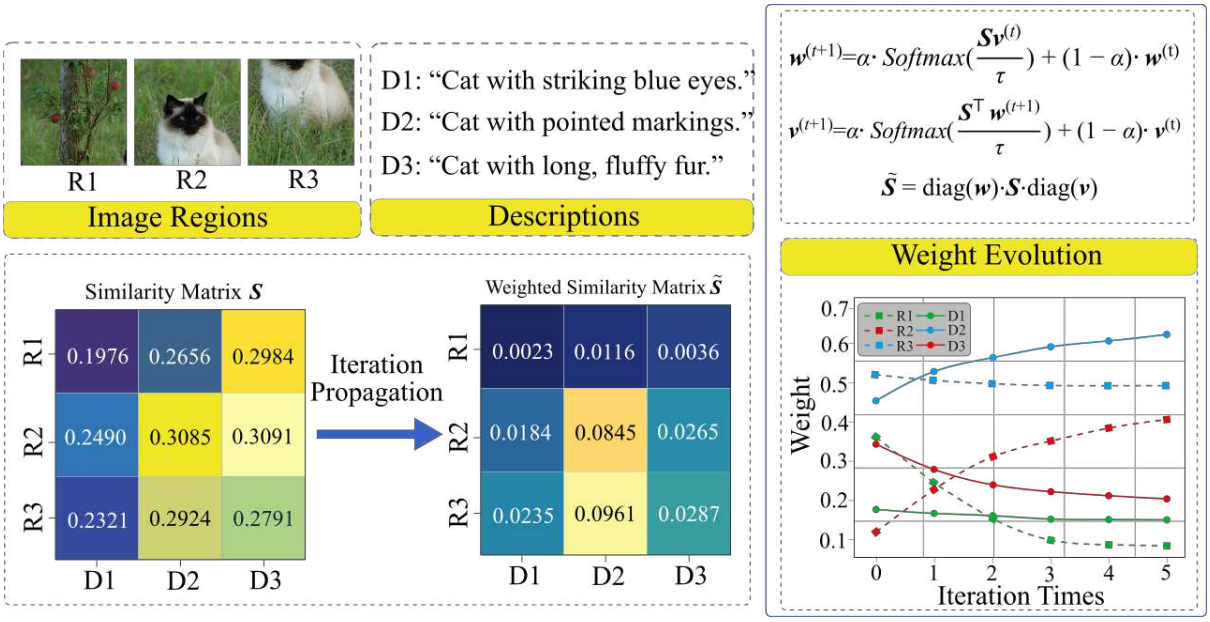


Fig. 6 Bidirectional iterative propagation and weight evolution on the region-description bipartite graph. Weights are initially uniformly distributed; through bidirectional information passing, the weights of highly discriminative regions and precise fine-grained descriptions are synergistically enhanced, while those of irrelevant descriptions and background regions are suppressed, ultimately converging to the optimal semantic matching distribution.

Prediction output and algorithm. The final predicted class is obtained by applying softmax normalization to the alignment scores of all candidate classes $c \in C_{cand}$:

$$P(c \mid x) = \frac{exp(S_{align}(\boldsymbol{x},c))}{\Sigma_{\square\,,\,\in C_{cand}} exp(S_{align}(\boldsymbol{x},c'))}\ , \tag{9}$$

$$\hat{c} = arg\ max_c P(c \mid x) \tag{10}$$

Attention-guided sampling and iterative propagation are not independent modules but form a structural positive complementarity. Attention-guided sampling suppresses background noise and focuses on discriminative regions, improving the signal-to-noise ratio of the similarity matrix and providing a clean initial bipartite graph for iterative propagation. Iterative propagation, in turn, leverages this high-quality input and, through bidirectional weight cross-reinforcement, amplifies the matching relationships between discriminative regions and their corresponding descriptions, thereby fully exploiting the fine-grained information furnished by the sampling stage. This architectural design allows the gains of the two modules to reinforce each other, producing a synergistic effect that surpasses the sum of their independent contributions. The complete algorithm is summarized in **Algorithm~1**.

Algorithm 1: ALDA: Attention-Guided Local Dynamic Alignment

Require: Image $\boldsymbol{x}$, class sets V, LLM descriptions $\{D_c\}$, CLIP encoders f, g, DINO

Extract DINO fused attention $\boldsymbol{A}_{enhanced}$

Compute $Conc(\boldsymbol{x})$, dynamic sampling

Sample N' points $\{(u,v)\}$, get $\{\boldsymbol{x}_i\}_{i=1}^{N'}$

Extract features: $\{f(\boldsymbol{x}_i)\}_{i=1}^{N'}$,$\{f(\boldsymbol{x})\}$

Select top-k candidates C_{cand}

for each $c \in C_{cand}$ **do**:

  Extract description features $\{g(\boldsymbol{d}_j)\}$

  Compute similarity $\boldsymbol{S}$

  Initialize $\boldsymbol{w}^{(0)}$, $\boldsymbol{v}^{(0)}$

  for each $t=0$ **to** $T-1$ **do**:

    $\boldsymbol{w}^{(t+1)} \leftarrow \alpha \cdot Softmax(\boldsymbol{S}\boldsymbol{v}^{(t)}/\tau)+(1-\alpha) \cdot \boldsymbol{w}^{(t)}$

    $\boldsymbol{v}^{(t+1)} \leftarrow \alpha \cdot Softmax(\boldsymbol{S}^{\top}\boldsymbol{w}^{(t+1)}/\tau)+(1-\alpha) \cdot \boldsymbol{v}^{(t)}$

  end for

  Compute $S_{align}(\boldsymbol{x},c) \leftarrow \boldsymbol{w}^{\top}\boldsymbol{S}\boldsymbol{v}$

end for

$\hat{c} \leftarrow \arg\max_{c} P(c|\boldsymbol{x})$

return $\hat{c}$

4 Experiments

4.1 Experimental setup

Datasets. We evaluate zero-shot classification on six standard datasets: ImageNet [22] is used for recognizing daily objects, CUB [23]] is used for fine-grained classification of birds, Oxford Pets [24] is used for recognizing common animals, DTD [25] is used for texture recognition, Food101 [26] is specifically designed for food classification, and Place365 [27] is applied to scene recognition. These datasets cover diverse domains including objects, fine-grained categories, textures, food, and scenes, enabling a comprehensive assessment of generalization ability.

Out-of-distribution generalization is evaluated on four ImageNet variants: ImageNet-V2 [28]

(newly collected natural images), ImageNet-R [29] (art, cartoon, sketch, and other diverse stylized images), ImageNet-S [30] (black-and-white sketch images), and ImageNet-A [31] (natural adversarial examples).

Baseline methods. The compared methods include: CLIP [1] (using "a photo of {class}" as the prompt), CLIP-E [1] (ensembling multiple manually designed templates), DCLIP [8] (GPT-3 generated descriptions, averaging similarities), CuPL [11] (optimized description generation template), WaffleCLIP [17] (using random strings as descriptions), WCA [12] (random cropping + weighted aggregation), ProAPO [32] (evolutionary algorithm based progressive prompt optimization), and ABS [14] (DINO attention-guided cropping + global semantic preservation). All compared methods use the same CLIP backbone networks (ViT-B/32, ViT-B/16, ViT-L/14) and are evaluated under the zero-shot setting without any training or fine-tuning.

*Implementation details.*We use DINO ViT-B/16 to extract attention heatmaps. Descriptions are generated by GPT-3 with CuPL's generation templates, producing 50 descriptions per class. The number of sampled regions is N=40, the number of iterations T=2, the number of candidate classes top-k=30, the temperature parameter τ_0=0.02, and the momentum coefficient α=0.4. All experiments are run on an RTX 4090 GPU.

4.2 Zero-shot classification results

Tabel 1 presents the zero-shot classification accuracy on six datasets. ALDA consistently outperforms mainstream baseline methods across all datasets and backbone networks. On the ViT-B/16 backbone, ALDA achieves an average accuracy of 69.17%, which is 0.60% higher than ABS (68.57%) and 1.29% higher than WCA (67.88%). Notably, on the fine-grained bird classification dataset CUB, the improvement is 1.61%; on the texture classification dataset DTD, the improvement is 1.38%, validating the advantages of adaptive scaling and iterative propagation in capturing local discriminative details.

Tabel 1 Zero-shot classification accuracy (%) on six standard datasets. All methods use CLIP pre-trained weights without any additional training or fine-tuning. Results are grouped by backbone: ViT-B/32, ViT-B/16, and ViT-L/14. ALDA achieves competitive performance across all groups, with an average accuracy of 69.17% on ViT-B/16.

Method	ImageNet	CUB	Oxford Pets	Food101	Place365	DTD	Average
	ViT-B / 32 with pre-trained weights from CLIP						
CLIP [1]	62.05	51.21	85.04	82.60	38.51	42.93	60.39
CLIP-E [1]	63.37	52.74	87.38	83.93	39.28	43.83	61.75
DCLIP [8]	63.01	52.69	84.46	84.12	39.90	44.20	61.39
WaffleCLIP [17]	63.30	52.04	85.50	83.98	39.47	44.98	61.54
CuPL [11]	64.37	49.76	87.03	84.20	39.08	47.50	61.99
WCA [12]	66.84	56.91	89.89	**86.40**	40.66	49.39	65.01
ProAPO [32]	64.70	55.40	**91.00**	85.30	**43.30**	**53.50**	65.53
ABS [14]	67.74	57.42	90.39	85.66	41.22	51.65	65.68
ALDA(ours)	**68.04**	**58.34**	90.08	86.28	41.91	52.77	**66.24**
	ViT-B/16 with pre-trained weights from CLIP						
CLIP [1]	66.74	56.01	88.14	88.40	39.27	42.98	63.59
CLIP-E [1]	68.37	56.16	89.10	88.83	40.30	45.27	64.67
DCLIP [8]	68.04	57.08	87.52	88.85	40.34	46.17	64.66
WaffleCLIP [17]	68.12	56.89	86.51	89.06	40.76	44.68	64.33
CuPL [11]	69.61	56.42	91.14	88.98	39.83	50.53	66.08
WCA [12]	71.08	59.78	92.23	**90.01**	41.43	52.79	67.88
ProAPO [32]	69.90	59.04	92.70	89.50	**43.80**	**55.80**	68.46
ABS [14]	71.92	61.01	92.64	89.69	41.88	54.26	68.57
ALDA(ours)	**72.08**	**62.62**	**92.78**	**89.70**	42.21	55.64	**69.17**
	ViT-L/14 with pre-trained weights from CLIP						
CLIP [1]	73.48	62.12	93.24	92.55	39.63	52.61	68.93
CLIP-E [1]	75.52	62.53	93.62	93.07	40.55	55.43	70.12
DCLIP [8]	75.03	63.26	93.30	93.03	40.55	55.05	70.03
WaffleCLIP [17]	75.31	62.27	91.55	93.33	40.89	54.31	69.61
CuPL [11]	76.62	62.15	94.33	93.37	40.77	60.59	71.30
WCA [12]	77.32	65.24	94.66	**93.96**	42.23	61.78	72.53
ProAPO [32]	76.80	65.10	94.60	93.90	**44.40**	**64.80**	73.27
ABS [14]	77.62	67.06	**94.88**	93.05	42.27	61.80	72.78
ALDA(ours)	**78.10**	**67.95**	94.74	93.94	42.59	62.50	**73.30**

As shown in Tabel 1, ProAPO [32] achieves an average accuracy of 68.46% on the same ViT-B/16 backbone, especially excelling on datasets such as DTD (55.80%) and Place365 (43.80%). This performance stems from its evolutionary search over prompts, which can find a combination of static descriptions that is optimal for each dataset. In contrast, ALDA, under a strict zero-shot setting

(without using any training sample to evaluate or filter prompts), obtains a 69.17% average accuracy, outperforming all compared methods including ProAPO. More importantly, the performance patterns of ALDA and ProAPO reflect methodological differences: ProAPO's advantages are more pronounced in scenarios where class descriptions themselves already provide strong discriminative power (e.g., textures, scenes), whereas ALDA exhibits greater advantages on fine-grained tasks that require dynamic discrimination based on specific image content. For instance, on CUB, ALDA (62.62%) significantly outperforms ProAPO (59.04%). This demonstrates that, under a completely training-free constraint, image-content-driven dynamic alignment can achieve fine-grained matching with greater robustness and generalization ability than the best statically searched prompts.

4.3 Out-of-distribution generalization results

To comprehensively evaluate the cross-domain robustness of the model, we conduct experiments on four out-of-distribution variants of ImageNet (ImageNet-V2/R/S/A). The results are shown in Tabel 2. ALDA achieves excellent overall generalization performance, with an average accuracy of 65.88%. Its advantage is most significant on natural adversarial examples (ImageNet-A), reaching 62.00%, which is a 5.87% improvement over WCA and also surpasses ABS, verifying the strong suppression of noise perturbations enabled by bidirectional iterative alignment.

Tabel 2 Generalization performance (%) on four ImageNet out-of-distribution datasets (ViT-B/16). "Tuned" indicates models fine-tuned on ImageNet and tested on the target datasets.

Method	Tuned?	Source	Target				Average
		ImageNet	ImageNet-V2	ImageNet-R	ImageNet-S	ImageNet-A	
CLIP[1]	×	66.74	60.80	74.00	47.80	47.80	59.43
CoCoOp[33]	√	71.02	64.10	76.20	48.80	50.60	62.14
KgCoOp[34]	√	71.20	64.10	76.70	48.97	50.69	62.33
TPT[35]	√	69.70	64.30	73.90	46.40	53.60	61.58
ProGrad[36]	√	72.24	64.70	74.50	47.90	49.30	61.73
TDA[37]	√	69.51	64.67	**80.24**	**50.54**	60.11	65.01
CuPL[11]	×	69.61	63.30	77.10	48.80	50.80	61.92
WCA[12]	×	71.08	64.71	78.06	<u>50.18</u>	56.13	64.03
LaZSL[13]	×	69.20	63.30	75.60	48.20	56.20	62.50
ABS[14]	×	<u>71.92</u>	<u>66.19</u>	<u>79.57</u>	**50.54**	<u>61.80</u>	**66.00**
ALDA(ours)	×	**72.08**	**66.52**	78.89	49.93	**62.00**	<u>65.88</u>

On the artistic stylized dataset ImageNet-R, ALDA performs slightly worse than ABS. This can

be explained by the visualization in Fig. 7. ALDA's adaptive scaling tends to over-focus on local noisy patterns, weakening the capture of global semantic information and leading to matching deviations, which reflects the inherent limitation of local alignment strategies in scenarios dominated by global semantics.

In the remaining out-of-distribution scenarios, ALDA performs comparably to ABS with good cross-domain adaptability. Overall, ALDA exhibits prominent robustness in most OOD scenarios, and the performance drop in heavily stylized scenarios is explainable, while its generalization ability surpasses numerous mainstream methods.

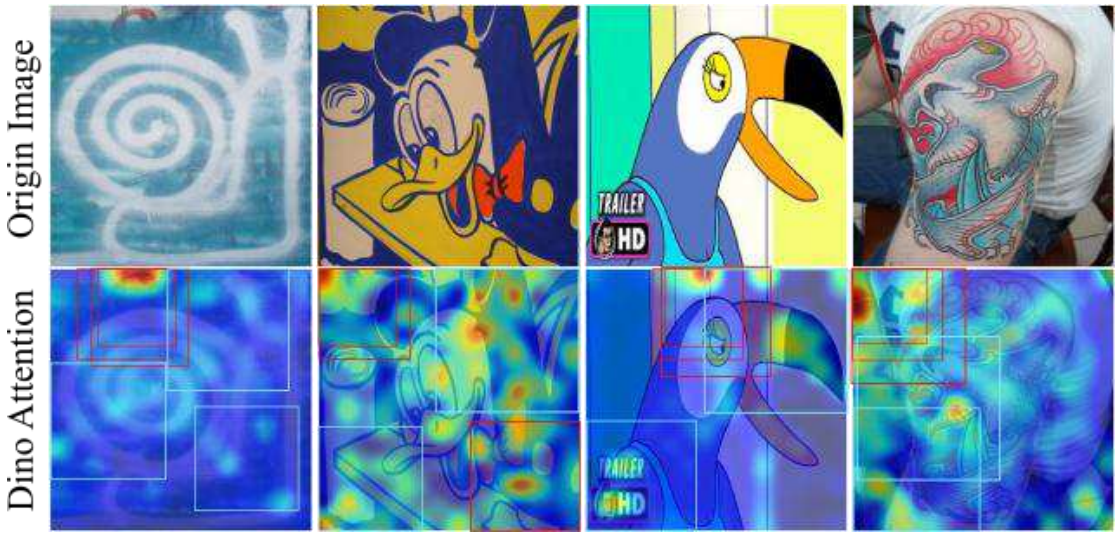


Fig. 7 Visualization of sampling on ImageNet-R. The adaptive scale sampling of ALDA can easily over-focus on distorted local noise regions, weakening global semantic information capture and causing feature matching bias.

4.4 Ablation study

To verify the necessity and effectiveness of each core component in ALDA, we conduct a removal-based ablation study. Taking the full ALDA as the baseline, we remove or replace individual core modules in a unified framework. All experiments are performed on ViT-B/16 using the ImageNet, CUB, and DTD datasets. The results are shown in Tabel 3.

Removing "attention-guided spatial sampling (AGS)" and replacing it with WCA random sampling leads to a performance drop of 0.48%, indicating that DINO self-supervised attention effectively suppresses background noise and focuses on discriminative regions. Removing "adaptive scale conditioning (ASC)" and using a fixed scale results in a drop of 1.32%, validating the critical role of dynamic scaling in capturing fine-grained features. Replacing "iterative propagation alignment

(IPA)" with WCA's original weighting strategy causes a substantial performance drop, demonstrating that bidirectional weight evolution more accurately models the complex matching relationships between regions and descriptions. Replacing it with ABS's soft matching strategy also leads to a drop of 0.89%, fully proving the superiority of bidirectional iterative co-alignment over unidirectional matching mechanisms.

Tabel 3 Ablation study results of ALDA's core components (ViT-B/16, average accuracy % on ImageNet/CUB/DTD). AGS: attention-guided Spatial Sampling; ASC: adaptive scale conditioning; IPA: iterative propagation alignment. Δ indicates the performance drop relative to the full ALDA. The results confirm that all three modules are necessary, with adaptive scaling and bidirectional iterative alignment contributing most significantly.

Method	ImageNet	CUB	DTD	Average	Δ vs (Avg)
ALDA(full)	**72.08**	**62.62**	**55.64**	**63.45**	—
w/o AGS → Random	71.75	62.15	55	62.97	−0.47
w/o ASC → Fixed	70.47	61.03	54.89	62.13	−1.32
w/o IPA → WCA	71.01	62.37	51.70	61.69	−1.76
w/o IPA → ABS	71.98	61.18	54.52	62.56	−0.89
Random + Fixed + WCA	69.53	61.62	51.22	60.79	−2.66

In summary, all ablation results consistently indicate that attention-guided sampling, adaptive scale conditioning, and iterative propagation co-alignment together constitute the core of ALDA, each is indispensable and brings independent yet complementary performance gains.

4.5 Parameter sensitivity analysis

Effect of the number of sampled regions N. Fig. 8(a) shows the impact of N from 10 to 60 on ImageNet accuracy. Performance improves as N increases, saturating at N=40 (72.08%), and remains nearly flat at N=50 (72%). Therefore, $N=40$ is set by default, striking a balance between performance and efficiency. Sampling 40 local image regions is sufficient to cover the discriminative information in an image; gains from more regions are negligible.

Effect of the number of iterations T. Fig. 8(b) shows that performance converges at T=2; accuracy does not further improve at T=3, and further increasing T leads to no further gain or even a decline, possibly due to weight collapse. This indicates that effective weight co-adaptation is achieved after just one update of bidirectional propagation, and two iterations are sufficient for convergence.

Effect of the number of candidate classes k. Fig. 8(c) shows that accuracy peaks at top-k=30. Increasing the number of candidate classes not only increases computational overhead but also no longer improves classification accuracy. Fig. 8(c) shows that accuracy peaks at top-k=30. Increasing the number of candidate classes not only increases computational overhead but also no longer improves classification accuracy.

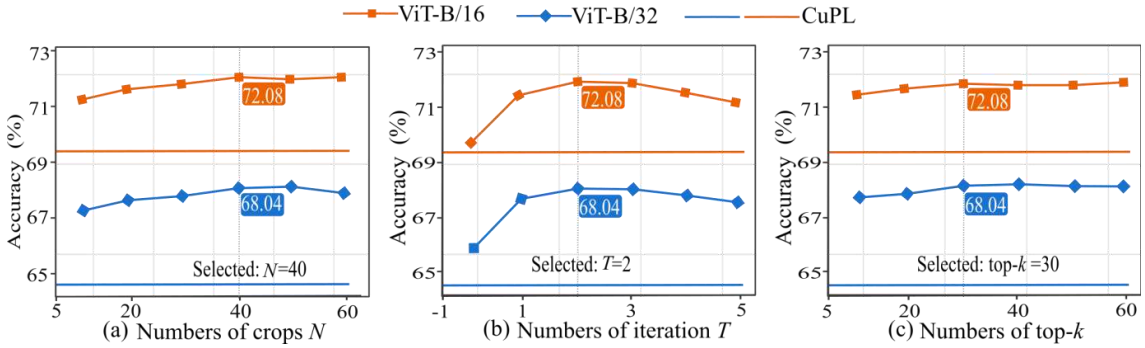


Fig. 8 Parameter sensitivity analysis of ALDA (ViT-B/16, ImageNet). (a) Number of sampled regions N: performance saturates at N=40. (b) Number of iterations T: convergence is reached at T=2, with no further gain from more iterations. (c) Number of candidate classes k: top-k=30 achieves the optimal accuracy-efficiency trade-off.

4.6 Qualitative analysis and visualization

To intuitively demonstrate the working mechanisms and synergistic effects of each core module in ALDA, this section presents a qualitative visualization analysis. Fig. 9 systematically reveals the internal operational logic of ALDA from three dimensions. Fig. 9(a) takes a fine-grained bird image as an example and illustrates the spatial distribution of attention-guided adaptive sampling: DINO can localize the main object in the image, ensuring fine-grained capture of local discriminative features, while background regions are handled at larger scales to preserve necessary context. The sampling boxes are highly concentrated on semantically salient parts, effectively avoiding the background noise introduced by random cropping.

Fig. 9(b) reveals the evolution of the similarity matrix throughout iterative propagation. Initially, the weight distributions are relatively uniform. As information bidirectionally propagates on the region-description bipartite graph, the system automatically locks onto the matching pairs between discriminative descriptions and corresponding image regions, whose weights gradually increase, while the weights of generic descriptions are suppressed. This soft assignment process precisely characterizes the complex one-to-many semantic correspondences between local regions and fine-grained

descriptions.

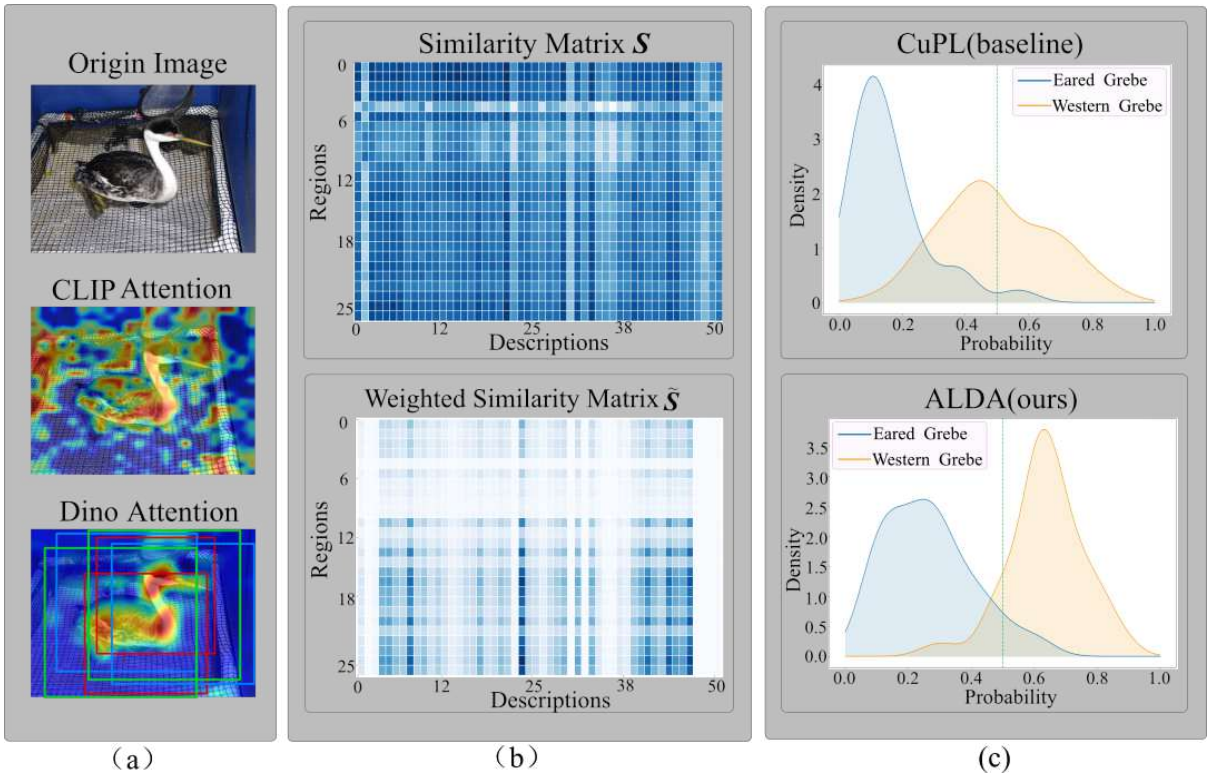


Fig. 9 Qualitative visualization and mechanism verification of ALDA. (a) Attention-guided adaptive sampling: high-attention regions use smaller scales to focus on details, while background regions use larger scales to retain context. (b) Bidirectional iterative propagation: weights automatically focus on discriminative matching pairs. (c) Fine-grained classification on CUB: ALDA significantly increases the embedding distance between confusable classes, resulting in a clearer decision boundary.

Fig. 9(c) compares the embedding distributions of the baseline method and ALDA for the confusable classes "Pied-billed Grebe" and "Black-necked Grebe" on the CUB dataset. Under the baseline method, the features of the two classes heavily overlap with a blurred decision boundary. In contrast, ALDA, through the positive synergy of attention sampling providing high-SNR candidate regions and iterative propagation reinforcing high-quality matching pairs, achieves a clear separation of the embeddings, resulting in a well-defined decision boundary. The above visualization results are highly consistent with the synergistic gain observed in the ablation study, where the combined modules yield gains surpassing the sum of their individual contributions.

This fully validates that ALDA's closed-loop design, from "where to look, at what scale" to "what to look at, which description to trust", offers core advantages in enhancing fine-grained discriminative robustness and interpretability.

4.7 Computational efficiency analysis

Fig. 10 compares methods (WCA, ABS) and the baseline (CLIP) across three dimensions: accuracy vs. number of sampled regions, accuracy vs. inference time, and time ratio vs. number of sampled regions, highlighting the comprehensive advantages of ALDA. When N=40 on an RTX 4090 GPU, ALDA achieves 72.08% accuracy with a per-image inference time of only 0.1091 seconds and a time ratio of 8.3. Compared to WCA, ALDA introduces an additional 0.0108 seconds per image due to DINO and iterative propagation, yet brings a significant performance gain (1%). Compared to ABS (which has a time ratio of 15.6 at similar accuracy), ALDA achieves the optimal balance between accuracy and efficiency.

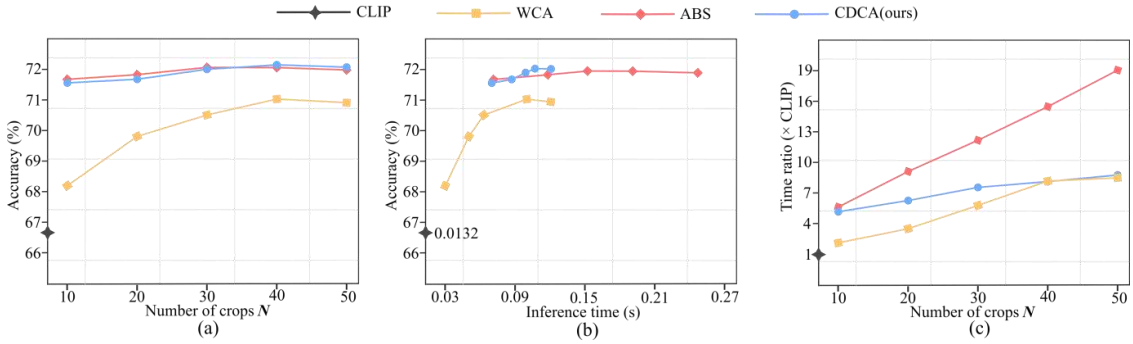


Fig. 10 Accuracy-inference efficiency trade-off (RTX 4090, ViT-B/16). (a) Accuracy as a function of the number of sampled regions. (b) Per-image inference time comparison. (c) Time ratio curve. ALDA achieves 72.08% accuracy at N=40 with only 0.1091s inference time per image, significantly faster than ABS while reaching the optimal accuracy-efficiency balance.

Tabel 4 Efficiency and cost characteristics of ALDA compared to representative methods (WCA, ABS, ProAPO). The table contrasts zero-shot constraints, pre-inference overhead, inference latency, and accuracy.

Method	**Zero-shot Constraint**	**Extra cost before inference**	**Per-image time (s)**	**ImageNet acc. (%)**
ALDA	Strict	None	0.1091	72.08
WCA	Strict	None	0.0983	71.08
ABS	Strict	None	0.2065	71.92
ProAPO	One-shot	~15min	0.×××[1]	69.90

[1]ProAPO does not report per-image inference time; its inference cost depends on the final number of selected descriptions and the aggregation protocol.

In terms of computational efficiency and deployment complexity, ALDA's advantages are even more pronounced. The evolutionary search process of ProAPO [32], although accelerated to approximately 15 minutes through subsampling strategies, still constitutes a mandatory pre-processing cost that must be incurred for every new task before inference can begin. Moreover, this process

requires storing a small set of labeled samples for each task. In contrast, ALDA operates in an out-of-the-box manner, requiring no pre-computation, proxy task training, or labeled samples. Its per-image inference time (0.1091 seconds) already encompasses all dynamic sampling and iterative alignment computations, with no upfront temporal overhead. Tabel 4 further analyzes the end-to-end deployment efficiency and cost characteristics of ALDA against the representative methods WCA, ABS, and ProAPO.

5 Conclusion

In this paper, we have addressed two key bottlenecks in zero-shot classification with vision-language models: the dynamics of spatial sampling and the bidirectional coupling of semantic matching. We propose Attention-Guided Local Dynamic Alignment (ALDA), which, under the constraints of no training, no fine-tuning, and preserving CLIP's zero-shot generalization ability, achieves more robust and interpretable fine-grained cross-modal semantic alignment through attention-driven adaptive multi-scale sampling and bidirectional iterative co-alignment between regions and descriptions. Extensive experiments on six standard datasets and four out-of-distribution generalization benchmarks demonstrate that ALDA achieves state-of-the-art performance on the ViT-B/16 backbone, with an average accuracy of 69.17%. Consistent gains are obtained in fine-grained classification and out-of-distribution robustness scenarios, with particularly significant improvements on challenging out-of-distribution datasets such as ImageNet-A.

ALDA reveals that attention-guided adaptive sampling and bidirectional iterative propagation alignment can form an effective complement in design: the former provides the necessary discriminative visual input for the latter, while the latter fully unleashes the fine-grained expressive power of the former, together constituting an effective paradigm for local-to-global cross-modal alignment in zero-shot scenarios. The method has limited advantages in artistic style image scenarios dominated by strong global semantics. Future work will explore: 1) extending ALDA to video understanding tasks, using temporal attention to guide cross-frame sampling; 2) incorporating a small amount of labeled samples to achieve semi-automatic prompt optimization.

Funding Declaration: This research was conducted without external funding.

References

1. Radford, A., Kim, J.W., Hallacy, C.: A. Ramesh, G. Goh, et al., Transferable visual models from natural language supervision, in: International conference on machine learning, 139, 8748-8763(2021).
2. Wang, Y., Li, K., Li, Y., et al.: Internvideo: General video foundation models via generative and discriminative learning, arXiv preprint arXiv:2212.03191(2022).
3. Li, K., He, Y., Wang, Y., et al.: Videochat: Chat-centric video understanding, Science China Information Sciences, 68(10): 200102(2025).
4. Li, K., Wang, Y., He, Y., et al.: Mvbench: A comprehensive multi-modal video understanding benchmark, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 22195-22206(2024).
5. Jia, C., Yang, Y., Xia, Y., et al.: Scaling up visual and vision-language representation learning with noisy text supervision, in: International conference on machine learning. PMLR, 4904-4916(2021).
6. Li, J., Li, D., Xiong, C., et al.: Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation, in: International conference on machine learning. PMLR, 12888-12900(2022).
7. Brown, T., Mann, B., Ryder, N., et al.: Language models are few-shot learners, Advances in neural information processing systems,vol. 33, 1877-1901(2020).
8. Menon, S., Vondrick, C.: Visual classification via description from large language models, arXiv preprint arXiv:2210.07183(2022).
9. Ren, Z., Su, Y., Liu, X.: ChatGPT-powered hierarchical comparisons for image classification, Advances in neural information processing systems, 36, 69706-69718(2023).
10. Achiam, J., Adler, S., Agarwal, S., et al.: Gpt-4 technical report, arXiv preprint arXiv:2303.08774(2023).
11. Pratt, S., Covert, I., Liu, R., et al.: What does a platypus look like? generating customized prompts for zero-shot image classification, in: Proceedings of the IEEE/CVF international conference on computer vision, 15691-15701(2023).
12. Li, J., Li, H., Erfani, S., et al.: Visual-text cross alignment: Refining the similarity score in vision-language models, arXiv preprint arXiv:2406.02915(2024).
13. Chen, S., Duan, B., Khan, S., et al.: Interpretable zero-shot learning with locally-aligned vision-language model, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 478-487(2025),.
14. Cai, L., Kang, J., Li, S., et al.: From local details to global context: Advancing vision-language models with attention-based selection, arXiv preprint arXiv:2505.13233(2025).
15. Zhu, Y., Ji, Y., Zhao, Z., et al.: Awt: Transferring vision-language models via augmentation, weighting, and transportation, Advances in Neural Information Processing Systems, 37, 25561-25591(2024).
16. Caron, M., Touvron, H., Misra, I., et al.: Emerging properties in self-supervised vision transformers, in: Proceedings of the IEEE/CVF international conference on computer vision, 9650-9660(2021).
17. Roth, K., Kim, J.M.: A. Koepke, et al., Waffling around for performance: Visual classification with random words and broad concepts, in: Proceedings of the IEEE/CVF international

conference on computer vision, 15746-15757(2023).
18. Noori, M., Cheraghalikhani, M.: A. Bahri, et al., Tfs-vit: Token-level feature stylization for domain generalization, Pattern Recognition, vol. 149, 110213(2024).
19. Ni, H., Li, Y., Gao, L., et al.: Part-aware transformer for generalizable person re-identification, in: Proceedings of the IEEE/CVF international conference on computer vision, 11280-11289(2023).
20. Liao, S., Shao, L., Transmatcher: Deep image matching through transformers for generalizable person re-identification, Advances in Neural Information Processing Systems, vol. 34, 1992-2003(2021).
21. Li, H., Zhang, C., Hu, Z., et al.: Interactive attack-defense for generalized person re-identification, Neural Networks, vol. 176, 106349(2024).
22. Deng, J., Dong, W., Socher, R., et al.: Imagenet: A large-scale hierarchical image database, in: 2009 IEEE conference on computer vision and pattern recognition, Ieee, 248-255(2009).
23. Wah, C., Branson, S., Welinder, P., et al.: The caltech-ucsd birds-200-2011 dataset(2011).
24. O.M. Parkhi, A. Vedaldi, A. Zisserman, et al., Cats and dogs, in :2012 IEEE conference on computer vision and pattern recognition. IEEE, pp.3498-3505(2012).
25. Cimpoi, M., Maji, S., Kokkinos, I., et al.: Describing textures in the wild, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 3606-3613(2014).
26. Bossard, L., Guillaumin, M., et al.: Food-101–mining discriminative components with random forests, in: European conference on computer vision, Cham: Springer International Publishing, 446-461(2014).
27. Zhou, B., Lapedriza, A., Khosla, A., et al.: Places: A 10 million image database for scene recognition, IEEE transactions on pattern analysis and machine intelligence, vol. 40, 1452-1464(2017).
28. Recht, B., Roelofs, R., Schmidt, L., et al.: Do imagenet classifiers generalize to imagenet?, in: International conference on machine learning. PMLR, 5389-5400(2019).
29. Hendrycks, D., Basart, S., Mu, N., 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, 8340-8349(2021).
30. Wang, H., Ge, S., Lipton, Z., et al.: Learning robust global representations by penalizing local predictive power, Advances in neural information processing systems, vol. 32(2019).
31. Hendrycks, D., Zhao, K., Basart, S., et al.: Natural adversarial examples, in: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 15262-15271(2021).
32. Qu, X., Gou, G., Zhuang, J., et al.: Proapo: Progressively automatic prompt optimization for visual classification, in: Proceedings of the Computer Vision and Pattern Recognition Conference, 25145-25155(2025).
33. Zhou, K., Yang, J., Loy, C.C., et al.: Conditional prompt learning for vision-language models, in: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp.16816-16825(2022).
34. Yao, H., Zhang, R., Xu, C.: Visual-language prompt tuning with knowledge-guided context optimization, in: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 6757-6767(2023).
35. Shu, M., Nie, W., Huang, D., et al.: Test-time prompt tuning for zero-shot generalization in vision-language models, Advances in Neural Information Processing Systems,vol. 35, 14274-14289(2022).

36. Zhu, B., Niu, Y., Han, Y., et al.: Prompt-aligned gradient for prompt tuning, in: Proceedings of the IEEE/CVF international conference on computer vision, 15659-15669(2023).
37. Karmanov, A., Guan, D., Lu, S., et al.: Efficient test-time adaptation of vision-language models, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 14162-14171(2024).
38. Palatucci, M., Pomerleau, D., Hinton, G.E., et al.: Zero-shot learning with semantic output codes, Advances in neural information processing systems(2009).
39. Xian, Y., Schiele, B., Akata, Z.: Zero-shot learning-the good, the bad and the ugly, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 4582-4591(2017).
40. Alayrac, J.B., Donahue, J., Luc, P. A.: Miech, I. Barr, Y. Hasson, K. Lenc, A. Mensch, K. Millican, M. Reynolds, et al., Flamingo: a visual language model for few-shot learning, Advances in Neural Information Processing Systems, vol. 35, 23716–23736(2022).
41. Zhou, K., Yang, J., Loy, C.C., et al.: Learning to prompt for vision-language models, International journal of computer vision, 130(9), 2337-2348(2022).
42. Feng, C., Zhong, Y., Jie, Z., et al.: Promptdet: Towards open-vocabulary detection using uncurated images, in: European conference on computer vision. Cham: Springer Nature Switzerland, 701-717(2022).
43. Du, Y., Wei, F., Zhang, Z., et al.: Learning to prompt for open-vocabulary object detection with vision-language model, in: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 14084-14093(2022).
44. Rombach, R., Blattmann, A., Lorenz, D., et al.: High-resolution image synthesis with latent diffusion models, in: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 10684-10695(2022).
45. Weng, Z., Yang, X., Li, A., et al.: Open-vclip: Transforming clip to an open-vocabulary video model via interpolated weight optimization, in: International conference on machine learning, 36978-36989(2023).