UVA: A UNIFIED VISION-LANGUAGE ALIGNMENT FRAMEWORK FOR FINE-GRAINED ZERO-SHOT AND FEW-SHOT CLASSIFICATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Vision-language models such as Contrastive Language-Image Pre-training (CLIP) achieve strong zero-shot recognition, yet they remain unreliable on fine-grained visual classification, where visually similar classes differ only in small, localised parts. Recent training-free methods address this by matching selected image regions against detailed class descriptions generated by large language models (LLMs). We show that this is insufficient: on CUB-200-2011, zero-shot region matching reaches only 61.17% accuracy, and accuracy falls to 0% on the images whose regions are least well grounded in their class descriptions. We introduce Unified Vision-Language Alignment (UVA), a framework built on a frozen CLIP ViT-B/16 that combines Attention-guided Bounding-box Selection (ABS), Bi-directional Fine-grained Text Alignment Description Refinement (BiFTA-DR), lightweight visual and textual residual adapters, and Weighted Cross-Alignment (WCA). The adapters are trained with a region-to-class-prototype contrastive objective that aligns local image regions with refined class descriptions. With 16 labelled images per class, UVA improves CUB-200-2011 accuracy from 56.02% (CLIP) to $80.29 \pm 0.31\%$ while training only 1.05M parameters, approximately 0.7% of the backbone. Across Oxford-IIIT Pet, ImageNet-100, and CUB-200-2011, the benefit of adaptation increases with the difficulty of the zero-shot task. Finally, we show that agreement between WCA and Localized Zero-Shot Learning (LaZSL), an independent optimal-transport matcher, provides an unsupervised confidence signal: predictions on which the two agree are 28–41 percentage points more accurate than those on which they disagree.

1 INTRODUCTION

Contrastive Language-Image Pre-training (CLIP) (Radford et al., 2021) learns a joint embedding space for images and text, which allows a new category to be recognised by comparing an image with a short textual prompt such as `a photo of a [CLASS]`. Fine-grained visual classification (FGVC), however, remains challenging: classes are visually similar, and the evidence that separates them, such as the colour of an eye-ring or the pattern of a wing bar, occupies only a small part of the image. On CUB-200-2011 (Wah et al., 2011), a 200-class bird benchmark, zero-shot CLIP with a ViT-B/16 image encoder reaches only 56.02% accuracy.

We identify three main bottlenecks:

1. **Global pooling.** CLIP represents the whole image with a single embedding, in which small discriminative details are averaged together with the background.
2. **Fine-grained attribute dilution.** A class name says little about the parts that distinguish it from similar classes. Descriptions generated by large language models (LLMs) add this detail, but uncurated sets contain many near-duplicates and some unreliable attributes.
3. **Local grounding breakdown.** Recent methods select informative image regions and match them against class descriptions: Attention-guided Bounding-box Selection (ABS) (Cai et al., 2025), Bi-directional Fine-grained Text Alignment Description Refinement (BiFTA-DR) (Sun et al., 2026), and Weighted Cross-Alignment (WCA) (Li et al., 2024). Together, ABS and WCA raise zero-shot accuracy on CUB-200-2011 to 61.17%. However, when we group test images

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

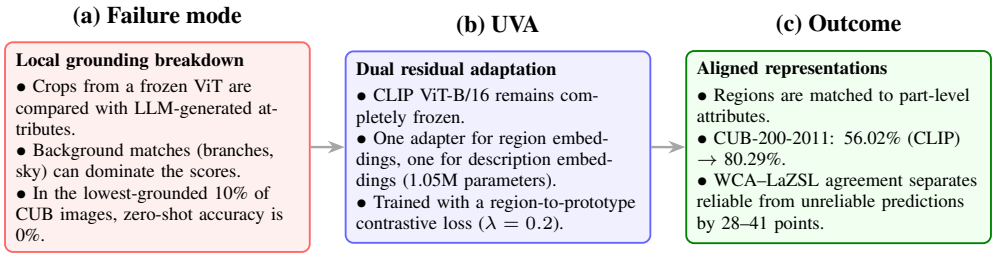


Figure 1: **Problem motivation and proposed resolution.** (a) In the frozen CLIP space, region and description embeddings are poorly aligned, so regions are often matched to generic background phrases. (b) UVA keeps the backbone frozen and adds one lightweight residual adapter per modality. (c) Adapted regions align with discriminative attributes, and agreement with an independent matcher provides a test-time confidence signal.

by the quality of their region-description grounding, accuracy in the lowest decile is 0% (Appendix D.4). Localisation alone is therefore not enough: even when the right region is selected, its embedding is often misaligned with the embeddings of the relevant descriptions.

Figure 1 summarises this failure mode and the approach we take to resolve it.

To address this misalignment, we propose Unified Vision-Language Alignment (UVA). UVA keeps the CLIP backbone frozen and learns two lightweight residual adapters from a small number of labelled images per class: a visual adapter that recalibrates region embeddings and a textual adapter that recalibrates description embeddings. The adapters are trained with a region-to-class-prototype contrastive objective, which pulls every region towards the mean description of its class and away from the prototypes of other classes. The adapted regions and descriptions are then scored with WCA.

Combining several stages naturally raises the question of whether each is necessary. Each component addresses a different bottleneck. ABS determines *where* to look, BiFTA-DR determines *which* descriptions to trust, the residual adapters make regions and descriptions comparable, and WCA determines *how much* each region-description match should contribute. The components are also complementary: a global adapter of the same form (CLIP-Adapter) reaches 66.30% on CUB-200-2011, whereas UVA reaches 80.29% by adapting region and description embeddings (Section 4). The overhead remains small: features are cached once, only 1.05M parameters are trained (18–26 minutes on one A100 GPU, versus about 110 minutes for full fine-tuning), and inference takes 38.5 ms per image.

We evaluate UVA on Oxford-IIIT Pet (Parkhi et al., 2012), ImageNet-100 (Deng et al., 2009), and CUB-200-2011, which cover increasingly fine-grained recognition settings. The gain from adaptation over zero-shot ABS+WCA grows accordingly: +0.33 points on Pets, +0.99 on ImageNet-100, and +19.12 on CUB-200-2011. We refer to this pattern as the *Difficulty-Scaling Rule*. In addition, agreement between WCA and the independent LaZSL verifier (Chen et al., 2025) provides an unsupervised confidence signal.

1.1 Main Contributions

- **Diagnosis of local representation misalignment.** We show that zero-shot fine-grained region matching fails mainly because region and description embeddings are misaligned, rather than because the regions are inaccurate; WCA accuracy falls to 0% where grounding is weakest.
- **Frozen-backbone residual calibration.** Lightweight visual and textual residual adapters, trained with a region-to-class-prototype contrastive objective, recalibrate local features without updating the backbone.
- **A unified fine-grained architecture.** ABS, BiFTA-DR, residual calibration, and WCA are integrated into a single framework.

- **Substantial fine-grained gains.** With 16 images per class, UVA reaches $80.29 \pm 0.31\%$ on CUB-200-2011, 24.27 points above CLIP and about 14 above CLIP-Adapter and Tip-Adapter; with one image per class it already surpasses 16-shot CLIP-Adapter.
- **The Difficulty-Scaling Rule.** The gain from adaptation increases monotonically as zero-shot accuracy decreases across our three benchmarks.
- **An unsupervised agreement gate.** WCA–LaZSL agreement separates reliable from unreliable predictions by 28–41 points without labels.

2 Related Work

Prompt learning. Prompt-learning methods replace CLIP's hand-crafted prompts with learned ones. CoOp (Zhou et al., 2022b) learns context vectors, CoCoOp (Zhou et al., 2022a) conditions them on the image, PLOT (Chen et al., 2023) aligns multiple prompts with local features through optimal transport, and MaPLe (Khattak et al., 2023) learns prompts in both branches. Most still compare one global image embedding with one text embedding per class.

Feature adapters. CLIP-Adapter (Gao et al., 2024) appends a small residual multi-layer perceptron (MLP) to the image or text encoder and blends its output with the original feature. Tip-Adapter (Zhang et al., 2022) constructs a key-value cache from the few-shot training features, and CaFo (Zhang et al., 2023) extends this cache with knowledge from several foundation models. UVA's adapters resemble CLIP-Adapter in form but recalibrate the embeddings of *image regions* and *class descriptions* rather than the global image embedding, and each region is trained against class prototypes built from the adapted descriptions.

Description-based classification. Classification by description (Menon & Vondrick, 2023), which we refer to as CLIP-D, and Customized Prompts via Language models (CuPL) (Pratt et al., 2023) query an LLM for visual attributes of each class and average their text embeddings. Follow-up differential descriptions (Esfandiarpoor & Bach, 2024) generate descriptions that explicitly contrast commonly confused classes. Because raw description sets are often redundant, we incorporate BiFTA-DR (Sun et al., 2026), which combines cosine deduplication with relevance ranking.

Localised alignment. MaskCLIP (Dong et al., 2023) improves CLIP's local features during pre-training through masked self-distillation. ABS (Cai et al., 2025) uses the self-attention of the ViT to extract informative bounding boxes, and WCA (Li et al., 2024) scores these crops against attribute phrases with a weighted cross-modal similarity. GOAL (Choi et al., 2025) studies global-local object alignment, and LaZSL (Chen et al., 2025) performs localised zero-shot matching through optimal transport. These methods are either training-free or fine-tune the backbone. UVA shows that frozen cross-modal representations remain misaligned on fine-grained tasks even when features are localised, and that a small trained calibration of both modalities closes much of this gap.

3 Methodology

Given an image x and a label space $\mathcal{C}$, UVA extracts localised regions, curates class descriptions, recalibrates both with residual adapters, and scores region-description correspondences with WCA (Figure 2). In parallel, an independent optimal-transport matcher provides a label-free confidence signal. Table 3 (Appendix A) links each contribution to the component that implements it and the experiment that supports it. Throughout, N denotes the number of regions per image, K the number of retained descriptions per class, and $D = 512$ the embedding dimension. All embeddings are ℓ_2-normalised. A complete list of symbols is given in Appendix A, and full pseudocode in Algorithms 1 and 2.

3.1 Frozen Vision-Language Feature Extraction

UVA keeps every parameter of the pre-trained CLIP ViT-B/16 model frozen, which preserves its general-purpose representation. Let $f_{\text{img}}(\cdot)$ and $f_{\text{text}}(\cdot)$ denote the image and text encoders. For an image or image region v and a text string t, the normalised embeddings are

$$r(v) = \frac{f_{\text{img}}(v)}{\|f_{\text{img}}(v)\|_2} \in \mathbb{R}^D, \qquad e(t) = \frac{f_{\text{text}}(t)}{\|f_{\text{text}}(t)\|_2} \in \mathbb{R}^D. \tag{1}$$

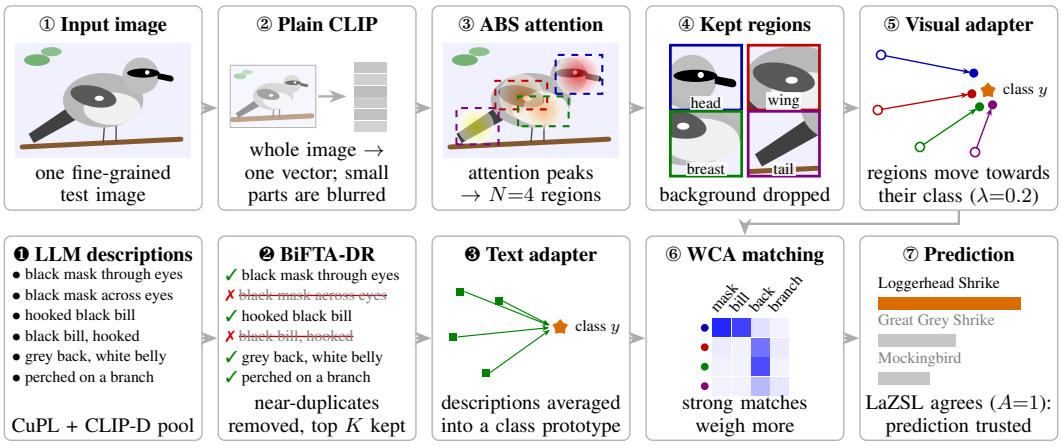


Figure 2: **How an image and its class descriptions change at each stage of UVA** (illustrative example). Plain CLIP ② compresses the whole image into one vector; ABS ③④ keeps only informative regions, and BiFTA-DR ❷ removes duplicate descriptions. The two adapters ⑤ ❸, the only trained parts, align regions and descriptions before WCA ⑥ predicts the class ⑦; LaZSL runs in parallel as an agreement check.

Because the encoders are never updated, these embeddings are computed once and cached (Section 5.1).

3.2 Attention-Guided Visual Region Selection (ABS)

Pooling the whole image into one vector dilutes small, discriminative features. To isolate them, ABS (Cai et al., 2025) uses the self-attention of the class token (`[CLS]`) in the final layer of the Vision Transformer (ViT). Let $\mathbf{A} \in \mathbb{R}^{H \times W}$ be this attention map over the $H \times W$ grid of image patches, normalised so that its entries sum to one. Local maxima of $\mathbf{A}$ serve as region centres. Starting from each centre, a bounding region is expanded to neighbouring patches in order of their attention until the cumulative attention inside the region reaches the ABS attention-energy threshold θ_{attn}. Because the threshold applies to the accumulated attention mass rather than to individual patches, the size of each region adapts to the spatial extent of the attended object or part, without a fixed box size. We use the ABS cumulative-energy criterion of Cai et al. (2025).

ABS yields N regions $\{\text{crop}_1, \ldots, \text{crop}_N\}$ per image (we use $N = 4$). Each region is resized to 224×224 and encoded, giving the set of region embeddings

$$\mathcal{Z}_v = \{r_1, \ldots, r_N\}, \qquad r_i = r(\text{crop}_i). \tag{2}$$

An optional view-refinement filter (BiFTA-VR) keeps crop_i only if its intersection-over-union (IoU) with every kept region is below $\theta_{\text{IoU}} = 0.5$, the conventional non-maximum-suppression threshold. It changes accuracy by at most 0.14 points (Section 4), because ABS regions are seeded at distinct attention peaks and rarely overlap, so we do not use it in the final model.

3.3 Textual Description Refinement (BiFTA-DR)

For each class c, we pool LLM-generated descriptions from CuPL (Pratt et al., 2023) and CLIP-D (Menon & Vondrick, 2023):

$$\mathcal{T}_c = \{d_{c,1}, d_{c,2}, \ldots, d_{c,M_c}\}, \tag{3}$$

where M_c is the number of raw descriptions for class c. These pools contain many near-synonymous sentences, which over-weight a few generic attributes and dilute the influence of the discriminative ones. Following BiFTA-DR (Sun et al., 2026), a description is retained only if it is not a near-duplicate of another description of the same class:

$$\text{keep}(d_{c,i}) \iff \max_{j \neq i} e(d_{c,i})^\top e(d_{c,j}) < \epsilon, \qquad \epsilon = 0.99. \tag{4}$$

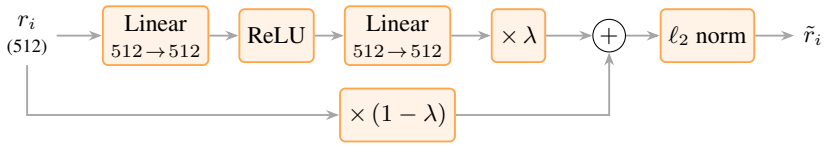


Figure 3: **Structure of a residual adapter** (A_v shown; A_t is identical with separate weights). The residual path keeps a fraction $1 - \lambda$ of the frozen CLIP embedding.

The threshold is deliberately conservative. It removes exact and near-verbatim paraphrases, which are frequent in LLM output, while keeping sentences that differ in a single attribute, such as "red crown" and "red nape". A lower threshold would risk merging such descriptions and losing the fine distinctions that FGVC depends on. The retained descriptions are then ranked by cross-entropy relevance to the baseline class prompt used by the CLIP baseline, and the top $K = 30$ are kept, giving the curated set $\mathcal{T}_c^*$ with embeddings $d_{c,k}^* = e(\cdot)$, $k = 1, \ldots, K$.

3.4 Lightweight Residual Representation Adaptation

Region and description embeddings come from different encoders and very different inputs, and our diagnosis (Section 1) shows that on fine-grained data they are often misaligned. UVA therefore introduces two residual adapters, a visual adapter $A_v : \mathbb{R}^D \to \mathbb{R}^D$ and a textual adapter $A_t : \mathbb{R}^D \to \mathbb{R}^D$, that recalibrate both modalities without modifying the backbone.

Adapter Architecture and Justification. Both adapters are two-layer MLPs with a ReLU non-linearity (Figure 3):

$$A(z) = \mathbf{W}_2 \,\mathrm{ReLU}(\mathbf{W}_1 z + \mathbf{b}_1) + \mathbf{b}_2, \qquad \mathbf{W}_1, \mathbf{W}_2 \in \mathbb{R}^{512\times512}, \;\; \mathbf{b}_1, \mathbf{b}_2 \in \mathbb{R}^{512}. \tag{5}$$

The hidden dimension equals the embedding dimension ($D_{\text{mid}} = 512$), and the two adapters have separate weights, for a total of 1.05M trainable parameters. We use two layers deliberately. A single linear map applies the same transformation to the entire embedding space and is too rigid to correct misalignments that differ between regions of that space. In our experiments, three or more layers overfit quickly under the few-shot budget ($K_{\text{shot}} \leq 16$). Two layers with a ReLU provide sufficient non-linear capacity while still generalising.

The adapter output is blended with the original embedding and projected back onto the unit sphere:

$$\tilde{r}_i = \frac{\lambda A_v(r_i) + (1-\lambda) r_i}{\|\lambda A_v(r_i) + (1-\lambda) r_i\|_2}, \tag{6}$$

$$\tilde{d}_{c,k} = \frac{\lambda A_t(d_{c,k}^*) + (1-\lambda) d_{c,k}^*}{\|\lambda A_t(d_{c,k}^*) + (1-\lambda) d_{c,k}^*\|_2}, \tag{7}$$

where $\lambda = 0.2$ is a fixed residual blend factor. With this value, 80% of each adapted feature comes from the frozen CLIP embedding and 20% from the learned correction. The adapter therefore refines CLIP's representation rather than replacing it, and the pre-trained features dominate early in training. It gave the highest accuracy in our sweep (Section 5.2). Appendix B proves that if the adapter output has norm at most B and $\lambda < 1/(1+B)$, every adapted feature keeps a strictly positive cosine similarity with its frozen counterpart (for $B = 1$, whenever $\lambda < 0.5$).

Contrastive Training Objective. For each class c, a prototype is formed by averaging its adapted description embeddings:

$$\mu_c = \frac{1}{K} \sum_{k=1}^{K} \tilde{d}_{c,k}, \qquad \hat{\mu}_c = \frac{\mu_c}{\|\mu_c\|_2}. \tag{8}$$

The adapters are trained in the 16-shot setting with a region-to-class-prototype Information Noise-Contrastive Estimation (InfoNCE) objective. For an image with ground-truth label $y \in \mathcal{C}$ and adapted regions $\{\tilde{r}_i\}_{i=1}^N$,

$$\mathcal{L}_{\text{contrastive}} = -\frac{1}{N} \sum_{i=1}^{N} \log \frac{\exp(\tilde{r}_i^\top \hat{\mu}_y / \tau_{\text{train}})}{\sum_{c\in\mathcal{C}} \exp(\tilde{r}_i^\top \hat{\mu}_c / \tau_{\text{train}})}, \tag{9}$$

where $\tau_{\text{train}} = 0.07$. Each region inherits its image's label. Because the prototypes are built from adapted descriptions, the loss trains both adapters jointly: A_v moves regions towards their class prototype, and A_t moves each prototype towards the regions of its class. No gradient reaches the backbone.

3.5 Weighted Cross-Modal Matching (WCA)

At inference, WCA (Li et al., 2024) scores every adapted region against every adapted description. For region i and description k of class c, WCA computes a temperature-scaled similarity and a softmax weight over the joint class-description space:

$$\text{sim}(i,c,k) = \frac{\tilde{r}_i^\top \tilde{d}_{c,k}}{\tau_{\text{WCA}}}, \qquad p(i,c,k) = \frac{\exp(\text{sim}(i,c,k))}{\sum_{c' \in \mathcal{C}} \sum_{k'=1}^{K} \exp(\text{sim}(i,c',k'))}, \qquad \tau_{\text{WCA}} = 0.03, \tag{10}$$

so that the descriptions that best match a region dominate its contribution. The weighted similarities are summed over descriptions and regions to obtain the class score and prediction:

$$s(i,c) = \sum_{k=1}^{K} p(i,c,k)\,\text{sim}(i,c,k), \qquad S_{\text{WCA}}(c) = \sum_{i=1}^{N} s(i,c), \qquad \hat{y}_{\text{WCA}} = \arg\max_{c \in \mathcal{C}} S_{\text{WCA}}(c). \tag{11}$$

3.6 Unsupervised Multi-Metric Agreement Gate

To estimate the reliability of a prediction without ground-truth labels, UVA runs LaZSL (Chen et al., 2025) as an independent verifier. Whereas WCA aggregates softmax-weighted similarities, LaZSL treats the regions of an image and the descriptions of a class as two distributions and computes the minimum cost of transporting one onto the other, the 1-Wasserstein or Earth Mover's Distance. The two predictions are compared:

$$A = \mathbb{1}\left[\hat{y}_{\text{WCA}} = \hat{y}_{\text{LaZSL}}\right]. \tag{12}$$

Agreement ($A = 1$) means that two scoring rules with different mathematical formulations support the same class. Disagreement ($A = 0$) is used as a test-time flag for ambiguous or unusual inputs. We additionally monitor the ratio of the two highest WCA scores, $\rho = S^{(1)}_{\text{WCA}}/(S^{(2)}_{\text{WCA}} + \epsilon_0)$ with $\epsilon_0 = 10^{-5}$, as an interpretable margin: a large ρ indicates a clear winner, and $\rho \approx 1$ indicates a near tie.

4 Experiments

4.1 Experimental Setup

Benchmarks and Selection Rationale. We evaluate UVA on three benchmarks chosen to span three levels of zero-shot difficulty. This range is what allows us to study when local adaptation is beneficial.

- **Oxford-IIIT Pet** (Parkhi et al., 2012): 37 cat and dog breeds. Zero-shot CLIP already reaches 88.20%, so this near-ceiling benchmark tests whether local adaptation harms performance when global features are already sufficient.
- **ImageNet-100**: a 100-class subset of ImageNet (Deng et al., 2009), consisting mainly of mammals, reptiles, arthropods, and marine life (Appendix C). With a zero-shot accuracy of 79.47%, it represents a moderate, broad-domain setting.
- **CUB-200-2011** (Wah et al., 2011): 200 North American bird species. Global CLIP reaches only 56.02%, and correct classification depends on part-level morphological cues.

Implementation Details. All models use the OpenAI CLIP ViT-B/16 backbone. In the 16-shot setting, 16 labelled images per class are used for training and 4 per class for validation, and the full test split is used for evaluation. Few-shot results are reported as mean $\pm$ standard deviation over independent random draws of the training images. Adapters are trained with AdamW (learning rate 10^{-4}, weight decay 10^{-2}, batch size 64) for 30 epochs on Oxford-IIIT Pet and CUB-200-2011,

Table 1: **Top-1 classification accuracy (%) with a frozen CLIP ViT-B/16.** Few-shot results for UVA are mean ± std over independent draws of 16 training images per class. CLIP-E and CLIP-D were evaluated on ImageNet-100, and view refinement (VR) on Oxford-IIIT Pet and CUB-200-2011; "–" marks settings that were not evaluated.

Regime	Method	Oxford-IIIT Pet	CUB-200-2011	ImageNet-100
Zero-shot	CLIP baseline (Radford et al., 2021)	88.20	56.02	79.47
	CLIP-E (prompt ens.)	–	–	78.92
	CLIP-D (Menon & Vondrick, 2023)	–	–	79.78
	CuPL (Pratt et al., 2023)	91.71	54.59	80.81
	BiFTA-DR (Sun et al., 2026)	91.63	57.39	80.67
	ABS + WCA (Cai et al., 2025; Li et al., 2024)	92.94	61.17	82.72
	ABS + WCA + VR	92.80	61.22	–
16-shot	CoOp (Zhou et al., 2022b)	89.12	63.30	80.15
	CLIP-Adapter (Gao et al., 2024)	92.45	66.30	82.30
	Tip-Adapter (Zhang et al., 2022)	92.60	65.50	82.55
	UVA (ours)	**93.27 ± 0.27**	**80.29 ± 0.31**	**83.71**

retaining the checkpoint with the lowest validation loss. On ImageNet-100 we use early stopping with a patience of 5 epochs, and the best checkpoint typically occurs between epochs 13 and 18. Because ImageNet-100 has about 1,170 training images per class (versus about 30 for CUB-200-2011), we draw its training and validation images from a pool of 30 per class to keep region-feature precomputation tractable. The same hyperparameters are used for all three datasets without per-dataset tuning; they are listed, with the rationale for each, in Appendix C.

4.2 Main Results and Component Ablations

Table 1 reports the progression from zero-shot CLIP (88.20%, 79.47%, and 56.02% on Oxford-IIIT Pet, ImageNet-100, and CUB-200-2011) through each component to the few-shot methods. ABS+WCA improves all three datasets, to 92.94% (+4.74), 82.72% (+3.25), and 61.17% (+5.15) respectively. Matching localised regions exposes evidence that the global embedding blurs.

Description and View Refinement (BiFTA-DR / BiFTA-VR). Filtering redundant descriptions with BiFTA-DR helps on CUB-200-2011, where accuracy rises from 54.59% with raw CuPL descriptions to 57.39% (+2.80). On the coarser datasets the effect is negligible (Oxford-IIIT Pet: 91.71%→91.63%; ImageNet-100: 80.81%→80.67%). The difference is explained by the description pools themselves (Appendix D.7). For bird species, the LLM repeatedly produces near-synonymous descriptions of generic features, and about 37% of the pool is removed as redundant. These duplicates over-weight a few non-discriminative attributes in the class prototype, so removing them improves accuracy. For pet breeds, the descriptions are more distinct and only about 8% are removed, which leaves the prototypes essentially unchanged. IoU-based view refinement (BiFTA-VR) has only a marginal effect (Oxford-IIIT Pet: 92.94%→92.80%; CUB-200-2011: 61.17%→61.22%). ABS already seeds its regions at distinct attention peaks, so few overlapping regions remain to be filtered.

Residual Adaptation and the Difficulty-Scaling Rule. Residual adaptation produces the largest gain of any component. On CUB-200-2011, UVA reaches $80.29 \pm 0.31\%$ (best draw 80.60%), 19.12 points above zero-shot ABS+WCA and 24.27 above CLIP. On ImageNet-100 the gain is 0.99 points (82.72%→83.71%), and on Oxford-IIIT Pet 0.33 points on average (+0.06 to +0.60 across draws), so adaptation never degrades the zero-shot pipeline. The adapters are also highly sample-efficient: with a single labelled image per class, UVA already reaches 68.40% on CUB-200-2011, above CLIP-Adapter trained on 16 (Appendix D.6).

These results form the **Difficulty-Scaling Rule** shown in Figure 4: the gain from adaptation increases monotonically with the difficulty of the zero-shot task. Two factors account for this. The first is headroom: on Oxford-IIIT Pet, zero-shot ABS+WCA already reaches 92.94%. The second is where the discriminative evidence lies. Pet breeds and most ImageNet-100 categories differ in global shape, texture, and colour, which the frozen embedding already captures. Bird species differ

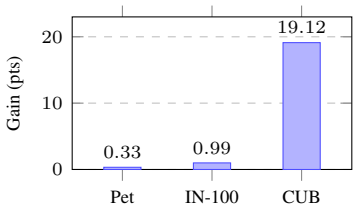


Figure 4: **The Difficulty-Scaling Rule.** Mean 16-shot gain of UVA over zero-shot ABS+WCA on Oxford-IIIT Pet, ImageNet-100 (IN-100), and CUB-200-2011.

Table 2: **Unsupervised confidence diagnostic.** Accuracy (%) when WCA and LaZSL agree or disagree, and the gap in points.

Setting	Agree	Acc. agree	Acc. disagree	Gap
Zero-shot (pooled)	∼91.2	86.40	58.10	28.30
Pet (adapted)	97.98	94.24	59.46	34.78
ImageNet-100 (adapted)	95.43	85.59	44.44	41.15

in small parts that are only visible in individual regions, and these region embeddings are exactly where the misalignment occurs. The smaller gain on ImageNet-100 may also partly reflect its capped training pool of 30 images per class. With three benchmarks we regard it as an empirical regularity rather than a law, but it gives practical guidance: local adaptation is most valuable when zero-shot accuracy is low and classes are separated by small parts.

UVA also outperforms all few-shot baselines, most clearly on CUB-200-2011, where CLIP-Adapter (66.30%) and Tip-Adapter (65.50%) trail UVA by about 14 points. Both methods adapt only the global `[CLS]` embedding, in which small parts are already mixed with the background before adaptation takes place. UVA instead recalibrates each region separately, so a region that shows the bill can be aligned with bill-related descriptions. CoOp, which learns only the text prompt and leaves image features unchanged, is the weakest few-shot method on CUB-200-2011 (63.30%). On fine-grained data, the limiting factor thus lies on the visual, local side of the matching problem.

The components are therefore complementary: the regions expose the relevant evidence, and the adapters make it comparable to the class descriptions (Appendix D.8 collects all CUB-200-2011 results).

4.3 Confidence Diagnostics and Error Calibration

The score ratio ρ averages 12.0 for correct predictions but only 1.3 for incorrect ones, whose top scores are nearly tied.

Table 2 reports the agreement-based diagnostic. After adaptation, WCA and LaZSL agree on 97.98% of Oxford-IIIT Pet and 95.43% of ImageNet-100 test instances. Agreeing instances are far more accurate than disagreeing ones (94.24% vs. 59.46% on Pets; 85.59% vs. 44.44% on ImageNet-100). The gap already exists zero-shot (28.30 points, pooled), so the signal stems from the complementarity of the two scoring rules rather than from the adapters. Disagreement thus isolates a small, error-dense subset (2–5% of instances) without labels.

Physical Reasoning for the Agreement Separation. The accuracy gap follows from the different ways in which the two matchers aggregate evidence. *Softmax attention versus optimal mass transport:* WCA's unconstrained softmax lets a single salient but misleading region, such as foliage or bark, inflate the score of a wrong class, whereas LaZSL's transport plan conserves mass across *all* regions and descriptions, limiting the influence of any single region. *Geometric disagreement under ambiguity:* on clean, prototypical images both converge on the same label, but under occlusion, clutter, or unusual appearance the locally dominated WCA score and the globally constrained transport plan tend to diverge, so disagreement ($A = 0$) concentrates on the hardest instances.

5 Detailed Analysis and Discussion

5.1 Computational Overhead and Efficiency

The two residual adapters contain 1.05M trainable parameters, approximately 0.7% of the 149.62M parameters of CLIP ViT-B/16. Because the backbone is frozen, all embeddings are computed once

and cached, and the adapters are optimised directly on the cached tensors. Training takes 18–26 minutes on a single A100 GPU, compared with approximately 110 minutes for full fine-tuning (Appendix D.1). At inference, ABS region extraction and multi-region cross-alignment increase latency from 14.2 ms to 38.5 ms per image. The increase stems from the N extra region encodings and the $O(N|\mathcal{C}|K)$ comparisons in WCA, not from the adapters.

5.2 Adapter and Description-Pool Sensitivity to λ and K

We vary the residual blend factor $\lambda \in \{0.0, 0.1, 0.2, 0.5, 1.0\}$ and the number of retained descriptions $K \in \{10, 20, 30, 40, 50\}$ on CUB-200-2011 (Appendix D.2). At $\lambda = 0$ the model reduces to zero-shot ABS+WCA (61.17%). A blend of $\lambda = 0.1$ already recovers most of the gain (78.45%), so a modest correction resolves much of the misalignment. Accuracy peaks at $\lambda = 0.2$ (80.60%) and declines for larger values ($\lambda = 0.5$: 77.20%; $\lambda = 1.0$: 74.12%). At $\lambda = 1.0$ the frozen embedding is discarded, and an adapter trained on 16 images per class cannot reproduce CLIP's knowledge; the adapter works best as a correction rather than a replacement, consistent with the anchoring bound (Appendix B). For the description pool, accuracy rises from 77.10% at $K = 10$ to 80.60% at $K = 30$ as prototypes average over more attributes, then plateaus (80.45% at $K = 50$) as the added descriptions become lower-ranked and more generic.

5.3 Limitations

- **Few-shot variance across random draws.** Performance varies with the choice of training images (Oxford-IIIT Pet: 93.00–93.54%; CUB-200-2011: 79.98–80.60%). We therefore report means and standard deviations.
- **Benchmark scope.** We evaluate three benchmarks and one backbone; full ImageNet-1k and shifted variants such as ImageNet-A/R would further test robustness.
- **Diagnostic scope on large label spaces.** The agreement gate is evaluated on Oxford-IIIT Pet and ImageNet-100; its behaviour on open-vocabulary label spaces remains to be studied.
- **Region-level supervision.** Regions inherit their image's label, so background-dominated regions add label noise.
- **Fixed region selection.** ABS and BiFTA-VR use fixed criteria; region selection that adapts to object scale could improve efficiency.

6 Conclusion

We presented Unified Vision-Language Alignment (UVA), a framework for fine-grained visual classification that keeps the CLIP ViT-B/16 backbone entirely frozen. Our starting point was a diagnosis: on fine-grained data, selecting the right regions is not sufficient, because their embeddings are often misaligned with those of the class descriptions, driving zero-shot accuracy to 0% on the least well-grounded CUB-200-2011 images. UVA addresses it with complementary components: ABS selects informative regions, BiFTA-DR removes redundant descriptions, two lightweight residual adapters trained with a region-to-class-prototype objective align regions with descriptions, and WCA scores the aligned representations.

With 1.05M trainable parameters and under half an hour of training on one GPU, UVA raises CUB-200-2011 accuracy from 56.02% to $80.29 \pm 0.31\%$, about 14 points above global adapters such as CLIP-Adapter and Tip-Adapter. It also improves on its zero-shot counterpart on the easier benchmarks. The size of the improvement follows the Difficulty-Scaling Rule: it is largest where zero-shot accuracy is lowest and classes differ in small, local details. The adapters work best as a small correction to CLIP's representation, as the residual anchoring bound makes precise. Finally, agreement between WCA and LaZSL provides a label-free confidence signal that separates reliable from unreliable predictions by 28–41 points.

More broadly, our results suggest that much of the information needed for fine-grained recognition is already present in frozen vision-language models and can be recovered by selecting better local evidence, refining class descriptions, and explicitly aligning the two modalities, without fine-tuning the backbone. Future work includes larger and distribution-shifted benchmarks, region-level supervision, and learned, scale-adaptive region selection.

LLM USAGE STATEMENT

We used large language model tools during the preparation of this manuscript for code debugging, formatting assistance, and editorial suggestions on grammar, clarity, and readability. Their role was limited to language editing and implementation assistance. All scientific contributions, including the theoretical formulations, experimental designs, empirical analyses, and conclusions, are the original work of the authors, and LLM use did not alter the scientific content or its interpretation. The authors reviewed all AI-assisted content and take full responsibility for this paper.

REPRODUCIBILITY STATEMENT

All experiments use a frozen CLIP ViT-B/16 backbone. The model architecture, training configuration, and loss are described in Section 3 and Appendix A. Benchmark splits, preprocessing, and evaluation protocols are specified in Section 4.1 and Appendix C, and all hyperparameters are listed in Table 6. The formal claim is proved in Appendix B. Source code, execution environments, description pools, and evaluation configurations are available in an anonymised repository at `https://anonymous.4open.science/r/Unified-Framework-Project-6448/` and also submitted as supplementary material with the paper.

REFERENCES

Lincan Cai, Jingxuan Kang, Shuang Li, Wenxuan Ma, Binhui Xie, Zhida Qin, and Jian Liang. From local details to global context: Advancing vision-language models with attention-based selection. In *Proceedings of the 42nd International Conference on Machine Learning (ICML)*, 2025.

Guangyi Chen, Weiran Yao, Xiangchen Song, Xinyue Li, Yongming Rao, and Kun Zhang. PLOT: Prompt learning with optimal transport for vision-language models. In *International Conference on Learning Representations (ICLR)*, 2023.

Shiming Chen, Bowen Duan, Salman Khan, and Fahad Shahbaz Khan. Interpretable zero-shot learning with locally-aligned vision-language model. *arXiv preprint arXiv:2506.23822*, 2025.

Hyungyu Choi, Young Kyun Jang, and Chanho Eom. GOAL: Global-local object alignment learning. *arXiv preprint arXiv:2503.17782*, 2025.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. ImageNet: A large-scale hierarchical image database. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 248–255, 2009.

Xiaoyi Dong, Jianmin Bao, Yinglin Zheng, Ting Zhang, Dongdong Chen, Hao Yang, Ming Zeng, Weiming Zhang, Lu Yuan, Dong Chen, Fang Wen, and Nenghai Yu. MaskCLIP: Masked self-distillation advances contrastive language-image pretraining. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10995–11005, 2023.

Reza Esfandiarpoor and Stephen H. Bach. Follow-up differential descriptions: Language models resolve ambiguities for image classification. In *International Conference on Learning Representations (ICLR)*, 2024.

Peng Gao, Shijie Geng, Renrui Zhang, Teli Ma, Rongyao Fang, Yongfeng Zhang, Hongsheng Li, and Yu Qiao. CLIP-Adapter: Better vision-language models with feature adapters. *International Journal of Computer Vision*, 132(2):581–595, 2024.

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 (CVPR)*, pp. 19113–19122, 2023.

Jinhao Li, Haopeng Li, Sarah M. Erfani, Lei Feng, James Bailey, and Feng Liu. Visual-text cross alignment: Refining the similarity score in vision-language models. In *Proceedings of the 41st International Conference on Machine Learning (ICML)*, pp. 28205–28224, 2024.

Sachit Menon and Carl Vondrick. Visual classification via description from large language models. In *International Conference on Learning Representations (ICLR)*, 2023.

Omkar M. Parkhi, Andrea Vedaldi, Andrew Zisserman, and C. V. Jawahar. Cats and dogs. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2012.

Sarah Pratt, Ian Covert, Rosanne Liu, and Ali Farhadi. 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 (ICCV)*, 2023.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning (ICML)*, pp. 8748–8763, 2021.

Yuhao Sun, Chengyi Cai, Jiacheng Zhang, Zesheng Ye, Xingliang Yuan, and Feng Liu. Let's roll a BiFTA: Bi-refinement for fine-grained text-visual alignment in vision-language models. *Transactions on Machine Learning Research*, 2026.

Catherine Wah, Steve Branson, Peter Welinder, Pietro Perona, and Serge Belongie. The Caltech-UCSD birds-200-2011 dataset. Technical Report CNS-TR-2011-001, California Institute of Technology, 2011.

Renrui Zhang, Wei Zhang, Rongyao Fang, Peng Gao, Kunchang Li, Jifeng Dai, Yu Qiao, and Hongsheng Li. Tip-Adapter: Training-free adaption of CLIP for few-shot classification. In *European Conference on Computer Vision (ECCV)*, 2022.

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 (CVPR)*, 2023.

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 (CVPR)*, 2022a.

Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Learning to prompt for vision-language models. *International Journal of Computer Vision*, 130(9):2337–2355, 2022b.

A ALGORITHMIC DESCRIPTION OF UVA

This appendix collects the notation used throughout the paper (Table 4) and links each contribution in Section 1.1 to the component that implements it and the experiment that supports it (Table 3). It then presents the two procedures that define UVA, each followed by a step-by-step explanation: inference with the parallel optimal-transport verifier (Appendix A.1, Algorithm 1) and few-shot training of the residual adapters (Appendix A.2, Algorithm 2).

Table 3: **Mapping of contributions to method components and supporting evidence.**

Contribution	Method	Evidence
Diagnosis of local misalignment	Sec. 1	Fig. 8; Table 9 ($\lambda = 0$ vs. 0.2)
Frozen-backbone residual calibration	Sec. 3.4	Tables 1, 9; Fig. 11
Unified fine-grained architecture	Secs. 3.2–3.5	Table 1; Table 10
Fine-grained gains	Sec. 4	Table 1
Difficulty-Scaling Rule	Sec. 4	Fig. 4
Unsupervised agreement gate	Sec. 3.6	Table 2

Table 4: **Notation.**

Symbol	Meaning
$\mathcal{C}$, $\lvert\mathcal{C}\rvert$	label space and number of classes
x, y	input image and its ground-truth label
f_{img}, f_{text}	frozen CLIP image and text encoders
$r(\cdot)$, $e(\cdot)$	ℓ_2-normalised image and text embeddings
D	embedding dimension (512)
$\mathbf{A}$, $H \times W$	`[CLS]` attention map and patch grid of the ViT
θ_{attn}	ABS cumulative attention-energy threshold
θ_{IoU}	IoU threshold of the optional BiFTA-VR filter (0.5)
N	number of regions per image (4)
r_i, $\tilde{r}_i$	embedding of region i before and after the visual adapter
$\mathcal{T}_c$, M_c	raw description pool of class c and its size
ϵ	BiFTA-DR deduplication threshold (0.99)
$\mathcal{T}_c^*$, K	curated description set of class c and its size (30)
$d_{c,k}^*$, $\tilde{d}_{c,k}$	k-th curated description of class c, before and after A_t
A_v, A_t	visual and textual residual adapters
λ	residual blend factor (0.2)
μ_c, $\hat{\mu}_c$	class prototype before and after normalisation
τ_{train}, τ_{WCA}	temperatures of the training loss (0.07) and of WCA (0.03)
$\text{sim}(i,c,k)$, $p(i,c,k)$	scaled similarity and WCA weight of region i and description (c,k)
$S_{\text{WCA}}(c)$	WCA score of class c
A	agreement indicator between WCA and LaZSL
ρ, ϵ_0	top-two score ratio and its stabilising constant (10^{-5})
B	upper bound on the adapter output norm (Appendix B)

Figure 5 gives a block-level view of the architecture that complements the illustrated overview in Figure 2.

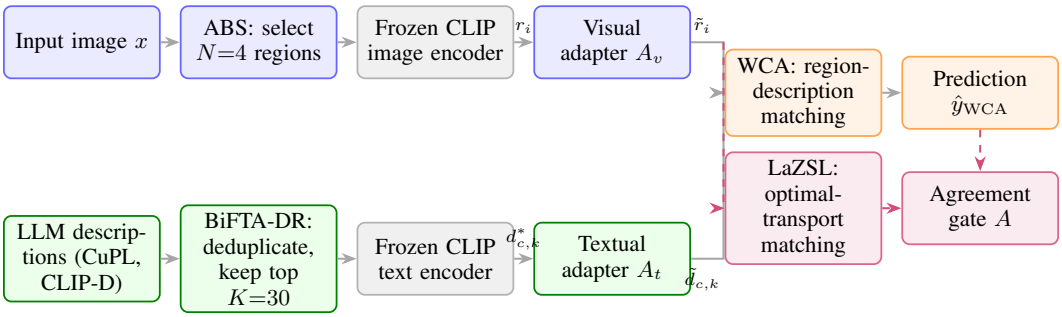


Figure 5: **Block-level architecture of UVA.** ABS extracts N localised regions from the input image, and BiFTA-DR retains K curated descriptions per class from the pooled CuPL and CLIP-D descriptions. Both are encoded by the frozen CLIP encoders (grey) and recalibrated by the residual adapters, the only trained components. WCA scores every adapted region against every adapted description. In parallel (dashed), LaZSL makes an independent prediction, and their agreement A serves as an unsupervised confidence gate.

A.1 Inference with the Parallel Verification Gate

Algorithm 1 describes how UVA classifies a single test image once the adapters have been trained. It has four blocks: region selection, residual adaptation, weighted cross-alignment, and parallel verification.

Algorithm 1 UVA Inference with Parallel Verification Gate

Require: image x; label space $\mathcal{C}$; frozen f_{img}; adapters A_v, A_t; curated $\{\mathcal{T}_c^*\}$; blend factor λ; temperature τ_{WCA}
Ensure: predicted class $\hat{y}_{\text{WCA}}$, class scores $S_{\text{WCA}}(\cdot)$, agreement gate A

Region selection (ABS)
$\{\text{crop}_i\}_{i=1}^N \leftarrow \textsc{AbsRegions}(x)$
$r_i \leftarrow f_{\text{img}}(\text{crop}_i)/\|f_{\text{img}}(\text{crop}_i)\|_2$ for $i = 1, \dots, N$
Residual adaptation
$\tilde{r}_i \leftarrow \textsc{Blend}(r_i, A_v, \lambda)$ for $i = 1, \dots, N$
$\tilde{d}_{c,k} \leftarrow \textsc{Blend}(d^*_{c,k}, A_t, \lambda)$ for all c, k
Weighted cross-alignment (WCA)
$\text{sim}(i, c, k) \leftarrow \tilde{r}_i^\top \tilde{d}_{c,k}/\tau_{\text{WCA}}$
$p(i, c, k) \leftarrow \text{softmax}_{(c,k)}(\text{sim}(i, c, k))$ ▷ joint class-description space
$S_{\text{WCA}}(c) \leftarrow \sum_{i=1}^N \sum_{k=1}^K p(i, c, k)\,\text{sim}(i, c, k)$
$\hat{y}_{\text{WCA}} \leftarrow \arg\max_c S_{\text{WCA}}(c)$
Parallel optimal-transport verification (LaZSL)
$\hat{y}_{\text{LaZSL}} \leftarrow \textsc{EarthMoversDistanceMatching}(\{\tilde{r}_i\}, \{\tilde{d}_{c,k}\})$ ▷ independent verifier
$A \leftarrow \mathbb{1}[\hat{y}_{\text{WCA}} = \hat{y}_{\text{LaZSL}}]$ ▷ unsupervised gate
return $\hat{y}_{\text{WCA}}, S_{\text{WCA}}(\cdot), A$

function $\textsc{Blend}(z, A, \lambda)$ ▷ residual blend, Eqs. (6)–(7)
  $u \leftarrow \lambda A(z) + (1 - \lambda)z$
  return $u/\|u\|_2$
end function

Offline preprocessing (before Algorithm 1). Description refinement does not depend on the test image, so it is computed once per dataset and cached. For each class c, a candidate $d_{c,i}$ is discarded if $\max_{j \neq i} e(d_{c,i})^\top e(d_{c,j}) \geq \epsilon$ with $\epsilon = 0.99$. The remaining descriptions are ranked by cross-entropy relevance to the baseline class prompt, and the top $K = 30$ form $\mathcal{T}_c^*$. Their frozen embeddings $d^*_{c,k}$ are the textual input to the algorithm.

Region selection (lines 1–2). ABSREGIONS extracts the [CLS] attention map $\mathbf{A}$ from the final ViT block and uses its local peaks as region centres. Each region is grown from its centre until its accumulated normalised attention reaches θ_{attn}, so region size adapts to the spatial extent of the attended object or part rather than being fixed in advance. The $N = 4$ regions are resized to 224×224, passed through the frozen image encoder, and ℓ_2-normalised (line 2). This step accounts for most of the additional inference cost relative to CLIP, since it requires N extra forward passes of the image encoder.

Residual adaptation (lines 3–4). Each region embedding is passed through the visual adapter and blended with its original value using the BLEND function (lines 12–15), which implements Eqs. (6)–(7): the adapter output is weighted by λ, the frozen embedding by $1 - \lambda$, and the result is projected back onto the unit sphere. Line 4 applies the same operation to every curated description with the textual adapter. Because the descriptions are fixed, the adapted description embeddings $\tilde{d}_{c,k}$ can be computed once after training and reused for every test image, so at test time only the N region embeddings pass through an adapter.

Weighted cross-alignment (lines 5–8). Line 5 computes the temperature-scaled similarity between every adapted region and every adapted description. Line 6 normalises these similarities with a softmax over all class-description pairs, separately for each region, so that each region distributes a unit of weight across the descriptions it matches best. Line 7 forms the class score as the weighted sum of similarities over all regions and all descriptions of that class, and line 8 selects the highest-scoring class. The cost of this block is $O(N|\mathcal{C}|K)$ dot products, which is small compared with the region encoding in line 2.

Parallel verification (lines 9–11). LaZSL receives the same adapted regions and descriptions and predicts a class by optimal transport, that is, by finding the class whose description set can be matched to the image's region set at the lowest Earth Mover's cost (line 9). Line 10 compares the two predictions to obtain the gate A. The algorithm returns the WCA prediction, the full score vector $S_{\text{WCA}}(\cdot)$, from which the margin ρ can be computed, and the gate. The prediction itself is always $\hat{y}_{\text{WCA}}$; the gate does not change it but indicates whether it should be trusted ($A = 1$) or flagged for review ($A = 0$).

A.2 FEW-SHOT TRAINING OF THE RESIDUAL ADAPTERS

Algorithm 2 trains the two residual adapters from the 16-shot support set, and Figure 6 shows the same loop as a flowchart.

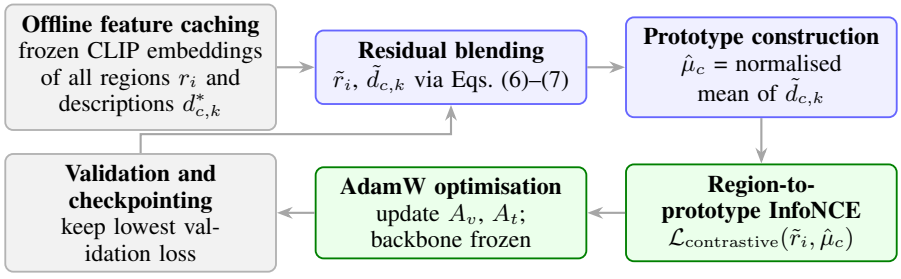


Figure 6: **Execution flowchart for UVA few-shot training (Algorithm 2).** Embeddings are precomputed once with the frozen CLIP backbone. In each epoch, the cached embeddings pass through the residual adapters, class prototypes are rebuilt, and the adapters are updated with the region-to-prototype InfoNCE objective. The validation loss is evaluated after every epoch, and the best checkpoint is retained.

Algorithm 2 UVA Adapter Training (16-shot, Frozen Backbone)

Require: support set $\mathcal{S}$ (16 shots per class); validation set $\mathcal{V}$ (4 shots per class); curated $\{\mathcal{T}_c^*\}$; epochs E; temperature τ_{train}
Ensure: adapters A_v, A_t at the best validation checkpoint
1: Precompute and cache $\{r_i\}$ for all support and validation images and $\{d^*_{c,k}\}$ for all classes
2: Initialise A_v, A_t; $\mathcal{L}^\star_{\text{val}} \leftarrow \infty$
3: **for** epoch $= 1$ to E **do**
4: **for** each mini-batch $\mathcal{B} \subset \mathcal{S}$ **do**
5: $\tilde{r}_i \leftarrow \textsc{Blend}(r_i, A_v, \lambda)$; $\tilde{d}_{c,k} \leftarrow \textsc{Blend}(d^*_{c,k}, A_t, \lambda)$
6: $\hat{\mu}_c \leftarrow \textsc{Normalize}\big(\frac{1}{K}\sum_k \tilde{d}_{c,k}\big)$ ▷ class prototypes
7: $\mathcal{L} \leftarrow -\frac{1}{N}\sum_i \log \frac{\exp(\tilde{r}_i^\top \hat{\mu}_y/\tau_{\text{train}})}{\sum_c \exp(\tilde{r}_i^\top \hat{\mu}_c/\tau_{\text{train}})}$
8: update A_v, A_t by AdamW on $\nabla\mathcal{L}$ ▷ backbone receives no gradient
9: **end for**
10: $\mathcal{L}_{\text{val}} \leftarrow \mathcal{L}$ evaluated on $\mathcal{V}$
11: **if** $\mathcal{L}_{\text{val}} < \mathcal{L}^\star_{\text{val}}$ **then**
12: save A_v, A_t; $\mathcal{L}^\star_{\text{val}} \leftarrow \mathcal{L}_{\text{val}}$
13: **end if**
14: *(ImageNet-100)* stop if the validation loss has not improved for 5 epochs
15: **end for**
16: **return** best validation checkpoint

Feature caching (line 1). Before training begins, every support and validation image is passed through ABS and the frozen image encoder, and every curated description through the frozen text encoder. The resulting embeddings r_i and $d^*_{c,k}$ are stored. Since the backbone is never updated, these embeddings stay valid for the whole of training, and no image is passed through CLIP again. This is the main reason UVA trains in 18–26 minutes rather than the approximately 110 minutes needed for full fine-tuning.

Initialisation (line 2). The adapters A_v and A_t are initialised, and the best validation loss is set to infinity so that the first epoch always produces a checkpoint. Because the residual blend keeps a fraction $1 - \lambda = 0.8$ of each frozen embedding, the adapted features start close to the original CLIP features, and training begins from a point near the zero-shot solution.

Adaptation and prototype construction (lines 5–6). For each mini-batch, the cached region embeddings are passed through the visual adapter, and all description embeddings through the textual adapter. The class prototypes $\hat{\mu}_c$ are then rebuilt as the normalised mean of each class's adapted descriptions. The prototypes are recomputed at every step rather than fixed, because the textual adapter changes during training and the prototypes must reflect its current state.

Loss and update (lines 7–8). Line 7 evaluates the region-to-prototype InfoNCE loss: for each region, the prototype of its image's class is the positive, and the prototypes of all other classes are the negatives. The loss is averaged over the N regions of each image and over the images in the mini-batch. Line 8 updates both adapters with AdamW. The gradient flows into A_v through the region embeddings and into A_t through the prototypes, so the two adapters are trained jointly towards a shared alignment. No gradient reaches the backbone.

Validation and checkpoint selection (lines 10–13). After each epoch, the same loss is computed on the held-out validation set of 4 images per class. If it is the lowest seen so far, the current adapters are saved. The returned model is therefore the checkpoint with the best validation loss rather than the final one, which guards against overfitting to the small support set. On Oxford-IIIT Pet and CUB-200-2011, training runs for 30 epochs, and the best checkpoints occur at epochs 28 and 27, respectively (Appendix D.3).

Early stopping (line 14). On ImageNet-100, training stops once the validation loss has not improved for 5 consecutive epochs. In practice the best checkpoint occurs between epochs 13 and 18, so early stopping shortens training without affecting the selected model.

B Theoretical Supplement: Residual Anchoring and the Role of λ

Let $r \in \mathbb{R}^D$ be a frozen, ℓ_2-normalised CLIP embedding ($\|r\|_2 = 1$), let $u = \lambda A(r) + (1-\lambda)r$, and let $\tilde{r} = u/\|u\|_2$. Assume that the adapter output is bounded, $\|A(r)\|_2 \le B$.

Proposition 1 (Residual anchoring). *If $\lambda B < 1 - \lambda$, then for every input r,*

$$\cos(\tilde{r}, r) \;=\; \frac{r^\top u}{\|u\|_2} \;\ge\; \frac{(1-\lambda) - \lambda B}{(1-\lambda) + \lambda B} \;>\; 0. \tag{13}$$

Proof. Since $\|r\|_2 = 1$, $r^\top u = (1-\lambda) + \lambda\, r^\top A(r)$. By the Cauchy–Schwarz inequality, $r^\top A(r) \ge -\|A(r)\|_2 \ge -B$, so $r^\top u \ge (1-\lambda) - \lambda B > 0$. By the triangle inequality, $\|u\|_2 \le (1-\lambda)\|r\|_2 + \lambda\|A(r)\|_2 \le (1-\lambda) + \lambda B$. Combining the two bounds yields the claim. □

Interpretation. The condition $\lambda B < 1 - \lambda$ is equivalent to $\lambda < 1/(1+B)$. For unit-scale adapter outputs ($B = 1$) it becomes $\lambda < 0.5$, and the lower bound takes the values in Table 5. At the chosen $\lambda = 0.2$, every adapted embedding retains a cosine similarity of at least 0.6 with its frozen counterpart, regardless of what the adapter learns. The adapted representation therefore cannot drift arbitrarily far from the pre-trained CLIP space, which protects the knowledge acquired during pre-training when only 16 labelled images per class are available. Beyond $\lambda = 0.5$ the guarantee becomes vacuous. This coincides with the region of the sweep in which few-shot accuracy degrades. The bound is a worst-case guarantee rather than a prediction of accuracy. It explains why small blend factors are safe, while the precise optimum at $\lambda = 0.2$ is determined empirically.

Table 5: **Worst-case anchoring bound of Proposition 1** at $B = 1$, alongside CUB-200-2011 16-shot accuracy from the λ sweep.

λ	0.0	0.1	**0.2**	0.5	1.0
Guaranteed $\cos(\tilde{r}, r) \ge$	1.000	0.800	**0.600**	*vacuous*	*vacuous*
CUB-200-2011 16-shot top-1 (%)	61.17	78.45	**80.60**	77.20	74.12

C Implementation Details and Hyperparameters

Adapter Architecture. The residual adapters A_v and A_t are two-layer MLPs, $A(z) = \mathbf{W}_2\,\mathrm{ReLU}(\mathbf{W}_1 z + \mathbf{b}_1) + \mathbf{b}_2$ with $\mathbf{W}_1, \mathbf{W}_2 \in \mathbb{R}^{512\times512}$, and the residual blend factor is fixed at $\lambda = 0.2$. Weights are not shared between the visual and textual adapters. Each adapter has $2 \times (512 \times 512 + 512) = 525{,}312$ parameters, giving 1.05M in total.

Optimisation. The adapters are trained with AdamW (learning rate 10^{-4}, weight decay 10^{-2}, batch size 64) for 30 epochs. The best checkpoints occur at epoch 28 on Oxford-IIIT Pet and epoch 27 on CUB-200-2011. ImageNet-100 uses early stopping with a patience of 5 epochs.

ImageNet-100 Benchmark Configuration. We evaluate on the `ambityga` split, available on Hugging Face as `ilee0022/ImageNet100`, which consists primarily of mammals, reptiles, arthropods, and marine life. To make region-feature precomputation tractable, the training pool is capped at 30 images per class before the support and validation sets are drawn, a 39× reduction in precomputation. The held-out test split is evaluated in full.

Table 6: **Complete hyperparameter listing for UVA.**

Stage	Hyperparameter	Value
Backbone	architecture	OpenAI CLIP ViT-B/16 (frozen)
Backbone	joint embedding dimension D	512
ABS	attention-energy threshold θ_{attn}	ABS cumulative-energy criterion
ABS	regions per image N	4
BiFTA-VR	IoU threshold θ_{IoU}	0.5 (not adopted)
BiFTA-DR	deduplication threshold ϵ	0.99
BiFTA-DR	retained descriptions K	30
BiFTA-DR	description sources	CuPL + CLIP-D
Adapter	hidden dimension D_{mid}	512
Adapter	non-linearity	ReLU
Adapter	residual blend factor λ	0.2
Adapter	trainable parameters	1.05M
Training	optimiser	AdamW
Training	learning rate / weight decay	10^{-4} / 10^{-2}
Training	batch size / InfoNCE temperature τ_{train}	64 / 0.07
Training	support / validation shots	16 / 4 per class
Scoring	WCA temperature τ_{WCA}	0.03
Diagnostic	margin constant ϵ_0	10^{-5}

Rationale for the Hyperparameter Values. Table 7 summarises how the main hyperparameters were chosen and where the supporting evidence appears in the paper.

Table 7: **Rationale for the main hyperparameter choices.**

Hyperparameter	Value	Rationale
Residual blend factor λ	0.2	Highest accuracy in the sweep over $\{0, 0.1, 0.2, 0.5, 1.0\}$ (Table 9). Keeps 80% of each frozen embedding, which guarantees a cosine similarity of at least 0.6 with the original feature (Proposition 1).
Retained descriptions K	30	Accuracy rises up to $K = 30$ and plateaus beyond it (Table 9); larger K only adds lower-ranked, generic descriptions.
Deduplication threshold ϵ	0.99	Removes exact and near-verbatim LLM paraphrases while keeping descriptions that differ in a single attribute (Section 3.3); removes 37% of CUB descriptions but only 8% of Pet descriptions (Figure 12).
IoU threshold θ_{IoU}	0.5	Conventional non-maximum-suppression threshold. The filter changes accuracy by at most 0.14 points and is therefore not adopted (Table 1).
Adapter depth	2 layers	A single linear layer is too rigid to correct misalignments that vary across the embedding space, while three or more layers overfit under the 16-shot budget (Section 3.4).
Adapter hidden size D_{mid}	512	Equal to the CLIP embedding dimension D, so each adapter maps the embedding space onto itself without a bottleneck.
Training temperature τ_{train}	0.07	The value used to initialise the temperature of CLIP's own contrastive loss (Radford et al., 2021), so the adapters are trained on the same similarity scale as the pre-trained model.
Support / validation shots	16 / 4	The 16-shot setting is the standard few-shot protocol for CLIP adaptation used by CoOp, CLIP-Adapter, and Tip-Adapter, which keeps our results comparable with these baselines.
Epochs	30	The best validation checkpoints occur at epoch 28 (Pet) and 27 (CUB), and the validation loss has levelled off by epoch 30 (Figure 7).
Early-stopping patience	5	On ImageNet-100 the best checkpoint occurs between epochs 13 and 18, so training can stop early without affecting the selected model.
ImageNet-100 training pool	30 / class	Keeps region-feature precomputation tractable (a 39× reduction) while providing enough images to draw 16 training and 4 validation shots per class.

D Additional Experimental Results

D.1 Computational Footprint

UVA's 1.05M trainable parameters represent approximately 0.7% of the CLIP ViT-B/16 model, and no backbone parameter is updated during adaptation. The training-time advantage over full fine-tuning comes from feature caching. Because the backbone is frozen, all region and description embeddings are precomputed once, and the adapters are optimised directly on the cached tensors (Algorithm 2). The increase in inference latency is due to the N additional region passes through the image encoder in ABS and the $O(N|\mathcal{C}|K)$ cross-alignment computation in WCA, not to the adapters themselves, whose cost is two small matrix multiplications per embedding.

Table 8: **Computational footprint and inference latency** on a single NVIDIA A100 GPU.

Method	Params	Backbone	Train time (16-shot)	Inference
CLIP ViT-B/16 (Radford et al., 2021)	0	frozen	0 min	14.2 ms
CoOp (Zhou et al., 2022b)	0.02M	frozen	∼12 min	14.8 ms
Full fine-tuning	149.62M	updated	∼110 min	14.2 ms
UVA (ours)	**1.05M**	**frozen**	**18–26 min**	**38.5 ms**

D.2 Residual Blend Factor λ and Description Depth K

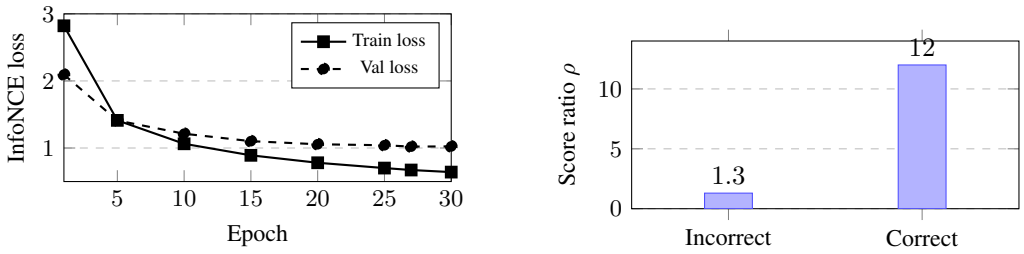
Table 9 reports the sensitivity sweeps on CUB-200-2011. At $\lambda = 0$ the model reduces to zero-shot ABS+WCA (61.17%). Introducing even a small residual blend recovers most of the available improvement, and performance peaks at $\lambda = 0.2$ (80.60%). Beyond $\lambda = 0.2$ accuracy declines, and for $\lambda \geq 0.5$, where the worst-case anchoring guarantee of Proposition 1 becomes vacuous, it falls to 77.20% and then 74.12%. For the description depth K, accuracy improves steadily up to $K = 30$ and then plateaus. Truncating the ranked pool therefore removes low-relevance descriptions without discarding useful ones, which keeps the class prototypes focused on discriminative attributes.

Table 9: **Sensitivity of CUB-200-2011 16-shot accuracy to λ and K.**

(a) Residual blend factor λ (fixed $K = 30$)					
λ	0.0 (zero-shot)	0.1	**0.2**	0.5	1.0 (pure adapter)
Top-1 acc. (%)	61.17	78.45	**80.60**	77.20	74.12
(b) Description depth K (fixed $\lambda = 0.2$)					
K	10	20	**30**	40	50
Top-1 acc. (%)	77.10	79.25	**80.60**	80.52	80.45

D.3 Training Dynamics and Calibration Signals

Contrastive adapter training on CUB-200-2011 converges stably (Figure 7a). The training loss decreases monotonically from 2.822 to 0.641, and the validation loss reaches its minimum of 1.021 at epoch 27 before levelling off. The gap between training and validation loss remains moderate, consistent with the small capacity of the adapters. The score ratio $\rho = S^{(1)}/S^{(2)}$ separates confident correct predictions ($\rho \approx 12.0$) from incorrect ones, whose two highest class scores are nearly tied ($\rho \approx 1.3$) (Figure 7b).



(a) CUB-200-2011 adapter convergence.

(b) Confidence score margin ρ.

Figure 7: **Training convergence and score separation.** (a) The contrastive loss on CUB-200-2011 converges stably, with the best validation loss at epoch 27. (b) The ratio of the two highest WCA scores is much larger for correct predictions than for incorrect ones.

D.4 Diagnostic Grounding Curve

Figure 8 reports zero-shot WCA accuracy on CUB-200-2011 across deciles of regional cross-modal grounding quality, that is, how well an image's regions are grounded in the class descriptions. Accuracy decreases steadily from 86.0% in the highest decile to 0.0% in the lowest. When the unadapted region embeddings lie far from the description embeddings, zero-shot matching fails completely, even though the regions themselves are selected by attention. This observation motivates targeted residual adaptation: the problem lies in the alignment between modalities rather than in the choice of regions, and the adapters are trained specifically to close this gap.

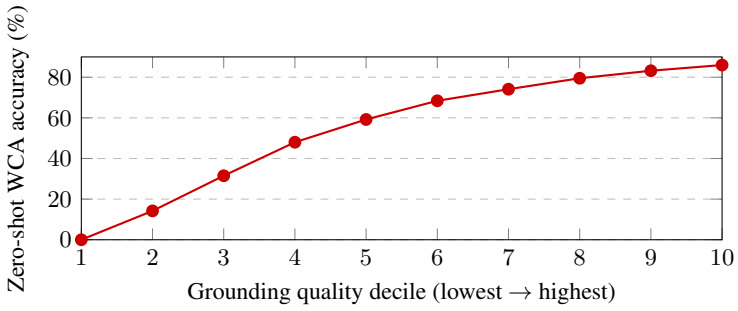


Figure 8: **Diagnostic grounding curve.** Zero-shot WCA accuracy across deciles of regional cross-modal grounding quality on CUB-200-2011. Accuracy falls to zero in the most poorly grounded decile.

D.5 Qualitative Grounding and Local Alignment Comparisons

Figure 9 summarises how an input image and the class descriptions are processed at each stage of UVA. It is a conceptual schematic of the inference pathway rather than an additional experimental result.

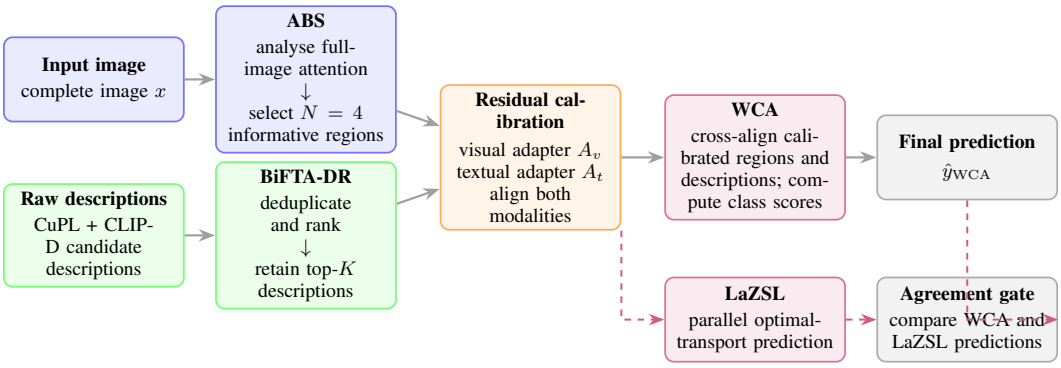


Figure 9: **Stage-by-stage processing pathway in UVA.** The input image is first processed by ABS, which uses the full-image attention map to extract the most informative regions. In parallel, BiFTA-DR removes redundant descriptions from the CuPL and CLIP-D pool and retains the most relevant subset. The regions and descriptions are then calibrated by the visual and textual residual adapters. WCA computes the cross-modal alignment scores and produces the class prediction, while LaZSL operates independently as an optimal-transport verifier, and the agreement between the two predictions serves as the test-time confidence signal.

To illustrate the grounding failure diagnosed in Section 1 and Figure 8, Figure 10 compares region-level matching under zero-shot ABS+WCA and under UVA. In zero-shot ABS+WCA, region embeddings frequently exhibit false-positive background entanglement. Regions containing non-diagnostic background, such as tree perches, water, or foliage, align with generic attribute phrases such as "perched on a branch" or "swimming in open water". Because WCA's softmax normalises over all classes, these high-scoring background matches dilute the logit of the true class. UVA's residual adapters pull foreground region embeddings towards their class prototypes through the InfoNCE objective. This suppresses background matches and re-centres the alignment on diagnostic morphological features such as bill curvature and crown markings. For an ambiguous input, the class scores remain nearly uniform and the two matchers disagree, so the prediction is flagged by the agreement gate.

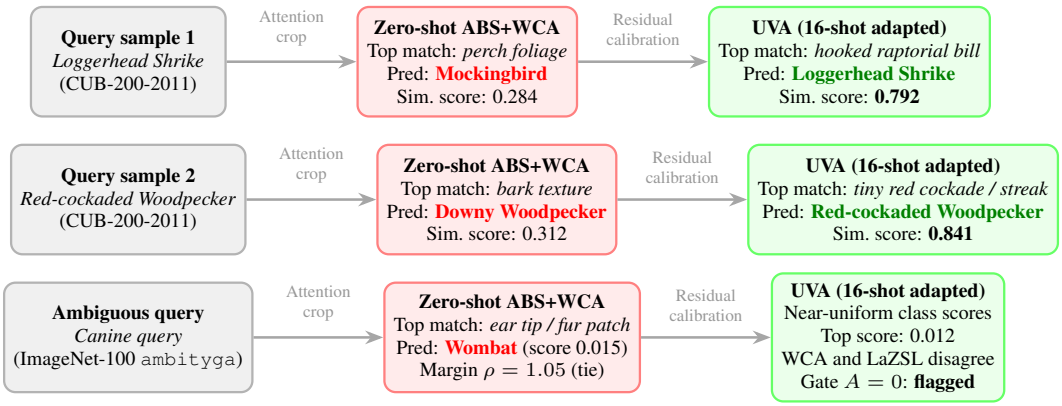


Figure 10: **Qualitative cross-modal matching comparisons on challenging queries.** UVA's residual adapters shift region embeddings towards discriminative morphological descriptions, while ambiguous inputs retain flat class scores ($\rho \approx 1$) and are rejected by the agreement gate ($A = 0$).

D.6 Few-Shot Sample Efficiency and Scaling Dynamics

Figure 11 examines sample efficiency by varying the number of labelled images per class, $K_{\text{shot}} \in \{1, 2, 4, 8, 16\}$, on CUB-200-2011 and ImageNet-100. UVA is highly sample-efficient on CUB-200-2011. A single labelled image per class raises accuracy from 61.17% to 68.40% (+7.23 points), already exceeding 16-shot CLIP-Adapter (66.30%). Accuracy continues to rise with more shots,

reaching 80.60% at 16 shots, while CLIP-Adapter improves more slowly. That a single example per class is already so effective suggests that the misalignment corrected by the adapters is largely systematic and shared across classes, rather than specific to individual categories. On ImageNet-100 the curve is nearly flat (82.72% to 83.71%), consistent with the small overall gain predicted by the Difficulty-Scaling Rule.

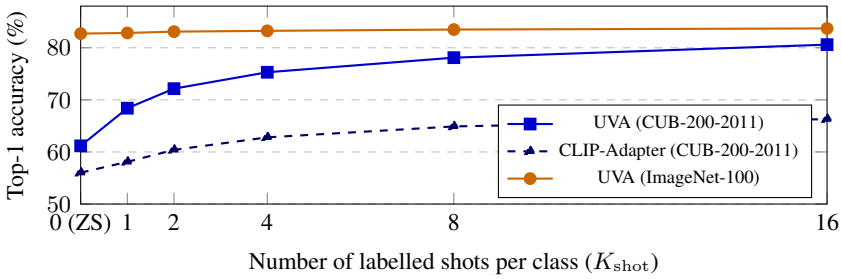


Figure 11: **Few-shot sample efficiency across support budgets** ($K_{\text{shot}} \in \{1, 2, 4, 8, 16\}$). With a single labelled image per class, UVA gains 7.23 points on CUB-200-2011 and already surpasses 16-shot CLIP-Adapter.

D.7 Semantic Pruning Density of BiFTA-DR

Figure 12 shows the average number of descriptions per class before refinement, after cosine deduplication ($\epsilon = 0.99$), and after top-K truncation ($K = 30$). For fine-grained bird species, LLMs repeatedly generate near-synonymous descriptions of generic features, and about 37% of the raw pool (58.4 to 36.8 descriptions per class) is removed as redundant. For the coarser pet breeds, descriptions are more distinct and deduplication removes only about 8% (42.1 to 38.4). This difference explains why BiFTA-DR improves accuracy on CUB-200-2011 but has little effect on Oxford-IIIT Pet (Section 4.2).

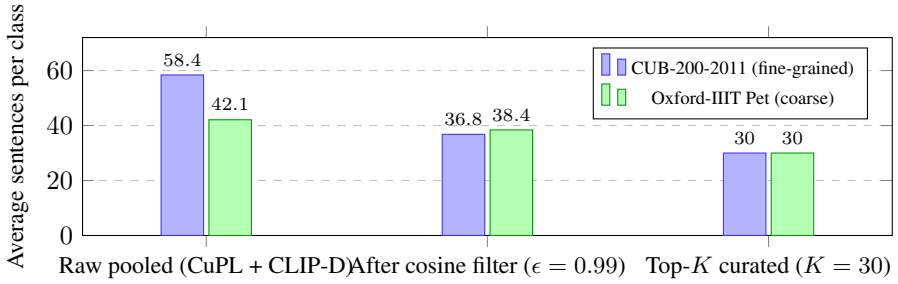


Figure 12: **Description corpus filtering density under BiFTA-DR.** Raw LLM description pools for fine-grained classes contain about 37% redundant sentences, which are removed by $\epsilon = 0.99$ cosine filtering.

D.8 Summary of Component Contributions on CUB-200-2011

Table 10 collects all CUB-200-2011 results reported in this paper in a single view, to make the contribution of each component explicit. Every entry is taken from Tables 1 and 9.

Table 10: **Contribution of each component on CUB-200-2011.** All results use a frozen CLIP ViT-B/16. Changes are measured against the row given in the last column.

Regime	Configuration	Acc. (%)	Change	Relative to
Zero-shot	CLIP baseline (global embedding, class prompt)	56.02	–	–
Zero-shot	CuPL descriptions	54.59	−1.43	CLIP baseline
Zero-shot	+ BiFTA-DR refinement	57.39	+2.80	CuPL
Zero-shot	ABS regions + WCA	61.17	+5.15	CLIP baseline
Zero-shot	+ BiFTA-VR filter	61.22	+0.05	ABS + WCA
16-shot	Global adapter (CLIP-Adapter)	66.30	+10.28	CLIP baseline
16-shot	UVA, $\lambda = 0.1$	78.45	+17.28	ABS + WCA
16-shot	UVA, $\lambda = 0.2$ (best draw)	80.60	+19.43	ABS + WCA
16-shot	UVA, $\lambda = 0.2$ (mean over draws)	80.29	+19.12	ABS + WCA

Three observations follow. First, raw LLM descriptions alone do not help on CUB-200-2011 (54.59% versus 56.02% for the class prompt), and they become useful only after redundancy is removed. Second, region-level matching is the largest zero-shot improvement, but it leaves accuracy at 61.17%. Third, adapting the global embedding (CLIP-Adapter) recovers 10.28 points over CLIP, whereas adapting region and description embeddings recovers 19.12 points over the zero-shot regional pipeline. Where the adaptation is applied matters, and the regional pipeline is what makes the adapters effective.