KERNEL-BASED STEERING OF CLIP WITH VISION-LANGUAGE MODEL PREFERENCES

Anonymous authors
Paper under double-blind review

ABSTRACT

Large vision-language models (VLMs) can judge visual similarity, but their judgments are not directly available as compact image embeddings for efficient comparison. We study how to transfer these preferences into CLIP while retaining its image–text capabilities. We introduce **ASK**, a kernel-based steering method that learns from elicited pairwise judgments without accessing teacher embeddings or collecting new human similarity annotations. **ASK** constructs positive semidefinite target kernels within small image groups and combines visual kernel matching with an image–text distributional anchor. Low-rank adapters jointly update the visual and text encoders while regularizing predictions toward frozen CLIP. After adaptation, retrieval uses CLIP image embeddings and cosine similarity, with no VLM calls. Experiments across five image domains, four CLIP backbones, and six judges evaluate teacher agreement, retrieval, and recognition retention. For ViT-B/16, mean retrieval mAP on classes excluded from adaptation increases from 53.8 to 75.0, compared with 71.7 for DINOv2 targets with KL anchoring. Mean zero-shot accuracy with jointly adapted encoders increases from 61.8% to 62.4%, averaged over 12 benchmarks and the five adaptation domains. Prompting provides an additional capability: selecting which visual distinctions the student learns. Human-annotated evaluations across four datasets support this criterion-specific control.

1 INTRODUCTION

Large vision-language models (VLMs), such as GPT (OpenAI, 2026) and Qwen (Qwen, 2026a), can judge visual similarity through natural-language interactions (Wu et al., 2024; Xiong et al., 2025). Conversational interfaces return responses to image pairs rather than reusable embeddings for cosine comparison. CLIP provides compact embeddings and a shared image–text space (Radford et al., 2021), but its pretrained similarity need not agree with a VLM's judgments (Tong et al., 2024; Vaze et al., 2023). We ask: *How can a compact CLIP model learn a large VLM's visual similarity preferences while retaining its pretrained image–text capabilities?*

Our starting point is relational transfer, which learns relationships among examples without matching the teacher's feature coordinates (Park et al., 2019; Tung & Mori, 2019). KUEA (Gong et al., 2025) aligns CLIP with a vision teacher such as DINOv2 (Oquab et al., 2024), using an L2 anchor on CLIP's original visual features. We instead obtain the target relationships by asking a VLM to compare images. This requires no internal representations, making hosted VLMs usable as teachers.

We introduce Alignment from Solicited Kernels (**ASK**), a method that transfers these judgments into CLIP's embedding geometry. **ASK** scores image pairs within small groups and converts their scores into positive semidefinite target kernels. Projection ensures that independently elicited scores form a valid kernel. Low-rank adaptation (Hu et al., 2022) then fits the visual encoder to the resulting similarity structure. Because matching image–image relationships alone does not constrain their correspondence with text, we combine kernel alignment with an image–text distributional anchor. The anchor jointly updates adapters in both encoders while regularizing predictions over shared text prompts toward frozen CLIP. After adaptation, images can be encoded independently and gallery embeddings cached for cosine retrieval. A new query therefore requires one CLIP image encoding and comparisons with the stored embeddings, avoiding repeated VLM calls for candidate pairs. The jointly adapted encoders also support zero-shot recognition.

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.

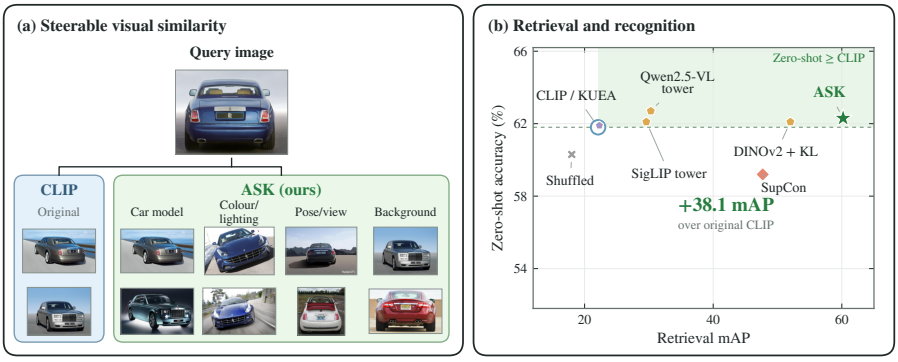


Figure 1: **Transferring VLM similarity preferences into CLIP while retaining recognition. (a) Criterion-specific retrieval.** For one selected car query, columns show the exact top-two cosine matches from CLIP and **ASK**, ranked top to bottom. Separate **ASK** encoders emphasize car model; colour and lighting; pose, viewpoint, and framing; or background. Retrieval uses image embeddings alone, with no VLM calls. Additional examples and selection details are in Appendix F.2. **(b) Retrieval and recognition together.** Eight CLIP ViT-B/16 methods on Dogs (Table 2) compare unseen-breed retrieval mAP with mean zero-shot accuracy over 12 benchmarks. The green region shows retrieval gains while matching or exceeding CLIP's recognition accuracy. Feature labels identify teachers, including SigLIP (Zhai et al., 2023); CLIP and KUEA nearly coincide.

We measure agreement with held-out VLM judgments, retrieval on classes excluded from adaptation, and retention of zero-shot recognition across five domains, four CLIP backbones, and six judges (Tables 1–3). Class labels define retrieval relevance independently of the judge. For ViT-B/16, mean retrieval mAP increases from 53.8 to 75.0, compared with 68.3 for supervised contrastive learning and 71.7 for DINOv2+KL. Mean zero-shot accuracy with jointly adapted encoders increases from 61.8% to 62.4%, averaged over 12 benchmarks and the five adaptation domains. Figure 1(b) illustrates the combination of improved retrieval and retained recognition.

Language provides an additional advantage: prompts can specify which distinctions the student should learn. Asking the same teacher to emphasize different properties yields target relationships, which **ASK** transfers into separate CLIP encoders. For the same car query in Figure 1(a), colour-aligned retrieval favours blue cars, whereas viewpoint-aligned retrieval favours rear views. We test criterion specificity on shared images and evaluate attribute, viewpoint, pose, and background retrieval against human annotations across four datasets. Annotations are reserved for evaluation.

Conditional similarity networks and instruction-conditioned embedders support different notions of similarity (Veit et al., 2017; Hsieh et al., 2025; Zhang et al., 2024; Jiang et al., 2025). **ASK** incorporates the criterion during adaptation, so deployment uses CLIP embeddings without an instruction or judge call. New criteria require supervision and adaptation (Appendix B).

Our contributions are: **(I) Transferring VLM preferences into CLIP.** We learn visual similarity from elicited pairwise judgments through batchwise kernel alignment, without requiring teacher embeddings or new human similarity annotations. **(II) Joint visual and textual adaptation.** We combine visual kernel matching with an image–text distributional anchor, adapting both encoders to support preference alignment and retention of pretrained recognition capabilities. **(III) Evaluating alignment, utility, and control.** We assess teacher fidelity, unseen-class retrieval, and zero-shot recognition across domains, backbones, and judges, and demonstrate prompt-selected criterion control through independent human-annotated evaluations.

2 Method

We introduce **ASK**, a framework for adapting CLIP's visual similarity to pairwise judgments from a vision-language model. **ASK** elicits pairwise scores within image batches, converts these scores into positive semidefinite target kernels, and aligns the visual encoder to their similarity structure.

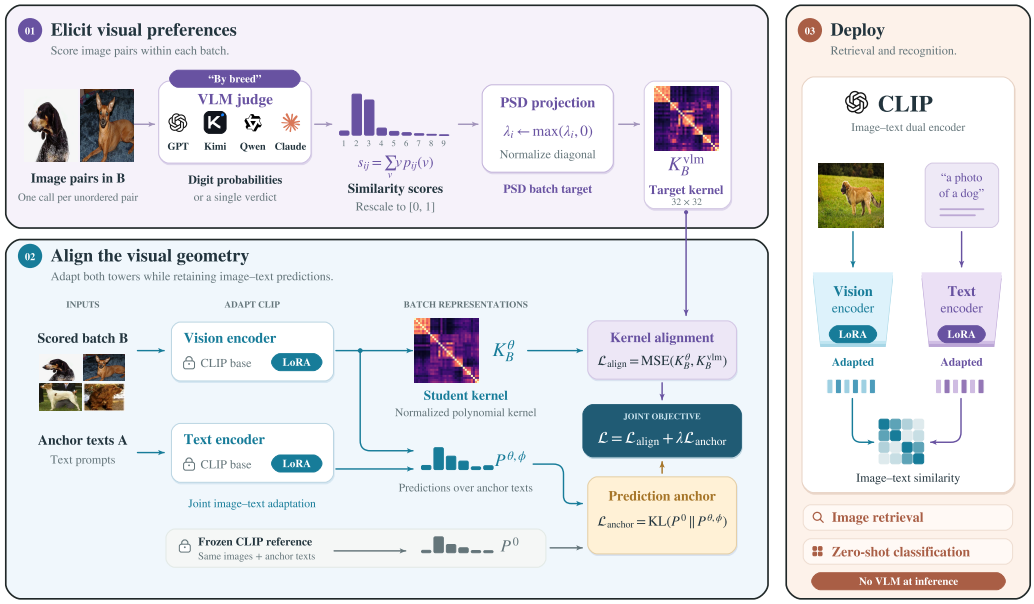


Figure 2: **ASK transfers a VLM's similarity preferences into CLIP for retrieval and recognition. (1) Elicit supervision.** A prompted VLM scores image pairs within small groups. Digit probabilities, when available, yield expected scores; symmetrization, positive-semidefinite (PSD) projection, and diagonal normalization produce a target kernel for each group. Neither teacher embeddings nor new human similarity annotations are required. **(2) Adapt both encoders.** Visual kernel matching teaches the image encoder the requested similarity structure. A KL anchor keeps image–text predictions over shared text prompts close to frozen CLIP's predictions, updating LoRA adapters in both towers while pretrained weights remain fixed. **(3) Deploy CLIP.** Both adapted encoders are retained: image–image cosine similarity supports retrieval, and image–text similarity supports zero-shot classification. Deployment requires no VLM calls; a new criterion requires new judgments and adaptation. Heatmaps are measured; probability bars are illustrative.

An image–text distributional regularizer supports retention of zero-shot classification performance. Figure 2 summarizes the framework, and Algorithm 1 gives the training procedure.

2.1 Preliminaries: Kernels over Embeddings

Let f_θ denote CLIP's visual encoder (Radford et al., 2021), mapping an image I to an embedding $x = f_\theta(I) \in \mathbb{R}^d$. A positive semidefinite (PSD) kernel k defines a symmetric matrix $\mathbf{K} = [k(x_i, x_j)]_{ij} \succeq 0$ for any finite collection of embeddings. This matrix represents their pairwise similarity structure. We use the polynomial kernel

$$k_{\gamma,c}(x, y) = (\gamma\, x^\top y + c)^p, \qquad \gamma > 0, \quad c \geq 0, \tag{1}$$

where γ scales the inner product, c is an additive offset, and $p \geq 1$ is an integer degree. We normalize the kernel to unit diagonal: $\tilde{k}_{\gamma,c}(x, y) = k_{\gamma,c}(x, y)/\sqrt{k_{\gamma,c}(x, x)\, k_{\gamma,c}(y, y)}$. Here, x and y are unnormalized encoder outputs; the normalization acts on the kernel values. Following kernel alignment (Gong et al., 2025), we match pairwise similarity structure without requiring correspondence between teacher and student feature coordinates. We separately regularize image–text predictions to preserve alignment between visual and textual representations (Section 2.5).

2.2 Eliciting a Preference Kernel from a VLM

The teacher in our framework is not an embedding model but a conversation. Given a pair of images (I_i, I_j) and a natural-language comparison prompt q (for instance, "judge similarity by breed, build, coat, and markings"), we present both images to an instruction-tuned VLM and ask it to answer with a single digit on a V-point scale.

Algorithm 1 **ASK**: Alignment from Solicited Kernels

Data: Image groups $\{\mathcal{G}_g\}_{g=1}^{G}$ of size b; comparison prompt q; anchor texts $\mathcal{A}$.
Models: VLM judge $\mathcal{J}$; frozen CLIP encoders $(f_{\theta_0}, g_{\phi_0})$.
Settings: Score scale V; kernel (γ, c, p); LoRA rank r; steps T; anchor weight λ; logit scale τ.

1. Construct teacher targets

Query $\mathcal{J}$ under q once per distinct unordered within-group pair; store scores s_{ij}.
for each group $\mathcal{G}_g$ **do**
  Form symmetric $\mathbf{S}_g$ with unit diagonal and off-diagonal entries $(s_{ij} - 1)/(V - 1)$.
  $\mathbf{A}_g \leftarrow \Pi_+(\mathbf{S}_g)$; $D_g \leftarrow \mathrm{diag}\big((\mathbf{A}_g)_{11}, \ldots, (\mathbf{A}_g)_{bb}\big)$
  Store $\mathbf{K}_{\mathcal{G}_g}^{\mathrm{vlm}} \leftarrow D_g^{-1/2} \mathbf{A}_g D_g^{-1/2}$.
end for

2. Adapt CLIP

Initialize (f_θ, g_ϕ) from $(f_{\theta_0}, g_{\phi_0})$ with rank-r LoRA in both towers.
Keep pretrained weights and kernel parameters fixed; shuffle the group order.
for $t = 1, \ldots, T$ **do**
  Take the next group as B; reshuffle after each pass through the groups.
  $\mathbf{K}_B^\theta \leftarrow [\tilde{k}_{\gamma,c}(f_\theta(I_i), f_\theta(I_j))]_{i,j=1}^{b}$
  Compute P_i^0 and $P_i^{\theta,\phi}$ over $\mathcal{A}$ for each $I_i \in B$ using Equation 6.
  $\mathcal{L} \leftarrow \mathcal{L}_{\mathrm{align}}(\theta; B) + \lambda\, \mathcal{L}_{\mathrm{anchor}}(\theta, \phi; B)$ (Equation 8)
  Update only the LoRA parameters in (θ, ϕ) using gradients of $\mathcal{L}$.
end for
return f_θ and g_ϕ

Expected-digit scoring. Selecting the most probable digit discards the probability assigned to alternative ratings. When token probabilities are available, we instead summarize the model's next-token distribution over the rating digits by its expectation,

$$s_{ij} = \sum_{v=1}^{V} v \cdot p_{ij}(v), \qquad p_{ij}(v) = \frac{\exp z_{ij}(v)}{\sum_{u=1}^{V} \exp z_{ij}(u)}, \tag{2}$$

where $z_{ij}(v)$ is the logit of digit token v. If the API returns only the top-k token log-probabilities, we retain the returned rating digits and divide their probabilities by their sum before computing the expectation. No additional teacher query is needed. Each unordered image pair is scored once. We map its score to $[0, 1]$ and assign it to both corresponding off-diagonal entries. Setting the diagonal to one yields the symmetric elicited matrix $\mathbf{S}$.

Prompt-dependent similarity. The comparison prompt q can name a category, such as dogs or cars, to elicit within-category similarity, or specify an attribute such as colour or pose. This covers both domain-focused alignment and attribute-focused steering. For C prompts, we construct C sets of groupwise target kernels, each used to train a separate encoder. Holding the images and groups fixed isolates how the requested comparison changes the learned geometry.

Batchwise scoring. The alignment loss requires only within-group similarities. For each group of b images, we elicit at most $\binom{b}{2}$ judgments and construct a $b \times b$ target, avoiding a full dataset-wide similarity matrix. Our experiments cache scores for G predefined groups and reuse them across training steps without additional teacher queries (Algorithm 1).

2.3 From elicited similarities to PSD kernels

For a batch B, the elicited matrix $\mathbf{S}_B$ is symmetric with unit diagonal, but its pairwise judgments need not be consistent with a PSD kernel. This differs from a Gram matrix constructed from teacher embeddings, which is PSD by construction. The student's polynomial kernel is also PSD under the constraints in Equation 1, and diagonal normalization preserves this property. An indefinite target therefore cannot be matched exactly by the student.

We construct a PSD target for each batch by clipping negative eigenvalues. Let $\mathbf{S}_B = U_B \Lambda_B U_B^\top$, with eigenvalues $\lambda_1, \ldots, \lambda_b$. Its nearest PSD matrix in Frobenius norm is (Higham, 1988)

$$\mathbf{A}_B = \Pi_+(\mathbf{S}_B) = U_B \,\mathrm{diag}(\max\{\lambda_1, 0\}, \ldots, \max\{\lambda_b, 0\})\, U_B^\top. \tag{3}$$

We then restore unit self-similarity by diagonal normalization. Writing D_B for the diagonal matrix with $(D_B)_{ii} = (\mathbf{A}_B)_{ii}$, the training target is

$$\mathbf{K}_B^{\text{vlm}} = D_B^{-1/2} \mathbf{A}_B D_B^{-1/2}. \tag{4}$$

This normalization preserves PSD and gives unit diagonal. The nearest-PSD property applies to $\mathbf{A}_B$ before normalization; the final target is not necessarily the nearest unit-diagonal PSD matrix. Both operations are applied independently within each scored group.

2.4 Alignment objective

For a batch $B = \{I_1, \ldots, I_b\}$, let $\mathbf{K}_B^\theta = [\tilde{k}_{\gamma,c}(f_\theta(I_i), f_\theta(I_j))]_{ij}$ be the normalized kernel matrix of the adapted visual encoder. We align it with the PSD teacher target $\mathbf{K}_B^{\text{vlm}}$ by minimizing the mean squared difference between their entries:

$$\mathcal{L}_{\text{align}}(\theta; B) \;=\; \frac{1}{b^2} \left\| \mathbf{K}_B^\theta - \mathbf{K}_B^{\text{vlm}} \right\|_F^2. \tag{5}$$

In our experiments, the alignment objective averages this loss over the scored groups (Section 2.2). We train rank-r LoRA adapters (Hu et al., 2022) in the attention projections of both the visual and text encoders, keeping all pretrained weights frozen. The alignment loss updates the visual adapters, while the image–text regularizer in Section 2.5 updates both visual and text adapters.

2.5 Distributional anchoring

Matching image–image similarities alone does not constrain the visual embeddings' relationship to CLIP's text embeddings. To support retention of zero-shot predictions, we regularize image–text distributions against the frozen model. Empirically, **ASK** can also improve zero-shot accuracy over untrained CLIP (Section 3.2). Let g_ϕ denote the adapted text encoder and $\mathcal{A} = \{T_1, \ldots, T_M\}$ a set of anchor texts. For image I_i, the adapted model defines a distribution over these texts:

$$P_i^{\theta,\phi} = \text{softmax}\Big(\tau \left[\hat{f}_\theta(I_i)^\top \hat{g}_\phi(T_m) \right]_{m=1}^M \Big), \tag{6}$$

where hats denote ℓ_2 normalization and $\tau > 0$ is a fixed logit scale. Let P_i^0 be the same distribution computed using the frozen encoders f_{θ_0} and g_{ϕ_0}. For a batch B of b images, the anchor loss is

$$\mathcal{L}_{\text{anchor}}(\theta, \phi; B) = \frac{1}{b} \sum_{i=1}^{b} \text{KL}\Big(P_i^0 \,\|\, P_i^{\theta,\phi} \Big), \tag{7}$$

and the full objective is

$$\mathcal{L}(\theta, \phi; B) = \mathcal{L}_{\text{align}}(\theta; B) + \lambda \, \mathcal{L}_{\text{anchor}}(\theta, \phi; B). \tag{8}$$

Here, $\lambda \geq 0$ controls the strength of the anchor. The KL term couples the two encoders through their predictions on $\mathcal{A}$, with gradients flowing through both adapted towers. At inference, zero-shot classification compares normalized image and prompt embeddings from the jointly adapted encoders f_θ and g_ϕ, while image retrieval uses cosine similarity between visual embeddings. The anchor texts and training settings are specified in Section 3.1; comparisons of regularizers and KL directions are given in Section 3.3.

3 Experimental Results

3.1 Setup

Data. We evaluate five fine-grained domains: ImageNet dog breeds (Deng et al., 2009), Oxford-IIIT Pets (Parkhi et al., 2012), Describable Textures (Cimpoi et al., 2014), Stanford Cars (Krause et al., 2013), and Oxford Flowers (Nilsback & Zisserman, 2008). Each domain contains 25 adaptation classes with 40 training and 10 held-out fidelity-test images per class. We elicit domain-specific similarities and additionally score dogs and cars under multiple comparison criteria. For **ASK**, SupCon, and DINOv2+KL, we select hyperparameters by retrieval mAP on a separate Dogs validation set and

Table 1: **Supervision across five adaptation domains.** CLIP ViT-B/16 with label, feature, or judgment supervision. KUEA uses an L2 feature anchor; other adapted methods use KL. Mean averages the five domains. Blue/red indicates improvement/decline relative to untrained CLIP; intensity is scaled separately per metric. Bold marks the best score.

	Untrained CLIP			SupCon			KUEA			Qwen2.5-VL tower			DINOv2 + KL			**ASK (ours)**		
Domain	Fid.	mAP	ZS	Fid.	mAP	ZS	Fid.	mAP	ZS	Fid.	mAP	ZS	Fid.	mAP	ZS	Fid.	mAP	ZS
Dogs	0.281	22.1	61.8	0.466	47.7	59.2	0.286	22.3	61.9	0.397	30.3	**62.7**	0.539	52.0	62.1	**0.794**	**60.2**	62.3
Textures	0.371	44.1	61.8	0.366	42.6	58.6	0.373	44.5	61.9	0.447	48.7	62.2	0.323	49.1	62.1	**0.706**	**54.5**	**62.9**
Flowers	0.679	77.4	61.8	0.440	83.3	60.7	0.681	77.8	61.8	0.686	83.5	61.7	0.486	**86.8**	61.6	**0.798**	85.8	**62.1**
Cars	0.616	67.2	61.8	0.434	81.1	61.3	0.618	67.4	61.9	0.651	68.3	61.7	0.581	83.9	61.9	**0.859**	**85.1**	**62.1**
Pets	0.582	58.2	61.8	0.471	87.0	60.7	0.589	58.7	61.9	0.556	67.0	62.2	0.548	86.9	62.2	**0.917**	**89.3**	**62.6**
Mean	0.506	53.8	61.8	0.435	68.3	60.1	0.509	54.1	61.9	0.548	59.5	62.1	0.495	71.7	62.0	**0.815**	**75.0**	**62.4**

reuse the selected settings across domains. Dataset splits, scoring budgets, and evaluation protocols appear in Appendix C. We also evaluate attribute, viewpoint, pose, and background retrieval using human annotations from four datasets: CUB-200-2011 (Wah et al., 2011), PASCAL3D+ (Xiang et al., 2014), COCO-Stuff (Caesar et al., 2018), and AP-10K (Yu et al., 2021) (Appendix D.5).

Baselines. We compare untrained CLIP (Radford et al., 2021) with label supervision (SupCon; Khosla et al., 2020), class-label kernels (Appendix E.1), and feature-based kernel targets from frozen encoders, including DINOv2 (Oquab et al., 2024), SigLIP (Zhai et al., 2023), and the Qwen2.5-VL vision tower (Qwen, 2025; Bai et al., 2025). Feature-target comparisons retain our KL anchor; KUEA (Gong et al., 2025) uses its L2 feature anchor. We also evaluate shuffled VLM scores and linear or MLP heads trained on frozen CLIP features. Adaptation comparisons use the same training images, with changes to supervision, anchoring, or adaptation stated in each table. SupCon combines a supervised contrastive image loss with KL anchoring (Appendix C.3).

Training. Unless stated otherwise, we align CLIP ViT-B/16 to Qwen3.5-397B-A17B judgments (Qwen, 2026a) using rank-32 LoRA in both towers for 3,000 steps, keeping pretrained weights frozen. The default uses expected-digit scores, PSD-projected targets, a normalized cubic kernel, and a KL anchor over ImageNet class-name prompts with weight selected on the Dogs validation set. At evaluation, we retain both the adapted image and text encoders. We also compare GPT-5.6 Sol (OpenAI, 2026), Kimi-K3 (Moonshot AI, 2026), Claude Sonnet 5 (Anthropic, 2026), Qwen3.8-27B (Qwen, 2026b), and Gemma 4 12B (Google, 2026; Gemma Team, 2026) as judges under a matched 31-group budget. Full settings are in Appendix C.

Metrics. *Fidelity* (Fid.) is Spearman correlation between the student's normalized kernel similarities and raw VLM scores on held-out image pairs from the adaptation classes. *Retrieval* (mAP) ranks images by cosine similarity on classes excluded from adaptation, using class labels to define relevance independently of the judge. CUB attribute retrieval uses nDCG@10 ($\times 100$), with human annotations defining graded relevance. *Zero-shot accuracy* (ZS) averages top-1 accuracy over 12 benchmarks using the jointly adapted encoders, measuring retention of CLIP's classification ability. Unless stated otherwise, results are averaged over three random seeds.

3.2 Main results

Judgments versus labels and features. Table 1 compares supervision sources using CLIP ViT-B/16 and the same training images. SupCon, Qwen2.5-VL tower, DINOv2+KL, and **ASK** use the same LoRA configuration and KL anchoring objective; KUEA uses DINOv2 targets with its L2 feature anchor. Feature-based targets use the Qwen2.5-VL-7B vision tower or DINOv2, while **ASK** uses elicited judgments. KUEA remains close to untrained CLIP on fidelity and retrieval. **ASK** achieves the highest fidelity in all five domains and the highest retrieval mAP in four, with DINOv2+KL leading on Flowers. Across domains, mean mAP reaches 75.0, compared with 71.7 for DINOv2+KL and 68.3 for SupCon. Mean zero-shot accuracy is 62.4 for **ASK**, 60.1 for SupCon, and 61.8 for untrained CLIP. **ASK** achieves higher average retrieval mAP than the evaluated feature-target baselines, while retaining zero-shot performance.

Consistency across CLIP backbones. Table 2 repeats the comparison on four CLIP vision-transformer backbones (Radford et al., 2021; Dosovitskiy et al., 2021) using the same dog training set. **ASK** leads fidelity and retrieval at every scale, exceeding DINOv2+KL by 6.5–10.5 mAP

Table 2: **Supervision across four CLIP backbones.** Models are adapted on Dogs. Blue/red indicates improvement/decline relative to each backbone's untrained CLIP; intensity is scaled separately per metric. Bold marks the best score per column.

Method	ViT-B/32			ViT-B/16			ViT-L/14			ViT-L/14-336		
	Fid.	mAP	ZS	Fid.	mAP	ZS	Fid.	mAP	ZS	Fid.	mAP	ZS
Untrained CLIP	0.288	18.7	**60.6**	0.281	22.1	61.8	0.283	30.5	66.1	0.264	30.0	66.9
Shuffled teacher	0.221	16.5	58.2	0.203	18.0	60.3	0.253	24.3	65.2	0.206	22.9	65.5
SupCon (Khosla et al., 2020)	0.513	39.3	56.2	0.466	47.7	59.2	0.503	65.6	66.4	0.435	63.4	65.5
KUEA (Gong et al., 2025)	0.291	18.8	**60.6**	0.286	22.3	61.9	0.286	30.9	66.1	0.269	30.4	67.0
DINOv2 + KL	0.533	40.7	59.3	0.539	52.0	62.1	0.521	64.1	67.0	0.518	65.8	67.8
SigLIP tower	0.378	23.6	60.5	0.377	29.6	62.1	0.380	38.1	66.8	0.370	38.8	67.4
Qwen2.5-VL tower	0.397	24.9	60.4	0.397	30.3	**62.7**	0.398	39.1	66.8	0.390	40.4	**68.0**
ASK (ours)	**0.779**	**51.2**	59.2	**0.794**	**60.2**	62.3	**0.805**	**70.9**	**67.2**	**0.810**	**72.3**	67.6

Table 3: **Effect of the VLM judge.** Rows identify the training judge; fidelity columns identify the evaluation judge. All use 31 training groups. Blue/red indicates improvement/decline relative to untrained CLIP, scaled per metric. Bold marks column bests.

Judge	Verdicts		Student fidelity to judge						Student	
	Logp.	Bits	Q3.5-397B	GPT	Kimi	Sonnet	Q3.8-27B	G4-12B	mAP	ZS
Qwen3.5-397B	✓	2.22	**0.79**	0.76	0.80	0.74	**0.78**	0.68	**59.66**	**62.00**
GPT-5.6 Sol	✓	2.20	0.75	**0.79**	0.79	0.73	0.77	0.68	58.94	61.95
Kimi-K3	✓	2.34	0.77	0.77	**0.84**	0.74	**0.78**	0.67	58.96	61.85
Claude Sonnet 5	×	2.32	0.77	0.76	0.79	**0.78**	0.75	0.65	58.24	61.91
Qwen3.8-27B	✓	1.68	0.74	0.75	0.78	0.69	0.76	0.67	58.94	**62.00**
Gemma 4 12B	✓	2.26	0.72	0.72	0.74	0.65	0.74	**0.77**	55.94	61.81
Untrained CLIP			0.25	0.29	0.21	0.27	0.17	0.23	22.06	61.81

points. On ViT-L/14, SupCon reaches 65.6 mAP, exceeding DINOv2+KL at 64.1. Zero-shot accuracy improves over untrained CLIP on three backbones, with a 1.4-point decrease on ViT-B/32. Shuffling the judgments reduces all three metrics below untrained CLIP at every scale, showing that the structure of the elicited similarities matters. The contrast between KUEA and DINOv2+KL also highlights the role of the anchor, examined in the ablations.

Robustness to the choice of judge. Table 3 compares six judges using 31 training groups and the same 4,950 held-out pairs. Logp. indicates whether token log-probabilities support expected-digit scoring; Bits is the entropy of rounded verdicts. Students agree strongly with their training judge (0.76–0.84 fidelity) and other judges (0.65–0.80), suggesting substantial shared similarity structure. Retrieval improves from 22.06 mAP for untrained CLIP to 55.94–59.66 across judges, while zero-shot accuracy remains at 61.81–62.00. Qwen3.5-397B gives the highest retrieval score, but the gains are not restricted to this judge or to models exposing token log-probabilities.

Geometry under breed alignment. Figure 3(a,b) shows the effect of breed alignment on 250 held-out dog images. The aligned CLIP kernel recovers the teacher's breed structure, increasing the mean similarity gap between same-breed and different-breed pairs from 0.093 to 0.352, compared with 0.408 for the teacher.

Criterion retrieval against human annotations. For CUB, we train three criterion-specific students on the same 992 images using Qwen3.8-27B judgments. Figure 3(c) evaluates cosine retrieval against human annotations for wing colour, bill shape, and breast pattern in CUB-200-2011. Relevance depends on shared attributes, and queries retrieve only other species. These labels do not supervise **ASK**. Baselines include CLIP, DINOv2+KL, instruction-conditioned VLM2Vec (Jiang et al., 2025) and MagicLens (Zhang et al., 2024), and human-aligned AligNet (Muttenthaler et al., 2025). Matching **ASK** encoders lead all three criteria, exceeding DINOv2+KL, the strongest baseline, by 3.03, 1.73, and 2.94 nDCG@10 points, respectively. They also outperform both mismatched encoders (Table 10(a)), supporting criterion specificity independently of the VLM judge. **ASK** and DINOv2+KL use CUB adaptation images; the other learned baselines use released checkpoints. Appendix D.5 gives evaluation details and adds species-disjoint CUB and three-dataset transfer results (Figures 4–5). **ASK** also outperforms the baselines on all three transfer datasets.

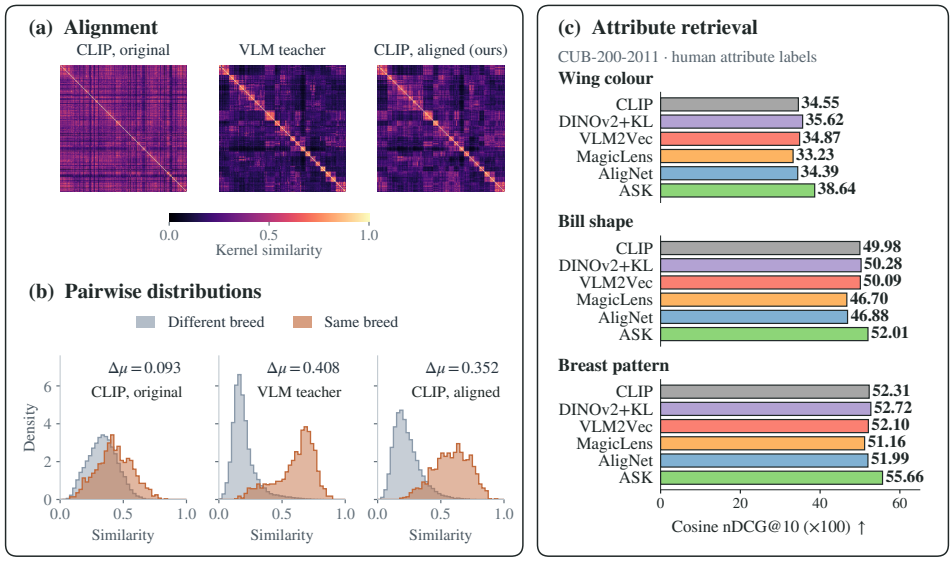


Figure 3: **Visual geometry and criterion-specific retrieval.** **(a)** Original CLIP, the VLM breed teacher, and breed-aligned CLIP on 250 held-out dog images. Normalized kernels share breed ordering and a $[0, 1]$ colour scale; the teacher kernel is PSD-projected. **(b)** Corresponding similarity distributions for same-breed and different-breed pairs, each normalized independently; $\Delta\mu$ is the same-breed mean minus the different-breed mean. **(c)** Cosine retrieval on CUB-200-2011, with human annotations defining relevance for wing colour, bill shape, and breast pattern. Queries retrieve images of other species; bars report nDCG@10 $\times 100$ (higher is better). Each **ASK** bar uses the corresponding criterion-trained encoder. Human attributes do not supervise **ASK**; Appendix D.5 details evaluation and baseline settings.

Table 4: **Cross-domain retrieval and criterion specificity on fixed images.** **(a)** Training domain versus retrieval domain, using unseen-class labels. **(b)** Training criterion versus held-out teacher, on fixed dog image splits. DINOv2+KL and Shuffled are dog-trained; Shuffled is a negative control with permuted breed targets. Entries are gains over CLIP (absolute scores first). Blue/red denotes improvement/decline; bold marks column bests.

(a) Cross-domain retrieval (ΔmAP)						**(b) Criterion specificity** (ΔSpearman)					
Elicited on	Dogs	Textures	Flowers	Cars	Pets	Asked about	Breed	Colour/ lighting	Pose/view	Coat tex.	Backgr.
Untrained	*22.1*	*44.1*	*77.4*	*67.2*	*58.2*	*Untrained*	*0.28*	*0.34*	*0.49*	*0.19*	*0.27*
Dogs	**+38.1**	+6.2	**+8.9**	+14.7	+29.0	Breed	**+0.51**	+0.07	−0.25	+0.53	−0.12
Textures	+2.4	**+10.4**	+4.0	+4.1	+3.8	Colour/lighting	+0.20	**+0.41**	−0.10	+0.16	+0.18
Flowers	+7.3	+4.6	+8.4	+7.4	+9.9	Pose/view	+0.02	+0.02	**+0.26**	+0.00	+0.05
Cars	+10.9	+3.2	+7.0	**+17.9**	+12.8	Coat tex.	+0.41	−0.04	−0.28	**+0.68**	−0.17
Pets	+35.5	+4.3	+7.8	+13.2	**+31.1**	Backgr.	−0.10	+0.10	−0.22	−0.10	**+0.65**
DINOv2+KL	+29.9	+3.5	+5.9	+10.6	+23.8	DINOv2+KL	+0.26	−0.03	−0.22	+0.30	−0.08
Shuffled (neg.)	−4.1	−4.4	−14.3	−6.1	−14.3	Shuffled (neg.)	−0.08	−0.16	−0.23	−0.03	−0.12

Transfer and criterion specificity. Table 4(a) measures cross-domain class retrieval: all 25 combinations improve, with cross-domain gains of 2.4–35.5 mAP points. Panel (b) fixes the dog images: each matching student outperforms every mismatched student and both controls against its teacher. This within-column comparison supports specificity. Off-diagonal declines alone do not: breed alignment lowers pose agreement by 0.25, versus 0.23 for shuffled supervision. Shuffled is a single negative control using permuted breed targets. All adapted models use 93 groups; Dogs and Breed share a checkpoint. Appendix D.3 separately tests transfer of the same criteria to non-dog images.

3.3 Ablations

PSD projection and supervision targets. Table 5(a) shows four cases from Table 18. Under matched 31-group raw/PSD training, projection improves retrieval by 0.17–1.45 mAP points and

Table 5: **Ablations of the alignment recipe.** (a) Raw minus PSD for the four cases with highest negative spectral mass (Neg.%). (b) Supervision targets under matched MSE, KL, and LoRA; mAP on unseen breeds. (c,d) KL-anchored heads versus LoRA. (e) Anchor regularisation. (f) Training objective. Blue/red indicates improvement/decline relative to PSD in (a), LoRA in (c,d), and the default in (b,e,f); bold marks metric bests.

(a) Raw vs. PSD teacher

Judge	Neg.%	ΔFid	ΔmAP	ΔZS
Gemma 4 12B	6.4	−0.009	−1.43	−0.41
GPT-5.6	3.8	−0.005	−0.53	−0.29
Qwen3.5	3.5	+0.002	−1.45	−0.26
Sonnet 5	2.6	−0.001	−0.17	−0.23

(b) Supervision target

Target	Unseen mAP	ZS
Binary labels	43.11	59.83
WordNet	52.01	60.70
VLM class-pair mean	57.17	61.77
ASK	**60.16**	**62.28**

(c) Heads vs. LoRA: Fid

Criterion	Linear	MLP	**LoRA**
Breed	0.622	0.734	**0.794**
Colour/lighting	0.491	0.673	**0.747**
Pose/view	0.564	0.674	**0.748**
Coat texture	0.699	0.826	**0.874**
Background	0.730	0.887	**0.919**

(d) Heads vs. LoRA: ΔmAP

Retrieval on	Linear	MLP	**LoRA**
Unseen breeds	+13.9	+22.8	**+38.1**
Non-dog	+1.3	+1.9	**+4.9**
Other domains	+5.5	+8.2	**+14.7**

(e) Anchor regulariser

Regulariser	w	Fid	mAP	ZS
KL (ASK)	.2	0.794	**60.16**	**62.28**
None	—	**0.796**	59.41	56.67
Feature L2	1	0.286	22.29	61.91
Text L1 only	.5	**0.796**	59.30	56.70
KL, text frozen	.2	0.752	49.71	61.35

(f) Training objective

Objective	Fid	mAP	ZS
Cosine MSE	0.756	58.35	**62.57**
Cubic MSE (ASK)	0.794	**60.16**	62.28
Cubic ranking	**0.799**	54.71	62.23

zero-shot accuracy by 0.23–0.41 points across these cases. Gemma 4 12B has the largest negative mass (6.4%) and gains 0.009 fidelity, 1.43 mAP, and 0.41 ZS. Qwen car-pose targets have 3.5% negative mass: projection gains 1.45 car-model mAP and 0.26 ZS, with a 0.002 fidelity decrease. These results support PSD correction as an effective part of alignment when elicited targets have substantial negative spectral mass. Panel (b) fixes kernel MSE, KL, and LoRA: binary labels, WordNet (Miller, 1995), and VLM class-pair means yield 43.11, 52.01, and 57.17 unseen-breed mAP, versus 60.16 for individual judgments. Appendix E.1 gives the full comparison.

Frozen heads versus encoder adaptation. Table 5(c,d) compares linear and MLP heads with LoRA using the same criterion-specific targets, cubic objective, and KL anchor. Across five criteria, mean fidelity is 0.816 for LoRA, 0.621 for the linear head, and 0.759 for the MLP. The breed-trained models improve unseen-breed mAP over CLIP by 38.1, 13.9, and 22.8 points, respectively. Both heads also improve non-dog and cross-domain retrieval, while LoRA achieves larger gains. The MLP matches LoRA in visual parameter count; these comparisons use a shared learning rate. Full retrieval results and the protocol are in Appendix E.5.

Preserving zero-shot performance. Table 5(e) shows that removing the KL anchor leaves fidelity nearly unchanged (0.796 versus 0.794) but lowers zero-shot accuracy from 62.28 to 56.67. An L1 penalty on text-feature drift alone provides little protection (56.70 ZS). An L2 penalty on image-feature drift at the tested weight of 1.0 reduces fidelity to 0.286. Freezing the text tower during KL-anchored training lowers retrieval by 10.45 mAP points. Additional regularisers and weight sweeps are reported in Appendix E.4.

Kernel form and sensitivity. Table 5(f) compares cosine MSE, normalized-cubic MSE, and normalized-cubic ranking with the same LoRA student, PSD targets, and KL anchor. Fidelity uses each training similarity against raw held-out judgments, with no link fitted on test targets. Cubic MSE reaches 0.794 Spearman and 60.16 mAP, versus 0.756 and 58.35 for cosine MSE; their zero-shot means are 62.28 and 62.57. Ranking raises Spearman slightly to 0.799 but lowers mAP to 54.71. Appendix E.6 additionally evaluates every model with cosine: the cubic-MSE model retains 0.794 Spearman, while Pearson changes from 0.840 to 0.792. These comparisons use a shared learning rate and anchor weight. Sensitivity and reduced-query supervision are examined in Appendices E.7 and D.10, respectively.

4 Conclusion

ASK transfers VLM similarity preferences into CLIP through positive semidefinite kernel matching and joint image–text adaptation. Across five domains, four backbones, and six judges, it improves unseen-class retrieval while an image–text anchor supports recognition retention. Retrieval uses image embeddings and cosine similarity, without VLM calls. Prompt-selected criteria additionally control the learned similarity, with human-annotated evaluations supporting attribute, viewpoint, pose, and background retrieval across four datasets, including species excluded from adaptation.

AI USE STATEMENT

Large language models assisted with writing and refining the manuscript. The authors reviewed all text and take full responsibility for the paper's content.

REFERENCES

Anthropic. 2026. Claude Sonnet 5. https://www.anthropic.com/news/claude-sonnet-5. Accessed 2026-09-24.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, Jiabo Ye, Xi Zhang, Tianbao Xie, Zesen Cheng, Hang Zhang, Zhibo Yang, Haiyang Xu, and Junyang Lin. Qwen2.5-VL Technical Report. arXiv preprint arXiv:2502.13923, 2025. URL https://arxiv.org/abs/2502.13923.

Holger Caesar, Jasper Uijlings, and Vittorio Ferrari. COCO-Stuff: Thing and stuff classes in context. In *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1209–1218, 2018. URL https://github.com/nightrome/cocostuff.

Tsai-Shien Chen, Aliaksandr Siarohin, Gordon Guocheng Qian, Kuan-Chieh Jackson Wang, Egor Nemchinov, Moayed Haji-Ali, Riza Alp Guler, Willi Menapace, Ivan Skorokhodov, Anil Kag, Jun-Yan Zhu, and Sergey Tulyakov. Omni-Attribute: Open-vocabulary attribute encoder for visual concept personalization. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8194–8204, 2026. URL https://snap-research.github.io/omni-attribute/.

Gong Cheng, Junwei Han, and Xiaoqiang Lu. Remote sensing image scene classification: Benchmark and state of the art. *Proceedings of the IEEE*, 105(10):1865–1883, 2017. doi: 10.1109/JPROC.2017.2675998.

Mircea Cimpoi, Subhransu Maji, Iasonas Kokkinos, Sammy Mohamed, and Andrea Vedaldi. Describing textures in the wild. In *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3606–3613, 2014.

Timothée Darcet, Maxime Oquab, Julien Mairal, and Piotr Bojanowski. Vision transformers need registers. In *International Conference on Learning Representations*, pp. 2632–2652, 2024.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. ImageNet: A large-scale hierarchical image database. In *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 248–255, 2009. doi: 10.1109/cvpr.2009.5206848.

Keyan Ding, Kede Ma, Shiqi Wang, and Eero P. Simoncelli. Image quality assessment: Unifying structure and texture similarity. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(5):2567–2581, 2022. doi: 10.1109/TPAMI.2020.3045810.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2021.

Li Fei-Fei, Rob Fergus, and Pietro Perona. Learning generative visual models from few training examples: An incremental Bayesian approach tested on 101 object categories. In *IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pp. 178, 2004. doi: 10.1109/cvpr.2004.383.

Stephanie Fu, Netanel Tamir, Shobhita Sundaram, Lucy Chai, Richard Zhang, Tali Dekel, and Phillip Isola. DreamSim: Learning new dimensions of human visual similarity using synthetic data. In *Advances in Neural Information Processing Systems*, volume 36, pp. 50742–50768, 2023. doi: 10.52202/075280-2208.

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. doi: 10.1007/s11263-023-01891-x.

Gemma Team. Gemma 4 Technical Report. arXiv preprint arXiv:2607.02770, 2026. URL https://arxiv.org/abs/2607.02770.

Shizhan Gong, Yankai Jiang, Qi Dou, and Farzan Farnia. Kernel-based unsupervised embedding alignment for enhanced visual representation in vision-language models. In *International Conference on Machine Learning*, pp. 19912–19931. PMLR, 2025.

Ian J. Goodfellow, Dumitru Erhan, Pierre Luc Carrier, Aaron Courville, Mehdi Mirza, Ben Hamner, Will Cukierski, Yichuan Tang, David Thaler, Dong-Hyun Lee, Yingbo Zhou, Chetan Ramaiah, Fangxiang Feng, Ruifan Li, Xiaojie Wang, Dimitris Athanasakis, John Shawe-Taylor, Maxim Milakov, John Park, Radu Ionescu, Marius Popescu, Cristian Grozea, James Bergstra, Jingjing Xie, Lukasz Romaszko, Bing Xu, Zhang Chuang, and Yoshua Bengio. Challenges in representation learning: A report on three machine learning contests. *arXiv preprint arXiv:1307.0414*, 2013. URL https://arxiv.org/abs/1307.0414.

Google. 2026. Gemma 4 12B. https://ai.google.dev/gemma/docs/core/model_card_4. Accessed 2026-09-25.

Patrick Helber, Benjamin Bischke, Andreas Dengel, and Damian Borth. EuroSAT: A novel dataset and deep learning benchmark for land use and land cover classification. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 12(7):2217–2226, 2019. doi: 10.1109/jstars.2019.2918242.

Dan Hendrycks, Kevin Zhao, Steven Basart, Jacob Steinhardt, and Dawn Song. Natural adversarial examples. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 15262–15271, 2021.

Nicholas J Higham. Computing a nearest symmetric positive semidefinite matrix. *Linear Algebra and its Applications*, 103:103–118, 1988. doi: 10.1016/0024-3795(88)90223-6.

Cheng-Yu Hsieh, Pavan Kumar Anasosalu Vasu, Fartash Faghri, Raviteja Vemulapalli, Chun-Liang Li, Ranjay Krishna, Oncel Tuzel, and Hadi Pouransari. FocalLens: Instruction tuning enables zero-shot conditional image representations. *arXiv preprint arXiv:2504.08368*, 2025. URL https://arxiv.org/abs/2504.08368.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, pp. 12513–12525, 2022.

Mohammad Jalali, Bahar Dibaei Nia, and Farzan Farnia. Towards an explainable comparison and alignment of feature embeddings. In *International Conference on Machine Learning*, pp. 26757–26796. PMLR, 2025.

Ziyan Jiang, Rui Meng, Xinyi Yang, Semih Yavuz, Yingbo Zhou, and Wenhu Chen. VLM2Vec: Training vision-language models for massive multimodal embedding tasks. In *International Conference on Learning Representations*, pp. 1255–1279, 2025.

Muhammad Uzair Khattak, Hanoona Rasheed, Muhammad Maaz, Salman Khan, and Fahad Shahbaz Khan. MaPLe: Multi-modal prompt learning. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 19113–19122, 2023a.

Muhammad Uzair Khattak, Syed Talal Wasim, Muzammal Naseer, Salman Khan, Ming-Hsuan Yang, and Fahad Shahbaz Khan. Self-regulating prompts: Foundational model adaptation without forgetting. In *IEEE/CVF International Conference on Computer Vision*, pp. 15190–15200, 2023b.

Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. Supervised contrastive learning. In *Advances in Neural Information Processing Systems*, volume 33, pp. 18661–18673, 2020.

Simon Kornblith, Mohammad Norouzi, Honglak Lee, and Geoffrey Hinton. Similarity of neural network representations revisited. In *International Conference on Machine Learning*, pp. 3519–3529. PMLR, 2019.

Jonathan Krause, Michael Stark, Jia Deng, and Li Fei-Fei. 3D object representations for fine-grained categorization. In *IEEE International Conference on Computer Vision Workshops*, pp. 554–561, 2013.

Alex Krizhevsky. Learning multiple layers of features from tiny images. Technical report, University of Toronto, 2009. URL https://www.cs.toronto.edu/~kriz/learning-features-2009-TR.pdf.

Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. G-Eval: NLG evaluation using GPT-4 with better human alignment. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 2511–2522, 2023. doi: 10.18653/v1/2023.emnlp-main.153.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, pp. 4061–4078, 2019.

George A. Miller. WordNet: A lexical database for English. *Communications of the ACM*, 38(11): 39–41, 1995. doi: 10.1145/219717.219748.

Moonshot AI. 2026. Kimi K3. https://huggingface.co/moonshotai/Kimi-K3. Accessed 2026-09-24.

Lukas Muttenthaler, Klaus Greff, Frieda Born, Bernhard Spitzer, Simon Kornblith, Michael C. Mozer, Klaus-Robert Müller, Thomas Unterthiner, and Andrew K. Lampinen. Aligning machine and human visual representations across abstraction levels. *Nature*, 647:349–355, 2025. doi: 10.1038/s41586-025-09631-6.

Maria-Elena Nilsback and Andrew Zisserman. Automated flower classification over a large number of classes. In *Indian Conference on Computer Vision, Graphics and Image Processing*, pp. 722–729, 2008. doi: 10.1109/icvgip.2008.47.

OpenAI. 2026. GPT-5.6 Sol. https://developers.openai.com/api/docs/models/gpt-5.6-sol. Accessed 2026-09-24.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy V. Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Mahmoud Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Hervé Jegou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. DINOv2: Learning robust visual features without supervision. *Transactions on Machine Learning Research*, 2024.

Wonpyo Park, Dongju Kim, Yan Lu, and Minsu Cho. Relational knowledge distillation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3967–3976, 2019.

Omkar M. Parkhi, Andrea Vedaldi, Andrew Zisserman, and C. V. Jawahar. Cats and dogs. In *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3498–3505, 2012. doi: 10.1109/cvpr.2012.6248092.

Qwen. 2025. Qwen2.5-VL-7B. https://huggingface.co/Qwen/Qwen2.5-VL-7B-Instruct. Accessed 2026-09-24.

Qwen. 2026a. Qwen3.5-397B-A17B. https://huggingface.co/Qwen/Qwen3.5-397B-A17B. Accessed 2026-09-25.

Qwen. 2026b. Qwen3.8-27B. https://huggingface.co/Qwen/Qwen3.8-27B. Accessed 2026-09-24.

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 *International Conference on Machine Learning*, pp. 8748–8763. PMLR, 2021.

Christian Schlarmann, Naman Deep Singh, Francesco Croce, and Matthias Hein. Robust CLIP: Unsupervised adversarial fine-tuning of vision embeddings for robust large vision-language models. In *International Conference on Machine Learning*, pp. 43685–43704. PMLR, 2024.

Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *Annual Meeting of the Association for Computational Linguistics*, pp. 2556–2565, 2018. doi: 10.18653/v1/P18-1238.

Shengbang Tong, Zhuang Liu, Yuexiang Zhai, Yi Ma, Yann LeCun, and Saining Xie. Eyes wide shut? exploring the visual shortcomings of multimodal LLMs. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9568–9578, 2024.

Frederick Tung and Greg Mori. Similarity-preserving knowledge distillation. In *IEEE/CVF International Conference on Computer Vision*, pp. 1365–1374, 2019.

Sagar Vaze, Nicolas Carion, and Ishan Misra. GeneCIS: A benchmark for general conditional image similarity. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6862–6872, 2023.

Bastiaan S. Veeling, Jasper Linmans, Jim Winkens, Taco Cohen, and Max Welling. Rotation equivariant CNNs for digital pathology. In *Medical Image Computing and Computer Assisted Intervention*, pp. 210–218, 2018. doi: 10.1007/978-3-030-00934-2_24.

Andreas Veit, Serge Belongie, and Theofanis Karaletsos. Conditional similarity networks. In *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 830–838, 2017.

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. URL https://www.vision.caltech.edu/datasets/cub_200_2011/.

Haohan Wang, Songwei Ge, Zachary Lipton, and Eric P. Xing. Learning robust global representations by penalizing local predictive power. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

Mitchell Wortsman, Gabriel Ilharco, Jong Wook Kim, Mike Li, Simon Kornblith, Rebecca Roelofs, Raphael Gontijo Lopes, Hannaneh Hajishirzi, Ali Farhadi, Hongseok Namkoong, and Ludwig Schmidt. Robust fine-tuning of zero-shot models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7959–7971, 2022.

Haoning Wu, Zicheng Zhang, Weixia Zhang, Chaofeng Chen, Liang Liao, Chunyi Li, Yixuan Gao, Annan Wang, Erli Zhang, Wenxiu Sun, Qiong Yan, Xiongkuo Min, Guangtao Zhai, and Weisi Lin. Q-Align: Teaching LMMs for visual scoring via discrete text-defined levels. In *International Conference on Machine Learning*, pp. 54015–54029. PMLR, 2024.

Youqi Wu, Mohammad Jalali, and Farzan Farnia. KODA: Contrastive representation comparison and alignment for vision-language foundation models. In *International Conference on Machine Learning*, 2026.

Yu Xiang, Roozbeh Mottaghi, and Silvio Savarese. Beyond PASCAL: A benchmark for 3D object detection in the wild. In *IEEE Winter Conference on Applications of Computer Vision*, pp. 75–82, 2014. doi: 10.1109/wacv.2014.6836101. URL https://cs.stanford.edu/groups/cvgl/projects/pascal3d.html.

Tianyi Xiong, Xiyao Wang, Dong Guo, Qinghao Ye, Haoqi Fan, Quanquan Gu, Heng Huang, and Chunyuan Li. LLaVA-Critic: Learning to evaluate multimodal models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13618–13628, 2025.

Hang Yu, Yufei Xu, Jing Zhang, Wei Zhao, Ziyu Guan, and Dacheng Tao. AP-10K: A benchmark for animal pose estimation in the wild. In *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*, volume 1, 2021. URL https://github.com/AlexTheBad/AP-10K.

Maxime Zanella and Ismail Ben Ayed. Low-rank few-shot adaptation of vision-language models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pp. 1593–1603, 2024.

Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid loss for language image pre-training. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 11975–11986, October 2023.

Kai Zhang, Yi Luan, Hexiang Hu, Kenton Lee, Siyuan Qiao, Wenhu Chen, Yu Su, and Ming-Wei Chang. MagicLens: Self-supervised image retrieval with open-ended instructions. In *International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 59403–59420. PMLR, 2024.

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*, pp. 493–510, 2022. doi: 10.1007/978-3-031-19833-5_29.

Richard Zhang, Phillip Isola, Alexei A. Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 586–595, 2018.

Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Learning to prompt for vision-language models. *International Journal of Computer Vision*, 130(9):2337–2348, 2022. doi: 10.1007/s11263-022-01653-1.

A APPENDIX

CONTENTS

B Related work **16**

C Experimental setup details **17**

C.1 Datasets and splits . . . 17

C.2 Training and evaluation protocols . . . 18

C.3 Supervised contrastive baseline . . . 19

D Additional results **20**

D.1 Per-benchmark zero-shot accuracy . . . 20

D.2 Retrieval and fidelity on unseen classes . . . 21

D.3 Criterion transfer beyond dogs . . . 22

D.4 Criterion specificity on car images . . . 22

D.5 Criterion retrieval with human annotations . . . 23

D.6 Per-benchmark retention across adaptation domains . . . 25

D.7 Judge agreement and order consistency . . . 25

D.8 Judge comparisons across backbones . . . 27

D.9 Teacher-scoring cost and student training time . . . 28

D.10 Reducing teacher API calls . . . 29

E Ablation studies **30**

E.1 Class labels versus VLM judgments . . . 30

E.2 Effect of PSD projection . . . 31

E.3 Verdict extraction . . . 31

E.4 Regulariser families and weights . . . 32

E.5 Frozen-feature heads versus encoder adaptation . . . 33

E.6 Kernel form . . . 34

E.7 Sensitivity to training choices . . . 35

F Additional visual results **36**

F.1 Kernel geometry . . . 36

F.2 Criterion-steered retrieval examples . . . 39

F.3 Top retrievals across judge-trained models . . . 39

F.4 Comparing fixed candidates across criteria . . . 42

G Elicitation prompts **43**

G.1 Dog similarity criteria . . . 43

G.2 Primary criteria for other domains . . . 44

G.3 Car-specific criterion prompts . . . 45

G.4 CUB attribute scoring prompts . . . 46

B Related work

Visual representations and perceptual similarity. CLIP learns visual representations through image–text contrastive supervision (Radford et al., 2021), while DINOv2 learns transferable visual features through self-supervision (Oquab et al., 2024). Strong recognition performance does not ensure sensitivity to every visual distinction: Tong et al. (2024) identify image pairs that CLIP represents similarly despite differences relevant to visual reasoning. Perceptual metrics address a related problem from human judgments. LPIPS calibrates deep features for perceptual comparison (Zhang et al., 2018); DISTS combines structure and texture similarity while tolerating texture resampling (Ding et al., 2022); and DreamSim learns from human comparisons of synthetic image triplets (Fu et al., 2023). AligNet distils synthetic targets from a human-aligned teacher into criterion-independent visual representations (Muttenthaler et al., 2025). These works motivate measuring more than category identity. ASK learns from a selected VLM's criterion-dependent judgments. Its fidelity measures teacher agreement, while human-annotated attribute retrieval independently evaluates the selected visual distinctions; class-based retrieval and zero-shot classification assess category retrieval and recognition retention.

Conditional similarity and instruction-guided representations. Conditional Similarity Networks learn masks over embedding dimensions to represent different notions of similarity (Veit et al., 2017). GeneCIS evaluates whether a model can retrieve images according to changing object and attribute conditions, and learns conditional similarity from image-caption data (Vaze et al., 2023). FocalLens uses instruction tuning to produce image representations conditioned on natural-language instructions (Hsieh et al., 2025). MagicLens learns image-and-instruction embeddings from web image pairs and synthesized instructions (Zhang et al., 2024). Omni-Attribute learns open-vocabulary attribute embeddings using image pairs annotated with shared and differing attributes, supporting retrieval and visual concept personalization (Chen et al., 2026). VLM2Vec takes a broader approach, converting VLMs into instruction-aware multimodal embedding models through contrastive training (Jiang et al., 2025). ASK shares the goal of controlling which visual information determines similarity, but incorporates the criterion through teacher supervision during adaptation. Each resulting encoder can embed an image without an accompanying instruction; changing to a previously untrained criterion requires further adaptation.

Model-based evaluation and continuous score extraction. Language and multimodal models can provide evaluation signals as well as generated answers. G-Eval combines prompted evaluation with probability-weighted rating scores for text generation (Liu et al., 2023). Q-Align trains multimodal models on discrete quality levels and converts their output probabilities into continuous visual scores (Wu et al., 2024). LLaVA-Critic learns to evaluate multimodal responses and provides feedback for preference learning (Xiong et al., 2025). These approaches establish precedents for using model judgments and for extracting finer-grained scores from discrete responses. ASK applies probability-weighted digit scores, when available, to criterion-specific image pairs and uses the resulting similarities to supervise an encoder. Our contribution is the transfer of these judgments into visual geometry, rather than a new principle for converting rating probabilities into scalar scores.

Relational distillation and representation comparison. Relational knowledge distillation transfers distances and angles between examples (Park et al., 2019), and similarity-preserving distillation matches pairwise activation similarities (Tung & Mori, 2019). Both show that a student can learn a teacher's relationships without matching its feature coordinates. Related tools compare representation structure: centered kernel alignment measures agreement between representational similarity matrices (Kornblith et al., 2019); SPEC identifies differences in sample clustering through spectral analysis (Jalali et al., 2025); and KODA discovers subsets organized differently across vision-language representations (Wu et al., 2026). ASK adopts this relational perspective, using pairwise geometry as the target of adaptation. The target is supplied by an elicited criterion-dependent judgment, rather than derived solely from another encoder's features.

Kernel alignment with elicited targets. KUEA aligns CLIP's visual geometry to a vision teacher such as DINOv2 through kernel matching (Gong et al., 2025). Its objective updates the visual encoder and anchors its features to frozen CLIP with an L2 penalty, without using text during adaptation. ASK builds on this foundation in two respects. First, its target is elicited from a VLM under a chosen criterion, allowing the same image set to support different target geometries without access to teacher embeddings. Second, its retention objective acts on image–text prediction distributions and

jointly updates adapters in both CLIP encoders. The text supervision is a shared prompt bank; it does not require a newly annotated image–caption dataset. A feature Gram matrix is PSD by construction, whereas elicited scores need not be, so we apply classical spectral PSD projection (Higham, 1988) and diagonal normalization within each scored group. The methodological contribution is the use of elicited criterion-dependent targets with joint distributional anchoring; kernel matching and PSD projection themselves are established tools. Our DINOv2+KL control uses the same anchor and student configuration to distinguish the supervision source from the change in regularization relative to KUEA.

Efficient adaptation of vision-language models. CoOp learns text context vectors for recognition (Zhou et al., 2022), and MaPLe couples learned prompts across the vision and language branches (Khattak et al., 2023a). CLIP-Adapter learns residual feature adapters (Gao et al., 2024), while Tip-Adapter constructs a cache from labeled examples for few-shot classification (Zhang et al., 2022). LoRA provides low-rank parameter updates (Hu et al., 2022), and CLIP-LoRA applies them to few-shot vision-language adaptation (Zanella & Ben Ayed, 2024). ASK uses LoRA as an adaptation mechanism, with a different objective: reproducing criterion-specific pairwise similarities while retaining useful image–text predictions. The criterion is expressed in the teacher prompt, rather than supplied as a class prediction target.

Retaining pretrained capabilities. PromptSRC regularizes prompt adaptation against pretrained CLIP to reduce forgetting (Khattak et al., 2023b). WiSE-FT interpolates pretrained and fine-tuned weights to improve robustness under distribution shift (Wortsman et al., 2022). Robust CLIP adapts the visual encoder through an unsupervised adversarial objective, illustrating another setting in which visual features must change while remaining useful in a vision-language system (Schlarmann et al., 2024). ASK regularizes image–text predictions during similarity alignment, so both visual and textual adapters participate in optimization. Our regularizer ablation shows higher fidelity with trainable text adapters than with a frozen text tower at each tested weight of the default KL anchor. Both adapted encoders are retained at deployment. Zero-shot evaluation measures retention of recognition capabilities after joint adaptation, separately from agreement with the VLM judge.

C Experimental setup details

C.1 Datasets and splits

Image sources. We use ImageNet dog breeds (Deng et al., 2009), Describable Textures (DTD; Cimpoi et al., 2014), Oxford Flowers (Nilsback & Zisserman, 2008), Stanford Cars (Krause et al., 2013), and Oxford-IIIT Pets (Parkhi et al., 2012). Dog images are drawn from the ImageNet training partition. The other four domains use the CoOp dataset splits (Zhou et al., 2022). Retrieval galleries contain 20 images per unseen class: 25 classes each for Dogs, Flowers, and Cars (500 images per domain), 20 for DTD (400 images), and 12 for Pets (240 images). A separate non-dog ImageNet gallery uses 50 classes with four images each (200 images) to evaluate transfer of dog-trained models.

Adaptation and held-out fidelity. Each domain contains 25 adaptation classes, with 40 training images and 10 held-out images per class, giving 1,000 training and 250 held-out images. Training and held-out images are disjoint, but their classes are shared. Fidelity therefore measures agreement with the teacher on new images of the adaptation classes. The full held-out set contains $\binom{250}{2} = 31{,}125$ unordered off-diagonal pairs. Matched cross-judge comparisons instead use a shared 100-image test subset, giving 4,950 pairs.

Validation split. For Dogs, we additionally sample 10 images from each of the 25 adaptation breeds, giving 250 validation images. These share classes with the adaptation set but are image-disjoint from training, the held-out fidelity set, and all reported evaluation galleries and benchmark sets. Validation images are excluded from teacher-target construction, kernel fitting, and adapter training, and are not added to training after hyperparameter selection. We provide their image identifiers in the experimental manifests.

Retrieval on unseen classes. The retrieval sets use classes excluded from adaptation in the corresponding domain. Each image queries the remaining images in its set, with relevance defined by

class-label agreement rather than teacher scores. Here, "unseen" refers to our adaptation data, not to CLIP's pretraining corpus.

For a query q in a set of N images, let $r_q(k)$ indicate whether the image at rank k has the same class label, and let $R_q = \sum_{k=1}^{N-1} r_q(k)$. We compute

$$\mathrm{AP}(q) = \frac{1}{R_q} \sum_{k=1}^{N-1} r_q(k) \frac{\sum_{j=1}^{k} r_q(j)}{k}, \qquad \mathrm{mAP} = \frac{100}{N} \sum_{q=1}^{N} \mathrm{AP}(q). \tag{9}$$

The query contributes neither a ranked result nor a relevant item: $R_q = 19$ in the five domain galleries and $R_q = 3$ in the non-dog gallery. Reported mAP values are percentages and gains are percentage-point differences. Paired permutation tests use 10,000 random sign flips of the query-level AP differences from CLIP and compare absolute mean differences.

Criteria on shared images. The main criteria concern breed, build, coat and markings for dogs; breed and coat for pets; pattern structure for textures; body style and model for cars; and petal shape and arrangement for flowers. Additional criteria cover whole-image colour and lighting; pose, viewpoint, and framing; coat texture (dogs); and background. "Colour/lighting" and "Pose/view" are compact labels for the first two composite criteria, not isolated coat/paint colour or body pose. Within each domain, criterion comparisons reuse the same images and splits, so changing the criterion changes the supervision rather than the image set. Verbatim scoring prompts are provided in Appendix G.

Zero-shot evaluation. ZS is the unweighted mean top-1 accuracy on ImageNet (Deng et al., 2009), CIFAR-10 and CIFAR-100 (Krizhevsky, 2009), Caltech-101 (Fei-Fei et al., 2004), FER2013 (Goodfellow et al., 2013), Oxford-IIIT Pets (Parkhi et al., 2012), DTD (Cimpoi et al., 2014), RESISC45 (Cheng et al., 2017), EuroSAT (Helber et al., 2019), PCam (Veeling et al., 2018), ImageNet-Sketch (Wang et al., 2019), and ImageNet-O (Hendrycks et al., 2021). The reported scores use the jointly adapted visual and text encoders, measuring zero-shot recognition after joint image–text adaptation.

C.2 Training and evaluation protocols

Feature teachers. DINOv2+KL uses the frozen DINOv2 ViT-L/14 with four register tokens (Oquab et al., 2024; Darcet et al., 2024), loaded from `facebookresearch/dinov2` through the Torch Hub entry point `dinov2_vitl14_reg`, with weights `dinov2_vitl14_reg4_pretrain.pth`. We extract the final layer-normalized [CLS] token. RGB images are bicubically resized to a shorter side of 224 pixels, centre-cropped to 224×224, and normalized with ImageNet mean $(0.485, 0.456, 0.406)$ and standard deviation $(0.229, 0.224, 0.225)$. The SigLIP (Zhai et al., 2023) feature-teacher control uses `google/siglip-so400m-patch14-384`, extracting `pooler_output` from `SiglipVisionModel`. Its supplied `SiglipImageProcessor` resizes RGB images to 384×384 and normalizes each channel with mean and standard deviation 0.5. Both feature teachers remain frozen.

Scoring budget. Training images are arranged into stratified groups of 32, with eight classes and four images per class. Each group contains $\binom{32}{2} = 496$ unordered pairs. The default Qwen3.5-397B protocol uses 93 groups, giving 46,128 pair slots and 44,450 distinct pairs in the dog run; the difference reflects pairs repeated across groups. Other judges use 31 groups (15,376 pairs), and matched cross-judge comparisons use this same budget for Qwen3.5-397B. Reduced-budget experiments reuse the first 31 or 62 groups of the default protocol.

Kernel fitting. The default targets are PSD-projected teacher similarities. We fit the cubic link parameters by least squares on frozen-encoder features before adaptation, then keep them fixed during training and evaluation. The link is fitted separately for each teacher and backbone. Kernel normalization follows the method section. Fidelity compares the student's normalized kernel with the named teacher's raw scores over unordered off-diagonal held-out pairs. Every arm is evaluated against unprojected judgments, so PSD projection changes the training targets rather than the evaluation target.

Shuffled-score control. We permute each group's off-diagonal entries of the dog primary-criterion target, preserving their distribution while destroying the correspondence between image pairs and judgments. Table 4 uses this single dog-trained negative control across all evaluation columns. It

checks the importance of meaningful pair supervision; it does not isolate criterion-specific suppression. Criterion specificity is assessed by comparing matching and mismatched students against the same held-out teacher.

Optimization. For **ASK**, we train rank-32 LoRA adapters with scaling $\alpha = 64$ in the query, key, value, and output projections of both CLIP towers, keeping pretrained weights frozen. Training uses AdamW (Loshchilov & Hutter, 2019) for 3,000 steps, each using one scored group of 32 images, with 200 warm-up steps. For **ASK**, SupCon, and DINOv2+KL, we search learning rates in $\{5\text{e}{-3}, 1\text{e}{-4}, 5\text{e}{-5}, 1\text{e}{-5}\}$ and KL weights in $\{0.01, 0.1, 0.2, 1.0\}$, selecting each method and backbone separately on the same Dogs validation set. **ASK** tuning uses the default Qwen3.5-397B breed judgments. Each configuration is trained for 3,000 updates and evaluated using its final checkpoint. Selection maximizes validation retrieval mAP: each validation image queries the other 249 images by cosine similarity, with the nine images of the same breed defining relevance. The selected settings are reused across the other domains. Reported fidelity, retrieval, and zero-shot test scores do not enter selection. The KL anchor uses a bank of 1,000 ImageNet class-name prompts. In the CLIP vision-transformer backbone comparison (Radford et al., 2021; Dosovitskiy et al., 2021), ViT-B/32 and ViT-B/16 use the full bank at each step, while ViT-L/14 and ViT-L/14-336 sample 200 prompts per step. Experiments were run on four NVIDIA RTX A6000 GPUs.

Geometry visualization protocol. Figure 3(a,b) uses the same 250 held-out dog images, ordered by breed. VLM targets are PSD-projected kernels. Original and aligned CLIP use the same fitted cubic link on unnormalized visual features, followed by diagonal normalization. Histograms include all 31,125 unordered off-diagonal pairs, with shared bins and axis limits and independently normalized same-breed and different-breed densities. The reported $\Delta\mu$ is the mean same-breed similarity minus the mean different-breed similarity. Panel (c) evaluates cosine retrieval against human attribute annotations on CUB-200-2011, following Appendix D.5.

Ablation protocols. Unless varied, ablations use learning rate $1\text{e}{-5}$ and KL weight 0.2, without separate tuning for each ablation. Table 5(a) uses 31 groups per teacher and criterion, matched within each raw/PSD pair. Dog-breed fidelity uses the same 100 test images (4,950 pairs); Qwen car-pose fidelity uses 250 test images (31,125 pairs). Retrieval uses 500 images from 25 unseen breeds or car-model classes, respectively. Neg.% is the fraction of absolute eigenvalue mass on negative eigenvalues, averaged over training groups. Table 18 includes all six comparisons. Panels (b–f) use 93 groups; their default uses Qwen3.5-397B expected-digit scores, PSD targets, and the default KL anchor with weight 0.2. Panel (b) compares binary labels, WordNet, and VLM class-pair means with individual judgments under shared kernel MSE, KL, and LoRA (Table 17). Verdict extraction is examined in Appendix E.3. In (c), each criterion has separately trained linear, MLP, and LoRA models, all with the same KL anchor. The MLP has 2.36M visual parameters, matching visual LoRA, while the linear head has 0.26M; each model also trains 1.57M text parameters. Panel (d) evaluates the breed-trained models and reports mAP gains over untrained CLIP; "Other domains" averages Textures, Flowers, Cars, and Pets. The default breed LoRA checkpoint is shared across panels (b–f). Panel (e) retains selected anchor ablations at the listed weights; the full regulariser sweep is in Table 19. Panel (f) compares cosine MSE, normalized-cubic MSE, and normalized-cubic ranking with the same LoRA and KL settings. Its fidelity uses each training similarity with fixed training-fitted parameters. The full head and kernel comparisons are in Tables 20 and 21.

C.3 Supervised contrastive baseline

SupCon uses the same training images and 93 groups as the corresponding ASK run, with rank-32 LoRA in both towers, AdamW, 200 warm-up steps, and 3,000 updates. The supervised contrastive loss uses temperature 0.07, treats other images of the same class as positives, and excludes self-pairs. It replaces the visual kernel-regression objective while retaining KL anchoring.

Both methods use the same prompt construction, group sampler, optimizer, and scheduling implementation. The anchor uses all 1,000 ImageNet class-name prompts per step on ViT-B/16 and ViT-B/32, and the corresponding ASK run's 200-prompt subsampling on ViT-L/14 and ViT-L/14-336. In the non-ImageNet domains, labels determine the contrastive positives; they do not select prompts from the full anchor bank. The schedule follows the reference implementation's update ordering, including its initial full-learning-rate update before warm-up.

Table 6: **Per-benchmark zero-shot accuracy across CLIP backbones.** Top-1 accuracy (%); Mean is the unweighted average over 12 benchmarks. Blue/red cells indicate gains/losses relative to untrained CLIP within each backbone, scaled separately per column across backbones; bold marks the best result per column.

	Method	IN-1k	C-10	C-100	Caltech	FER	Pets	DTD	RESISC	EuroSAT	PCam	IN-Sk.	IN-O	Mean
ViT-B/32	Untrained CLIP	63.5	**89.7**	63.3	**81.6**	41.4	87.3	**44.3**	53.6	50.4	**62.3**	42.3	47.8	**60.6**
	Shuffled teacher	62.6	87.6	59.3	80.6	36.7	86.0	42.6	51.0	42.0	62.1	41.4	47.0	58.2
	SupCon	58.7	83.5	57.5	79.1	**45.8**	79.9	40.9	50.7	38.9	57.4	38.9	43.5	56.2
	KUEA	63.5	89.6	63.4	**81.6**	41.5	87.4	44.0	53.6	51.4	62.1	**42.4**	47.3	**60.6**
	DINOv2 + KL	63.5	87.6	62.0	81.0	37.4	88.0	42.2	53.4	46.7	60.7	42.0	47.2	59.3
	SigLIP tower	**63.6**	89.0	63.0	81.3	39.3	**88.2**	43.6	**54.3**	**52.3**	61.4	42.3	**48.2**	60.5
	Qwen2.5-VL tower	63.4	88.9	**64.0**	81.0	39.9	88.0	44.1	53.3	50.5	61.3	42.2	47.7	60.4
	ASK (ours)	62.4	87.6	61.9	80.3	38.0	86.7	42.2	53.4	48.4	60.7	41.5	47.3	59.2
ViT-B/16	Untrained CLIP	68.3	90.0	65.6	82.2	46.3	89.0	44.9	58.2	55.9	50.7	48.3	42.3	61.8
	Shuffled teacher	66.9	86.8	62.7	81.5	45.9	85.6	44.9	54.3	51.9	**54.2**	47.4	41.2	60.3
	SupCon	63.7	87.5	62.2	79.7	41.9	84.8	43.3	57.8	53.2	53.0	44.6	39.0	59.2
	KUEA	68.3	90.0	65.8	82.0	46.6	88.9	44.8	58.1	56.2	51.2	48.3	42.6	61.9
	DINOv2 + KL	68.5	89.5	66.3	82.2	45.9	**91.0**	43.8	58.1	57.1	50.9	**48.4**	43.1	62.1
	SigLIP tower	**68.6**	89.6	65.9	**83.2**	46.1	89.9	44.9	58.3	56.8	50.5	48.0	43.1	62.1
	Qwen2.5-VL tower	68.3	**90.4**	**67.1**	82.5	**47.8**	90.4	**45.1**	57.5	58.1	53.7	48.1	**43.2**	**62.7**
	ASK (ours)	67.7	89.2	66.6	82.7	44.9	90.3	43.9	**60.4**	**59.0**	51.6	47.8	43.1	62.3
ViT-L/14	Untrained CLIP	75.0	95.2	71.1	83.3	50.0	93.2	55.2	63.3	62.6	52.0	59.6	32.2	66.1
	Shuffled teacher	73.7	94.6	67.2	83.9	48.3	91.5	55.2	61.7	60.8	53.6	59.1	32.3	65.2
	SupCon	72.9	92.1	70.4	83.7	49.9	91.5	**56.3**	64.8	62.0	**63.1**	58.1	31.5	66.4
	KUEA	75.1	95.2	71.3	83.4	49.8	93.2	55.2	63.4	62.3	52.1	59.6	32.5	66.1
	DINOv2 + KL	**75.6**	95.1	73.2	**84.8**	48.6	**94.1**	56.1	64.2	65.4	54.4	**60.1**	32.8	67.0
	SigLIP tower	75.3	**95.4**	72.5	84.0	49.3	93.7	56.0	64.0	63.3	54.9	59.7	33.0	66.8
	Qwen2.5-VL tower	75.1	95.2	73.4	84.4	**51.0**	93.4	55.7	63.4	65.1	51.6	59.9	33.5	66.8
	ASK (ours)	75.2	95.0	**74.5**	84.7	47.7	93.9	56.1	**64.9**	**67.4**	54.2	59.8	**33.6**	**67.2**
ViT-L/14-336	Untrained CLIP	75.8	94.5	71.1	83.4	49.0	93.7	55.6	63.7	61.5	60.7	61.0	32.8	66.9
	Shuffled teacher	75.0	93.4	66.2	83.6	50.2	91.7	55.3	62.1	57.6	59.7	60.1	31.4	65.5
	SupCon	73.0	90.5	67.4	83.9	45.2	89.2	55.6	**67.1**	62.2	61.8	58.8	30.8	65.5
	KUEA	75.9	94.4	71.5	83.5	49.5	93.7	55.4	63.9	61.3	60.9	61.0	33.0	67.0
	DINOv2 + KL	**76.4**	94.2	74.2	**84.3**	46.5	93.9	**57.2**	65.6	64.2	**63.1**	**61.4**	32.5	67.8
	SigLIP tower	76.1	**94.6**	72.5	83.5	48.9	93.9	56.5	63.9	63.9	60.7	61.0	**33.4**	67.4
	Qwen2.5-VL tower	76.0	94.5	73.9	83.7	**51.1**	93.7	57.0	64.3	65.3	61.9	61.3	33.3	**68.0**
	ASK (ours)	76.0	94.2	**75.0**	**84.3**	47.1	**94.0**	56.8	66.8	**65.6**	57.5	61.0	**33.4**	67.6

SupCon and DINOv2+KL follow the validation protocol in Appendix C.2, using the same search grid, validation images, and selection metric as **ASK**, with hyperparameters selected separately for each method. Every candidate uses its final 3,000-update checkpoint; intermediate checkpoints are not selected. Fidelity uses the corresponding ASK link fitted on training data, while retrieval uses cosine similarity with the query excluded.

D Additional results

D.1 Per-benchmark zero-shot accuracy

Evaluation. Table 6 expands the ZS averages in Table 2 for dog-trained models. ImageNet uses 10 images per class across 1,000 classes; the other 11 benchmarks use their full test splits. C-10/100 denote CIFAR-10/100, FER denotes FER2013, IN-Sk. denotes ImageNet-Sketch, and IN-O denotes ImageNet-O. The methods retain their respective targets and regularisers from the main comparison. Colour intensity is scaled separately for each benchmark and the mean, using a shared scale across backbones.

Retention varies across benchmarks. **ASK** improves mean zero-shot accuracy on ViT-B/16, ViT-L/14, and ViT-L/14-336, but reduces it on ViT-B/32. The mean hides consistent trade-offs: CIFAR-10 and FER2013 decline on all four backbones, while CIFAR-100, Caltech-101, Pets, RESISC45, EuroSAT, and ImageNet-O improve on the three larger backbones. For ViT-B/16, EuroSAT rises from 55.9 to 59.0 and RESISC45 from 58.2 to 60.4, while CIFAR-10 falls from 90.0 to 89.2. On

Table 7: **Retrieval and fidelity on unseen classes.** CLIP ViT-B/16. Retrieval uses 500 unseen-breed and 200 non-dog images; scores are percentages and Δ is the mAP gain over CLIP. Fidelity is Spearman ρ against Qwen3.8-27B on 250 unseen-breed images. All uses every pair; Within uses pairs of the same breed; Between uses pairs of different breeds. Blue/red indicates improvement/decline from CLIP, scaled per metric column; bold marks the best score within each block.

		In domain: 25 unseen breeds					Out of domain: 50 non-dog classes					Unseen-breed fidelity (ρ)		
	Method	R@1	R@5	P@5	mAP	Δ	R@1	R@5	P@5	mAP	Δ	All	Within	Between
Target	Untrained CLIP	47.0	82.2	38.8	22.1	—	77.0	93.5	44.1	70.0	—	0.268	0.231	0.245
	Shuffled teacher	36.6	76.0	28.8	18.0	−4.1	73.5	93.0	43.1	67.7	−2.3	0.239	0.227	0.211
	SupCon	69.4	87.2	62.0	47.7	+25.6	78.5	92.0	43.9	71.3	+1.3	0.546	0.471	0.508
	KUEA	47.4	82.2	38.9	22.3	+0.2	78.0	93.5	44.0	70.1	+0.1	0.271	0.232	0.247
	DINOv2 + KL	75.2	92.0	67.5	52.0	+29.9	**81.0**	95.0	**47.1**	74.8	+4.8	0.560	0.536	0.523
	SigLIP tower	59.6	86.2	46.3	29.6	+7.5	77.0	94.0	44.3	70.3	+0.3	0.356	0.291	0.326
	Qwen2.5-VL tower	59.6	87.0	47.0	30.3	+8.2	79.5	94.5	44.5	71.6	+1.5	0.401	0.357	0.370
	ASK	**78.6**	**93.6**	**74.0**	**60.2**	+38.1	80.5	**95.5**	46.7	**74.9**	+4.9	**0.756**	**0.657**	**0.734**
ASK by judge	Qwen3.5-397B	**78.2**	**93.0**	**73.9**	**59.7**	+37.6	78.5	94.0	**46.9**	**74.7**	+4.7	**0.765**	0.657	**0.745**
	GPT-5.6 Sol	77.8	91.6	72.6	58.9	+36.9	78.0	95.0	46.2	73.7	+3.7	0.725	0.610	0.701
	Kimi-K3	77.8	92.2	72.3	59.0	+36.9	78.0	94.5	45.8	73.2	+3.2	0.764	**0.668**	0.743
	Claude Sonnet 5	76.0	91.8	72.5	58.2	+36.2	**79.0**	94.5	46.0	73.3	+3.3	0.750	0.665	0.729
	Qwen3.8-27B	77.0	92.6	72.6	58.9	+36.9	78.5	94.5	45.5	73.1	+3.1	0.745	0.660	0.723
	Gemma 4 12B	77.0	91.8	70.4	55.9	+33.9	77.5	**95.5**	45.3	72.0	+1.9	0.698	0.602	0.671

ViT-L/14-336, PCam falls from 60.7 to 57.5 despite the higher overall mean. Thus, average retention does not imply unchanged accuracy on every benchmark.

D.2 Retrieval and fidelity on unseen classes

Evaluation. Table 7 reports cosine retrieval using CLIP ViT-B/16 on 500 images from 25 unseen dog breeds and 200 images from 50 non-dog ImageNet classes. Each image queries all other images in its set, and matching class labels define relevance. R@k is the percentage of queries with at least one relevant image among the first k results; P@5 is the mean fraction of relevant images in the first five. mAP averages query-level average precision over the full ranking. The upper block compares supervision sources, with **ASK** using the default 93-group Qwen protocol. The SupCon row uses the baseline configuration described in Appendix C.3, and the same final checkpoint as Table 2. The lower block compares **ASK** students trained with different judges under the matched 31-group protocol.

Retrieval and transfer. In the upper block, **ASK** reaches 60.2 mAP on unseen breeds, compared with 52.0 for DINOv2+KL and 47.7 for SupCon. The corresponding P@5 scores are 74.0, 67.5, and 62.0, showing that the gain also appears among the first few neighbours. On non-dog classes, **ASK** and DINOv2+KL perform similarly (74.9 and 74.8 mAP). SupCon also improves on untrained CLIP, reaching 71.3 versus 70.0 mAP. Under the matched scoring budget, all six judges improve retrieval over CLIP on both sets, with mAP gains of 33.9–37.6 points on unseen breeds and 1.9–4.7 on non-dog classes. For each unseen-breed gain, none of 10,000 paired sign-flip permutations is as extreme as the observed difference. The non-dog gains are not uniformly statistically significant: Gemma 4 12B's 1.9-point gain has $p = 0.116$.

Fidelity on unseen breeds. The three fidelity columns in Table 7 evaluate agreement with a common Qwen3.8-27B judge under the unchanged dog-similarity rubric. A deterministic subset of 10 images per unseen breed is selected before scoring, giving 250 images. All uses all 31,125 unordered, distinct image pairs. Within uses the 1,125 pairs whose images belong to the same breed; Between uses the 30,000 pairs whose images belong to different breeds. Entries are Spearman correlations with raw expected-digit judgments. All saved encoders use the same normalized-cubic link fitted on the original ASK training data; neither weights nor link parameters are updated on these images or judgments. Every student is compared with the same evaluation judge, irrespective of its training judge.

Graded similarity beyond the adaptation breeds. **ASK** reaches 0.756 overall agreement, compared with 0.268 for CLIP and 0.560 for DINOv2+KL. Its within-breed and between-breed correlations are 0.657 and 0.734, compared with 0.536 and 0.523 for DINOv2+KL. Computing the correlation separately within each breed and averaging the 25 values gives 0.652 for **ASK**, 0.299 for CLIP, and 0.543 for DINOv2+KL. This additional calculation holds the breed fixed, separating within-breed variation from differences in mean similarity across breeds. These measurements assess agreement with the common evaluation VLM on new breeds.

D.3 Criterion transfer beyond dogs

Holding the criterion fixed across domains. Table 8 evaluates the five dog-trained criterion encoders and two controls from Table 4 on the 200 non-dog images used for class retrieval in Table 7. Here the targets are raw Qwen3.5-397B judgments under the unchanged colour/lighting, pose/view, and background prompts. This measures whether the requested notion of similarity transfers to new image classes; background agreement, for example, need not preserve class identity. Each encoder retains its training-fitted, diagonally normalized cubic similarity, with no adaptation or link fitting on non-dog images; CLIP uses the breed link. Each criterion uses the same 19,900 unordered pairs from 50 classes. These images share neither paths nor identical file contents with the adaptation images.

Table 8: **Criterion transfer beyond dogs.** Spearman fidelity to raw Qwen3.5-397B judgments on 200 non-dog images (50 classes; 19,900 pairs). Dog-trained encoders receive no non-dog adaptation, and each column reuses its dog scoring prompt unchanged. Each model keeps its training-fitted cubic link; CLIP uses the breed link. Blue/red shading denotes improvement/decline relative to CLIP, scaled within the table; bold marks column bests.

Training criterion / model	Colour/lighting	Pose/view	Background
Untrained CLIP	*0.343*	*0.464*	*0.535*
Breed	0.302	0.390	0.472
Colour/lighting	**0.428**	0.451	0.576
Pose/view	0.332	**0.492**	0.552
Coat texture	0.320	0.412	0.498
Background	0.387	0.413	**0.795**
DINOv2+KL	0.342	0.431	0.494
Shuffled (neg.)	0.323	0.435	0.489

Transfer varies across criteria. The matching student leads every teacher column, exceeding CLIP, the mismatched students, and both controls. Background fidelity rises from 0.535 to 0.795, colour/lighting from 0.343 to 0.428, and pose/view from 0.464 to 0.492. Transfer is strongest for background and more modest for pose/view. Cosine similarity preserves the matching-model ordering, with corresponding fidelities of 0.795, 0.427, and 0.491. Criterion specificity on dogs therefore coexists with transfer beyond dogs. These measurements assess fidelity to the training judge on new classes; they do not provide independent human or attribute-ground-truth validation.

D.4 Criterion specificity on car images

Comparing criteria on fixed images. Table 9 complements the car retrieval examples by evaluating every encoder against every criterion. Qwen3.5-397B scores the same 250 held-out images under all four prompts. These images come from the 25 adaptation classes and are disjoint from the 1,000 training images. Fidelity uses all 31,125 unordered pairs, raw judgments, and each encoder's fixed training-fitted cubic link. Models are evaluated at the final 3,000-step checkpoint. Their scoring budgets differ, so this comparison assesses criterion fidelity rather than the effect of budget. Colour/lighting concerns the whole image, while pose/view includes the camera angle and composition (Appendix G).

Matching criteria yield the highest fidelity. Each matching student leads its teacher column: 0.859 for car model, 0.704 for colour/lighting, 0.756 for pose/view, and 0.813 for background. This ordering also holds under cosine similarity. The car retrieval examples are thus supported by quantitative

Table 9: **Criterion specificity on cars.** Spearman fidelity to raw Qwen3.5-397B judgments on the same 250 held-out car images (25 adaptation classes; 31,125 pairs). Rows identify car-trained encoders and columns the evaluation criteria. Each model keeps its training-fitted cubic link; CLIP uses the car-model link. Blue/red shading denotes improvement/decline relative to CLIP, scaled within the table; bold marks column bests.

Training criterion / model	Car model	Colour/lighting	Pose/view	Background
Untrained CLIP	*0.616*	*0.232*	*0.317*	*0.081*
Car model	**0.859**	0.227	0.384	0.051
Colour/lighting	0.409	**0.704**	0.257	0.231
Pose/view	0.497	0.196	**0.756**	0.096
Background	0.209	0.303	0.154	**0.813**
DINOv2+KL	0.597	0.190	0.310	0.057

criterion-specific agreement on held-out images. As with the dog comparisons, agreement with the training judge does not establish independent criterion-retrieval accuracy.

D.5 Criterion retrieval with human annotations

Human attributes and adaptation. We evaluate wing colour, bill shape, and breast pattern using existing image-level human annotations from CUB-200-2011 (Wah et al., 2011). Three CLIP ViT-B/16 students use the same 992 uniformly sampled training images and 44,702 Qwen3.8-27B judgments per criterion, with the kernel objective, PSD processing, and KL anchor described in the method. Prompts compare the visible attribute while ignoring species and unrelated appearance. Each request contains two training images and the corresponding criterion prompt; Appendix G.4 gives the complete prompts. Human attribute labels do not enter student training. DINOv2+KL uses the same images with criterion-independent feature targets.

Cosine retrieval across species. Evaluation uses the 5,794 official test images. For criterion c, $A_c(x)$ contains attributes annotated as present with certainty at least 3; images with empty sets are excluded. Each query retrieves eligible images of other species, giving 4,985, 5,266, and 5,001 queries for wing colour, bill shape, and breast pattern. Relevance is the Jaccard overlap $r_c(x, y) = |A_c(x) \cap A_c(y)|/|A_c(x) \cup A_c(y)|$. We report nDCG@10 with linear relevance gain and logarithmic rank discount, normalized by each query's ideal ranking. All methods use cosine similarity and identical eligible galleries; neither a fitted kernel link nor VLM judgments enter evaluation.

Baseline representations. Alongside CLIP (Radford et al., 2021) and DINOv2+KL (Oquab et al., 2024), two training-free controls use CLIP's original text encoder to describe attribute values: *text subspace* projects image embeddings onto the span of centred text vectors, and *text profile* uses image–text similarities as coordinates. Neither fits image labels. VLM2Vec (Jiang et al., 2025) (Qwen2-VL-2B) embeds images with task instructions; MagicLens (Zhang et al., 2024) (CLIP-B/16) learns image-and-instruction embeddings from web pairs and synthesized instructions. AligNet (Muttenthaler et al., 2025) learns criterion-independent representations from a human-aligned teacher's synthetic targets; we use its released CLIP-B's 512-dimensional image-encoder output. Each released baseline retains its native preprocessing and receives no weight adaptation on our evaluation datasets.

Selecting instruction settings. Following the initial baseline evaluation, we select CUB instruction settings by mean attribute nDCG@10 on 512 unused training images, disjoint from adaptation and test images. We compare detailed, short, and positive prompts, gallery conditioning, and VLM2Vec resolution. The selected native-resolution prompts are "Represent the [attribute] of the bird in this image" for VLM2Vec and "Find a bird with the same [attribute]" for MagicLens, applied to both query and gallery. Human validation labels guide selection; revised test scores do not.

Criterion accuracy. In Table 10(a), the matching **ASK** student leads each column and exceeds both mismatched students. Gains over CLIP are 4.09, 2.03, and 3.35 points. Gains over VLM2Vec are 3.78, 1.92, and 3.56, with paired 95% intervals [2.92, 4.70], [0.90, 2.91], and [2.43, 4.67], resampling query-species clusters with fixed checkpoints and galleries. **ASK** and DINOv2+KL use CUB adaptation, whereas the other learned baselines use released weights. Panel (b) compares

Table 10: **Criterion retrieval and baseline checks on CUB-200-2011.** Cosine nDCG@10 (×100) with cross-species galleries. **(a)** Model comparison; each **ASK** row is one fixed criterion-trained student. **(b)** Instruction settings and AligNet output spaces. Symmetric conditioning uses the same instruction for query and gallery; *marks validation-selected settings used in (a), where AligNet uses its image encoder. Blue/red denotes improvement/decline from CLIP in (a) or each block's first row in (b), scaled per column; bold marks column bests in (a) and block bests in (b).

(a) Attribute retrieval

Model	Wing colour	Bill shape	Breast pattern
Untrained CLIP	34.55	49.98	52.31
DINOv2+KL	35.62	50.28	52.72
CLIP text subspace	35.42	38.94	45.97
CLIP text profile	34.44	38.25	45.91
VLM2Vec	34.87	50.09	52.10
MagicLens	33.23	46.70	51.16
AligNet: base CLIP-B	31.91	46.86	51.25
AligNet: aligned CLIP-B	34.39	46.88	51.99
ASK: wing colour	**38.64**	47.26	51.90
ASK: bill shape	35.02	**52.01**	52.56
ASK: breast pattern	34.77	47.89	**55.66**

(b) Baseline checks

Configuration	Wing colour	Bill shape	Breast pattern
VLM2Vec			
Detailed, symmetric	33.79	45.96	50.86
Positive, symmetric*	**34.87**	**50.09**	**52.10**
MagicLens			
Detailed, query only	17.58	34.44	45.45
Short, symmetric*	33.23	**46.70**	51.16
Empty text on both sides	**34.87**	46.64	**51.27**
AligNet CLIP-B			
Base: image encoder	31.91	46.86	51.25
Aligned: image encoder	**34.39**	**46.88**	**51.99**
Base: triplet head	31.80	46.33	51.22
Aligned: triplet head	34.31	46.86	51.98

instruction settings and AligNet output spaces; MagicLens's instruction-free embeddings remain competitive with its selected configuration.

Transfer to species excluded from adaptation. We repeat CUB adaptation using 992 images from 100 species, reserving the other 100 species for evaluation. Fresh **ASK** and DINOv2+KL adapters start from original CLIP. The 2,903 held-out-species test images yield 2,526 colour, 2,634 shape, and 2,485 pattern queries under the same relevance and cross-species gallery rules. Figure 4 shows gains of 3.26, 1.58, and 2.55 points over DINOv2+KL; matching students also exceed both mismatched students, with all paired species-bootstrap intervals above zero. Species separation concerns adaptation, not foundation-model pretraining.

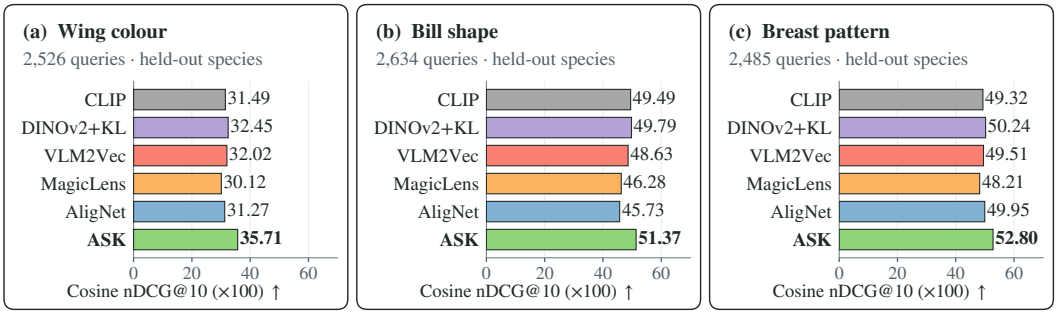


Figure 4: **Attribute retrieval on species excluded from adaptation.** **ASK** and DINOv2+KL adapt on 100 CUB species and evaluate on the other 100, with cross-species galleries. Each bar reports cosine nDCG@10 for one fixed encoder; **ASK** uses the matching criterion. VLM2Vec and MagicLens retain the earlier CUB settings, selected on validation images spanning all 200 species; their prompt selection is not species-disjoint.

Viewpoint, background, and pose on new datasets. Figure 5 evaluates existing car Pose/view and dog background and Pose/view encoders without updating their weights. PASCAL3D+ (Xiang et al., 2014) supplies 86 clear car images from the PASCAL validation split; within-category relevance is $\max(0, 1 - d_{\text{az}}/90°)$, where d_{az} is circular azimuth error. COCO-Stuff (Caesar et al., 2018) supplies 3,701 non-dog validation images with at least 20% stuff and 5% foreground coverage in the

CLIP crop. Relevance is Jaccard overlap of stuff classes covering at least 1% of the crop; galleries exclude the query's dominant foreground class. AP-10K (Yu et al., 2021) supplies 517 non-dog pose queries from 33 species in test split 1. We use single-animal images with at least ten visible keypoints, subtract the bounding-box centre, and divide coordinates by the longer box side. Within-species galleries control species identity; relevance is $\exp[-d^2/(2\cdot 0.1^2)]$, where d^2 is mean squared distance over at least eight jointly visible keypoints. These annotations define retrieval relevance, not direct human similarity votes; stuff overlap measures background composition.

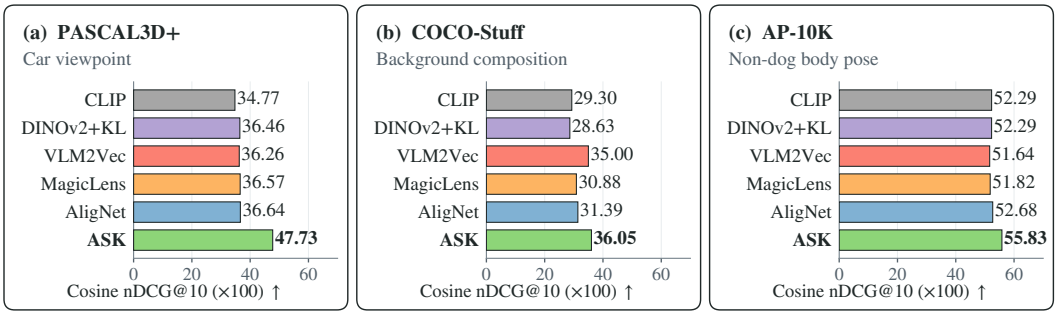


Figure 5: **Criterion transfer evaluated with independent annotations. (a)** Car viewpoint from PASCAL3D+ azimuth annotations. **(b)** Background composition from COCO-Stuff masks. **(c)** Non-dog body pose from AP-10K keypoints. Each bar reports cosine nDCG@10 for one fixed encoder, with no weight updates on these datasets. VLM2Vec and MagicLens use separate validation images to select instructions. **ASK** uses the car Pose/view, dog background, and dog Pose/view encoders, respectively; DINOv2+KL uses the corresponding adaptation domain.

Transfer comparisons. For each new dataset, VLM2Vec and MagicLens receive the same search budget of five prompt families and two gallery-conditioning formats, using 274 PASCAL, 512 COCO, and 768 AP-10K training images disjoint from evaluation and **ASK** adaptation. Selection averages nDCG@10 across vehicle categories for PASCAL, uses background retrieval for COCO, and averages non-dog pose and background for AP-10K. All models then use fixed weights and cosine retrieval. **ASK** exceeds the strongest plotted baseline by 11.09, 1.05, and 3.16 points on car viewpoint, background, and non-dog pose. Transfer is not uniform: on PASCAL buses and motorbikes, **ASK** scores 57.14 and 33.37 versus VLM2Vec's 57.85 and 33.50.

D.6 Per-benchmark retention across adaptation domains

Evaluation scope. Table 11 breaks down the 12-benchmark means in Table 1, using the evaluation protocol of Table 6. Each block changes the adaptation domain; the untrained ViT-B/16 baseline is shared. These results measure retention after adaptation. In particular, Pets evaluation after Pets adaptation and DTD evaluation after Textures adaptation use the same source datasets, so these entries should not be interpreted as transfer to an unseen domain. Dog adaptation also uses ImageNet classes; the ImageNet column measures retention on that benchmark.

Domain-dependent changes. **ASK** improves the 12-benchmark mean over CLIP in all five domains, but the gains are not uniform across benchmarks. Texture adaptation raises FER2013 accuracy from 46.3 to 50.1 and ImageNet-O from 42.3 to 44.2. Car adaptation raises CIFAR-100 from 65.6 to 67.4 and RESISC45 from 58.2 to 59.7, while lowering EuroSAT from 55.9 to 52.8. Pet adaptation improves Caltech-101 from 82.2 to 83.1 and RESISC45 to 60.0, but slightly lowers EuroSAT and ImageNet-Sketch. These variations show why the benchmark-level results complement the aggregate retention score.

D.7 Judge agreement and order consistency

Evaluation. Table 12 compares the six judges on the same 100 held-out dog images (4,950 pairs). Cross-judge agreement is Spearman correlation between their scores. To assess sensitivity to image order, 1,000 pairs are re-scored with the two images swapped. We report Spearman and Pearson self-agreement and the mean absolute score change $|\Delta|$ on the 1–9 scale. The final columns give

Table 11: **Per-benchmark accuracy after domain adaptation.** CLIP ViT-B/16; Mean averages the 12 benchmarks. Blue/red indicates improvement/decline from untrained CLIP, scaled per column across domains. Bold marks the highest adapted score within each domain when it exceeds CLIP. Same-domain evaluations are discussed in the text.

	Method	IN-1k	C-10	C-100	Caltech	FER	Pets	DTD	RESISC	EuroSAT	PCam	IN-Sk.	IN-O	Mean
	Untrained CLIP	68.3	90.0	65.6	82.2	46.3	89.0	44.9	58.2	55.9	50.7	48.3	42.3	61.8
Dogs	SupCon	63.7	87.5	62.2	79.7	41.9	84.8	43.3	57.8	53.2	53.0	44.6	39.0	59.2
	KUEA	68.3	90.0	65.8	82.0	46.6	88.9	44.8	58.1	56.2	51.2	48.3	42.6	61.9
	Qwen2.5-VL tower	68.3	**90.4**	**67.1**	82.5	**47.8**	90.4	**45.1**	57.5	58.1	**53.7**	48.1	**43.2**	**62.7**
	DINOv2 + KL	**68.5**	89.5	66.3	82.2	45.9	**91.0**	43.8	58.1	57.1	50.9	**48.4**	43.1	62.1
	ASK	67.7	89.2	66.6	**82.7**	44.9	90.3	43.9	**60.4**	**59.0**	51.6	47.8	43.1	62.3
Textures	SupCon	63.9	87.0	58.6	80.3	47.2	86.4	**46.7**	53.6	49.5	50.6	43.3	36.5	58.6
	KUEA	68.3	90.0	65.5	82.2	46.5	88.9	45.3	58.4	56.1	50.1	48.2	42.6	61.9
	Qwen2.5-VL tower	68.5	**90.6**	66.3	82.0	47.9	88.9	45.4	58.1	55.3	52.2	48.0	43.5	62.2
	DINOv2 + KL	68.4	90.3	65.8	82.1	47.6	88.8	45.0	57.8	55.6	52.0	47.9	43.6	62.1
	ASK	**68.6**	90.4	**67.2**	82.1	**50.1**	**89.4**	46.1	**59.0**	**57.5**	**52.6**	48.2	**44.2**	**62.9**
Flowers	SupCon	67.8	89.1	64.4	**82.6**	45.3	89.1	44.4	56.4	52.1	47.5	47.9	42.4	60.7
	KUEA	68.3	89.9	65.4	82.2	**46.5**	88.8	44.9	58.2	**56.3**	**51.0**	48.3	42.2	61.8
	Qwen2.5-VL tower	68.4	90.2	65.8	81.9	46.1	89.0	44.8	58.9	55.6	48.5	48.4	43.2	61.7
	DINOv2 + KL	68.5	90.0	65.1	82.1	45.1	88.8	44.1	59.2	55.9	48.8	48.4	**43.5**	61.6
	ASK	**68.9**	**90.4**	**66.5**	82.4	46.1	**89.7**	44.3	**59.3**	55.8	50.0	**48.7**	**43.5**	**62.1**
Cars	SupCon	67.8	89.5	66.0	80.4	**51.8**	88.0	44.6	58.8	52.1	46.5	47.4	42.3	61.3
	KUEA	68.3	90.0	65.6	82.1	46.4	89.0	45.1	58.2	**56.4**	51.0	48.3	42.2	61.9
	Qwen2.5-VL tower	68.2	90.5	65.9	82.0	45.3	89.0	44.9	58.2	53.6	**51.3**	48.2	**42.9**	61.7
	DINOv2 + KL	**68.4**	**90.9**	66.7	**82.4**	46.0	88.9	45.1	59.4	54.0	50.0	**48.6**	42.4	61.9
	ASK	**68.4**	90.6	**67.4**	81.5	48.8	88.7	**45.2**	**59.7**	52.8	50.3	**48.6**	**42.9**	**62.1**
Pets	SupCon	64.4	89.1	63.3	80.7	47.9	87.4	44.1	58.0	51.3	**55.7**	45.1	40.9	60.7
	KUEA	68.3	90.1	65.6	82.2	46.6	89.0	44.7	58.3	**56.1**	50.8	48.2	42.6	61.9
	Qwen2.5-VL tower	68.1	**90.5**	**67.3**	81.9	46.4	90.1	**45.7**	57.1	54.3	54.2	48.3	42.7	62.2
	DINOv2 + KL	68.2	89.5	65.4	82.1	45.2	**91.7**	44.6	58.1	55.3	55.5	48.1	42.2	62.2
	ASK	**68.4**	90.1	66.9	**83.1**	**48.3**	89.9	45.3	**60.0**	55.5	52.3	48.2	**42.8**	**62.6**

Table 12: **Agreement and order consistency of VLM judges.** Pairwise agreement is Spearman correlation. Blue intensity encodes absolute agreement on a shared scale; bold marks the highest cross-judge value per column. Self-agreement uses swapped image order; Ratio is student fidelity divided by self-Spearman agreement.

Judge	Agreement with judge						Self, swapped order			Own student	
	Q3.5-397B	GPT	Kimi	Sonnet	Q3.8-27B	G4-12B	ρ_S	ρ_P	$\lvert\Delta\rvert$	Fid	Ratio
Qwen3.5-397B	—	0.753	0.799	0.706	0.819	0.702	0.837	0.921	0.35	0.791	94%
GPT-5.6 Sol	0.753	—	0.811	0.725	0.770	0.679	0.901	0.926	0.32	0.794	88%
Kimi-K3	0.799	**0.811**	—	**0.750**	**0.829**	0.719	**0.948**	**0.975**	**0.20**	**0.838**	88%
Claude Sonnet 5	0.706	0.725	0.750	—	0.690	0.605	0.820	0.912	0.33	0.778	**95%**
Qwen3.8-27B	**0.819**	0.770	**0.829**	0.690	—	**0.738**	0.829	0.858	0.87	0.760	92%
Gemma 4 12B	0.702	0.679	0.719	0.605	0.738	—	0.811	0.834	0.57	0.770	95%

each student's fidelity to its own judge and its ratio to that judge's self-Spearman agreement. The ratio is descriptive: order consistency is not a strict upper bound on student fidelity, and the two correlations use different pair sets.

Shared structure and variability. Cross-judge correlations range from 0.605 for Sonnet–Gemma 4 12B to 0.829 for Kimi–Qwen3.8-27B, indicating shared but non-identical similarity judgments. Self-Spearman agreement ranges from 0.811 to 0.948. Kimi has the highest order consistency and the smallest mean absolute score change (0.20), whereas Qwen3.8-27B changes by 0.87 on average. Student fidelity is 88–95% of the corresponding self-agreement reference. These measurements help contextualize fidelity differences across judges without treating any judge as noise-free ground truth.

Table 13: **VLM judges across CLIP backbones.** All students use 31 scoring groups. Spearman shading compares ρ_S with Base ρ_S in the same judge row. These judge-specific CLIP baselines replace a single value in the Untrained CLIP row. Retrieval and zero-shot shading compare with that backbone's Untrained CLIP row. Blue/red denotes improvement/decline, scaled per metric; Pearson is unshaded. Bold marks column bests within each backbone.

		Fidelity			Retrieval	Zero-shot	
	Judge	ρ_S	ρ_P	Base ρ_S	mAP	12 sets	IN-1k
ViT-B/32	Untrained CLIP	—	—	—	18.7	**60.6**	**63.5**
	Qwen3.5-397B	0.761	0.769	0.230	48.0	58.1	61.9
	GPT-5.6 Sol	0.745	0.733	0.274	47.0	57.2	61.6
	Kimi-K3	**0.815**	**0.813**	0.182	**49.1**	57.7	61.7
	Claude Sonnet 5	0.719	0.752	0.251	48.0	58.0	61.7
ViT-B/16	Untrained CLIP	—	—	—	22.1	61.8	**68.3**
	Qwen3.5-397B	0.791	0.838	0.251	**59.7**	**62.0**	67.5
	GPT-5.6 Sol	0.794	0.804	0.287	58.9	61.9	67.0
	Kimi-K3	**0.838**	**0.878**	0.209	59.0	61.9	67.2
	Claude Sonnet 5	0.778	0.828	0.269	58.2	61.9	67.2
ViT-L/14	Untrained CLIP	—	—	—	30.5	66.1	75.0
	Qwen3.5-397B	0.815	0.873	0.240	71.1	**67.2**	**75.0**
	GPT-5.6 Sol	0.805	0.822	0.274	**71.4**	67.0	75.0
	Kimi-K3	**0.869**	**0.903**	0.191	70.7	67.2	74.6
	Claude Sonnet 5	0.800	0.856	0.254	69.5	66.8	75.0
ViT-L/14-336	Untrained CLIP	—	—	—	30.0	66.9	**75.8**
	Qwen3.5-397B	0.827	0.877	0.243	71.9	**67.7**	75.8
	GPT-5.6 Sol	0.809	0.825	0.276	**72.5**	67.3	75.7
	Kimi-K3	**0.869**	**0.902**	0.190	71.7	67.3	75.7
	Claude Sonnet 5	0.800	0.854	0.246	70.6	67.1	75.7

D.8 JUDGE COMPARISONS ACROSS BACKBONES

Matched scoring budget. Table 13 compares four judges across four CLIP backbones using 31 groups (15,376 training pairs) per judge. Fidelity is evaluated on each judge's scores for the same 100 held-out images. Kernel parameters are fitted separately for each teacher–backbone combination. The base Spearman column reports each untrained backbone's agreement with the corresponding judge; this baseline differs across judges even for a fixed backbone. Retrieval and zero-shot evaluation use the datasets and scoring protocols above. The ViT-B/16 rows share their zero-shot results with Table 3. Pearson values are left unshaded because the table does not report their untrained baselines.

Fidelity and retrieval. Every judge improves Spearman fidelity and unseen-breed retrieval over the corresponding CLIP baseline at every scale. Kimi yields the highest own-judge fidelity for all four backbones, but does not consistently give the best retrieval. The highest mAP is obtained with Kimi on ViT-B/32, Qwen on ViT-B/16, and GPT on both ViT-L/14 variants. Own-judge fidelity measures agreement with different targets and should therefore not be treated as a common measure of teacher quality.

Retention across scales. Mean zero-shot accuracy exceeds CLIP for every judge on both large backbones: the gains are 0.7–1.1 points on ViT-L/14 and 0.2–0.8 on ViT-L/14-336. ImageNet accuracy differs from CLIP by at most 0.42 and 0.17 points, respectively. ViT-B/16 stays close to CLIP, whereas ViT-B/32 loses 2.6–3.4 points in mean accuracy. The 93-group results in Table 2 use different checkpoints; these comparisons do not isolate the effect of scoring budget.

Per-benchmark evaluation. Table 14 expands the 12-benchmark means in Table 13. ImageNet uses 10 images per class over 1,000 classes; the other benchmarks use their full test splits. Means are computed before rounding the per-benchmark values.

Table 14: **Per-benchmark zero-shot accuracy across judges and backbones.** Top-1 accuracy (%) for Table 13; Mean averages the 12 benchmarks. Blue/red indicates improvement/decline from each backbone's untrained CLIP, scaled per column across backbones. Bold marks the highest adapted-model score within each backbone when it exceeds CLIP.

	Judge	IN-1k	C-10	C-100	Caltech	FER	Pets	DTD	RESISC	EuroSAT	PCam	IN-Sk.	IN-O	Mean
ViT-B/32	Untrained CLIP	63.5	89.7	63.3	81.6	41.4	87.3	44.3	53.6	50.4	62.3	42.3	47.8	60.6
	Qwen3.5-397B	61.9	86.5	60.8	80.1	36.8	86.2	41.1	52.4	42.1	61.2	41.1	46.5	58.1
	GPT-5.6 Sol	61.6	85.6	58.6	79.9	32.6	85.6	41.7	52.6	40.9	59.4	41.0	46.7	57.2
	Kimi-K3	61.7	86.1	59.8	79.9	36.6	85.2	41.1	51.6	43.3	59.9	40.6	46.6	57.7
	Claude Sonnet 5	61.7	86.2	60.7	80.1	38.2	85.7	41.8	52.7	41.6	60.1	41.0	46.8	58.0
ViT-B/16	Untrained CLIP	68.3	90.0	65.6	82.2	46.3	89.0	44.9	58.2	55.9	50.7	48.3	42.3	61.8
	Qwen3.5-397B	67.5	88.5	**66.1**	82.2	44.1	**90.6**	44.1	61.1	55.8	53.4	47.5	**43.1**	**62.0**
	GPT-5.6 Sol	67.0	87.9	64.5	81.6	45.6	89.7	44.0	60.3	**59.1**	**54.0**	47.4	42.1	61.9
	Kimi-K3	67.2	88.2	65.2	81.7	45.4	90.0	43.5	60.8	58.3	51.8	47.4	42.6	61.9
	Claude Sonnet 5	67.2	88.5	65.4	82.0	44.5	90.2	43.8	**61.3**	56.9	53.1	47.5	42.5	61.9
ViT-L/14	Untrained CLIP	75.0	95.2	71.1	83.3	50.0	93.2	55.2	63.3	62.6	52.0	59.6	32.2	66.1
	Qwen3.5-397B	75.0	95.0	**74.4**	84.8	48.3	93.5	55.1	65.4	**68.1**	53.1	59.8	33.5	**67.2**
	GPT-5.6 Sol	75.0	94.6	71.4	84.7	47.2	93.2	55.2	65.1	67.4	**56.7**	**59.8**	33.5	67.0
	Kimi-K3	74.6	95.0	73.5	**84.9**	47.1	93.5	55.2	**65.6**	67.5	55.8	59.6	**33.6**	67.2
	Claude Sonnet 5	75.0	94.5	73.0	84.9	49.0	**93.6**	54.6	65.0	66.0	52.5	59.7	33.2	66.8
ViT-L/14-336	Untrained CLIP	75.8	94.5	71.1	83.4	49.0	93.7	55.6	63.7	61.5	60.7	61.0	32.8	66.9
	Qwen3.5-397B	75.8	94.2	**73.7**	84.7	47.9	93.1	55.1	**67.1**	**66.3**	**61.0**	**61.1**	32.5	**67.7**
	GPT-5.6 Sol	75.7	93.6	72.2	84.1	47.3	93.2	55.4	66.9	65.8	59.0	61.0	**33.4**	67.3
	Kimi-K3	75.7	94.1	73.4	**84.7**	47.0	93.2	54.9	66.9	65.7	57.4	60.9	33.4	67.3
	Claude Sonnet 5	75.7	93.6	72.0	84.3	46.4	93.0	55.3	66.9	65.0	58.7	61.0	32.9	67.1

Benchmark-dependent retention. All four ViT-B/32 students lose accuracy relative to CLIP on every benchmark. On both large backbones, every judge improves CIFAR-100, Caltech-101, RESISC45, and EuroSAT, while reducing CIFAR-10 and FER2013 accuracy. Thus, the higher aggregate accuracy of the large-backbone students coexists with losses on individual benchmarks.

D.9 Teacher-scoring cost and student training time

Teacher-scoring budget. Table 15 estimates OpenRouter cost and time for the default dog criterion: 44,450 distinct training-pair judgments for 1,000 images. Each request contains two images and a criterion prompt. Costs use the listed provider rates; times extrapolate published P50 latency plus output duration with ten concurrent requests. These are planning estimates, not measured scoring times; provider conditions affect realized time, and retries and fees are excluded.

Table 15: **OpenRouter teacher-scoring budget estimates.** For 44,450 judgments: 300–700 input tokens per pair (including both images), two output tokens without reasoning, and ten concurrent requests. Names follow the experiments; † marks Gemma 3 12B pricing used as a proxy for Gemma 4 12B.

Judge	Provider	$/M tokens input / output	P50 latency (s)	Output tokens/s	Est. cost ($)	Est. time (h)
Qwen3.5-397B	Alibaba Cloud Int.	0.39 / 2.34	1.11	61	5.4–12.3	1.4
GPT-5.6 Sol	OpenAI	2.00 / 10.00	3.58	49	27.6–63.1	4.5
Kimi-K3	Parasail	3.00 / 15.00	1.11	68	41.3–94.7	1.4
Claude Sonnet 5	Anthropic	2.00 / 10.00	2.30	63	27.6–63.1	2.9
Qwen3.8-27B	Parasail	0.24 / 2.20	1.35	58	3.4–7.7	1.7
Gemma 4 12B†	DeepInfra	0.05 / 0.15	0.75	14	0.7–1.6	1.1

Cost per criterion and reuse. The 250-image fidelity set adds 31,125 judgments, bringing the total to 75,575 (about 1.70× the training-only budget). Each new criterion requires teacher judgments and student adaptation; cached judgments are reusable across backbones and ablations. Sparse supervision retains most of the retrieval gain with fewer calls (Figure 6, Appendix D.10), reducing estimated cost and time proportionally.

Student training. On one RTX A6000, median times for 3,000 steps are 67, 78, 92, and 181 minutes for ViT-B/32, ViT-B/16, ViT-L/14, and ViT-L/14-336, respectively. These timestamp-based estimates summarize observed configurations, exclude teacher scoring, and retain durations between 20 and 400 minutes.

D.10 Reducing teacher API calls

Strong performance with fewer queries. Sparse supervision uses 14,320 distinct judgments instead of 44,450 (67.8% fewer), while retaining 59.41 mAP versus 60.16 for the full budget and 62.56 zero-shot accuracy versus 62.28 (Figure 6). Under the API assumptions in Appendix D.9, this requires about one third of the estimated teacher-scoring cost and time. Simply reducing the dense budget to 62 groups also retains 60.07 mAP with 30,174 judgments.

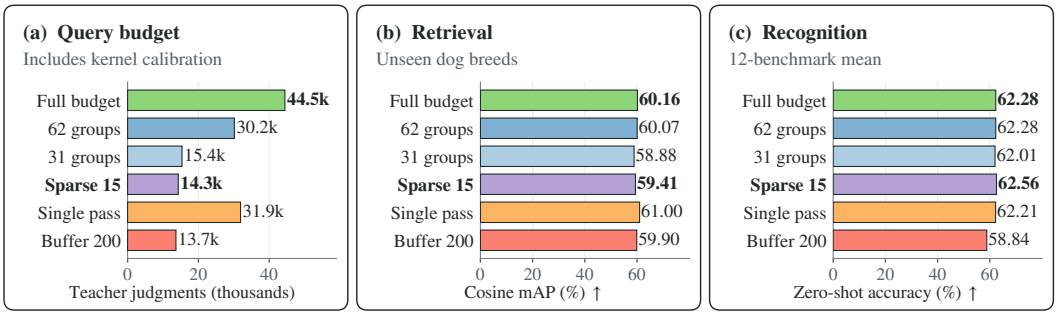


Figure 6: **Strong performance with fewer teacher API calls.** **(a)** Distinct judgments, including kernel calibration. **(b)** Cosine retrieval on unseen dog breeds. **(c)** Zero-shot recognition across 12 benchmarks. Full budget is the default 93-group baseline; Sparse 15 uses 15 pairs per block.

Baselines and variants. Full budget scores all 496 pairs within each of 93 groups of 32 images; the 62- and 31-group baselines use nested subsets. Sparse 15 selects 15 available pairs per random block. All cached pairs uses every available score (about 45 per block); wider blocks use 64 images and 30 pairs. Coverage selection chooses pairs to cover more distinct images. Sparse variants apply masked MSE only to observed, affine-scaled scores, without dense PSD projection or imputation.

Table 16: **Query budgets and performance.** Dog-trained ViT-B/16, 3,000 updates. Update pairs count distinct training judgments; total pairs also include the 9,920 calibration pairs, without double counting. Blue/red indicates improvement/decline from the full-budget baseline, scaled per metric; bold marks the best score per column.

Setting	Blocks	Pairs/block	Distinct judgments		Fidelity	Retrieval	Zero-shot
			Update pairs	Total pairs	ρ_S	mAP	12 sets
Full budget (dense)	93	496	44,450	44,450	**0.794**	60.16	62.28
Dense, 62 groups	62	496	30,174	30,174	0.790	60.07	62.28
Dense, 31 groups	31	496	15,376	15,376	0.781	58.88	62.01
Sparse 15 (random pairs)	400	15	5,624	14,320	0.780	59.41	**62.56**
All cached pairs per block	400	≈45	14,703	21,311	0.793	60.06	62.47
Wider blocks (64 images)	400	30	10,525	18,062	**0.794**	60.18	61.48
Single pass (each block once)	3,000	15	28,332	31,912	0.793	**61.00**	62.21
Sparse 15 (image coverage)	400	15	5,609	14,285	0.776	59.05	59.07
Sliding buffer (200 images)	400	15	4,857	13,661	0.749	59.90	58.84
Breed-sorted blocks	400	15	3,497	12,627	0.597	54.82	57.18
Same blocks, shuffled order	400	15	3,497	12,627	0.526	48.54	57.84

Streaming-style use. Single pass visits each block once, allowing images to recur, and reaches 61.00 mAP with 31,912 judgments. A sliding 200-image buffer retains 59.90 mAP but lowers zero-shot accuracy to 58.84. Breed-sorted blocks restrict batch diversity; shuffling their order preserves their composition. These cached-score replays support using sparse supervision when API budgets are limited; live acquisition remains untested.

E ABLATION STUDIES

E.1 CLASS LABELS VERSUS VLM JUDGMENTS

Table 17: **Class labels versus VLM judgments.** CLIP ViT-B/16 on Dogs, with shared kernel MSE and KL anchoring. Fidelity reports Spearman/Pearson agreement with raw held-out judgments. Retrieval mAP and the 12-benchmark zero-shot mean (ZS) are percentages. Blue/red indicates improvement/decline relative to untrained CLIP, scaled per column; bold marks the best score.

Supervision	Fidelity		Retrieval mAP		ZS
	Spearman	Pearson	Unseen breeds	Non-dog	
Untrained CLIP	0.281	0.280	22.06	70.01	61.81
Binary class labels	0.453	0.707	43.11	64.40	59.83
Grouped class labels	0.451	0.706	43.71	63.81	60.49
WordNet path similarity	0.444	0.707	52.01	69.41	60.70
Class-pair mean (VLM)	0.730	0.801	57.17	72.66	61.77
ASK (VLM judgments)	**0.794**	**0.840**	**60.16**	**74.91**	**62.28**

Class-derived similarities. Table 17 compares VLM judgments with class-derived targets under the same kernel objective. Unlike SupCon, this comparison retains kernel MSE as the image loss. Binary targets are $K_{ij} = \mathbf{1}[y_i = y_j]$. Grouped targets are

$$K_{ij} = \tfrac{1}{2}\mathbf{1}[y_i = y_j] + \tfrac{1}{2}\mathbf{1}[h(y_i) = h(y_j)],$$

where h assigns six fixed breed groups and treats each remaining class as its own group. Similarity is therefore 1 within a class, 0.5 between distinct classes in a group, and 0 otherwise. Both targets are PSD with unit diagonal. The groups are: Beagle, Bluetick Coonhound, Black and Tan Coonhound, and Redbone Coonhound; Toy and Miniature Poodle; Staffordshire Bull Terrier and American Staffordshire Terrier; Pekingese and Shih Tzu; Border and Norwich Terrier; and Kuvasz and Komondor.

Taxonomy-derived similarities. A graded label control uses WordNet 3.0 (Miller, 1995), mapping each ImageNet breed identifier directly to its noun synset. Its target is $K_{ij} = 1/(1 + d(y_i, y_j))$, where d is the shortest path through the hypernym hierarchy. This fixed path-similarity rule uses no VLM calls, manual breed groups, or held-out judgments. The 25 training breeds yield seven between-class distances (2–8 edges), giving scores from $1/9$ to $1/3$; same-class targets are 1. All image pairs with the same unordered breed labels share a target. The same training groups and PSD preprocessing are used; the resulting class kernel is already PSD, so projection changes the targets only at numerical precision. WordNet encodes semantic category relations, rather than genetic ancestry.

Averaging within class pairs. The class-pair-mean control replaces each raw training judgment with the mean over distinct scored training image pairs having the same unordered pair of class labels. Self-pairs are excluded from these means. The same affine scaling, unit self-similarity, and per-group PSD projection are then applied. All 325 class pairs are represented among the 44,450 scored training pairs; the 93 training groups are unchanged. This retains graded relationships between classes while removing the original variation among individual image pairs before PSD processing. No held-out ratings enter the averages or the link fit.

Shared adaptation protocol. All adapted models use the same 1,000 dog images and 93 groups, normalized-cubic kernel MSE, rank-32 LoRA in both towers, and the KL anchor of weight 0.2 over 1,000 ImageNet prompts. Optimization uses 3,000 AdamW updates, learning rate $1\mathrm{e}{-5}$, weight decay $1\mathrm{e}{-4}$, and 200 warm-up steps. Each target has its own cubic link fitted on frozen features from the first 20 training groups and fixed during adaptation. All models use the final checkpoint without selection on held-out data.

Fidelity compares each model's fixed link with raw VLM judgments on 250 held-out images of the adaptation classes; untrained CLIP uses the ASK link. This measures teacher imitation, although the label students do not train on those judgments. Cosine retrieval uses the same 500 unseen-breed images and 200 non-dog images as Table 7, excluding each query from its gallery and relevant-image count. Zero-shot accuracy is the unweighted mean over the same 12 benchmarks as Table 6.

Table 18: **PSD projection across judges and criteria.** Neg.% denotes negative spectral mass. Retrieval uses unseen breed labels for Dogs and unseen car-model labels for Cars. Raw-target cells show gains (blue) or losses (red) relative to PSD targets, scaled per metric; bold marks the higher score within each pair. Displayed ties are uncoloured.

Judge	Criterion	Neg.%	Fidelity		mAP		ZS	
			PSD	raw	PSD	raw	PSD	raw
Gemma 4 12B	Dogs / breed	6.4	**0.771**	0.761	**55.93**	54.49	**61.71**	61.30
GPT-5.6 Sol	Dogs / breed	3.8	**0.790**	0.785	**58.99**	58.45	**62.10**	61.81
Qwen3.5-397B	Cars / pose-view	3.5	0.758	**0.760**	**45.75**	44.30	**60.88**	60.63
Claude Sonnet 5	Dogs / breed	2.6	**0.777**	0.776	**58.57**	58.40	**62.07**	61.84
Qwen3.5-397B	Dogs / breed	1.8	**0.792**	0.791	**59.79**	59.57	62.26	**62.29**
Kimi-K3	Dogs / breed	0.7	**0.838**	**0.838**	59.35	**59.40**	**61.97**	61.85

Retrieval and recognition retention. Binary and grouped labels reach 43.11 and 43.71 unseen-breed mAP, compared with 60.16 for **ASK** and 22.06 for CLIP. Their non-dog mAP falls to 64.40 and 63.81, below CLIP's 70.01, while **ASK** reaches 74.91. Zero-shot means are 59.83 and 60.49 for the label targets, versus 62.28 for **ASK**. Thus, VLM judgments improve retrieval and retention over these two label-derived targets under the shared recipe. The class-pair-mean student reaches 57.17 unseen-breed mAP, 72.66 non-dog mAP, and 61.77 ZS, compared with 60.16, 74.91, and 62.28 for **ASK**. This comparison tests the effect of retaining individual image-pair judgments beyond their class-pair averages under the shared recipe; it does not establish optimal performance for all class-derived targets.

The WordNet control reaches 52.01 unseen-breed mAP, 69.41 non-dog mAP, and 60.70 ZS; its Spearman fidelity to the existing held-out VLM judgments is 0.444. This adds a query-free, graded taxonomy target to the binary and manually grouped label comparisons under the same adaptation recipe. No taxonomy metric or distance scale is selected using evaluation performance.

E.2 Effect of PSD projection

Matched comparison. Table 18 gives the absolute results underlying Table 5(a), together with the lower-negative-mass Kimi-K3 and Qwen dog-breed comparisons. Each raw/PSD pair uses the same 31 training groups, student, anchor, and training schedule. Dog-breed fidelity uses the shared 100-image holdout (4,950 pairs); car-pose fidelity uses 250 held-out car images (31,125 pairs). Retrieval uses 500 images from 25 unseen classes in the corresponding domain, with breed labels for Dogs and car-model labels for Cars. Thus, car retrieval measures preservation of model identity after pose alignment. Negative spectral mass is the mean, over training groups, of the fraction of absolute eigenvalue mass contributed by negative eigenvalues. The main table's deltas are calculated before rounding, so they can differ slightly from differences between the means displayed here.

Gains from correcting indefinite targets. Gemma 4 12B has the largest negative spectral mass (6.4%) and benefits on all three metrics: fidelity increases from 0.761 to 0.771, unseen-breed mAP from 54.49 to 55.93, and ZS from 61.30 to 61.71. GPT-5.6 Sol (3.8%) also improves on all three metrics, including 0.005 fidelity and 0.53 mAP. For Qwen car-pose targets (3.5%), PSD raises car-model mAP from 44.30 to 45.75 and ZS from 60.63 to 60.88, while fidelity decreases from 0.760 to 0.758. Sonnet (2.6%) gains 0.17 mAP and 0.23 ZS. The lower-negative-mass Qwen dog-breed (1.8%) and Kimi-K3 (0.7%) cases show smaller retrieval changes; Kimi's fidelity is unchanged at the displayed precision. The comparisons show that correcting negative spectral components can improve downstream performance, with the size and metric of the benefit depending on the teacher and criterion.

E.3 Verdict extraction

Comparison protocol. Figure 7 compares expected-digit scores, rounded expectations, and argmax digits derived from the same cached Qwen responses. All variants use the same 93 training groups,

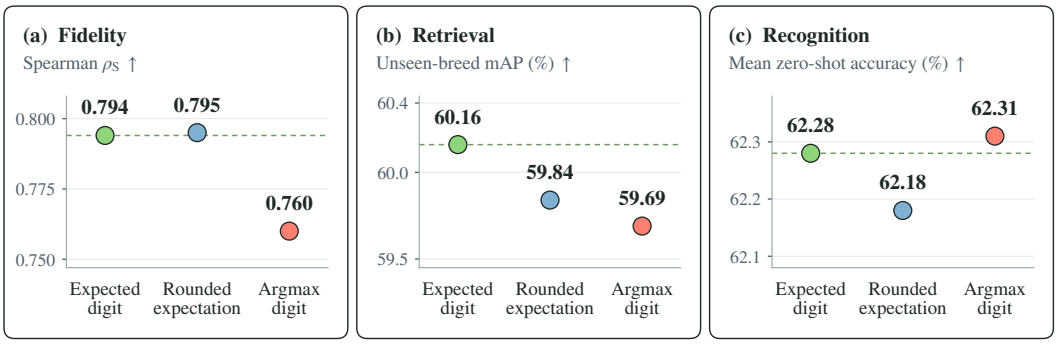


Figure 7: **Verdict extraction with Qwen3.5-397B.** Expected-digit scores (green), rounded expectations (blue), and argmax digits (coral). **(a)** Fidelity uses the normalized kernel against expected-digit judgments. **(b)** Cosine-retrieval mAP uses unseen-breed labels. **(c)** Zero-shot accuracy averages 12 benchmarks. Labels give exact reported values; dashed lines mark expected-digit results. Each panel uses its own zoomed vertical scale; higher is better.

PSD projection, kernel-fitting procedure, and optimization settings, including a KL weight of 0.2 and 1,000 anchor prompts per step. Fidelity uses each student's normalized kernel against the same expected-digit judgments on 250 held-out images. Retrieval uses cosine similarity on 500 images from 25 unseen breeds, excluding each query from its gallery; breed labels define relevance independently of the judge. ZS averages the 12 recognition benchmarks.

Fidelity, retrieval, and retention. Expected-digit scoring reaches 0.794 fidelity and 60.16 mAP, compared with 0.760 and 59.69 for argmax. Rounding the expectation gives similar fidelity (0.795), with 59.84 mAP. ZS is close across the three variants (62.18–62.31).

Distributional information and score precision. Taking the expectation combines support across rating levels; rounding that expectation can still produce a different digit from argmax. Thus, the expectation–argmax comparison and the expectation–rounded comparison test different choices. Moreover, PSD projection can produce continuous kernel entries from discrete input scores, so the ablation does not isolate the removal of ties in the final training targets.

E.4 Regulariser families and weights

Comparison setup. Table 19 expands the anchor comparison using dog-trained CLIP ViT-B/16, Qwen3.5-397B judgments, 93 groups. We sweep four weights for each main regulariser family; the grids differ because feature penalties and logit divergences have different scales. The single-setting block includes removing the anchor, anchoring text features, PromptSRC, and combined penalties. The final block replaces the judgment target with DINOv2 features while retaining the listed anchor. Fidelity is always measured against the VLM judgments.

Feature anchoring. Reducing the feature-L2 weight from 1.0 to 0.001 raises fidelity from 0.286 to 0.709 and mAP from 22.29 to 46.83, but both remain below the default KL configuration (0.794 and 60.16). Feature L1 at weight 0.01 approaches the default fidelity and retrieval, while its zero-shot mean is lower (58.73 versus 62.28). Thus, the tested feature penalties offer a less favourable balance of alignment and retention. With strong L2 anchoring, VLM and DINOv2 targets both yield 0.286 fidelity; at weight 0.001, their scores separate to 0.709 and 0.452. The anchor can therefore obscure differences between supervision sources.

KL strength and text adaptation. With a trainable text tower, increasing the weight of the default KL anchor from 0.01 to 1.0 raises mean zero-shot accuracy from 60.47 to 62.43, while reducing fidelity from 0.801 to 0.739 and mAP from 60.82 to 51.88. The default weight 0.2 retains most of the alignment gain with a 62.28 zero-shot mean. Adapting the text tower improves fidelity over freezing it at every tested weight of the default KL anchor. The two KL directions differ by at most 0.002 in fidelity, although their zero-shot means differ by up to 0.45 points.

Table 19: **Regulariser families and weights.** Dog-trained CLIP ViT-B/16. Here, P^0 and P denote the frozen and adapted image–text prediction distributions, respectively. Blue/red indicates improvement/decline from the default KL anchor, scaled per metric across all rows. Bold marks the best score within each block.

Regulariser	Weight	Fidelity		Retrieval	Zero-shot	
		ρ_S	ρ_P	mAP	12 sets	IN-1k
Default: $\mathrm{KL}(P^0 \| P)$**, weight** 0.2**, text trainable (ASK)**		0.794	0.840	60.16	62.28	67.70
Untrained CLIP		0.281	0.280	22.06	61.81	68.29
L2 on features (KUEA's anchor)	0.001	**0.709**	**0.733**	**46.83**	61.33	67.61
	0.01	0.391	0.398	26.72	**61.95**	**68.60**
	0.1	0.298	0.298	22.76	61.89	68.41
	1.0 (published)	0.286	0.285	22.29	61.91	68.35
L1 on features	0.001	**0.795**	**0.855**	59.65	57.31	61.01
	0.01	0.792	0.849	**59.87**	58.73	63.18
	0.1	0.760	0.786	51.15	61.24	66.98
	1.0	0.360	0.361	23.82	**61.70**	**68.31**
$\mathrm{KL}(P^0 \| P)$, text frozen	0.01	**0.772**	**0.840**	**56.26**	60.17	63.85
	0.1	0.760	0.818	52.15	61.39	66.65
	0.2	0.752	0.799	49.71	61.35	67.16
	1.0	0.671	0.636	34.03	**61.78**	**68.27**
$\mathrm{KL}(P \| P^0)$, text trainable	0.01	**0.801**	**0.857**	**60.78**	60.02	64.45
	0.1	0.799	0.849	60.51	61.93	66.92
	0.2	0.794	0.840	60.09	62.20	67.62
	1.0	0.741	0.749	52.15	**62.41**	**68.94**
$\mathrm{KL}(P^0 \| P)$, text trainable (ASK)	0.01	**0.801**	**0.857**	**60.82**	60.47	65.01
	0.1	0.798	0.849	60.43	62.21	67.13
	0.2 (default)	0.794	0.840	60.16	62.28	67.70
	1.0	0.739	0.747	51.88	**62.43**	**68.86**
Single settings	none	0.796	0.854	**59.41**	56.67	59.82
	text L1, 0.5	**0.796**	**0.855**	59.30	56.70	59.85
	KL 0.2 + text L1 0.5	0.783	0.828	57.67	**62.47**	68.20
	PromptSRC	0.775	0.819	56.54	62.25	68.22
	L2 1.0 + KL 0.2	0.285	0.283	22.26	61.89	**68.36**
DINOv2 teacher	L2, 0.001	0.452	0.530	36.15	61.63	68.01
	L2, 0.01	0.330	0.350	25.14	61.90	**68.59**
	L2, 0.1	0.290	0.292	22.63	61.91	68.40
	L2, 1.0 (KUEA)	0.286	0.285	22.27	61.91	68.32
	KL, 0.2 (ours)	**0.539**	**0.678**	**51.99**	**62.07**	68.45

Table 20: **Frozen-feature heads versus LoRA.** All models use KL anchoring. (a) Criterion-specific Spearman fidelity. (b) Retrieval gains of breed-trained models. Blue/red indicates improvement/decline relative to LoRA, scaled per panel; bold marks row bests.

(a) Fidelity to the criterion's own teacher				(b) ΔmAP over untrained CLIP			
Criterion	Linear	MLP	**LoRA**	Retrieval on	Linear	MLP	**LoRA**
Breed	0.622	0.734	**0.794**	Unseen breeds	+13.9	+22.8	**+38.1**
Colour/lighting	0.491	0.673	**0.747**	Non-dog	+1.3	+1.9	**+4.9**
Pose/view	0.564	0.674	**0.748**	Textures	+1.3	+2.2	**+6.2**
Coat texture	0.699	0.826	**0.874**	Flowers	+2.6	+3.2	**+8.9**
Background	0.730	0.887	**0.919**	Cars	+3.7	+5.5	**+14.7**
Mean	0.621	0.759	**0.816**	Pets	+14.6	+22.0	**+29.0**

E.5 Frozen-feature heads versus encoder adaptation

Matched training protocol. Table 20(a) trains separate linear, MLP, and LoRA models for breed, colour and lighting; pose, viewpoint, and framing; coat texture; and background. All models use

Table 21: **Training similarity and objective under matched LoRA adaptation.** All adapted rows share PSD targets and the KL anchor. Fidelity compares the listed similarity and deployed cosine similarity with the same raw held-out judgments, without test-time fitting. Untrained CLIP uses the default cubic link in the listed-similarity columns. mAP is cosine retrieval on unseen breeds; ZS is the 12-benchmark mean (both %). Blue/red indicates improvement/decline relative to cubic MSE; bold marks column bests.

Training similarity	Formula	Loss	Fidelity				mAP	ZS
			Listed similarity		Cosine			
			ρ_S	ρ_P	ρ_S	ρ_P		
Untrained CLIP	—	–	0.281	0.280	0.281	0.256	22.06	61.81
Dot product	$\gamma x_i^\top x_j + c$	MSE	0.696	0.677	0.688	0.672	40.50	60.63
Cosine	$\hat{x}_i^\top \hat{x}_j$	MSE	0.756	0.811	0.756	**0.811**	58.35	**62.57**
Normalized cubic (**ASK**)	$\tilde{k}_{\gamma,c}(x_i, x_j)$	MSE	0.794	**0.840**	0.794	0.792	**60.16**	62.28
Normalized cubic	$\tilde{k}_{\gamma,c}(x_i, x_j)$	Ranking	**0.799**	0.768	**0.799**	0.725	54.71	62.23
Unnormalized cubic	$(\gamma x_i^\top x_j + c)^3$	MSE	0.749	0.805	0.750	0.762	54.19	62.54

CLIP ViT-B/16, the same 1,000 dog images and 93 groups, the corresponding PSD-projected Qwen targets, normalized-cubic MSE, and the default KL anchor of weight 0.2 over 1,000 class-name prompts. Each criterion uses its training-fitted link, fixed during adaptation and evaluation. Training uses 3,000 AdamW updates, learning rate $1e{-5}$, weight decay $1e{-4}$, 200 warm-up steps, and rank-32 text LoRA. Existing anchored LoRA checkpoints and the completed anchored breed heads are reused; the other eight heads follow the same head-training implementation. All results use the final checkpoint.

Adaptation capacity. The identity-initialized 512×512 linear map has 262,144 visual parameters. The residual MLP $x + W_2\,\mathrm{GELU}(W_1 x + b_1) + b_2$ has hidden width 2,304 and a zero-initialized second layer; its 2,362,112 visual parameters closely match LoRA's 2,359,296. All models also train 1,572,864 text parameters. The visual backbone is frozen for both heads. Table 5(c,d) summarizes these same anchored models, with the four cross-domain retrieval gains averaged in panel (d).

Evaluation. Panel (a) reports Spearman correlation with each criterion's raw judgments on 250 held-out images (31,125 pairs); no link is fitted on test targets. Panel (b) reports cosine-retrieval mAP gains over untrained CLIP for the same breed-trained models: 500 unseen-breed images, 200 non-dog ImageNet images, 400 Textures, 500 Flowers, 500 Cars, and 240 Pets. Each query is excluded from its gallery and relevant-image count. The default 93-group LoRA checkpoint is shared with Table 21.

Fidelity and transfer. Mean fidelity is 0.816 for LoRA, 0.621 for the linear head, and 0.759 for the MLP. Unseen-breed mAP gains over CLIP are 38.1, 13.9, and 22.8 points, respectively. These experiments use a common learning rate and anchor weight; they do not establish per-architecture optimal performance.

E.6 Kernel form

Matched LoRA comparison. All adapted rows use CLIP ViT-B/16, the same 1,000 dog images, 93 PSD-projected Qwen training groups, rank-32 adapters in both towers, 3,000 updates, AdamW at $1e{-5}$ with 200 warm-up steps, and the default KL anchor of weight 0.2 over 1,000 class-name prompts. The training similarity or discrepancy changes between rows. Positive γ and nonnegative c are fitted on the first 20 training groups using frozen features and 600 Adam steps at learning rate $5e{-2}$, then fixed during adaptation. Dot-product and unnormalized-cubic scales are fitted for their own forms; the normalized-cubic ranking row uses the same fixed link as the MSE row. The completed cosine and cubic MSE controls supply those reference rows.

Objectives and evaluation. MSE averages the full group matrix, including its diagonal. Ranking compares all ordered pairs of off-diagonal image-pair scores, applies a hinge margin of 0.05, and ignores teacher-score gaps at most 10^{-3}. Fidelity reports Spearman and Pearson correlations between final-step checkpoint similarities and raw judgments on 31,125 held-out pairs. We report

Table 22: **Sensitivity to training choices.** One factor changes from the default at a time. Blue/red indicates improvement/decline from the default, scaled per metric; bold marks the best score per column.

Factor	Variant	Fidelity		Retrieval	Zero-shot	
		ρ_S	ρ_P	mAP	12 sets	IN-1k
Default recipe (ASK)		0.794	0.840	60.16	62.28	67.70
Kernel loss default: Frobenius	Spectral norm	**0.803**	0.856	60.75	61.60	66.47
	Top-4 eigenvalues	0.801	0.854	60.79	61.85	66.21
Pair budget default: 93 groups	31 groups (15,376 unique pairs)	0.781	0.822	58.88	62.01	67.71
	62 groups (30,174 unique pairs)	0.790	0.836	60.07	62.28	67.68
Anchor weight λ default: 0.2	$\lambda = 0.01$	0.801	**0.857**	**60.82**	60.47	65.01
	$\lambda = 0.1$	0.798	0.849	60.43	62.21	67.13
	$\lambda = 1$	0.739	0.747	51.88	**62.43**	**68.86**
Anchor text default: class prompts	1,000 Conceptual Captions	0.793	0.839	58.68	61.16	66.97
	50/50 prompts and captions	0.795	0.839	60.05	62.28	67.70
LoRA rank default: 32, α=2r	Rank 8	0.784	0.824	60.66	62.28	67.65
	Rank 64	0.797	0.848	58.44	61.44	67.26
	Rank 128	0.792	0.846	53.87	59.37	62.61

two readouts on the same embeddings: each row's listed similarity, with link parameters fixed from training, and cosine similarity for every row. Untrained CLIP uses the default normalized-cubic link for the first readout. No link is fitted on test targets. Retrieval uses cosine similarity on 500 images from 25 unseen breeds, with each query excluded. Zero-shot evaluation averages the same 12 benchmarks as Table 6. The experiments use a shared learning rate and loss weight; they do not constitute per-objective hyperparameter optimization.

Training objective and deployed similarity. With a positive offset and varying embedding norms, the normalized cubic kernel need not order candidates as cosine does. For cubic MSE, cosine fidelity is 0.794 Spearman and 0.792 Pearson, compared with 0.794 and 0.840 for the cubic kernel. Direct cosine MSE yields 0.756 Spearman and 0.811 Pearson under cosine, with 58.35 mAP versus 60.16 for cubic MSE. Thus cubic training improves cosine rank agreement and retrieval mAP in this comparison, while direct cosine training has higher cosine Pearson and a slightly higher zero-shot mean. The cubic-kernel Pearson score does not describe agreement under deployed cosine similarity.

E.7 Sensitivity to training choices

Protocol. Table 22 changes one factor at a time from the default dog-trained ViT-B/16 configuration with PSD-projected Qwen3.5-397B judgments. Unless varied, settings are 93 scoring groups, rank-32 LoRA, 3,000 steps, and default KL anchor weight 0.2 with class-name prompts. The 31- and 62-group budgets are nested subsets containing 15,376 and 30,174 distinct pairs. Evaluation uses the full 250-image fidelity set for both budgets.

Budget and rank. Reducing the budget from 93 to 31 groups lowers fidelity from 0.794 to 0.781 and mAP from 60.16 to 58.88. With 62 groups, these recover to 0.790 and 60.07. Rank 8 retains a 62.28 zero-shot mean and achieves 60.66 mAP, despite lower fidelity (0.784). Larger ranks do not consistently help: rank 128 yields 53.87 mAP and a 59.37 zero-shot mean. The default is therefore a practical balance among the tested settings, rather than the best choice for every metric.

Loss and anchor. The spectral-norm loss raises fidelity to 0.803 but lowers the zero-shot mean to 61.60. Replacing class-name anchors with Conceptual Captions (Sharma et al., 2018) changes fidelity little (0.793) while reducing mAP to 58.68 and zero-shot accuracy to 61.16; mixing prompts and captions stays closer to the default. The anchor-weight sweep reproduces the alignment–retention trade-off discussed above.

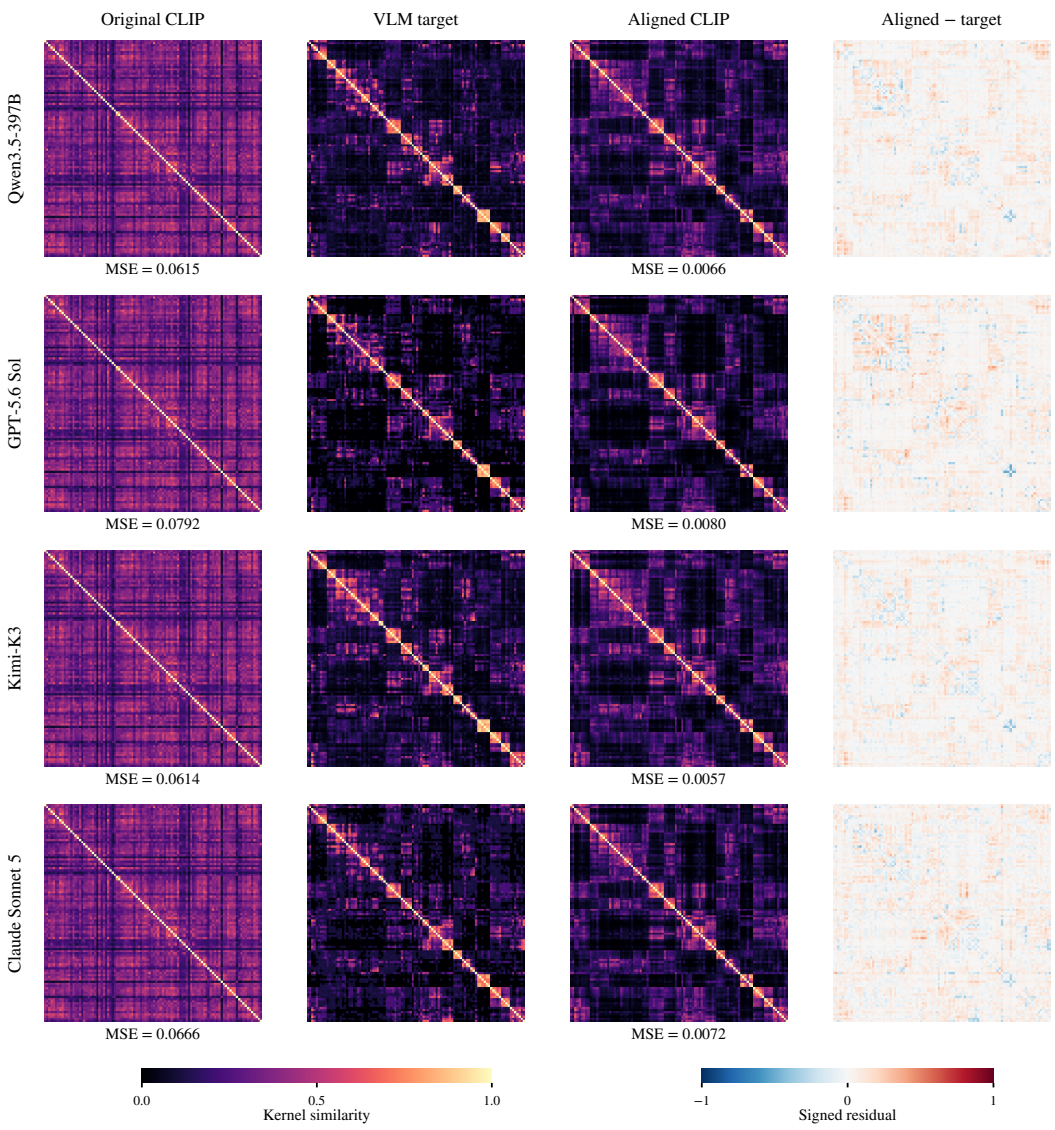


Figure 8: **Kernel alignment across VLM judges.** Original CLIP, teacher target, aligned CLIP, and signed residual on 100 held-out dog images. Each row uses its judge's saved link; residuals are aligned minus target.

F Additional visual results

F.1 Kernel geometry

These figures extend Figure 3 with judge comparisons, supervision baselines, and separately aligned models for all five criteria. All targets are PSD-projected. Encoder kernels use saved polynomial links on raw features followed by diagonal normalization, without test-time link fitting. Images have a fixed ordering by breed and image path. Kernel heatmaps share a $[0, 1]$ colour scale and signed residuals share $[-1, 1]$. MSE counts each unordered off-diagonal pair once; it is measured against the projected target, unlike fidelity to raw judgments.

Across judges. Figure 8 uses the same 100 held-out images and a matched 31-group training budget for all four judges. Alignment recovers the major structures of each target and reduces MSE relative to original CLIP. Target structure and scale vary by judge, so errors are best compared before and after alignment within each row.

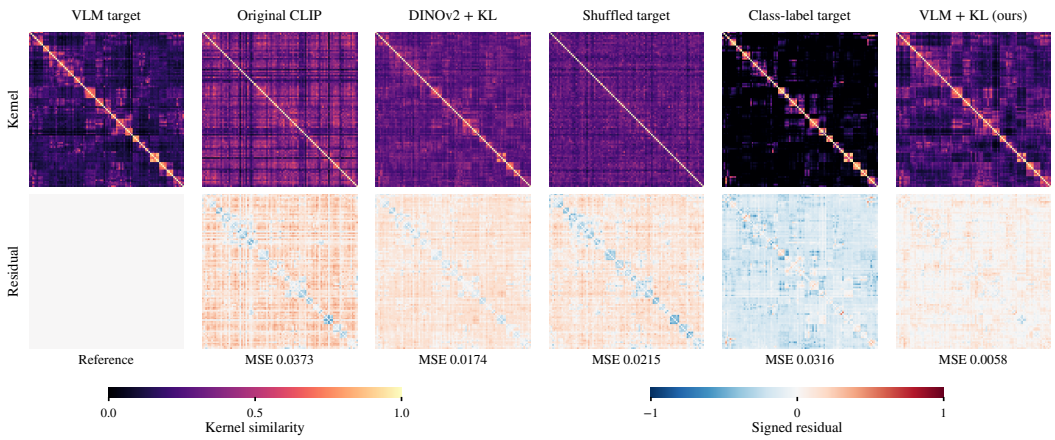


Figure 9: **Kernel geometry under alternative supervision.** Top: target and encoder kernels on 100 held-out dog images. Bottom: signed residuals relative to the Qwen breed target. The VLM-aligned model uses the 93-group protocol.

Across supervision sources. Figure 9 compares the stored supervision baselines on the same 100 images. The target is the corresponding submatrix of Qwen's 250-image projected breed kernel, and **ASK** uses the 93-group reference model. Each model uses its own saved link; original CLIP uses the VLM-alignment link. This comparison complements the matched-judge figure and is not a 31-group budget comparison. **ASK** has the smallest residual MSE among the displayed models; class-label supervision produces sharper blocks but does not reproduce the teacher's graded similarities.

Across criteria. Figure 10 compares each criterion's target with its own aligned model on 250 held-out images. It extends the breed example in Figure 3(a,b) to learned kernels and residuals for every criterion. The shared breed ordering makes breed and coat-texture blocks prominent; other criteria can express similarity through different patterns without forming the same blocks.

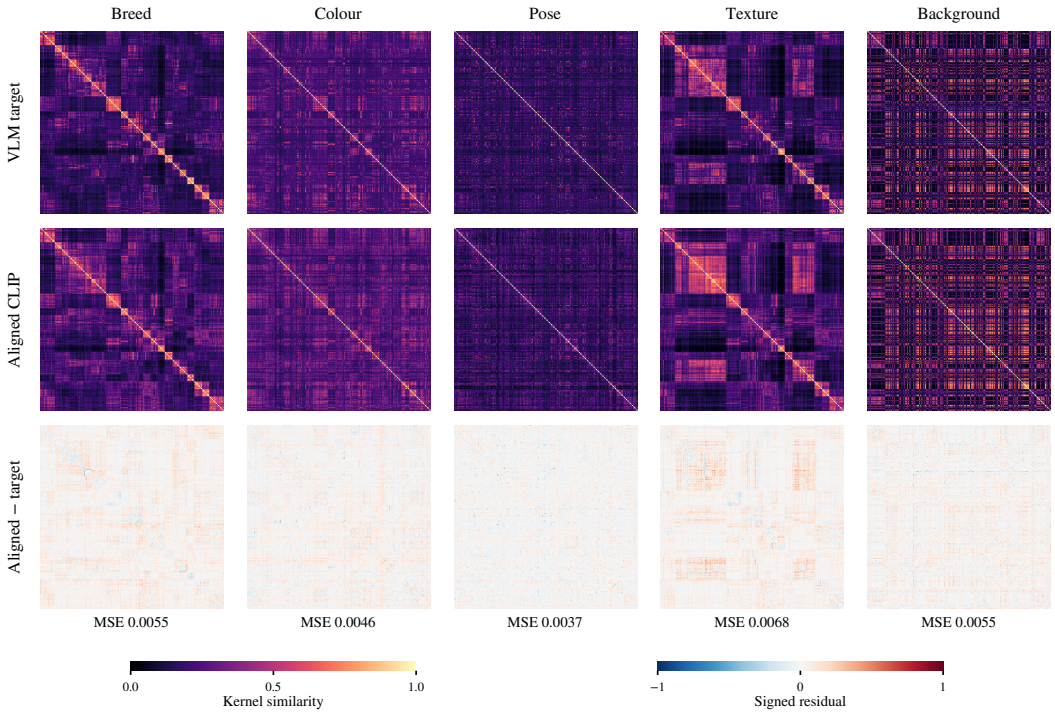


Figure 10: **Target and learned geometry for five criteria.** Rows show the VLM target, criterion-aligned CLIP, and signed residual on the same 250 held-out dog images. Columns retain the same image ordering across criteria.



Figure 11: **Dog retrieval under different criteria.** Each row shows the query (teal border) and its top three neighbours. Green/red borders indicate matching/different breeds only in the class-retrieval rows.

F.2 Criterion-steered retrieval examples

Figures 11 and 12 show two queries per domain, including background-based retrieval. Each row repeats the query and shows the exact top three cosine neighbours from the other 499 images. The queries were sampled from the existing example gallery, rather than the full dataset; these illustrations do not estimate typical retrieval performance. Teal borders mark queries. Green and red borders indicate matching and different class labels in the CLIP, DINOv2+KL, and breed/car-model rows. Other criteria have no correctness borders because class labels do not assess colour/lighting, pose/view, or background relevance.

The examples also show overlap between criteria. Several car neighbours recur across model, colour/lighting, and pose/view rows, while the dog colour and pose rows can retrieve different breeds. Background alignment can change the retrieved scene without preserving object identity. These comparisons illustrate the rankings produced by each encoder, including imperfect matches, rather than implying that every retrieved image satisfies its criterion.

F.3 Top retrievals across judge-trained models

Figure 13 compares original CLIP, dog-trained DINOv2+KL, and ASK encoders trained with Qwen3.5-397B, GPT-5.6 Sol, Kimi-K3, Claude Sonnet 5, or Gemma 4 12B. The five ASK models use the matched 31-group protocol. Each encoder ranks the same pool within Cars (500 images), Flowers (500), or Oxford-IIIT Pets (240) by cosine similarity, excluding the query. We select one readable query per domain with high disagreement among models' top-two results, without selecting for class accuracy or replacing neighbours. These selected examples illustrate different rankings rather than typical performance. All adapted encoders were trained on dogs, so the panels evaluate their behaviour on other image domains.

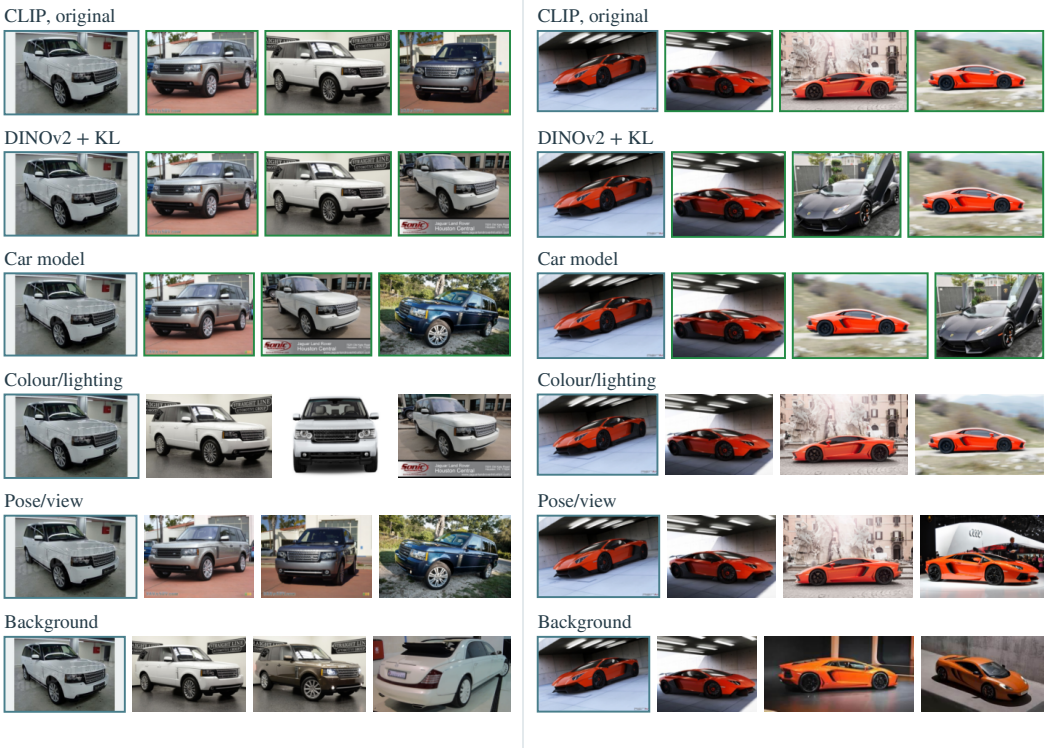


Figure 12: **Car retrieval under different criteria.** Each row shows the query (teal border) and its top three neighbours. Green/red borders indicate matching/different car classes only in the class-retrieval rows.

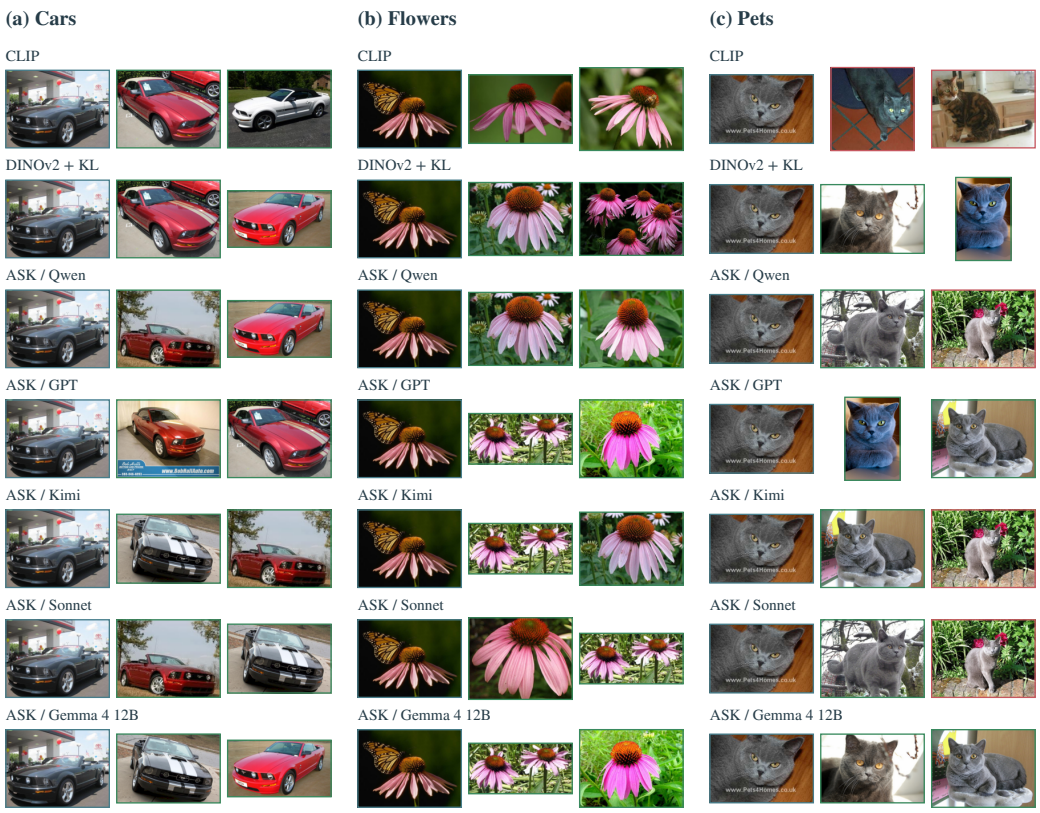


Figure 13: **Different models, different nearest neighbours.** Each row shows the query (teal) and the model's top two cosine retrievals. ASK rows identify the training judge, not direct retrieval by the VLM. Green/red borders indicate matching/different dataset class labels.

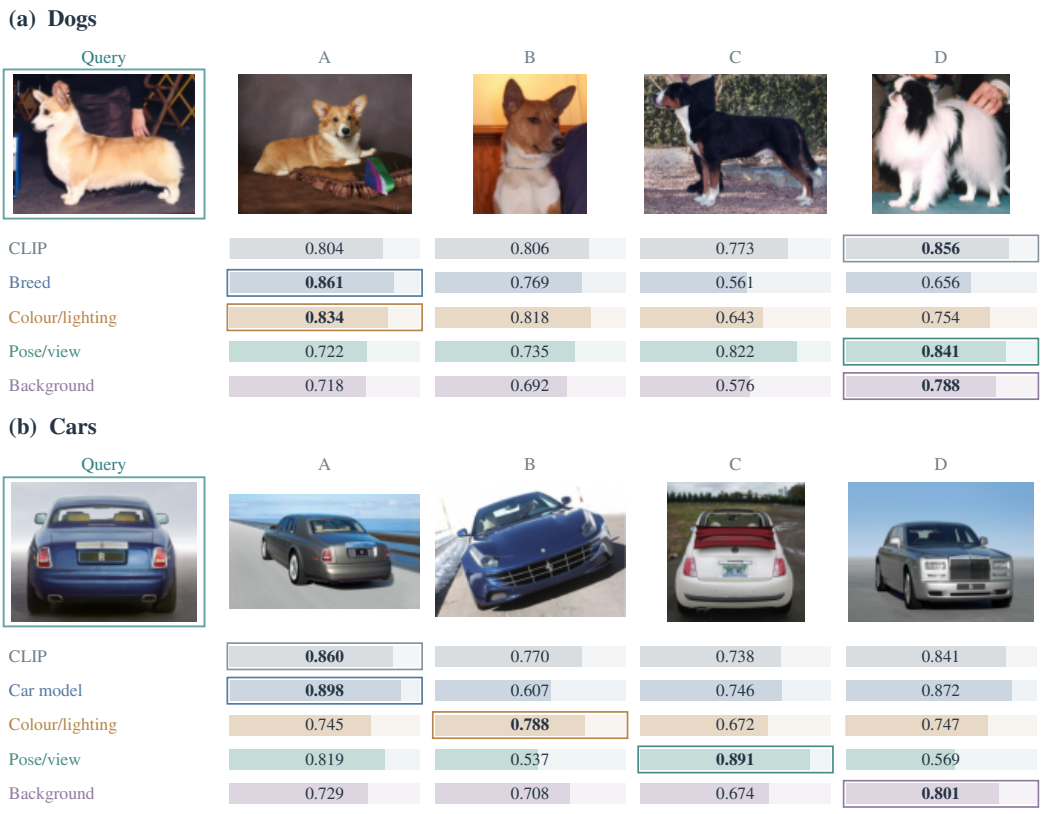


Figure 14: **Same candidates, different notions of similarity.** Scores are query–candidate cosine similarities. Outlined cells mark the highest score among the four displayed candidates in each row; colours identify criteria.

F.4 Comparing fixed candidates across criteria

Figure 14 holds the candidate photographs fixed while changing the similarity criterion. For cars, the highest-scoring candidate changes from A under model identity to B under colour and lighting, C under pose/view, and D under background. For dogs, breed and colour/lighting favour A, while pose and background favour D. CLIP is shown as a reference. These candidates were selected from highly ranked criterion retrievals to illustrate changes in preference, rather than to estimate retrieval accuracy. Scores are unadjusted cosine similarities; comparisons should be made within each row, since different encoders need not share a calibrated score scale.

G ELICITATION PROMPTS

Each judge receives two images and one of the following prompts. We reproduce the scoring instructions from the experiment code, with the prompt choices checked against the saved run configurations. The requested response is a single digit from 1 to 9; score extraction is described in the experimental setup. The dog primary prompt is shared across judges.

G.1 DOG SIMILARITY CRITERIA

The figure labels are abbreviated descriptions of these instructions. In particular, the primary criterion includes breed, build, colour, and markings; the colour criterion covers the full image palette and lighting, rather than only the coat. Pose also includes viewpoint and framing.

Primary similarity (breed)

```
Both images show dogs.  Rate how similar they are on an integer scale
1-9, judging breed, build, coat colour and markings.
1 = clearly different breeds, nothing in common
3 = different breeds, loosely similar type
5 = related breeds or similar size/coat
7 = same or nearly the same breed
9 = same breed, near-identical individual
Spread your answers across the whole 1-9 range.  Reply with only the
single digit.
```

Colour and lighting

```
Rate how similar these two images are on an integer scale 1-9,
judging ONLY colour palette, lighting and overall tone.  Ignore what
the subject is.
1 = completely different palettes 5 = loosely similar tones 9 =
near-identical colour and lighting
Spread your answers across the whole 1-9 range.  Reply with only the
single digit.
```

Pose/view

```
Rate how similar these two images are on an integer scale 1-9,
judging ONLY the subject's pose, viewing angle and framing.  Ignore
breed, colour and setting.
1 = completely different pose and viewpoint 5 = loosely similar
framing 9 = near-identical pose and camera angle
Spread your answers across the whole 1-9 range.  Reply with only the
single digit.
```

Coat texture

```
Rate how similar these two images are on an integer scale 1-9,
judging ONLY coat texture and surface detail -- fur length, curliness,
wiriness, smoothness.  Ignore breed, colour, pose and setting.
1 = completely different coat texture 5 = loosely similar texture 9 =
near-identical coat texture
Spread your answers across the whole 1-9 range.  Reply with only the
single digit.
```

Background

```
Rate how similar these two images are on an integer scale 1-9,
judging ONLY the background and setting -- indoors/outdoors, grass,
snow, studio, street.  Ignore the animal entirely.
1 = completely different setting 5 = loosely similar setting 9 =
near-identical setting
Spread your answers across the whole 1-9 range.  Reply with only the
single digit.
```

G.2 PRIMARY CRITERIA FOR OTHER DOMAINS

The five-domain experiments use the dog primary prompt above and the following four domain-specific prompts.

Cars

```
Both images show cars.  Rate how similar they are on an integer scale
1-9, judging body style, proportions, model and era.
1 = clearly different vehicles 3 = different models, similar class
5 = related models or same segment 7 = same or nearly the same model
9 = same model, near-identical vehicle
Spread your answers across the whole 1-9 range.  Reply with only the
single digit.
```

Flowers

```
Both images show flowers.  Rate how similar they are on an integer
scale 1-9, judging petal shape, arrangement, colour and overall form.
1 = clearly different species, nothing in common 3 = loosely similar
form
5 = related species or similar structure 7 = same or nearly the same
species
9 = same species, near-identical bloom
Spread your answers across the whole 1-9 range.  Reply with only the
single digit.
```

Pets

```
Both images show pets.  Rate how similar they are on an integer scale
1-9, judging breed, build, coat colour and markings.
1 = clearly different breeds, nothing in common 3 = different breeds,
loosely similar type
5 = related breeds or similar size and coat 7 = same or nearly the
same breed
9 = same breed, near-identical individual
Spread your answers across the whole 1-9 range.  Reply with only the
single digit.
```

Textures

```
Both images show textures or patterns.  Rate how similar they are on
an integer scale 1-9, judging the pattern structure, regularity,
scale and surface character.
1 = completely different patterns 3 = loosely related structure 5 =
similar family of pattern
7 = same kind of pattern 9 = near-identical pattern
Spread your answers across the whole 1-9 range.  Reply with only the
single digit.
```

G.3 CAR-SPECIFIC CRITERION PROMPTS

The car colour prompt is identical to the colour prompt above. The pose prompt replaces "breed" with "model", and the background prompt uses vehicle-specific wording. Their complete instructions are given below.

Car pose/view

```
Rate how similar these two images are on an integer scale 1-9,
judging ONLY the subject's pose, viewing angle and framing.  Ignore
model, colour and setting.
1 = completely different pose and viewpoint 5 = loosely similar
framing 9 = near-identical pose and camera angle
Spread your answers across the whole 1-9 range.  Reply with only the
single digit.
```

Car background

```
Rate how similar these two images are on an integer scale 1-9,
judging ONLY the background and setting -- indoors/outdoors, studio,
street, showroom, landscape.  Ignore the vehicle entirely.
1 = completely different setting 5 = loosely similar setting 9 =
near-identical setting
Spread your answers across the whole 1-9 range.  Reply with only the
single digit.
```

G.4 CUB ATTRIBUTE SCORING PROMPTS

For the CUB experiment in Appendix D.5, Qwen3.8-27B receives two training images and one of the following criterion-specific prompts.

Wing colour

```
Rate how similar these two bird images are on an integer scale 1-9,
judging ONLY the intrinsic colours of the visible wing feathers.
Compare the main wing colours and their approximate proportions.
Ignore bird species, bill shape, breast markings, pose, background,
and differences in illumination.  If a wing is partly hidden, judge
only the visible evidence rather than inferring its colour from the
species.
1 = completely different wing colours
5 = some shared wing colours but substantial differences
9 = nearly identical wing colours and proportions
Use the full 1-9 range.  Reply with only the single digit.
```

Bill shape

```
Rate how similar these two bird images are on an integer scale 1-9,
judging ONLY the shape of the bill or beak.  Compare its outline,
curvature, taper, thickness relative to length, and tip shape.
Ignore bird species, absolute image scale, bill colour, feather
colours and patterns, pose, lighting, and background.  Judge the
visible bill rather than inferring its shape from the species.
1 = completely different bill shapes
5 = some shared shape features but substantial differences
9 = nearly identical bill shapes and proportions
Use the full 1-9 range.  Reply with only the single digit.
```

Breast pattern

```
Rate how similar these two bird images are on an integer scale 1-9,
judging ONLY the visible breast-feather pattern:  solid or unmarked,
spotted, striped or streaked, or divided into multiple colour patches.
Compare the kind, density, and arrangement of markings, ignoring the
actual hues.  Ignore bird species, wing appearance, bill shape, pose,
lighting, and background.  Judge the visible breast rather than
inferring its pattern from the species.
1 = completely different breast patterns
5 = related patterns with noticeable differences
9 = nearly identical breast patterns and marking arrangement
Use the full 1-9 range.  Reply with only the single digit.
```