

CAPT: Cross-modal Anchored Prompt Tuning for Universal Cross-Domain Retrieval Supplementary Material

Chanmin Park* Minyoung Kang*
POSCO DX

{prov169, mykang}@poscodx.com

* These authors contributed equally.

This supplement gives additional details that did not fit in the main paper. We first derive DSIP and PACE, including the convex-combination view of the generated prompt and the gradients of the two alignment losses (S1). We then analyze the feature space with t-SNE and DBSCAN (S4), inspect attention across domains (S5.1), and report additional ablations (S6). Unless noted otherwise, all numbers are from our own runs on the sketch holdout and are reported as mAP@200 / Prec@200.

1. Derivations

Table 1 collects the notation used throughout.

Table 1. Notation.

symbol	meaning
d_v, d_t	visual (768) / shared-embedding (512) width
C, N, L	seen classes / patch tokens / encoder layers
r, P_q	bottleneck width (64) / prompt tokens (8)
$A, \hat{A}$	frozen text anchors / projected anchors
$\hat{A}^\ell$	image-conditioned anchors at layer ℓ
Q_0, P^ℓ	static queries / generated prompt at layer ℓ
z_i	final ℓ_2 -normalized image feature of x_i
$\mu_c, \nu_{c,e}$	class / same-domain class prototype
s_i, g_i, w_i	confidence / domain gap / routing weight
$\sigma(\gamma)$	DSIP conditioning gate (≈ 0.05 init)

1.1. DSIP read-out and the language-manifold property

Let the frozen CLIP [6] text-class anchors be $A \in \mathbb{R}^{C \times d_t}$, projected to the bottleneck as $\hat{A} = AW_a \in \mathbb{R}^{C \times r}$. At layer ℓ the anchors are image-conditioned by a gated single-head attention over the projected

patches $\bar{H}^\ell = h_{1:N}^\ell W_h$:

$$\hat{A}^\ell = \tilde{A} + \underbrace{\sigma(\gamma) \operatorname{softmax}\left(\frac{\tilde{A}(\bar{H}^\ell)^\top}{\sqrt{r}}\right)}_{\text{image residual } \Delta^\ell} \bar{H}^\ell, \quad (1)$$

with a scalar gate $\sigma(\gamma)$ initialized to $\sigma(-3) \approx 0.05$. The P_q prompt tokens are then read out by multi-head attention [11] from static queries Q_0 and up-projected:

$$P^\ell = \text{MHA}_\ell(Q_0, \hat{A}^\ell, \hat{A}^\ell) W_u^\ell. \quad (2)$$

Proposition 1 (Language-manifold property). Before the up-projection W_u^ℓ , every prompt token produced by Eq. 2 lies in the convex hull $\operatorname{conv}\{\hat{a}_c^\ell\}$ of the image-conditioned anchors, and hence within an $\mathcal{O}(\sigma(\gamma))$ neighborhood of the convex hull of the frozen text anchors $\{\tilde{a}_c\}$.

Proof. A single attention head maps a static query q to $\sum_c \alpha_c \hat{a}_c^\ell$ with weights $\alpha_c = \operatorname{softmax}_c(q^\top \hat{a}_c^\ell / \sqrt{r})$, so $\alpha_c \geq 0$ and $\sum_c \alpha_c = 1$ [11]. A nonnegative, sum-to-one weighting is by definition a convex combination, which lies in the convex hull of its points [1, §2.1]; averaging the heads (again convex weights) preserves this membership. Substituting Eq. 1,

$$\hat{a}_c^\ell = \tilde{a}_c + \sigma(\gamma) \Delta_c^\ell, \quad \|\sigma(\gamma) \Delta_c^\ell\| \ll \|\tilde{a}_c\|, \quad (3)$$

so $\operatorname{conv}\{\hat{a}_c^\ell\}$ is an $\mathcal{O}(\sigma(\gamma))$ perturbation of $\operatorname{conv}\{\tilde{a}_c\}$ (measured $\sigma(\gamma)=0.067$).

What the convex hull means here. The convex hull of the anchors is the set of all weighted averages $\sum_c \alpha_c \hat{a}_c^\ell$ with $\alpha_c \geq 0$ and $\sum_c \alpha_c = 1$: every point reachable by blending class meanings, and nothing else. Because the read-out weights are a softmax they satisfy exactly these two conditions, so the prompt can interpolate between classes in any proportion but can never extrapolate: it cannot scale an anchor up, subtract one, or add

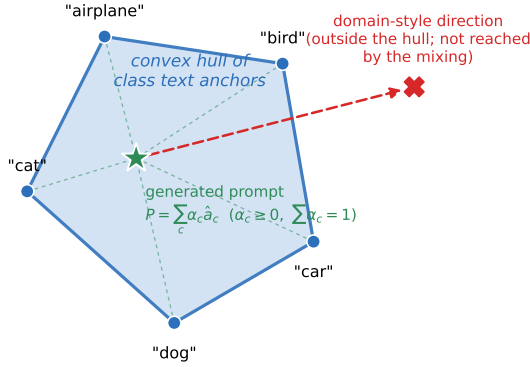


Figure 1. Language-manifold property (Prop. 1). The generated prompt is a convex combination of the class text anchors, a weighted average with nonnegative weights summing to one, so it stays inside their convex hull; a domain-style direction lies outside the hull, so the mixing alone does not produce it. Tab. 2 verifies this numerically at every layer: the prompts’ hull residual is 0.0000 while same-space controls sit at 1.0–1.4.

a direction that is not itself a blend of class anchors. A domain-style direction (“make it a sketch”) is not a weighted average of class meanings, so the read-out cannot emit it: the constraint Fig. 1 draws and Tab. 2 measures. The measured weights make this concrete: the read-out attention is uniform to three decimals (entropy = $\log 300$ at all three seeds, $\max \alpha = 1/300$), so the reported model sits at the center of the hull, an equal blend of the whole bank, shifted per image only by the gated conditioning term ($\sigma(\gamma)=0.067$).

Interpretation. The prompt is therefore pinned to the span of the class text anchors (Fig. 1): it can only be a mildly image-shifted mixture of class meanings, never a free source-domain direction. This is what keeps DSIP on CLIP’s semantic manifold; the property constrains the geometry of the read-out (the anchor mixing), not the content of any individual anchor. Empirically the property survives the learned projection: the post- W_u projection residual of the injected prompts off the conditioned anchor hull is 0.0000 at every one of the 12 layers (span residual ≈ 0.0003), versus ≈ 1.00 for CLIP patch tokens and 1.34–1.41 for norm-matched random vectors measured in the same space (per-layer values in Tab. 2); relative to the hull of the unconditioned frozen anchors the deviation is 0.07–0.33, the $\mathcal{O}(\sigma(\gamma))$ neighborhood of Prop. 1. The constraint is not specific to this model: across the other DomainNet holdouts and the Sketchy and TU-Berlin models the same residual stays below 2×10^{-4} , against the same 1.0–1.4 controls.

Table 2. Prop. 1, measured per layer (reported model, 8 prompt tokens $\times$ 450 images). Relative projection residual $\|p - \text{proj}(p)\|/\|p\|$ of the injected (post- W_u) prompts against the conditioned anchor hull, with the frozen-anchor hull and two same-space controls. The residual is 0.0000 at every layer and is domain-invariant (unseen sketch, gallery real, and training painting groups all $\leq 2 \times 10^{-7}$). The mixture itself is maximally spread: read-out attention entropy equals $\log 300$ to three decimal places at all three seeds, i.e. the prompt is a uniform convex combination of the entire bank ($\max \alpha = 1/300$), shifted by the gated image term ($\sigma(\gamma)=0.067$).

layer	prompt vs. cond. hull	vs. frozen hull	patch (ctrl.)	random (ctrl.)
1	0.0000	0.156	1.000	1.413
2	0.0000	0.069	1.000	1.408
3	0.0000	0.132	1.002	1.398
4	0.0000	0.143	1.002	1.384
5	0.0000	0.168	1.004	1.354
6	0.0000	0.132	1.008	1.373
7	0.0000	0.109	1.005	1.343
8	0.0000	0.100	1.000	1.369
9	0.0000	0.127	1.001	1.360
10	0.0000	0.188	0.998	1.371
11	0.0000	0.326	0.998	1.371
12	0.0000	0.246	1.000	1.367

1.2. PACE: routing and gradient complementarity

PACE mixes a contrastive “spread” (CDCM, a supervised InfoNCE [3, 10]) and a prototype “tighten” (CDPA, a prototype pull [4]) per sample. With class prototype μ_{y_i} and same-domain prototype ν_{y_i, e_i} , define confidence $s_i = z_i^\top \mu_{y_i}$ and the residual domain gap $g_i = z_i^\top \nu_{y_i, e_i} - z_i^\top \mu_{y_i}$; standardizing within the batch and routing by their difference gives a parameter-free per-sample weight

$$w_i = \text{sg}[\sigma(\tilde{s}_i - \tilde{g}_i)], \quad (4)$$

$$\mathcal{L}_{\text{PACE}}^i = (1 - w_i) \mathcal{L}_{\text{CDCM}}^i + w_i \mathcal{L}_{\text{CDPA}}^i.$$

A confident and already-aligned sample ($\tilde{s}_i$ high, $\tilde{g}_i$ low) gets $w_i \rightarrow 1$ and is tightened by CDPA. A domain-bound sample ($\tilde{g}_i$ high) gets $w_i \rightarrow 0$ and is pushed by CDCM. The same-domain prototype the router needs exists for $\sim 91\%$ of training samples; the remainder falls back to the confidence-only rule, so the gap term operates on the vast majority of the stream. Replacing the flat class prototype μ_{y_i} by a domain-balanced average changes each of the 18 sketch-holdout cells (six metrics $\times$ three seeds) by at most 0.0007, so we keep the simpler flat mean.

Why the two losses are complementary. For a unit-norm z_i , $\text{CDPA}=1 - z_i^\top \mu_{y_i}$ has gradient $\nabla_{z_i} \mathcal{L}_{\text{CDPA}}^i = -\mu_{y_i}$, a pure pull toward the class centroid (no negatives). CDCM is a multi-positive InfoNCE with hard-negative mining [7]; with temperature τ and softmax

Algorithm 1 CAPT training step

Require: batch $\{(x_i, y_i, e_i)\}$; frozen anchors A ; bank $\mathcal{B}$

- 1: $z_i \leftarrow \text{DSIP}(x_i)$ ▷ Eq. 1–2, all layers
- 2: $z_i \leftarrow \text{norm}(z_i + \text{MLP}(z_i))$ ▷ refinement head
- 3: $\mathcal{L}_{\text{ITC}} \leftarrow \text{CE}(z_i, u_{y_i})$ ▷ text targets u_c
- 4: enqueue (z_i, y_i, e_i) into $\mathcal{B}$
- 5: for each i do
- 6: $\mu_{y_i}, \nu_{y_i, e_i} \leftarrow$ prototypes from $\mathcal{B}$
- 7: $w_i \leftarrow \text{sg}[\sigma(\tilde{s}_i - \tilde{g}_i)]$
- 8: $\mathcal{L}_{\text{PACE}}^i \leftarrow (1 - w_i)\text{CDCM}_i + w_i\text{CDPA}_i$
- 9: end for
- 10: $\mathcal{L} \leftarrow \mathcal{L}_{\text{ITC}} + \lambda \text{mean}_i \mathcal{L}_{\text{PACE}}^i$; update trainable params

weights $p_a = \text{softmax}_a(z_i^\top z_a / \tau)$,

$$\nabla_{z_i} \mathcal{L}_{\text{CDCM}}^i = \frac{1}{\tau} \left(\underbrace{\sum_a p_a z_a}_{\text{push}} - \underbrace{\frac{1}{|\mathcal{P}|} \sum_{p \in \mathcal{P}} z_p}_{\text{pull}} \right). \quad (5)$$

CDPA alone would therefore collapse class centroids; CDCM (and the image-text loss below) supply the separating push, and PACE interpolates the two gradients per sample through w_i .

Full objective. The retrieval feature is also aligned to CLIP’s space by a standard CoOp-style [12] image-text contrastive loss $\mathcal{L}_{\text{ITC}}$ against learnable-context class text targets, giving $\mathcal{L} = \mathcal{L}_{\text{ITC}} + \lambda \text{mean}_i \mathcal{L}_{\text{PACE}}^i$ ($\lambda=0.5$). The text encoder stays frozen and is cached once, so it adds no test-time cost.

1.3. Training step

Algorithm 1 summarizes one optimization step; only the DSIP generator, the small text context, and the refinement head receive gradients.

2. Cross-Dataset Generalization

The main paper reports the standard in-benchmark results. Here we further test whether a DomainNet-trained model can transfer directly to new datasets. We train only on DomainNet and evaluate zero-shot on four unseen benchmarks (Sketchy, TU-Berlin, PACS, and Office-Home), with no retraining. For every method, we rebuild the class text for the new label set and run the same retrieval protocol over the full test gallery. CAPT reads its prompt from the new class anchors. The frozen CLIP backbone serves as the lower bound.

Correction to the submitted version. The submitted table gave ProS 0.1242 and DePro 0.3184 on Sketchy. Two faults produced those numbers. First, the baselines were transferred by loading their DomainNet checkpoints and dropping class-count-specific tensors, which leaves them without a usable class head, a crippled configuration, as one reviewer correctly identified.

Second, a default evaluation flag truncated the ProS gallery to 750 queries against 750 gallery images, making the task close to self-retrieval; its TU-Berlin entry returned identical mAP and precision, the signature of that degeneracy. We therefore rebuild the class head for ProS and DePro on the target label set and re-evaluate every method on the full gallery. Table 3 reports the corrected result, which is less favorable to us.

Table 3. Zero-shot cross-dataset transfer (DomainNet-trained, no fine-tuning, full gallery). Retrieval is sketch→photo, except Office-Home (†: clipart→real, as it has no sketch domain). All rows use the sketch holdout so that every method is read from a matched model. Bold = best per dataset. Sketchy and TU-Berlin use the fair class-head rebuild; for ProS we give both the transferred head and the rebuilt one. PACS and Office-Home admit no such rebuild in our harness, so those two blocks remain checkpoint-transfer only: they are reported for completeness and we draw no comparison from them.

dataset	method	mAP@200	Prec@200
DomainNet→Sketchy	frozen CLIP	0.2178	0.2354
	ProS (transferred head)	0.5854	0.5438
	ProS (rebuilt head)	0.4902	0.4083
	DePro	0.7072	0.6740
	CAPT (ours)	0.6941	0.6506
DomainNet→TU-Berlin	frozen CLIP	0.4374	0.3857
	ProS (transferred head)	0.6857	0.6682
	ProS (rebuilt head)	n/a	n/a
	DePro	0.6560	0.7258
	CAPT (ours)	0.7546	0.7390
DomainNet→PACS	frozen CLIP	0.6729	0.5514
	ProS	0.3391	0.1605
	DePro	0.5384	0.4731
	CAPT (ours)	0.8464	0.7908
DomainNet→Office-Home†	frozen CLIP	0.3334	0.1437
	ProS	0.2781	0.0891
	DePro	0.3630	0.1663
	CAPT (ours)	0.5749	0.2529

Once the baselines are given a fair class head and the full gallery, the result splits: DePro transfers better to Sketchy (0.7072 against our 0.6941), and CAPT transfers better to TU-Berlin (0.7546 against 0.6560, a margin seven times larger). We therefore withdraw the claim that CAPT dominates cross-dataset transfer, and we withdraw the submitted DePro figure of 0.3184, which understated a competitor. Cross-dataset transfer is benchmark-dependent; our claim is the in-benchmark UCDR setting of the main paper. What survives is narrower and still worth stating: because the prompt is read from class text anchors, a new dataset needs only its class names, so CAPT transfers without retraining at all, which is why it clears the frozen backbone by 0.18–0.48 mAP@200 on every dataset here.

Many Sketchy and TU-Berlin classes also appear in DomainNet, so the result above mixes new domains

with familiar classes. To test entirely new classes, we retrieve on the 69 TU-Berlin classes absent from DomainNet (Fig. 2), zero-shot. CAPT still retrieves well, e.g. armchair (10/10) and blimp (8/10), while rarer or shape-ambiguous classes (e.g. loudspeaker) fail; here neither the dataset nor the classes were seen in training.

3. Sensitivity to Training Epochs

Our main results fix one training epoch (epoch 4) for all five domains. Figure 3 tracks every metric over a dedicated 8-epoch run on the sketch and infograph holdouts under the constant learning rate we use throughout (no schedule); absolute values sit within the seed-level noise (± 0.002) of the main table. The two protocols the benchmark stresses, Unseen and Mixed, reach their best value at epoch 4 on sketch (Unseen 0.6931, Mixed 0.6136 mAP) and within 0.002 of the peak on infograph (its Unseen/Mixed top out one epoch earlier at 0.6210/0.5455 and read 0.6186/0.5402 at epoch 4). Past epoch 4 neither metric improves on either domain; both drift down by 0.005 to 0.011 through epoch 8. Sketch UdCDR slightly increases from 0.7924 at epoch 4 to 0.8012 at epoch 6, but this gain is within run-to-run noise and does not recur on infograph. We therefore stop every domain at the same epoch instead of tuning the epoch per domain or per metric. Epoch 4 is the last point before the generalized metrics start to decrease, and with a constant learning rate there is no late decay stage that would benefit from longer training.

A stronger backbone (ViT-B/16). On ViT-B/16 both methods peak at the first epoch and then decline (CAPT Mixed-mAP 0.7017 $\rightarrow$ 0.6938 over four epochs; DePro 0.6999 $\rightarrow$ 0.6905), so the ranking depends on the epoch-selection rule rather than on the methods: under each method’s own protocol (ours fixed at epoch 4, DePro’s own early stopping) DePro leads 5 of 6 metrics, at a matched epoch 1 the two split 3–3, and at a matched epoch 4 CAPT leads 5 of 6. We therefore do not claim that the B/32 ranking transfers to B/16; what the experiment shows is that CAPT scales with the stronger backbone (Unseen-mAP 0.6944 $\rightarrow$ 0.7635) and that our a-priori epoch-4 rule, calibrated on B/32, is not appropriate there.

Table 4. ViT-B/16, sketch holdout, seed 0 (mAP@200 / Prec@200).

model	epoch	Unseen	Mixed	U ^d CDR
CAPT B/16	1 (own best)	0.7635 / 0.7256	0.7017 / 0.6710	0.8421 / 0.5566
CAPT B/16	4 (a-priori rule)	0.7612 / 0.7210	0.6938 / 0.6637	0.8521 / 0.5597
DePro B/16	1 (its best-val)	0.7656 / 0.7252	0.6999 / 0.6673	0.8486 / 0.5606
DePro B/16	4	0.7591 / 0.7190	0.6905 / 0.6593	0.8515 / 0.5642

4. Feature-Space Analysis

On a fixed set of 450 images (3 unseen domains $\times$ 6 unseen classes, sketch holdout) we embed the frozen CLIP CLS feature, ProS, and CAPT. We quantify structure with a 5-fold linear class probe and DBSCAN cluster purity vs. class labels (adjusted Rand index):

	class-probe $\uparrow$	dom.-silh.	DBSCAN ARI $\uparrow$
CLIP (frozen)	0.978	0.055	0.000
ProS	0.982	0.020	0.621
CAPT	0.984	0.022	0.708

Both prompt-tuned methods produce clearer class structure than the frozen backbone: CLIP forms almost no class-pure density clusters (ARI 0.00), while ProS and CAPT reach 0.62 and 0.71 (Fig. 4). CAPT gives the highest ARI. Class-probe accuracy is near-saturated for all three methods, since the six classes are already mostly linearly separable in CLIP. The density-based ARI is therefore more informative here, because it measures the class structure used by cosine retrieval.

5. Qualitative Analysis

5.1. Attention

To show how the generated prompt acts, Figure 5 visualizes CAPT’s [CLS] $\rightarrow$ patch attention for one class across the six DomainNet domains and two further datasets it was never trained on (Sketchy, TU-Berlin). The inputs range from flat clipart and textured painting to abstract quickdraw strokes, real photos, and hand-drawn sketches from other benchmarks. Across these styles, the late-layer attention remains on the class object rather than the surrounding style or background.

5.2. Retrieval

The main paper gives a compact infograph retrieval comparison. Here we provide per-domain top-10 retrieval grids for all five held-out domains. Each grid uses a CAPT model trained with that domain unseen and queries the Real gallery over unseen classes. Green and red borders mark correct and incorrect classes. To avoid cherry-picking, each grid contains queries spread across the per-query accuracy range, from near-perfect rows (e.g. giraffe 10/10) to clear failures (e.g. skateboard 0/10, often confused with visually similar elongated objects). Across all five domains (Fig. 6), CAPT retrieves same-class images for many typical queries, while the failure rows show remaining errors on rare or shape-ambiguous categories.

Zero-shot retrieval on NOVEL classes (TU-Berlin classes absent from DomainNet)

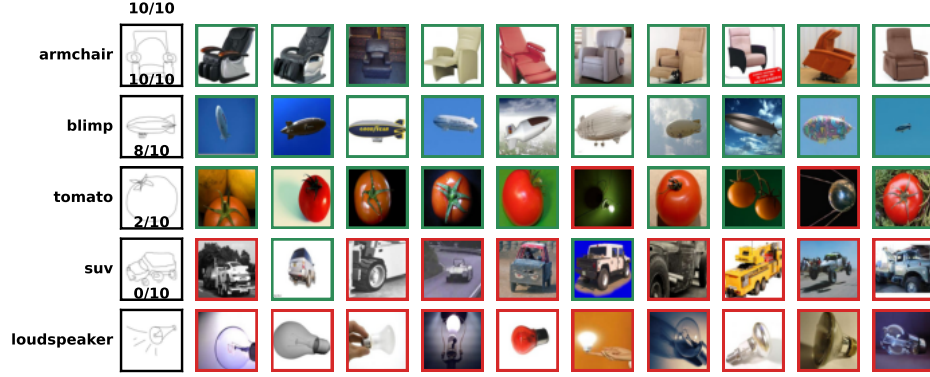


Figure 2. Zero-shot retrieval on novel classes. DomainNet-trained CAPT, no fine-tuning, retrieving over TU-Berlin classes that are absent from DomainNet (sketch query → photo gallery; green/red = correct/incorrect class). Rows span best to worst per-query accuracy.

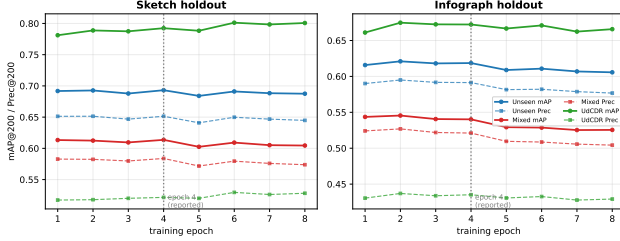


Figure 3. Per-epoch performance over an 8-epoch run (constant LR) for the sketch and infograph holdouts. Solid = mAP@200, dashed = Prec@200; the dotted line marks the reported epoch 4. Unseen and Mixed peak at (or within noise of) epoch 4 and do not improve afterward; UDCDR varies within ± 0.01 with no consistent trend.

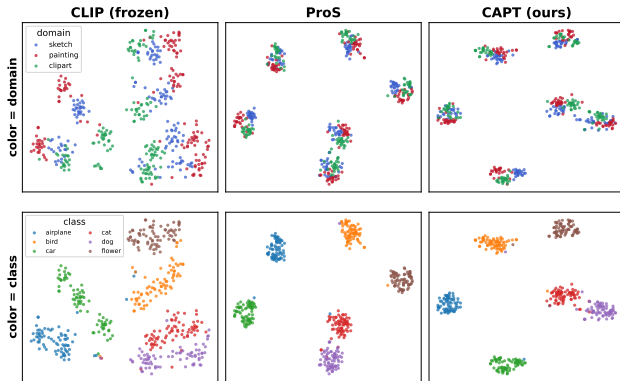


Figure 4. t-SNE of the feature space. 450 unseen-domain/class images by frozen CLIP, ProS, CAPT; top colored by domain, bottom by class.

6. Ablations Omitted from the Main Paper

For completeness, we report additional ablations omitted from the main paper for space. These results support the main design choices; when differences are within noise, we keep the simpler setting.

Component attribution (Tab. 12). Removing the prompt generator and learning the injected prompts directly, in the manner of VPT-Deep and at identical injection depth, costs 0.0072 Unseen-mAP ($p=0.0009$, 12 of 15 matched runs) over all five holdouts and three seeds; the larger share is the image-conditioning residual (0.0054, $p=0.0027$, $n=15$). The control learns the same 0.07M (73,728) injected prompt values, free rather than produced through our rank-64 read-out, and still loses. The generator carries more trainable parameters than this control (1.35M against 0.49M total trainable), but that is not the quantity the comparison turns on: every prompt DSIP injects is an up-projection of a 64-dimensional read-out and therefore lies in a rank-64 subspace of the 768-dimensional token space, a twelvefold reduction against the control’s unconstrained values (0.006M effective against 0.07M). In the space that actually reaches the transformer, ours is the more restricted model, and it leads on all six metrics. The two ingredients divide unevenly, and we state the division rather than credit it all to the constraint: image conditioning accounts for 0.0054 of the 0.0072, and the reparameterization on its own is a trade rather than a free win. Held static, so that only the rank-64 constraint distinguishes it from the control, it gains 0.0031 Unseen-mAP (10/11) but costs 0.0055 on U^dCDR-mAP (1/11, $p < 0.001$); only with the conditioning residual does the full model lead on every protocol. Substituting a frozen random basis of the same size for

CAPT [CLS] → patch attention for "airplane": DomainNet domains + other datasets

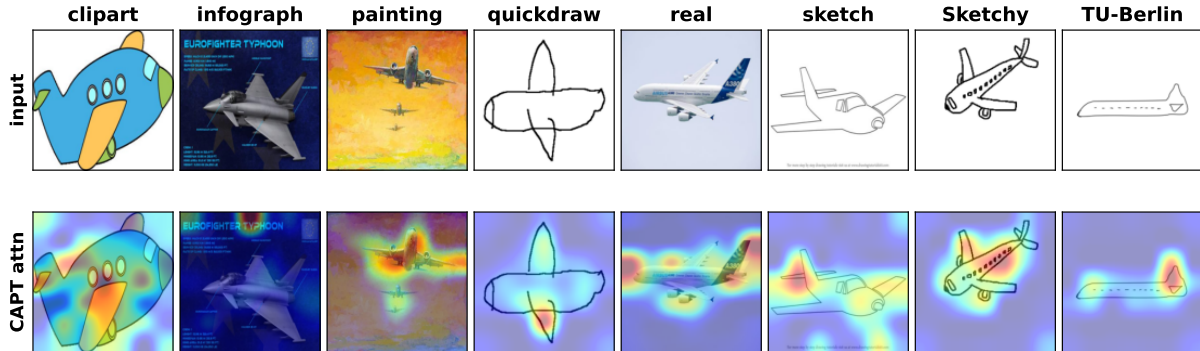


Figure 5. CAPT attention across domains and datasets. [CLS] → patch attention (jet overlay, bottom) for airplane on the six DomainNet domains and on Sketchy / TU-Berlin sketches (top: input). The attention concentrates on the object in every case, including the two datasets the model never saw during training.

the anchor bank moves Unseen-mAP by 0.0007, which three seeds do not resolve ($p = 0.2235$; Tab. 5). Since all four sources train identical parameters, swapping the anchors buys no capacity; class-name text is kept as the default because it adds CLIP’s class semantics at zero cost and is the source the nominal-scale analysis below rewards. What the measurement does resolve is the mechanism: on the generalization protocols DSIP’s benefit is the low-rank read-out from a frozen anchor bank, not prompt capacity. The representation-level view is the starkest form of this statement: swapping the semantic anchors for Gaussian noise perturbs the learned retrieval geometry less than re-running training with a different seed: linear CKA between the text- and random-anchor models is 0.990 for each of the three content-swap pairs (one per seed), against 0.965–0.971 for all six seed pairs of the same models (the two ranges do not overlap), and their k -NN retrieval neighborhoods overlap by 0.93 against a seed-to-seed baseline of 0.88; the same separation holds on the 45 classes unseen in training (1,200 probe images, domain- and class-stratified). Correcting Eq. 1’s attention to its nominal scale (in the reported model that softmax is uniform: entropy $\log 49$ to three decimals, so every anchor receives the same global patch average) makes the content term measurable: text anchors beat a matched random basis on 17 of 18 cells (3/3 runs; Unseen-mAP +0.0010, paired t $p=0.04$ uncorrected; sketch holdout, three seeds; U^dCDR precision is the one tied metric). The corrected variant itself matches the reported model on the generalization protocols (≈ -0.001) and gains 0.004–0.005 U^dCDR on average. On the clipart holdout the same correction leaves the two content arms within noise (0/6, 2/6, and 3/6 cells over three seeds; mean differences within ± 0.001); the content term we measure

Table 5. Anchor source at matched capacity (sketch hold-out; mean $\pm$ sd over three training seeds; confidence-only variant, identical parameterization; the vision arm trains the same 1,349,057 parameters). The four sources lie within one seed band of each other on every metric; pairwise tests against the class-name default do not separate (Unseen-mAP Welch $p=0.96$ words, $p=0.56$ random). The 0.0007 ($p=0.2235$) figure in the text is a paired comparison over three other holdouts (clipart, infograph, painting; nine runs), not the sketch cells shown here.

anchor source	Unseen		Mixed		U ^d CDR	
	mAP	prec	mAP	prec	mAP	prec
class-name text (default)	.6937 $\pm$.0023	.6518 $\pm$.0024	.6135 $\pm$.0026	.5832 $\pm$.0026	.7933 $\pm$.0011	.5217 $\pm$.0003
generic words	.6938 $\pm$.0025	.6518 $\pm$.0027	.6135 $\pm$.0027	.5833 $\pm$.0027	.7939 $\pm$.0010	.5221 $\pm$.0008
random basis	.6928 $\pm$.0009	.6509 $\pm$.0014	.6121 $\pm$.0015	.5817 $\pm$.0014	.7917 $\pm$.0012	.5214 $\pm$.0007
vision anchors (matched)	.6936 $\pm$.0021	.6519 $\pm$.0025	.6134 $\pm$.0024	.5834 $\pm$.0026	.7931 $\pm$.0010	.5219 $\pm$.0005

is therefore specific to the sketch holdout among those tested. This is an analysis variant, not the reported model; it bounds what the anchors’ content contributes once the read-out attends at nominal scale.

Is the anchor bank necessary? We remove it entirely: a variant reads the prompt from the image’s own patch tokens through the same attention, up-projection, and query (no class labels, no (C,512) bank; the anchor projection is inactive, all other trainable tensors unchanged). To keep the comparison single-variable the image-conditioning residual is disabled in both arms, so neither is the reported model. Over six matched runs (sketch and clipart holdouts, three seeds) the patch-read variant raises U^dCDR by 0.0122 mAP (6 of 6 runs, $p = 0.0002$) while moving five of six Mixed-mAP runs the other way (-0.0032 , $p = 0.08$). The representation-level probe locates the mechanism: the patch-read variant’s advantages (lower domain identifiability, tighter class structure) are present on the training classes and largely vanish on the 45 classes unseen in training.

Reading from a fixed external bank trades seen-class retrieval (-0.0122 U^dCDR-mAP) for the mixed-gallery protocol (5 of 6 Mixed-mAP runs), where gallery distractors from the training vocabulary must be rejected; on unseen-only galleries the two read-outs are indistinguishable (3 of 6).

CDCM bank: imbalance and staleness. The memory bank is a FIFO queue of $M=10,000$ features scored against the top- $K=256$ hardest negatives. Two properties are worth stating explicitly. First, the training split is mildly imbalanced rather than long-tailed: over the 300 training classes the per-class counts span 653–2317, a $3.5\times$ ratio with coefficient of variation 0.199, so no class is starved of bank entries. Second, staleness is bounded by construction: at batch size 60 the bank fills after 167 iterations, which is 0.57% of the 29,528 iterations of a four-epoch run, and every slot carries a validity flag so that unfilled slots never enter the loss. Entries are therefore at most one bank-length old for all but the first half-percent of training.

Transplanting DePro’s adaptation (Tab. 13). DePro’s `ln_trick` unfreezes LayerNorm and moves the retrieval token to CLS. CAPT already reads from CLS, so we match the first at DePro’s exact scope. Over five hold-outs and three seeds this moves the cell tally by one; unfreezing biases as well reaches 24/30 and brings the U^dCDR-mAP average to within a ninth of a seed standard deviation of DePro. In every configuration all twenty Unseen and Mixed cells remain CAPT’s, so the comparison does not depend on the frozen setting.

Seed-level anatomy of the bias-tuned comparison. Across all 72 seed-cells (12 completed runs $\times$ 6 metrics; a seed-cell is one run’s value on one metric, a finer unit than the seed-averaged cells above) CAPT+*bias* wins 59. Of the 13 losses, six are the quickdraw U^dCDR pair, the one genuine deficit (-0.007 to -0.022), while the remaining seven all lie within 0.005 of DePro (five within 0.0025, the smallest 0.0003), against winning margins of median $+0.008$ and up to $+0.018$. The win pattern is stable across seeds: every clipart seed loses only U^dCDR precision, and every quickdraw seed loses only the U^dCDR pair.

Corrected frozen control (Tab. 14). Our earlier frozen DePro disabled both parts of `ln_trick` and was therefore a different model. Retrained with the CLS read-out forced, so that LayerNorm is the only remaining difference, the baseline is stronger on every metric; the comparison in the main paper uses this corrected control.

Routing against published weighting schemes (Tab. 7). We plug five published weighting schemes into the same per-sample slot w_i , holding every other component fixed. Four of the five trail both per-sample rules

on every generalization metric: uncertainty weighting over all fifteen runs ($p=0.0005$, sign test corrected for the eight routing comparisons; 80 of 90 cells), and the scalar, GroupDRO, and the MLP on the sketch holdout (three seeds). Focal shaping, itself a fixed function of a stop-gradient statistic, lands inside the ± 0.002 family band over fifteen runs, consistent with the mechanism, not against it: what fails is letting the objective move the weight, not the shape of the fixed rule. Where an alternative edges ahead is U^dCDR precision (the collapsed MLP, the focal reshaping, and uncertainty weighting sit 0.0009–0.0017 above the combined rule there), and the mechanism is the diagnosis itself: each shifts weight onto the CDPA term, the loss that U^dCDR measures, and pays 0.003–0.017 Unseen-mAP for it (table values, seed 0). The failures are identifiable rather than incidental: the learned per-sample MLP, given exactly the two statistics our rule uses, converges to $w \equiv 1.0$ for every input, routing all samples to whichever loss is already smaller; uncertainty weighting [2] tilts its learned precisions 11.4:1 the same way; GroupDRO [8] concentrates its group weights on a single domain (0.99999). This is why w_i must be a fixed function of stop-gradient statistics: any weight the objective can reach, it drives toward the smaller term. (Prediction pre-registered in the training code before these runs.) The equal mix also fails off Domain-Net: on Sketchy and TU-Berlin it loses all 40 metric-seed cells to per-sample routing (10/10 runs, sign test $p=0.002$; five seeds), never ranking above third on any of the four metrics. Across the four further holdouts the reported rule wins 17 of 24 cells against GroupDRO, 22 of 24 against the learned scalar, and 22 of 24 against the MLP (arm at seed 0 versus our three-seed mean per holdout). The single above-band exception is GroupDRO on painting, which gains up to $+0.0036$ on the two generalization protocols while losing U^dCDR-mAP by 0.0048; the only other above-band exception is the MLP’s $+0.0018$ on painting U^dCDR-prec; every remaining above-band difference favors the reported rule. These conclusions do not depend on the shared four-epoch schedule: granting every rule, ours included, its per-run best checkpoint, Kendall’s weighting still loses 13 of 15 runs ($p=0.0074$; 73 of 90 cells), the learned scalar, GroupDRO and the MLP together win one run of nine, and the mean-matched focal variant stays inside the ± 0.002 band (-0.0001). The deficit is not the step size: over $\eta \in \{10^{-2}, 10^{-3}, 10^{-4}\}$ GroupDRO’s loss is unchanged (0.6891–0.6896 Unseen-mAP against our 0.6943), and at $\eta=10^{-4}$ the group weights stay spread ($q \in [0.05, 0.38]$) yet the deficit persists (seed 0, sketch). What fails is group-level reweighting itself, not its optimisation.

Two further probes fix the rule’s remaining degrees of freedom. Temperature. Replacing Eq. 7’s batch standardization by a fixed temperature is free while the temperature is right ($\tau=0.05$: $+0.0002$, $p=0.64$) and costs every cell once it is not: $\tau=0.2$ loses 18/18 cells (-0.0011 , $p=0.010$) and $\tau=0.4$ loses 18/18 (-0.0020 , $p=0.003$); three seeds, sketch holdout, confidence-only variant, paired t . The standardized rule has no such knob. Spread. With temperature and spread both matched (reported condition, fallback present), adding the gap direction moves the DomainNet run mean by -0.0005 (3/15 runs above zero, $p=0.05$; five holdouts, three seeds): on DomainNet, the reported rule’s advantage over every alternative in the main paper’s routing ablation comes from the per-sample assignment, not from a mean effect of the gap direction. Where the gap direction does pay is at the benchmark level, below.

When does the gap term win? Ranking the seven training settings by a quantity measurable before training, the frozen-CLIP domain gap $1 - \cos(\mu_q, \mu_g)$, the two largest-gap settings are exactly where the combined rule wins (Tab. 6): TU-Berlin (0.243; 4/5 runs, temperature-matched, the exception within ± 0.0004) and the quickdraw holdout (0.243; 4 of 5 runs over five seeds against the reported confidence rule, the exception within ± 0.0002 ; 2/3 against the temperature-matched variant). On the five smaller-gap settings the two rules’ per-run mean differences stay within 0.002. We report the ranking only: infograph (0.227) sits close behind, and we claim no threshold. The benchmark-dependent sign is itself systematic: across the two cross-dataset benchmarks (5 seeds each, temperature-matched), 9 of 10 runs fall on the side the frozen-gap ranking selects (sign test $p=0.02$). The rule was fixed on the first three seeds; the four runs added afterwards land 3 of 4 on the predicted side.

Table 6. Cross-benchmark rank of each routing rule: rank of the five-seed mean, per metric, among the four rules run on both cross-dataset benchmarks. The combined rule is the only one that never falls below second; every alternative falls to third or fourth on one of the two benchmarks.

rule	Sketchy (4 metrics)	TU-Berlin (4 metrics)	worst
combined (reported)	2 · 2 · 2 · 2	1 · 1 · 1 · 1	2
confidence-only	1 · 1 · 1 · 1	3 · 2 · 2 · 2	3
gap-only	4 · 4 · 4 · 4	2 · 3 · 3 · 3	4
fixed $w=0.5$	3 · 3 · 3 · 3	4 · 4 · 4 · 4	4

Where the two per-sample rules differ. At the level of individual queries the two rules are almost the same model: the mean cosine between a query’s two embeddings is 0.9964 (min 0.978). Their aggregate difference comes from rank-boundary reshuffling among contested queries (confidence-AP 0.2–0.8: -0.0053 at

Table 7. Published weighting schemes in the w_i slot. Table entries: sketch holdout, seed 0. Significance is assessed separately: uncertainty weighting and focal shaping over fifteen runs (five holdouts $\times$ three seeds); the scalar, GroupDRO and the MLP over three sketch seeds and seed 0 on the four further holdouts (MWN-style confirmed on seed 1). GroupDRO uses $\eta=0.01$; smaller steps leave it unchanged (see text).

rule in the w_i slot	Unseen		Mixed		U ^d CDR	
	mAP	prec	mAP	prec	mAP	prec
combined (ours, Eq. 7)	.6943	.6525	.6140	.5839	.7923	.5210
confidence-only (ablation)	.6956	.6538	.6158	.5856	.7934	.5219
focal reshaping [5]	.6913	.6498	.6109	.5810	.7934	.5227
GroupDRO [8]	.6896	.6490	.6084	.5793	.7879	.5208
learned global scalar	.6852	.6441	.6022	.5731	.7900	.5207
learned per-sample MLP (MWN-style [9])	.6774	.6380	.5957	.5668	.7870	.5227
uncertainty weighting [2]	.6771	.6373	.5952	.5664	.7880	.5219

0.2–0.5 and -0.0038 at 0.5–0.8 on the sketch holdout, against -0.0001 for already-failed queries; mean -0.0013 , CI $[-0.0016, -0.0010]$, $n = 9,729$); queries the confidence rule already fails or already solves are untouched. The reshuffling is locally concentrated: on the sketch holdout, three of 45 unseen classes account for 55% of the net difference, the worst class (peas) is the same in both seeds, and the class-level rank correlation across seeds is $\rho=0.55$ ($p=10^{-4}$); the remaining 35 classes sum to a net gain for the combined rule. The same anatomy holds where the combined rule wins: on the quickdraw holdout ($+0.0026$ overall, CI $[+0.0023, +0.0029]$, $n = 22,500$) the gains sit in the same contested band ($+0.0062$ at confidence-AP 0.2–0.5 and $+0.0094$ at 0.5–0.8, against $+0.0001$ for already-failed queries), are similarly concentrated (three classes carry 57% of the net gain, as three carried 55% of the sketch-holdout loss), and the two embeddings again agree to cosine 0.995. Where the gap term acts is invariant (contested queries, a handful of classes, a fixed representation), and only the sign of its effect depends on the holdout.

Refinement-layer attribution (Tab. 8). The main-paper build-up (Tab. 5) spans two code revisions, so its final-row difference is not the refinement layer; here we isolate the layer properly, twice. Top: a matched re-evaluation (same codebase, same evaluation path, seed 0) of the model with and without the layer on all five holdouts (this evaluation path differs from Tab. 1, so the two tallies are not comparable). The layer moves the aggregate against our corrected DePro reimplementations from 14/30 to 22/30 cells; the largest effect is infograph ($+0.0224$ Unseen-mAP, $0/6 \rightarrow 4/6$), on quickdraw its $+0.0023$ is inside the seed band and we claim nothing there. Bottom: the same ablation in the current code generation, paired over five holdouts $\times$ three seeds: significant on all six metrics. The layer is neces-

sary for the aggregate lead but not its source: a head-only model reaches 0.6217 on sketch against DePro’s 0.6843, and transplanted into DePro the same head moves it by at most 0.0036 with mixed sign, against +0.0101 inside CAPT.

Table 8. Refinement layer, isolated. Top: matched re-evaluation (seed 0, mAP@200, same codebase and eval path for both configs; DePro = our corrected reimplementation). Bottom: current-generation paired ablation over five holdouts $\times$ three seeds (15 matched runs).

holdout	Unseen	without layer		Unseen	with layer	
		Mixed	U ^d CDR		Mixed	U ^d CDR
sketch	.6858	.6118	.7784	.6956	.6158	.7936
clipart	.7939	.7412	.8504	.7991	.7439	.8506
painting	.7685	.7217	.7588	.7729	.7218	.7716
infograph	.5989	.5264	.6506	.6213	.5441	.6750
quickdraw	.3233	.2574	.3013	.3256	.2539	.3170
cells won vs. DePro		14/30			22/30	
Current generation, paired over 15 matched runs (with – without):						
Δ mAP		Unseen +0.0100 (15/15, $p < 10^{-4}$)	Mixed +0.0055 (13/15, $p = 0.002$)		U ^d CDR +0.0148 (15/15, $p < 10^{-4}$)	

6.1. Read-out attention heads

The number of heads in the MHA read-out has little effect; we use 4.

Table 9. Number of MHA read-out heads (full model, sketch holdout). Bold = default.

heads	Unseen	Mixed	UdCDR
2	.6945/.6527	.6142/.5840	.7924/.5211
4	.6944/.6527	.6141/.5841	.7926/.5212
8	.6942/.6524	.6139/.5838	.7920/.5209
16	.6946/.6528	.6144/.5843	.7922/.5210

6.2. CDCM denominator

CDCM restricts its denominator to the positives and the top- $K=256$ hardest negatives (cosine < 0.9). A full-bank InfoNCE denominator is slightly worse, so we keep hard-negative mining.

Table 10. CDCM denominator (full model). Bold = best per column.

denominator	Unseen	Mixed	UdCDR
full-bank InfoNCE	.6913/.6497	.6095/.5797	.7917/.5216
hard-neg (ours)	.6944/.6525	.6141/.5839	.7923/.5211

6.3. Class prototype

Computing μ_c as a flat class mean over the bank vs. a domain-balanced mean makes no significant difference (within noise), so we keep the simpler flat mean.

Table 11. Class prototype μ_c (full model, epoch 2). Bold = default.

prototype	Unseen	Mixed	UdCDR
flat class mean	.6918/.6513	.6133/.5828	.7811/.5172
domain-balanced	.6942/.6524	.6140/.5839	.7924/.5210

Table 12. Component attribution over three seeds. Each row removes one part of CAPT and reports the resulting loss in Unseen mAP@200, paired over matched (hold-out, seed) runs; the pairs column gives each comparison’s coverage (the headline row spans all five holdouts). The two indented rows are separate single-variable arms, not an exact decomposition. The read-out mechanism dominates the anchor-content term at this sample size. The control that removes the generator learns the same 0.07M (73,728) injected prompt values, free rather than produced through a rank-64 read-out, so the improvement comes from how the prompts are produced, not from extra capacity.

what is removed	Δ Unseen-mAP	paired p	wins	pairs
prompt generator (VPT-Deep control)	+0.0072	0.0009	12/15	15
of which: image-conditioning	+0.0054	0.0027	13/15	15
rank-64 constraint alone (static generator)	+0.0031	n/a	10/11	11
anchor content (text $\rightarrow$ random)	+0.0007	0.2235	4/9	9

6.4. Baseline reproduction

DePro, frozen backbone. Removing DePro’s Layer-Norm fine-tuning (matching our frozen setting) lowers its five-domain average to Unseen 0.6132 / Mixed 0.5426 / UdCDR-mAP 0.6765 (main paper Table 2). DePro, TU-Berlin LR sweep. Given the same TU-Berlin learning-rate sweep as CAPT (same environment), DePro’s best mAP@all / Prec@100 is 0.6629 / 0.7203, below CAPT’s three-seed mean (0.6641 ± 0.0024 / 0.7305 ± 0.0022); the mAP margin sits within the seed band, the precision margin does not.

ProS, Sketchy gallery. On the standard 100-image-per-class subset, our ProS reimplementation matches the published Sketchy mAP@200 (0.6993 vs. 0.6991); the main paper reports the full extended-gallery protocol.

7. Reproducibility

Table 16 lists all hyperparameters and Table 17 the benchmark statistics used in evaluation.

7.1. Computational complexity

CAPT adds only a small overhead over a frozen CLIP forward. The DSIP generator, per injected layer, computes (i) two bottleneck projections (AW_a , $H^\ell W_h$ to width $r=64$), (ii) a gated conditioning attention of cost $\mathcal{O}(CNr)$ over C anchors and N patches, and (iii) a read-out attention of cost $\mathcal{O}(P_q Cr)$ for $P_q=8$ queries; all are small relative to a ViT-B/32 block, whose self-

Table 13. Transplanting DePro’s backbone adaptation into CAPT (five holdouts $\times$ seeds 0, 1, 2). In_trick bundles unfreezing LayerNorm with moving the read-out to CLS; CAPT already reads from CLS, so we match only the first, at DePro’s exact scope (65,536 affinities across both towers). Cells are counted against DePro[†]; all 20 Unseen and Mixed cells are CAPT’s in every row. Our rows are three-seed means; the DePro[†] reference is a single run per holdout, and the seed σ of U^dCDR-mAP ranges over 0.0011–0.0058 across holdouts (mean 0.0032), so the residual 0.0004 is well inside it.

setting	trainable	cells vs. DePro [†]	U ^d CDR cells
CAPT (frozen)	1.35M	21/30	1/10
+ In_trick (LayerNorm, +65,536)	1.41M	22/30	2/10
+ LayerNorm and biases (+203,776)	1.55M	24/30	4/10
five-holdout U ^d CDR-mAP average: frozen 0.6845, +LayerNorm 0.6853, +LayerNorm & biases 0.6877 vs. DePro [†] 0.6881.			

Table 14. Corrected frozen control. Our earlier frozen DePro disabled both parts of In_trick, so it was a different model. Retrained on all five holdouts with the CLS read-out forced (LayerNorm the only remaining difference), the baseline becomes stronger, and CAPT still wins every Unseen and Mixed cell.

frozen DePro control	sketch Unseen-mAP	CAPT cells won
earlier (both parts disabled)	0.6257	n/a
corrected (CLS forced)	0.6626	20/20 Unseen+Mixed, 24/30 total

Table 15. Seed replication of the reported model (combined rule, five holdouts $\times$ three seeds; mean $\pm$ std, mAP@200/Prec@200). The three-seed means deviate from the reported Table 1 cells by at most 0.0028, and 24 of 30 cells lie within ± 0.002 ; the largest deviations sit on quickdraw, the domain with the largest seed spread. Table 1 is not a seed artifact. The conclusions are also robust to checkpoint selection: under per-run best-checkpoint (early-stopping) selection the read-out gain strengthens to +0.0070 ($p=0.0011$, 11/12).

holdout	Unseen		Mixed		U ^d CDR	
	mAP	prec	mAP	prec	mAP	prec
clipart	.7996 $\pm$.0017	.7593 $\pm$.0015	.7445 $\pm$.0012	.7109 $\pm$.0014	.8603 $\pm$.0027	.5496 $\pm$.0011
sketch	.6926 $\pm$.0020	.6508 $\pm$.0021	.6120 $\pm$.0024	.5818 $\pm$.0023	.7926 $\pm$.0014	.5211 $\pm$.0005
infograph	.6180 $\pm$.0017	.5912 $\pm$.0014	.5407 $\pm$.0016	.5220 $\pm$.0016	.6764 $\pm$.0049	.4358 $\pm$.0029
painting	.7734 $\pm$.0007	.7225 $\pm$.0003	.7215 $\pm$.0010	.6778 $\pm$.0006	.7718 $\pm$.0017	.4862 $\pm$.0007
quickdraw	.3306 $\pm$.0055	.3036 $\pm$.0048	.2573 $\pm$.0028	.2434 $\pm$.0025	.3182 $\pm$.0056	.1275 $\pm$.0026

attention is $\mathcal{O}((N+1)^2 d_v)$. The prompt enlarges the token sequence by only $P_q=8$ tokens (from $N+1$ to $N+1+P_q$), a $\sim 4\%$ length increase at $N=49$. Trainable parameters are 1.25–1.35M across the three benchmarks ($\sim 7\%$ of ProS/DePro’s 18.5M); the count varies only with the number of classes in the text head. At inference, the text encoder is not run (anchors are cached); a query is one frozen-CLIP forward with the generated prompt followed by a cosine search, with no test-time adaptation, momentum, or transductive step. The anchor count itself is not a cost driver: growing the

Table 16. CAPT hyperparameters (all domains share these defaults).

group	value
prompt	$P_q=8$, bottleneck $r=64$, heads 4, all $L=12$ layers
text	CoOp context length 4 + per-class shift
refine	residual MLP, width 256 (~ 0.26 M)
PACE	$\lambda=0.5$, router $\tau=0.1$, signal = conf.+gap
CDCM	temp. 0.07, top- $K=256$ neg., cap $\eta=0.9$, bank $M=10,000$
optim	Adam, lr 10^{-4} (constant), batch 60, 4 epochs, seed 0
input	224×224 , RandAugment (2, 9), ViT-B/32 frozen

Table 17. Evaluation benchmarks.

benchmark	protocol	gallery
DomainNet	245/55/45 cls, leave-1-domain-out	Real
Sketchy	ECCV’18 unseen split	full ext. photo (10453)
TU-Berlin	30-class test split	27989 images

bank from 50 to 300 class anchors changes throughput by under 1% (159.2 $\rightarrow$ 157.9 img/s) with peak memory unchanged (1357 MB).

All models train for 4 epochs at constant learning rate 10^{-4} (Adam), batch 60, seed 0, on one 80GB GPU. The DomainNet source domain is quickdraw for the sketch holdout and sketch otherwise; Real is always the gallery domain and is included in training, as in the standard protocol. With both CLIP encoders frozen, only ~ 1.35 M parameters update on DomainNet (1,349,057; 1,248,705 on Sketchy, 1,308,097 on TU-Berlin); the text encoder is run once to cache anchors and is not used at test time, so inference is a single frozen-CLIP forward with the generated prompt plus a cosine search (no test-time adaptation). We provide code to reproduce the reported numbers.

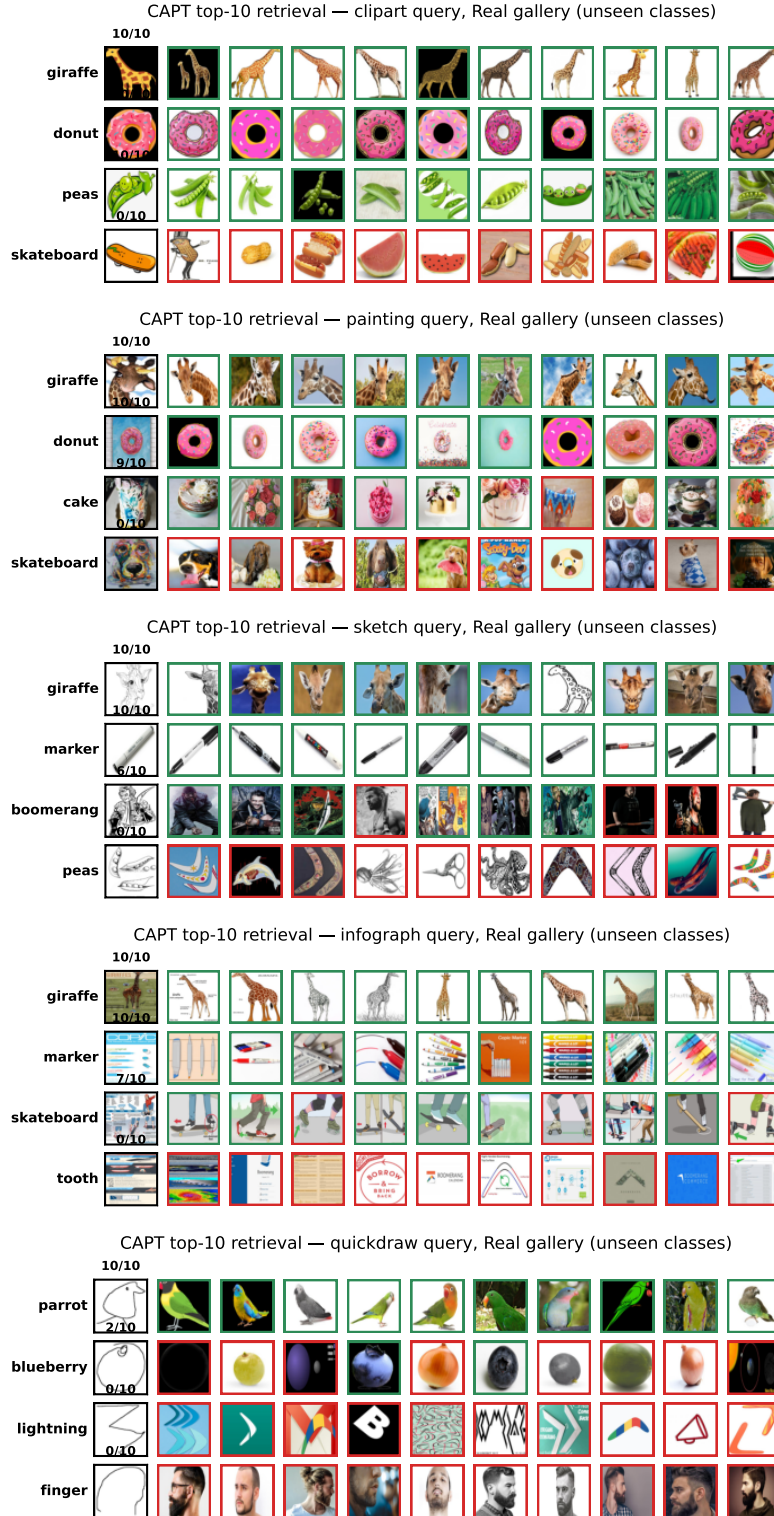


Figure 6. Top-10 retrieval across all five held-out domains (clipart, painting, sketch, infograph, quickdraw → Real gallery, unseen classes). Green/red borders mark correct/incorrect class; each panel shows a spread of queries from high to low per-query accuracy. CAPT returns predominantly same-class images, with failures concentrated on rare or shape-ambiguous categories (e.g. quickdraw lightning, finger).

References

- [1] Stephen Boyd and Lieven Vandenberghe. *Convex Optimization*. Cambridge University Press, 2004. [1](#)
- [2] Alex Kendall, Yarin Gal, and Roberto Cipolla. Multi-task learning using uncertainty to weigh losses for scene geometry and semantics. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7482–7491, 2018. [7](#), [8](#)
- [3] 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 (NeurIPS)*, pages 18661–18673, 2020. [2](#)
- [4] Junnan Li, Pan Zhou, Caiming Xiong, and Steven C.H. Hoi. Prototypical contrastive learning of unsupervised representations. In *International Conference on Learning Representations (ICLR)*, 2021. [2](#)
- [5] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *ICCV*, 2017. [8](#)
- [6] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *Proceedings of the International Conference on Machine Learning (ICML)*, pages 8748–8763, 2021. [1](#)
- [7] Joshua Robinson, Ching-Yao Chuang, Suvrit Sra, and Stefanie Jegelka. Contrastive learning with hard negative samples. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2021. [2](#)
- [8] Shiori Sagawa, Pang Wei Koh, Tatsunori B Hashimoto, and Percy Liang. Distributionally robust neural networks for group shifts: On the importance of regularization for worst-case generalization. In *ICLR*, 2020. [7](#), [8](#)
- [9] Jun Shu, Qi Xie, Lixuan Yi, Qian Zhao, Sanping Zhou, Zongben Xu, and Deyu Meng. Meta-weight-net: Learning an explicit mapping for sample weighting. In *NeurIPS*, 2019. [8](#)
- [10] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. In *arXiv:1807.03748*, 2018. [2](#)
- [11] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *NeurIPS*, 2017. [1](#)
- [12] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Learning to prompt for vision-language models. *International Journal of Computer Vision (IJCV)*, 130(9):2337–2348, 2022. [3](#)