TOWARD ROBUST HUMAN-FREE IMAGE ANNOTATION WITH THOUSANDS OF CANDIDATE CLASSES

Anonymous authors
Paper under double-blind review

ABSTRACT

Vision-language models (VLMs) can reduce manual image labeling, but annotation becomes difficult when the label space grows, and the target data are imbalanced. Question-answering VLMs are a natural choice, yet their computational cost increases and their annotation accuracy decreases as the candidate vocabulary grows. On ImageNet subsets, accuracy drops from 73.9% at 200 classes to 49.5% at 1,000. Retrieval-based VLMs such as Contrastive Language-Image Pre-training (CLIP) and Sigmoid Loss for Language-Image Pre-training (SigLIP) scale more efficiently. However, standard zero-shot inference still treats each image independently. We therefore ask: *can the unlabeled target pool itself improve annotation without target labels or knowledge of the true class distribution*? Our analysis shows that class names and target images provide complementary information: the class name is a stable semantic *anchor*, while the pool reveals target-specific visual structure and class prevalence. Standard pseudo-label prototyping only partly exploits this signal, because confidence filtering biases the visual representative and uniform balancing fails under class imbalance. We propose *robust prototyping*, a closed-form semantic–visual calibration that assigns images under an estimated class prior, builds a whitened visual head from all assigned images, and fuses it with the semantic anchor. Without target-pool labels, our method improves zero-shot retrieval on all 18 CLIP/SigLIP encoder-dataset pairs, outperforms the label-free prototype baseline on 15 of 18, and achieves the best results among the evaluated training-free methods on the *long-tailed* ImageNet-LT and Places-LT benchmarks, with only seconds of computation after feature encoding.

1 INTRODUCTION

Large-scale visual datasets such as ImageNet were built in an *ontology-first* manner: the class vocabulary was defined first, images were retrieved for each class, and human annotation was then used to clean and verify the resulting data (Deng et al., 2009; Russakovsky et al., 2015). We consider a different setting. Imagine asking an agent to "build a classifier for every object in this building"—or in a garden, greenhouse, or production line. A robot or camera system can explore the environment and collect images, but those images still need to be organized into classes before they can be used to train a classifier. Unlike class-by-class web retrieval, the composition of this target pool is not controlled in advance: some classes may appear frequently, some only a few times, and some candidate classes may not appear at all. The annotation problem is therefore both *many-class* and *distribution-unknown*. We focus on this first step of the larger autonomous learning pipeline: *given an unlabeled target pool and a fixed candidate vocabulary, can foundation models annotate the pool accurately and at scale without target labels?*

A straightforward solution is to use a multimodal large language model as a *question-answering (QA) annotator*. Models such as *Large Language and Vision Assistant* (LLaVA) (Liu et al., 2023; 2024a) and Qwen-VL (Bai et al., 2025) can be given the candidate vocabulary and asked to choose a class. This works well for small candidate sets: Qwen3-VL-8B reaches 96% on a 10-class Tiny-ImageNet subset. As the candidate vocabulary grows, however, computational cost increases while annotation accuracy decreases (Appendix B). On ImageNet-1k, annotating 1.28M images over 1,000 classes took 202 GPU-hours with a 2B QA model, compared with 6 hours for a retrieval encoder, while accuracy dropped from 73.9% at 200 classes to 49.5% at 1,000 (Fig. 4c). Much of the class information is still present before the language-model decision: on 200-class TinyImageNet, a linear

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.

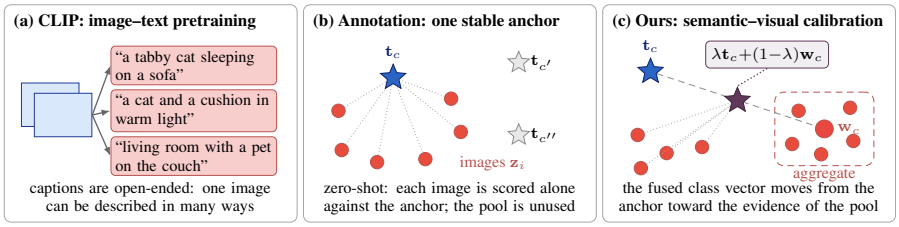

Figure 1: **Why labeling is special, and what we exploit.** (a) CLIP-style pretraining sees open-ended captions; annotation fixes a vocabulary. (b) Each class thus has a stable anchor $\mathbf{t}_c$, but zero-shot retrieval scores each image alone. (c) Ours aggregates the images assigned to class c into a whitened head $\mathbf{w}_c$ and scores with $\lambda\mathbf{t}_c + (1-\lambda)\mathbf{w}_c$ (assignment: Figure 3).

probe on the visual tokens received by LLaVA-1.6-7B reaches 78%, while the model itself answers correctly on 38% of images.

Retrieval-based VLMs such as *Contrastive Language–Image Pre-training* (CLIP) (Radford et al., 2021) and *Sigmoid Loss for Language–Image Pre-training* (SigLIP) (Zhai et al., 2023; Tschannen et al., 2025) are a more direct fit for this closed-set task. They encode each image and class name once and rank labels by image–text similarity; enlarging the vocabulary adds only similarity computation over more class embeddings rather than repeatedly processing a longer candidate list through an autoregressive model. On the same ImageNet subsets, SigLIP2 loses 8.7 points from 200 to 1,000 classes, compared with 24.4 points for the QA model. Labeling also has a useful structure. CLIP-style models are trained on image–text pairs whose descriptions are open-ended, while annotation fixes the output vocabulary. Each class name therefore provides a stable *semantic anchor* shared across the target pool (Fig. 1a,b).

Standard zero-shot retrieval uses this anchor but ignores the target pool as a whole. Class c is represented by a fixed text embedding $\mathbf{t}_c$, and each image embedding $\mathbf{z}_i$ is classified independently. The anchor captures *what the class means*, but not *how it appears in this target pool*. For SigLIP2, the cosine between a class anchor and the mean image embedding of the same class is only 0.11–0.16, a class-level manifestation of the modality gap (Liang et al., 2022). An oracle analysis shows that this visual information matters. Using ground truth only for diagnosis, the mean image embedding of each class reaches 84.7% on TinyImageNet with SigLIP2, compared with 77.1% for the text anchor; fusing the two reaches 86.0% and outperforms either component on every tested backbone–dataset configuration. Thus, the class name and target images provide complementary information. Without labels, the simplest approximation is *prototyping*: assign images by the arg-max prediction of the anchors, average each pseudo-class, and fuse the prototype with the anchor. This already adds 3–4 points on balanced TinyImageNet for every backbone.

This recipe is fragile in our setting. First, zero-shot arg-max can severely distort the class marginal: although TinyImageNet is balanced, predicted class sizes range from 15 to 2,657 images, a 177-fold spread. Uniform balancing does not solve this under true imbalance: on SUN397 it reduces pseudo-label purity from 75% to 66% and leaves InMaP 5.8 points below zero-shot with SigLIP2; under 100:1 imbalance, our uniform-prior variant loses nearly all its gain. Second, confidence filtering raises pseudo-label purity but biases the visual representation and hurts even with perfect labels. Third, raw class means are poorly conditioned: their pairwise cosine is 0.71–0.88 because a shared direction dominates the embeddings. We therefore develop *robust prototyping*, a label-free semantic–visual calibration with three closed-form steps (Fig. 1c): marginal-constrained assignment using an estimated target prior, a whitened visual head built from all assigned images, and fusion with the semantic anchor. It requires no target labels or fine-tuning and adds only seconds after ImageNet-1k encoding. Our setting is related to transductive CLIP adaptation (Zanella et al., 2024; Qian et al., 2023; Qiu & Ren, 2026), but differs in its focus on annotation with the entire target pool, hundreds to thousands of candidate classes, and an unknown class distribution that may include absent classes.

Across datasets, backbones, vocabulary sizes, and class distributions, the same calibration consistently improves retrieval-based annotation without target-pool labels. With three scalars fixed in advance, it beats zero-shot on all 18 CLIP-family encoder–dataset pairs: with SigLIP2, TinyImageNet improves from 77.1% to 82.2%, SUN397 from 75.1% to 80.3%, and ImageNet-1k from

85.5% to 86.9%. It also outperforms the label-free prototype on 15/18 pairs. Under imbalance and missing classes, it beats zero-shot in all 162 stress-test cells, including 100:1 imbalance and only 25% of candidate classes present, and performs best among the evaluated training-free methods on ImageNet-LT and Places-LT. Selecting the three scalars on a labeled quarter adds 1.4 points on average and beats the prototype on all 24 encoder–dataset pairs, including DINOv2 with SigLIP2 pseudo-labels. The principle is simple: *class names tell us what classes mean, while the unlabeled target pool reveals how they appear and how often they occur; an annotator should use both without assuming a balanced pool.*

Contributions.

- *QA versus retrieval.* We show that QA-based annotation becomes more costly, less accurate, and more sensitive to candidate presentation as the vocabulary grows, motivating retrieval-based matching for many-class closed-set annotation.
- *Semantic and visual complementarity.* We show that class-name anchors and target visual structure are complementary, and identify four failures of standard prototyping: distorted marginals, confidence-selection bias, uniform-prior failure under imbalance, and poorly conditioned class means.
- *Robust prototyping.* We develop a label-free, closed-form calibration combining marginal-constrained assignment with an estimated prior, a whitened head over all assigned images, and semantic-anchor fusion. We evaluate it on six pools up to ImageNet-1k, three CLIP/SigLIP backbones, DINOv2 transfer, vocabulary scaling, imbalance, missing classes, training-free baselines, and computational cost.

2 Diagnosis: What Limits Many-Class Annotation

We first diagnose the main limitations of current annotators and standard prototyping, which motivate the design in Section 3.

Setting and notation. Given an unlabeled pool $\mathcal{D} = \{\mathbf{x}_i\}_{i=1}^N$ and a fixed vocabulary $\mathcal{Y} = \{y_1, \dots, y_C\}$, a retrieval-based VLM produces normalized image embeddings $\mathbf{z}_i$ and class-name embeddings $\mathbf{t}_c$. With $Z \in \mathbb{R}^{N\times D}$, $T \in \mathbb{R}^{C\times D}$ and $S = ZT^\top$, zero-shot annotation is $\hat{y}_i = \arg\max_c S_{ic}$. We call $\mathbf{t}_c$ the *semantic anchor*; a vector estimated from target images of class c is a *target visual representative*. *Purity* is pseudo-label accuracy and *marginal* is the vector of assigned class sizes. Experimental splits and full notation are given in Section 4.1 and Appendix C.

Finding 1: QA annotation scales poorly with the vocabulary. On identical images and candidate sets, QA annotation becomes both more expensive and less accurate as C grows (Figure 4c; Appendix B). On ImageNet-1k, a 2B QA model requires 202 GPU-hours, versus six hours for SigLIP2 feature extraction. From 200 to 1,000 classes, QA accuracy drops by 24 points, compared with nine for SigLIP2. Diagnostic experiments further show sensitivity to list length, candidate position, and answer priors, while substantial class information remains in the visual features. Retrieval therefore provides a more practical base annotator and, importantly, reusable embeddings from which target-pool structure can be estimated.

Finding 2: semantic anchors and target visual evidence are complementary. Using ground truth only for diagnosis, we average fit-split images of each class into an oracle representative $\mathbf{p}_c^\star$ (Figure 2). It outperforms the text anchor on five of six datasets, while fusion outperforms both components on every tested configuration, including Food-101 where the visual representative alone is worse. For SigLIP2, the anchor and oracle representative are also geometrically distinct, with cosine 0.11–0.16. The gain persists with more data: with 250 oracle images per class, fusion still adds 1.4 points (Appendix H). Thus, the class name captures semantic meaning while the target pool contributes complementary visual evidence.

Finding 3: standard prototyping fails in three ways. A standard label-free recipe takes zero-shot predictions as pseudo-labels, keeps confident images, and averages them into class prototypes. We observe three failures. *(a) Distorted marginals.* TinyImageNet is exactly balanced, yet zero-shot

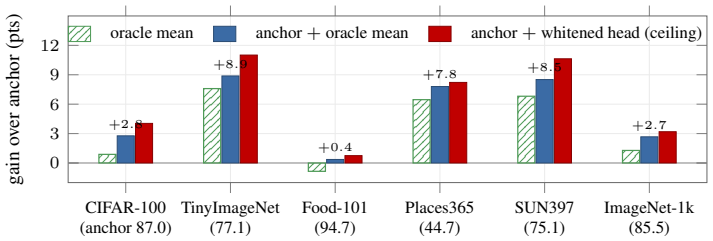

Figure 2: **Semantic and visual evidence are complementary.** Oracle class means improve over text anchors on five of six pools, while their fusion improves over both on every tested pool.

predictions assign 15–2,657 images per class, a 177× spread, despite 77% purity. *(b) Confidence filtering biases the representative.* Top-margin selection raises purity but lowers final accuracy relative to a random subset of equal size. The same holds with perfect labels ($p \leq 2.7 \times 10^{-4}$ over 10 caches), showing that the loss is selection bias rather than label noise alone (Appendix I). In contrast, averaging all assigned images attenuates pseudo-label noise: a 23% label error changes the representative by only $5.5°$ ($1 - \cos = 0.5\%$). *(c) Uniform balancing fails under imbalance.* Forcing equal class sizes raises TinyImageNet purity from 77% to 83%, but lowers SUN397 purity from 75% to 66%. Under 100:1 imbalance, our uniform-prior variant loses nearly all of its gain (Section 4.5). The target marginal must therefore be estimated rather than assumed.

Finding 4: class means are poorly conditioned. The target embeddings are strongly anisotropic: the leading eigenvalue of their second moment accounts for 48–75% of the spectral mass across our six pools. Consequently, raw class means are highly aligned (mean pairwise cosine 0.71–0.88 for SigLIP2). Reading the same class statistic through the target second moment, $(G + \alpha I)^{-1}\mathbf{S}_c$, suppresses the shared direction and yields nearly orthogonal class vectors ($|\cos| \leq 0.01$; Appendix F). With identical ground-truth labels, this whitened head improves average accuracy from 77.0% to 78.8% over 18 encoder–pool pairs; with zero-shot labels it adds 1.0 point. Together, these findings suggest a simple design: repair the assignment marginal without assuming balance, retain all assigned images, condition their class statistics using the target pool, and fuse the resulting visual evidence with the semantic anchor.

Table 1: From diagnosis to design: each method component addresses a failure identified in Section 2. *Evidence* reports controlled comparisons that isolate each design choice, while *Theory* points to the corresponding analysis in Appendix F.

Failure (finding in §2)	Design decision (in §3)	Evidence	Theory
arg-max distorts class sizes (3a)	balanced assignment (2a)	purity 77.1 to 82.9% (Fig. 3)	Prop. 1
uniform balance breaks (3c)	estimated prior (2a)	162/162 stress cells (§4.5)	–
filtering biases the mean (3b)	keep every image	random subset wins (App. I.2)	Lem. 1, 2
class means are collinear (4)	whitened head (2b)	+1.8 with true labels (§4.4)	Prop. 2, 4
anchor and pool complement (2)	fusion with anchor (3)	beats both inputs (Fig. 2)	Thm. 1

3 Method: Robust Prototyping by Semantic–Visual Calibration

The prototype recipe of Finding 3 belongs to a family of procedures that assign images to classes, summarize each class, and fuse the summary with the anchor. Our annotator keeps this structure but replaces each step that the diagnosis showed to fail; Table 1 pairs every design decision with the failure it addresses and the experiment that isolates it, and Figure 3 gives an overview. Stage 1 is the scoring step, in which the anchors organize the pool through $S = ZT^\top$, computed once. Stage 2 recovers the visual evidence without labels: it repairs the assignment marginal onto an estimated prior (Findings 3a and 3c), keeps every image (Finding 3b), and reads the class statistics through the second moment of the pool (Finding 4). Stage 3 fuses this evidence with the anchor (Finding 2).

Stage 2a: balanced assignment repairs label a marginal. The arg-max of S yields class sizes that reflect how each anchor attracts images rather than the actual composition of the pool. We replace it with a *balanced* assignment: the coupling P that is closest in KL divergence to $\mathrm{softmax}(S/\varepsilon)$

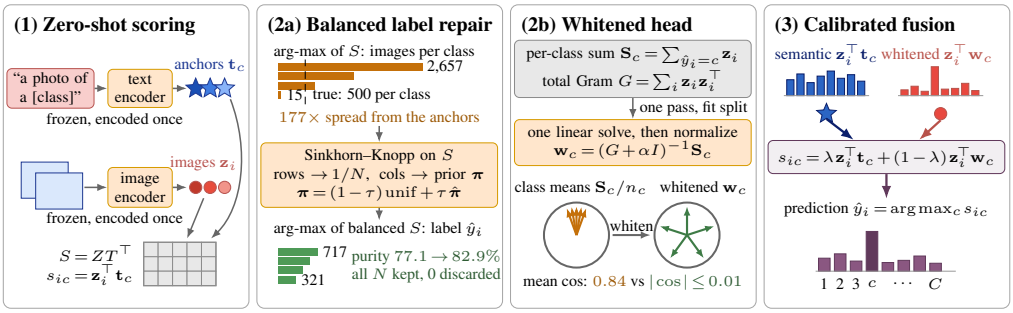

Figure 3: **Overview of the annotator.** (1) Frozen encoders embed the class names (anchors $\mathbf{t}_c$) and the pool ($\mathbf{z}_i$) once, giving $S = ZT^\top$. (2a) Balanced transport of S onto an estimated class prior replaces the arg-max and keeps every image (TinyImageNet, uniform prior: class sizes go from 15–2,657 to 321–717 and purity from 77.1% to 82.9%). (2b) A single closed-form solve yields the whitened head $\mathbf{w}_c$, whose vectors are near-orthogonal even where the class means are not. (3) The head score is fused with the anchor score.

among couplings whose rows sum to $1/N$ and whose columns sum to a class prior $\boldsymbol{\pi}$,

$$P^\star = \arg\min_{P\mathbf{1}=\mathbf{1}/N,\; P^\top\mathbf{1}=\boldsymbol{\pi}} \mathrm{KL}\big(P \,\|\, \mathrm{softmax}(S/\varepsilon)\big), \qquad \hat{y}_i = \arg\max_c P^\star_{ic}, \tag{1}$$

which we compute with 50 Sinkhorn–Knopp iterations (Cuturi, 2013) in log space on the full $N \times C$ matrix, at a fixed temperature $\varepsilon^{-1} = 100$, following self-labeling (Asano et al., 2020; Caron et al., 2020) and transductive CLIP adaptation (Qian et al., 2023; Qiu & Ren, 2026). The prior is $\boldsymbol{\pi}_\tau = (1-\tau)\,\mathrm{uniform} + \tau\,\hat{\boldsymbol{\pi}}$, where $\hat{\boldsymbol{\pi}}$ is the arg-max marginal: $\tau = 0$ enforces exact balance and $\tau = 1$ retains the estimated proportions. This prior determines robustness. On the balanced pools the uniform prior works best, but on SUN397 it lowers purity to 66% where the estimated prior reaches 77%, and under an imposed 100:1 imbalance the estimated prior preserves a gain that the uniform prior loses (Section 4.5). No image is discarded. Equation 1 is the information projection of the zero-shot posterior onto the couplings with the prescribed marginal (Proposition 1); with the uniform prior on TinyImageNet it reduces the spread of class sizes from 177× to 2.2×. With τ=1 the target marginal coincides with the arg-max marginal, so on a balanced pool the repair only reassigns images within it (purity 77.5% versus 77.1%) and most of the label-free gain comes from the head; the estimated prior matters on imbalanced pools (Section 4.5).

Stage 2b: the whitened head uses every assigned image. Given the labels $\hat{y}$, the obvious representative is the class mean. We use instead the ridge solution of the one-hot regression on the fit split,

$$\mathbf{w}_c = (G + \alpha I)^{-1}\mathbf{S}_c, \qquad G = \sum_i \mathbf{z}_i \mathbf{z}_i^\top, \qquad \mathbf{S}_c = \sum_{i:\hat{y}_i=c} \mathbf{z}_i, \tag{2}$$

with both sums over the fit split and each $\mathbf{w}_c$ then L_2-normalized. This is a *whitened* class mean: $\mathbf{S}_c/n_c$ is the mean of the n_c images assigned to class c, G is the label-free second moment of the fit split, and as $\alpha \to \infty$ the direction of $\mathbf{w}_c$ converges to the class mean (Proposition 2), so prototype averaging is one endpoint of the same family. Whitening addresses Finding 4 directly: it removes the shared direction that makes class means nearly collinear (mean pairwise cosine 0.71–0.88) and leaves representatives within 0.01 of orthogonal on every pool (Table 12; Proposition 4 explains the mechanism). The head is consistent for fixed regularization (Proposition 3) and outperforms a linear head trained for 300 epochs on the same pseudo-labels on 17 of 18 pairs, at 18× lower compute (Section 4.5).

Stage 3: fusion with the anchor. The final score combines anchor and representative,

$$s_{ic} = \lambda\, \mathbf{z}_i^\top \mathbf{t}_c + (1-\lambda)\, \mathbf{z}_i^\top \mathbf{w}_c, \qquad \hat{y}_i^{\mathrm{final}} = \arg\max_c s_{ic}. \tag{3}$$

The anchor encodes what the class name means, which the pool alone cannot recover (Food-101 in Figure 2); the head encodes how the class appears in this pool. Under the assumptions of Appendix F, the λ that minimizes a mean-squared surrogate of the risk satisfies $\mathrm{odds}(\lambda^\star) = A + B/n$, an inverse-sample-size law with a bias floor (Theorem 1) that fits the measured accuracy optimum over $n \in [1, 250]$ images per class with $R^2 = 0.97$; the optimum is broad and interior for every pair.

Table 2: Test accuracy (%, mean over 5 seeds, std ≤ 0.7 in all cells, Appendix I.1; fit-split protocol, policies of Section 4.1). Bold: best per column; blue: gain over zero-shot. Prototype rows are label-free (validated λ: within 0.2). † DINOv2 uses SigLIP2 pseudo-labels and anchors.

Encoder	method	CIFAR-100	TinyIN	Food-101	SUN397	Places365	ImageNet-1k	mean
SigLIP2	zero-shot	87.03	77.07	94.72	75.05	44.74	85.50	77.35
	prototype (label-free)	88.16	80.67	94.84	78.65	47.21	86.74	79.38
	ours (label-free)	88.66	82.24	95.25	80.28	46.94	86.88	80.04
	ours (validated)	**90.18** +3.15	**85.33** +8.26	**95.40** +0.69	**80.75** +5.70	**48.76** +4.01	**87.76** +2.26	**81.36** +4.01
CLIP ViT-B/16	zero-shot	69.00	63.50	84.69	60.88	40.20	68.10	64.40
	prototype (label-free)	70.82	66.43	85.42	65.45	42.62	69.98	66.79
	ours (label-free)	72.18	67.17	86.15	67.68	42.52	69.48	67.53
	ours (validated)	**74.63** +5.63	**70.72** +7.21	**86.54** +1.85	**68.05** +7.17	**44.21** +4.01	**72.14** +4.04	**69.38** +4.99
CLIP ViT-L/14	zero-shot	77.06	74.50	90.47	66.47	40.87	76.15	70.92
	prototype (label-free)	79.33	77.95	91.09	70.80	43.68	78.69	73.59
	ours (label-free)	80.85	78.90	91.39	73.40	43.69	78.80	74.50
	ours (validated)	**82.18** +5.12	**80.94** +6.44	**91.70** +1.23	**73.42** +6.94	**44.89** +4.02	**80.46** +4.31	**75.60** +4.68
DINOv2-base†	prototype (validated λ)	90.42	84.38	94.68	78.35	47.65	86.30	80.30
	ours (validated)	**91.26**	**86.09**	**94.76**	**78.37**	**48.43**	**86.63**	**80.92**

Hyper-parameters, label-free policy, and cost. The method has three scalars, α, τ and λ. The *label-free* policy fixes $\alpha = 1$, uses the estimated prior ($\tau = 1$) because the class balance of the pool is unknown, and transfers λ from other pools, so it uses no label of the target pool. The *validated* policy, reported as a reference, selects all three on a labeled quarter of the target pool; the label-free policy trails it by 1.4 points (Section 4.2; both policies are defined in Section 4.1 and Appendix D). After encoding, the annotator consists of one $N \times C$ product, 50 Sinkhorn iterations, one pass to accumulate G and $\{\mathbf{S}_c\}$, and one $D \times D$ solve, about twelve seconds on ImageNet-1k.

4 Experiments

4.1 Setup

Pools and backbones. We use the training sets of CIFAR-100 (C=100), TinyImageNet (200), Food-101 (101), SUN397 (397, 24× imbalanced), Places365 (365) and ImageNet-1k (1000) (Krizhevsky, 2009; Le & Yang, 2015; Bossard et al., 2014; Xiao et al., 2010; Zhou et al., 2018; Deng et al., 2009), each encoded once with SigLIP2 ViT-gopt/16-384 (Tschannen et al., 2025), CLIP ViT-B/16 and ViT-L/14@336 ("ViT-L/14" below) (Radford et al., 2021), and DINOv2-base (Oquab et al., 2024), the last annotated with the labels of SigLIP2; the three text-aligned encoders form the "CLIP family" (18 encoder–dataset pairs).

Protocols. The annotator only sees the unlabeled pool and the class names. In the *transductive* protocol a set of images is labeled from those images alone, as in deployment and in the published methods. In the *fit-split* protocol, used for controlled comparisons, every class is divided 50/25/25 into fit, validation and test images (five seeds); balanced assignment runs on the whole pool without labels, heads are built on the fit images, and accuracy is measured on the test images.

Policies. The *label-free* policy uses no label of the target pool (α=1, τ=1, λ transferred from other pools); the *validated* policy selects (α, τ, λ) on the labeled validation images of the target pool and is a reference for what target labels would add. The baseline is the prototype recipe (arg-max labels, class mean, fusion) under the same two policies. We report means over seeds and win counts over pairs; Appendix C gives pool sizes, splits, hyper-parameters, hardware and significance tests, and Appendix D the protocol and policy behind every result.

4.2 Main results

Table 2 reports the main result. The validated annotator outperforms the prototype recipe on all 24 backbone–dataset pairs (the DINOv2 gains on SUN397 and Food-101 are within one standard deviation) and zero-shot by $+0.7$ to $+8.3$ points on the 18 CLIP-family pairs. The gain is largest where the anchors are weakest: on TinyImageNet ($+6.4$ to $+8.3$), where the arg-max marginal is most distorted, and on the imbalanced SUN397 ($+5.7$ to $+7.2$); it is smallest on Food-101, where zero-shot accuracy is already highest. On average the method captures 64% of the per-cache headroom to the oracle-label whitened head (Figure 11). The label-free policy, which assumes nothing about the

Table 3: Published training-free methods on identical caches, anchors and splits (test accuracy %): means over the six pools of Table 2, and ImageNet-LT and Places-LT (Liu et al., 2019) from the same caches. Transductive rows label the unlabeled test split (TDA as a stream); fit-split rows follow Table 2. All rows but the last (validated, a reference) use no label of the target pool; best label-free value per column in bold. ‡Released components, template prompt.

	six pools, mean			ImageNet-LT (256×, C=1000)			Places-LT (996×, C=365)		
Method	SigLIP2	ViT-B/16	ViT-L/14	SigLIP2	ViT-B/16	ViT-L/14	SigLIP2	ViT-B/16	ViT-L/14
zero-shot retrieval	77.39	64.40	70.94	86.65	69.05	77.36	50.09	43.28	43.55
TDA (Karmanov et al., 2024), streamed	78.71	66.30	72.34	86.89	69.62	77.87	49.86	43.13	42.80
TransCLIP (Zanella et al., 2024)	78.94	67.51	73.98	85.77	70.34	78.31	48.29	43.62	43.13
Frolic (Zhu et al., 2024), released core‡	79.36	67.41	73.65	86.79	68.90	77.95	46.67	38.01	38.91
InMaP (Qian et al., 2023), uniform prior	79.55	**68.44**	74.29	70.96	59.03	64.70	32.64	31.00	30.82
InMaP, estimated prior	78.56	65.73	72.76	87.37	71.19	79.14	51.93	45.30	46.35
OTTER (Shin et al., 2024), estimated prior	77.93	64.67	71.54	87.38	70.91	78.62	52.16	47.08	47.36
ours, transductive, estimated prior	80.01	67.59	74.33	87.96	72.35	80.19	53.06	48.70	48.87
ours, fit split, label-free	**80.07**	67.53	**74.50**	**88.25**	**72.42**	**80.73**	**53.19**	**48.83**	**49.33**
ours, fit split, validated	81.36	69.38	75.60	88.41	72.78	80.78	53.36	48.73	49.33

Table 4: Build-up on identical caches and splits (test accuracy %, mean over 3 CLIP-family backbones × 5 seeds; last columns: wins and paired-t p of the full method against each row over 18 pairs). All rows but zero-shot and "whitened head alone" fuse with the anchor (validated λ); arg-max/balanced rows use zero-shot/Stage-2a labels. Oracle ceilings: Table 21.

Method	TinyIN	CIFAR-100	SUN397	Food-101	Places365	ImageNet-1k	mean	wins	p
zero-shot (anchor only)	71.64	77.75	67.39	90.01	41.93	76.56	70.88	18/18	$1 \cdot 10^{-7}$
prototype recipe	75.14	79.46	71.65	90.53	44.54	78.52	73.31	18/18	$6 \cdot 10^{-7}$
+ top-25% margin filter	74.46	78.95	71.13	90.22	44.31	77.99	72.84	18/18	$2 \cdot 10^{-7}$
+ balanced labels (2a)	76.85	80.55	72.17	90.69	45.26	79.22	74.12	18/18	$2 \cdot 10^{-6}$
whitened head on arg-max (2b)	76.86	80.95	73.04	90.98	44.81	78.96	74.27	18/18	$7 \cdot 10^{-7}$
whitened head alone (2a + 2b, no fusion)	78.42	82.25	73.18	90.09	45.00	78.64	74.60	17/18	$5 \cdot 10^{-5}$
ours (2a + 2b + fusion)	**78.99**	**82.33**	**74.07**	**91.21**	**45.95**	**80.12**	**75.45**	–	–

class balance of the pool, beats zero-shot on all 18 pairs and the label-free prototype recipe on 15, and trails the validated policy by 1.4 points. The validated policy is a reference rather than a recommended use of labels: a whitened head trained on the same labeled quarter beats it on all 18 pairs, by 3.0 points (Figure 10). We therefore recommend the label-free configuration for deployment.

4.3 Comparison with published training-free methods

We also run the closest published training-free methods from their official code on our cached embeddings (Table 3): TransCLIP (Zanella et al., 2024), InMaP (Qian et al., 2023), Frolic (Zhu et al., 2024), TDA (Karmanov et al., 2024) and OTTER (Shin et al., 2024) (Section 5; reproduction details in Appendix I.12). Given the same unlabeled test split and, for the prior-based methods, an estimated prior, our closed-form annotator has the highest mean accuracy: it is ahead of InMaP, OTTER and TDA on all 18 encoder–pool pairs, of TransCLIP on 13 and of Frolic on 10 (Table 24), and InMaP trails by 1.6 points (the settings differ, Appendix I; Table 4 isolates the head). With its published uniform prior, InMaP is the strongest competitor on the near-balanced pools (best ViT-B/16 mean; ahead of our label-free policy on 15 of 18 pairs, the difference of about one point being due to the closed-form head, Table 4). However, the same uniform prior collapses on the imbalanced SUN397 pool (69.5% versus 75.3% zero-shot with SigLIP2, where ours reaches 80.3%) and loses 10 to 17 points on the long-tailed benchmarks, where every method with the estimated prior stays above zero-shot; ours is ahead on all six columns. OTTER shows the value of the prior alone: near zero-shot with the estimated marginal, 74.7 with the true test proportions (an oracle, not tabulated).

4.4 Ablation, composition, and scaling with the vocabulary

Table 4 isolates each design decision. *Filtering hurts* (row 3): keeping only the top-margin quarter lowers accuracy on all six datasets. *Balanced labels help both heads* (rows 4 and 7): +0.81 for the class mean and +1.18 for the whitened head. *The whitened head helps on its own* (row 5, +0.96 with identical labels), and *fusion adds a further* +0.85 (rows 6 and 7). With the label-free prior (τ=1), balanced labels add only +0.08 on these near-balanced pools, so the gain there comes from the head (Table 20). *The two ingredients compose*: on SigLIP2/TinyImageNet, balanced labels alone add 1.9 points, the whitened head alone 2.3, and both 4.3 (Figure 11); with oracle labels the whitened head

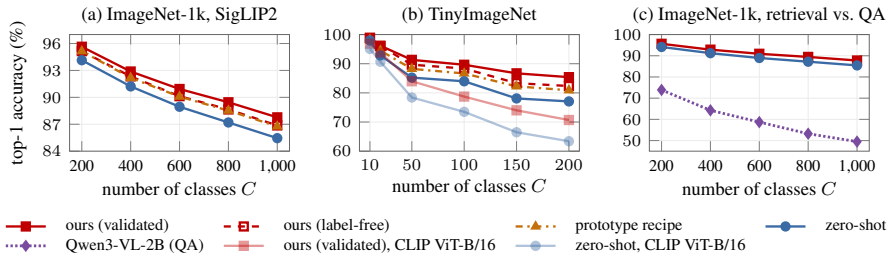

Figure 4: **The gain grows with the vocabulary.** Test accuracy on nested class subsets (3 seeds): (a) ImageNet-1k, SigLIP2; (b) TinyImageNet, SigLIP2 (faint curves: CLIP ViT-B/16); (c) ImageNet-1k on the full range, with the QA annotator (Qwen3-VL-2B).

raises the ceiling itself (77.0 to 78.8, Table 21), and the annotator built from pseudo-labels is ahead of a class mean built from *ground-truth* labels on 7 of 18 pairs. *Three plausible refinements do not help*: self-training on the labels of the annotator itself degrades monotonically over five rounds (79.8 to 78.6 on SigLIP2/TinyImageNet), soft weighting by confidence is neutral, and orthogonalizing the representative against the anchor is harmful (62.7 to 54.7 for CLIP ViT-B/16), so the text-aligned component of the representative carries signal rather than bias (Appendices I and H).

Scaling with the vocabulary. Figure 4 reruns the annotator on the nested class subsets of Section 2. Within a pool, the gain over zero-shot grows with the vocabulary: on ImageNet-1k from $+1.5$ to $+2.3$ points for SigLIP2 as C goes from 200 to 1,000 (larger for the CLIP encoders, Appendix I.11), and on TinyImageNet from $+0.8$ to $+8.3$ as C goes from 10 to 200. More competing classes make the distorted arg-max marginal and the collinear class means costlier; across pools, the accuracy of the anchors matters more than C. At $C{=}1000$ the annotator reaches 87.8%, whereas the QA annotator drops to 49.5%.

4.5 Imbalanced and incomplete pools: what the class prior does; cost

A real imbalanced pool. Figure 15 (Appendix I) splits the classes of SUN397 into thirds by size. The prototype recipe improves every third by a similar amount; the prior is what changes the picture. Forcing equal class sizes pushes majority-class images into minority classes (purity drops from 75% to 66%), which helps the tail third and, for SigLIP2, leaves the head third at the level of the prototype; the estimated prior does the opposite ($+2.7$ on the head third, -1.1 on the tail), and overall accuracy follows the head classes.

Two long-tailed benchmarks. The same pattern appears on ImageNet-LT and Places-LT (Table 3): InMaP, OTTER and our own uniform-prior variant fall below zero-shot (Table 25), whereas ours with the estimated prior is ahead of every published method on all six columns without labels; OTTER given the true test proportions, an oracle, is still 2 to 8 points higher (Appendix I.13). The gain is in overall accuracy, which the head classes dominate: macro accuracy stays within 1.4 points of the prototype recipe, and the tail third trails the uniform prior by 1.7 to 6.3 points (Figure 17); when rare classes matter, τ should be reduced.

A stress test. We subsample all 18 caches so that the largest class has $\rho \in \{1, 10, 100\}$ times as many images as the smallest and only a fraction $f \in \{25\%, 50\%, 100\%\}$ of the classes is present; the annotator still scores all C classes and is not told which are absent (3 seeds; Figure 14, Appendix I.7). With τ validated, the annotator beats zero-shot in all 162 condition–cache cells, with a mean gain of at least $+3.6$ points per condition (prototype recipe: at most $+2.4$), while the uniform prior loses most of its gain at $\rho{=}100$. The published methods behave the same way: with their uniform prior, InMaP, TransCLIP and Frolic lose up to 31 points on average when three quarters of the classes are missing, whereas with the estimated prior InMaP gains $+2.2$ on average and ours, label-free and transductive, $+3.9$, above zero-shot in all 162 cells (Table 26).

Cost and transfer. The closed-form head outperforms a linear head trained by SGD for 300 epochs on the same pseudo-labels, by 0.63 points on 17 of 18 caches at $18\times$ lower compute (Appendix I.9), and the whole annotator runs in about twelve seconds once the pool is encoded, which takes six GPU-hours for ImageNet-1k with SigLIP2 (Section 2). The labels of SigLIP2 can also supervise DINOv2-base, which has no text tower: its whitened head alone reaches 83.2% on TinyImageNet,

above the 77.1% zero-shot accuracy of the model that produced the labels, and fused with the anchors of SigLIP2 it reaches 86.1%, above the 85.3% of SigLIP2 itself (Table 2). The pseudo-labels thus carry more than the accuracy of the labeler itself, which a stronger visual representation can exploit (Appendix I).

5 Related Work

Foundation models as annotators. Retrieval-based VLMs (Radford et al., 2021; Zhai et al., 2023; Tschannen et al., 2025) rank images against class-name embeddings. QA-based VLMs (Liu et al., 2023; 2024a; Bai et al., 2025) answer a prompt that lists the candidates and thereby inherit the position and list-length effects documented for language models (Zheng et al., 2024).

Training-free and transductive adaptation of CLIP. TransCLIP (Zanella et al., 2024) fits a Gaussian mixture anchored to the text prior; Martin et al. (2024) fit a Dirichlet mixture on small batches; Frolic (Zhu et al., 2024) keeps the text embeddings as class means, estimates a covariance from the pool, and fuses logits; InMaP (Qian et al., 2023) learns vision-space proxies by gradient descent from Sinkhorn-balanced, thresholded pseudo-labels; IOTA (Qiu & Ren, 2026) refines prototypes from unbalanced-transport pseudo-labels; and the Gaussian-discriminant baseline of Wang et al. (2024) fits $\Sigma^{-1}\boldsymbol{\mu}_c$ heads. Gradient-based label-free tuning appears in UPL (Huang et al., 2022), LaFTer (Mirza et al., 2023), DPA (Ali et al., 2025) and Menghini et al. (2023). Sinkhorn–Knopp balancing (Cuturi, 2013) underlies self-labeling (Asano et al., 2020; Caron et al., 2020), whereas DebiasPL (Wang et al., 2022) corrects pseudo-label imbalance with an adaptive margin. Our setting is the one these methods share in zero-shot mode: no label of the target data (Appendix D). We share balanced assignment with InMaP and IOTA and score fusion with Frolic, but keep every image and solve once over the whole pool in closed form. The class prior is a further difference: TransCLIP, InMaP, Frolic and IOTA assume uniform proportions by default and OTTER (Shin et al., 2024) takes them from labels, whereas we estimate the prior from the unlabeled pool, motivated by the diagnosis in Section 2.

Label distribution and annotation. OTTER (Shin et al., 2024) transports the zero-shot scores of a test set onto a class marginal, in its experiments the true one; IOTA (Qiu & Ren, 2026) uses unbalanced transport and SOTA (Hu et al., 2026) transport across several foundation models. We use the same transport but estimate the marginal and add the whitened head. Annotation itself has been studied as budgeting (LabelBench, Zhang et al., 2024b), as VLM weak labels for active learning (Nguyen et al., 2026) and for complementary-label data at scale (Mai et al., 2025) (human-annotated counterparts: Wang et al., 2025), and as self-learning from class names (Todescato & Carbonera, 2025).

Test-time and few-shot adaptation with caches and prototypes. TDA (Karmanov et al., 2024), DMN (Zhang et al., 2024d), BoostAdapter (Zhang et al., 2024c), DPE (Zhang et al., 2024a), ECALP (Li et al., 2025), ProtoMM (Zhu et al., 2025) and PTA (Huang et al., 2026) adapt along a test stream using caches, graphs or online prototypes built from arg-max labels and confidence filtering (Section 2); TPT (Shu et al., 2022) tunes prompts per sample. Tip-Adapter (Zhang et al., 2022) adds the logits of a cache of few-shot labeled features to the zero-shot logits, the fusion closest to ours, which instead draws the visual side from the unlabeled pool.

6 Conclusion

Many-class annotation with unknown and imbalanced class frequencies is a ranking problem over a fixed vocabulary. Retrieval-based VLMs suit this setting better than QA-based models, whose computational cost increases and annotation accuracy decreases as the candidate set grows. Our central observation is that two complementary signals should be combined: class names provide stable semantic anchors, and the unlabeled pool reveals how the classes look and how often they occur. Standard prototyping discards part of this information through confidence filtering, uniform balancing, and poorly conditioned class means; our robust prototyping replaces these steps with marginal-constrained assignment under an estimated prior, a whitened head over all assigned images, and fusion with the anchor. Without target-pool labels or gradient training, it improves over zero-shot on all 18 CLIP-family encoder–dataset pairs, all 162 stress-test cells and both long-tailed benchmarks, in seconds after encoding.

ETHICS STATEMENT

All experiments use public benchmarks—CIFAR-100, TinyImageNet, Food-101, SUN397, Places365, ImageNet-1k, ImageNet-LT and Places-LT—and public model checkpoints. We collected no new data and involved no human annotators. Because the method automates part of the labeling process, its outputs should be treated as pseudo-labels rather than ground truth, and since they are derived from pretrained VLMs, they may inherit the biases of those models. We also note a specific failure mode under class imbalance: the estimated prior improves head-class accuracy at the expense of tail-class recall (Figure 17). In addition, the method always assigns each image to one of the candidate classes, even when none fits well. Rare classes and high-stakes cases should therefore be audited, with human review retained wherever an incorrect label could affect people. The computational footprint is modest: encoding a pool takes a few GPU-hours, and each annotation run after encoding takes seconds.

REPRODUCIBILITY STATEMENT

All encoders are public checkpoints and every result is a closed-form computation on cached embeddings. Appendix C lists datasets, checkpoints, prompts, splits, seeds and hardware, and Appendix E describes the code, which regenerates every table and figure.

AI USE STATEMENT

We used large language models solely to assist with writing and code maintenance. For writing, they were used to check typos, grammar, and phrasing. For code, they were used for minor cleaning, refactoring, and readability improvements, without contributing to the implementation of research methods or experimental design. They were not used to generate research ideas or design experiments; all research contributions were made entirely by the authors, who reviewed the AI-assisted outputs and took full responsibility for the content of the paper and the associated code.

REFERENCES

Guillaume Alain and Yoshua Bengio. Understanding intermediate layers using linear classifier probes. In *International Conference on Learning Representations, Workshop Track*, 2017.

Eman Ali, Sathira Silva, and Muhammad Haris Khan. DPA: Dual prototypes alignment for unsupervised adaptation of vision-language models. In *IEEE/CVF Winter Conference on Applications of Computer Vision*, 2025.

Yuki Markus Asano, Christian Rupprecht, and Andrea Vedaldi. Self-labelling via simultaneous clustering and representation learning. In *International Conference on Learning Representations*, 2020.

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, Wenbin Ge, Zhifang Guo, Qidong Huang, Jie Huang, Fei Huang, Binyuan Hui, Shutong Jiang, Zhaohai Li, Mingsheng Li, Mei Li, Kaixin Li, Zicheng Lin, Junyang Lin, Xuejing Liu, Jiawei Liu, Chenglong Liu, Yang Liu, Dayiheng Liu, Shixuan Liu, Dunjie Lu, Ruilin Luo, Chenxu Lv, Rui Men, Lingchen Meng, Xuancheng Ren, Xingzhang Ren, Sibo Song, Yuchong Sun, Jun Tang, Jianhong Tu, Jianqiang Wan, Peng Wang, Pengfei Wang, Qiuyue Wang, Yuxuan Wang, Tianbao Xie, Yiheng Xu, Haiyang Xu, Jin Xu, Zhibo Yang, Mingkun Yang, Jianxin Yang, An Yang, Bowen Yu, Fei Zhang, Hang Zhang, Xi Zhang, Bo Zheng, Humen Zhong, Jingren Zhou, Fan Zhou, Jing Zhou, Yuanzhi Zhu, and Ke Zhu. Qwen3-VL technical report. *arXiv preprint arXiv:2511.21631*, 2025.

Jean-David Benamou, Guillaume Carlier, Marco Cuturi, Luca Nenna, and Gabriel Peyré. Iterative Bregman projections for regularized transportation problems. *SIAM Journal on Scientific Computing*, 37(2):A1111–A1138, 2015.

Lukas Bossard, Matthieu Guillaumin, and Luc Van Gool. Food-101 – mining discriminative components with random forests. In *European Conference on Computer Vision*, pp. 446–461, 2014.

Mathilde Caron, Ishan Misra, Julien Mairal, Priya Goyal, Piotr Bojanowski, and Armand Joulin. Unsupervised learning of visual features by contrasting cluster assignments. In *Advances in Neural Information Processing Systems*, volume 33, 2020.

Marco Cuturi. Sinkhorn distances: Lightspeed computation of optimal transport. In *Advances in Neural Information Processing Systems*, volume 26, pp. 2292–2300, 2013.

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.

Zhanxuan Hu, Qiyu Xu, Yu Duan, Yonghang Tai, and Huafeng Li. SOTA: Self-adaptive optimal transport for zero-shot classification with multiple foundation models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2026.

Tony Huang, Jack Chu, and Fangyun Wei. Unsupervised prompt learning for vision-language models. *arXiv preprint arXiv:2204.03649*, 2022.

Zhaohong Huang, Yuxin Zhang, Wenjing Liu, Fei Chao, and Rongrong Ji. Prototype-based test-time adaptation of vision-language models. In *International Conference on Machine Learning*, 2026.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. Mistral 7b. *arXiv preprint arXiv:2310.06825*, 2023.

Adilbek Karmanov, Dayan Guan, Shijian Lu, Abdulmotaleb El Saddik, and Eric Xing. Efficient test-time adaptation of vision-language models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024.

Alex Krizhevsky. Learning multiple layers of features from tiny images. Technical report, University of Toronto, 2009.

Ya Le and Xuan Yang. Tiny ImageNet visual recognition challenge. Technical report, Stanford University, CS231N course report, 2015.

Yushu Li, Yongyi Su, Adam Goodge, Kui Jia, and Xun Xu. Efficient and context-aware label propagation for zero-/few-shot training-free adaptation of vision-language model. In *International Conference on Learning Representations*, 2025.

Weixin Liang, Yuhui Zhang, Yongchan Kwon, Serena Yeung, and James Zou. Mind the gap: Understanding the modality gap in multi-modal contrastive representation learning. In *Advances in Neural Information Processing Systems*, volume 35, 2022.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. In *Advances in Neural Information Processing Systems*, volume 36, 2023.

Haotian Liu, Chunyuan Li, Yuheng Li, Bo Li, Yuanhan Zhang, Sheng Shen, and Yong Jae Lee. LLaVA-NeXT: Improved reasoning, OCR, and world knowledge. https://llava-vl.github.io/blog/2024-01-30-llava-next/, 2024a.

Nelson F. Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. Lost in the middle: How language models use long contexts. *Transactions of the Association for Computational Linguistics*, 12:157–173, 2024b.

Ziwei Liu, Zhongqi Miao, Xiaohang Zhan, Jiayun Wang, Boqing Gong, and Stella X. Yu. Large-scale long-tailed recognition in an open world. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019.

Tan-Ha Mai, Nai-Xuan Ye, Yu-Wei Kuan, Po-Yi Lu, and Hsuan-Tien Lin. The unexplored potential of vision-language models for generating large-scale complementary-label learning data. In *Pacific-Asia Conference on Knowledge Discovery and Data Mining*, pp. 90–102, 2025.

Ségolène Martin, Yunshi Huang, Fereshteh Shakeri, Jean-Christophe Pesquet, and Ismail Ben Ayed. Transductive zero-shot and few-shot CLIP. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024.

Cristina Menghini, Andrew Delworth, and Stephen H. Bach. Enhancing CLIP with CLIP: Exploring pseudolabeling for limited-label prompt tuning. In *Advances in Neural Information Processing Systems*, volume 36, 2023.

M. Jehanzeb Mirza, Leonid Karlinsky, Wei Lin, Mateusz Kozinski, Horst Possegger, Rogerio Feris, and Horst Bischof. LaFTer: Label-free tuning of zero-shot classifier using language and unlabeled image collections. In *Advances in Neural Information Processing Systems*, volume 36, 2023.

Phuong Ngoc Nguyen, Kaito Shiku, Ryoma Bise, Seiichi Uchida, and Shinnosuke Matsuo. Leveraging vision-language models as weak annotators in active learning. In *IEEE International Conference on Image Processing*, 2026.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy 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.

Qi Qian, Yuanhong Xu, and Juhua Hu. Intra-modal proxy learning for zero-shot visual categorization with CLIP. In *Advances in Neural Information Processing Systems*, volume 36, 2023.

Shupeng Qiu and Chuan-Xian Ren. Inverse optimal transport for efficient adaptation of vision-language models. In *AAAI Conference on Artificial Intelligence*, volume 40, pp. 25018–25026, 2026.

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, 2021.

Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. ImageNet large scale visual recognition challenge. *International Journal of Computer Vision*, 115 (3):211–252, 2015.

Changho Shin, Jitian Zhao, Sonia Cromp, Harit Vishwakarma, and Frederic Sala. OTTER: Effortless label distribution adaptation of zero-shot models. In *Advances in Neural Information Processing Systems*, 2024.

Manli Shu, Weili Nie, De-An Huang, Zhiding Yu, Tom Goldstein, Anima Anandkumar, and Chaowei Xiao. Test-time prompt tuning for zero-shot generalization in vision-language models. In *Advances in Neural Information Processing Systems*, volume 35, 2022.

Matheus Vinícius Todescato and Joel Luís Carbonera. No labels needed: Zero-shot image classification with collaborative self-learning. In *IEEE International Conference on Tools with Artificial Intelligence*, 2025.

Joel A. Tropp. An introduction to matrix concentration inequalities. *Foundations and Trends in Machine Learning*, 8(1–2):1–230, 2015.

Michael Tschannen, Alexey Gritsenko, Xiao Wang, Muhammad Ferjad Naeem, Ibrahim Alabdulmohsin, Nikhil Parthasarathy, Talfan Evans, Lucas Beyer, Ye Xia, Basil Mustafa, Olivier Hénaff, Jeremiah Harmsen, Andreas Steiner, and Xiaohua Zhai. SigLIP 2: Multilingual vision-language encoders with improved semantic understanding, localization, and dense features. *arXiv preprint arXiv:2502.14786*, 2025.

Hsiu-Hsuan Wang, Mai Tan Ha, Nai-Xuan Ye, Wei-I Lin, and Hsuan-Tien Lin. CLImage: Human-annotated datasets for complementary-label learning. *Transactions on Machine Learning Research*, 2025.

Xudong Wang, Zhirong Wu, Long Lian, and Stella X. Yu. Debiased learning from naturally imbalanced pseudo-labels. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14647–14657, 2022.

Zhengbo Wang, Jian Liang, Lijun Sheng, Ran He, Zilei Wang, and Tieniu Tan. A hard-to-beat baseline for training-free CLIP-based adaptation. In *International Conference on Learning Representations*, 2024.

Jianxiong Xiao, James Hays, Krista A. Ehinger, Aude Oliva, and Antonio Torralba. SUN database: Large-scale scene recognition from abbey to zoo. In *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3485–3492, 2010.

Maxime Zanella, Benoît Gérin, and Ismail Ben Ayed. Boosting vision-language models with transduction. In *Advances in Neural Information Processing Systems*, volume 37, 2024.

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

Ce Zhang, Simon Stepputtis, Katia Sycara, and Yaqi Xie. Dual prototype evolving for test-time generalization of vision-language models. In *Advances in Neural Information Processing Systems*, volume 37, 2024a.

Jifan Zhang, Yifang Chen, Gregory Canal, Stephen Mussmann, Arnav M. Das, Gantavya Bhatt, Yinglun Zhu, Jeffrey Bilmes, Simon Shaolei Du, Kevin Jamieson, and Robert D. Nowak. LabelBench: A comprehensive framework for benchmarking adaptive label-efficient learning. *Journal of Data-centric Machine Learning Research*, 2024b.

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.

Taolin Zhang, Jinpeng Wang, Hang Guo, Tao Dai, Bin Chen, and Shu-Tao Xia. BoostAdapter: Improving vision-language test-time adaptation via regional bootstrapping. In *Advances in Neural Information Processing Systems*, volume 37, 2024c.

Yabin Zhang, Wenjie Zhu, Hui Tang, Zhiyuan Ma, Kaiyang Zhou, and Lei Zhang. Dual memory networks: A versatile adaptation approach for vision-language models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024d.

Zihao Zhao, Eric Wallace, Shi Feng, Dan Klein, and Sameer Singh. Calibrate before use: Improving few-shot performance of language models. In *International Conference on Machine Learning*, pp. 12697–12706, 2021.

Chujie Zheng, Hao Zhou, Fandong Meng, Jie Zhou, and Minlie Huang. Large language models are not robust multiple choice selectors. In *International Conference on Learning Representations*, 2024.

Bolei Zhou, Agata Lapedriza, Aditya Khosla, Aude Oliva, and Antonio Torralba. Places: A 10 million image database for scene recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 40(6):1452–1464, 2018.

Xingyu Zhu, Beier Zhu, Yi Tan, Shuo Wang, Yanbin Hao, and Hanwang Zhang. Enhancing zero-shot vision models by label-free prompt distribution learning and bias correcting. In *Advances in Neural Information Processing Systems*, volume 37, 2024.

Xingyu Zhu, Shuo Wang, Beier Zhu, Miaoge Li, Yunfan Li, Junfeng Fang, Zhicai Wang, Dongsheng Wang, and Hanwang Zhang. Dynamic multimodal prototype learning in vision-language models. In *IEEE/CVF International Conference on Computer Vision*, 2025.

APPENDIX

A LIMITATIONS

First, hyper-parameter selection still involves labels: the label-free policy transfers the fusion weight from other pools, and the validated policy tunes three scalars on a labeled quarter of the target pool. Second, all experiments use curated benchmarks with fixed vocabularies and ground truth; the long-tailed and stress-test experiments only approximate the real-world setting that motivates this work, and the method remains closed-set: every image is assigned to one of the given classes.

B QA-BASED VS. RETRIEVAL-BASED ANNOTATION

This appendix provides the evidence behind the choice, made in Section 1, of retrieval-based VLMs over question-answering (QA) VLMs as the base annotator. Its scope is deliberately narrow. We do not claim that QA-based VLMs are inferior in general; we ask which *inference mechanism* is better suited to closed-set annotation over a large fixed vocabulary, and why. All numbers are our own measurements on the annotation pools of the main paper.

B.1 EXPERIMENTAL PROTOCOL

QA-based annotator. An image $\mathbf{x}$, the candidate vocabulary $\mathcal{Y}$, and a question q are given to a multimodal language model, which generates an autoregressive answer that is then matched to a class name: $(\mathbf{x}, \mathcal{Y}, q) \rightarrow$ answer $\rightarrow \hat{y}$. We use two models. **Qwen3-VL** (the 2B and 8B Instruct checkpoints of the Qwen3-VL family (Bai et al., 2025)) receives the full, shuffled candidate list in the prompt (*"Examine the image carefully and identify which objects from the candidate list are present. Candidates ({labels}). From this list, choose only ONE label that is most likely present in the image. Do not include any label that is not in the candidate list."*) and uses greedy decoding with at most 15 new tokens and batch size 64; the answer is normalized to its first whitespace token (class names are written with underscores so that multi-word names survive this step). **LLaVA-1.6-Mistral-7B** (Liu et al., 2023; 2024a; Jiang et al., 2023) uses the best of seven prompts we tried (*"Identify the main foreground object, ignoring the background. Choose exactly one label from [{labels}]. Output only the exact label."*), greedy decoding, at most 15 new tokens, *anyres* tiling and fp16. All runs in this appendix were made on NVIDIA RTX 6000 Ada Generation GPUs (48 GB). For LLaVA, we additionally run *likelihood scoring*: the log-probability of every class name given the image and a fixed question, with no list in the prompt, optionally calibrated by subtracting the same scores computed on a blank image (Zhao et al., 2021).

Retrieval-based annotator. With image encoder f_I and text encoder f_T, we set $\mathbf{z} = f_I(\mathbf{x})$, $\mathbf{t}_c = f_T(y_c)$ and $\hat{y} = \arg\max_c \mathbf{z}^\top \mathbf{t}_c$. Checkpoints: SigLIP2 ViT-gopt/16 at 384 px (*timm/ViT-gopt-16-SigLIP2-384* via *open_clip*) (Tschannen et al., 2025), and OpenAI CLIP ViT-B/16 and ViT-L/14 (Radford et al., 2021). Class names are embedded with the single template "a photo of a {label}." (Appendix C), and images use the default preprocessing of each model. Embeddings are computed once and cached; every subsequent operation, including the method of the main paper, reuses the cache.

Pools and class subsets. The annotation pools are the *training* sets of TinyImageNet (100,000 images, 200 classes) (Le & Yang, 2015) and ImageNet-1k (1,281,167 images, 1,000 classes) (Deng et al., 2009; Russakovsky et al., 2015). Class subsets of size C are drawn once with a fixed seed and are nested (C=10 $\subset$ 20 $\subset$ 50 $\subset$ 100 $\subset$ 150 on TinyImageNet; 200 $\subset$ 400 $\subset$ 600 $\subset$ 800 on ImageNet-1k). Each subset keeps all training images of its classes, so N grows with C (256k images at C=200, 1.28M at C=1000). Both annotators see exactly the same images and the same candidate set at every C. QA runs use a single seed.

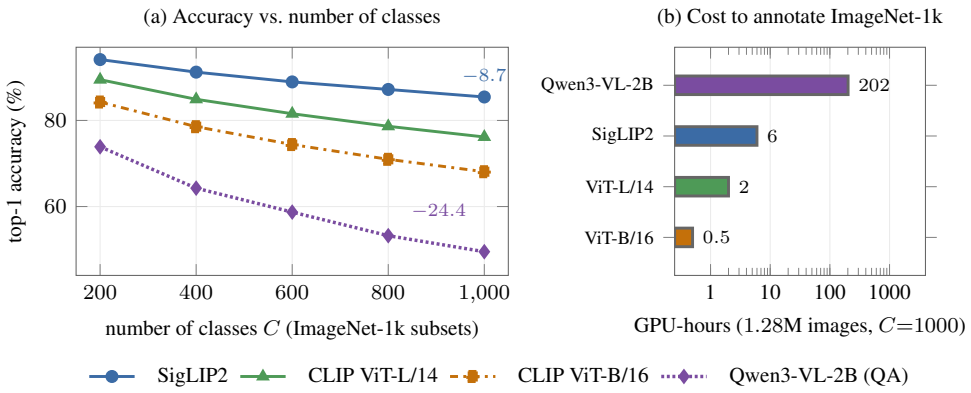

Figure 5: **Scaling behavior of QA-based and retrieval-based annotation.** (a) Top-1 accuracy on nested ImageNet-1k class subsets (training pools, same images for every annotator). The QA model (Qwen3-VL-2B) loses 24.4 points from 200 to 1,000 classes; SigLIP2 loses 8.7 and stays 20–36 points above it at every C; CLIP ViT-L/14 and ViT-B/16 lie in between. (b) Wall-clock GPU-hours to annotate the full 1.28M-image pool at C=1000 (log scale): 202 h for the QA model versus 0.5–6 h for the retrieval encoders, whose per-image cost does not depend on C. All timings are wall-clock times of our own runs on one NVIDIA RTX 6000 Ada (48 GB); the batch sizes of these timing runs were not recorded, so the QA figure is an upper bound on what a tuned pipeline would need.

B.2 Scaling with the number of candidate classes

Figure 5 addresses four questions. (The ImageNet QA curve comes from a single run of ours on a separate machine; the aggregate accuracies were carried over, but the per-image outputs were not.) On TinyImageNet, which is not shown in the figure, Qwen3-VL-8B reaches 96.0, 91.1, 80.9, 76.3 and 73.2% at C=10, 20, 100, 150 and 200 (the C=50 run did not complete), compared with 98.0, 92.9, 85.2, 84.0, 78.1 and 77.1 for SigLIP2 at C=10–200, 96.9, 93.3, 85.8, 82.0, 77.3 and 74.5 for CLIP ViT-L/14, and 95.1 down to 63.4 for ViT-B/16. *How fast does QA accuracy fall?* On ImageNet-1k the QA model drops from 73.9% at 200 classes to 49.5% at 1,000, a loss of 24.4 points, whereas SigLIP2 drops 8.7 points (94.1 to 85.5) over the same subsets. On TinyImageNet the two annotators decline at a similar rate (Qwen3-VL-8B from 96.0 to 73.2; SigLIP2 from 98.0 to 77.1), so the faster QA decline is specific to the thousand-class regime rather than to every vocabulary size, and we claim it only for ImageNet-scale vocabularies. *How stable is retrieval accuracy?* SigLIP2 stays within 9 points across a five-fold growth of the vocabulary and is the most accurate annotator at every C on both pools; the stronger 8B QA model matches CLIP ViT-L/14 on TinyImageNet but not SigLIP2. *How does cost scale?* Every QA query carries the full candidate list in its prompt, so the prompt length grows linearly with C. Annotating the 1.28M-image pool at C=1000 took 202 GPU-hours with the 2B model (about 1.8 images per second), against 6 h for SigLIP2 and 0.5 h for ViT-B/16, whose per-image cost is independent of C once the C class names have been embedded. *How much faster is retrieval?* Between 34× and 400× on this pool.

B.3 CLIP-initialized vision tower, different read-out: seven prompts for LLaVA-1.6

LLaVA-1.6-Mistral-7B reads a ViT-L/14 vision tower initialized from the very CLIP model we compare it with (the tower is fine-tuned during visual instruction tuning and passed through a two-layer projector into the language model). The two annotators therefore differ mainly in how the visual representation is read out: as a cosine against 200 class-name embeddings, or as a generated answer to a prompt that lists the 200 names. Table 5 lists the seven prompts we tried on the full 100,000-image TinyImageNet training pool (greedy decoding, candidate list shuffled per image, exact-match scoring after normalization), next to the zero-shot accuracy of the CLIP read-out with the single template `a photo of a {class}`.

Table 5: The CLIP ViT-L/14 tower readout by the cosine of CLIP, and its fine-tuned copy readout by LLaVA-1.6-Mistral-7B under seven prompts (TinyImageNet, 200 classes, 100k training images, top-1 accuracy %). {*labels*} is the shuffled comma-separated list of the 200 class names; the full string of prompt 7 is in the released code (Appendix E). OOLS: out-of-label-space answers (after case, whitespace, underscore and trailing-punctuation folding); refusal: answers that comment on image quality instead of naming a class. Retrieval has OOLS = 0 by construction.

Read-out	id	prompt	acc.	OOLS (%)	refusal (%)
CLIP ViT-L/14	–	*a photo of a* {*class*} (text encoder; no list)	**74.45**	0	0
LLaVA-1.6-7B	1	*A photo of a ____. Choose exactly one label from the following list: [{labels}]. Output only the chosen label.*	33.63	4.0	0.3
	2	*Which label belongs to this image? Answer the question with a single word from [{labels}].*	12.45	0.5	0.0
	3	*You are an image classifier. Identify the single main object in the image, ignoring the background. Choose exactly one label from these candidates: [{labels}]. Even if the image is small or unclear, choose the closest candidate. Return only the exact label as written in the candidate list. Do not explain, add punctuation, or repeat the candidates.*	35.32	8.9	0.5
	4	*Classify the main object into exactly one of these categories: [{labels}]. If uncertain, choose the closest category. Output only the exact category name.*	32.91	10.9	2.1
	5	*Identify the main foreground object, ignoring the background. Choose exactly one label from [{labels}]. Output only the exact label.*	**37.84**	7.1	0.1
	6	*[silent step-by-step analysis, then one label]*	32.97	21.6	11.3
	7	*[expert, four silent iterations, then one label]*	32.72	5.6	0.1

We draw three conclusions from this table. First, the ceiling is set by the read-out, not by the wording of the prompt: six of the seven prompts land within 32.7–37.8%, forty points below the cosine read-out of CLIP, and the one prompt that leaves this band (prompt 2) does so downward, because its terse question lets the answer prior take over (three answers account for 27,000 of the 100,000 images). Second, asking the model to reason silently (prompts 6 and 7) does not help; prompt 6 additionally produces 21.6% out-of-vocabulary answers and 11.3% refusals, and hedging clauses ("if uncertain", "even if unclear") raise the out-of-vocabulary rate to 9–11%. Third, the best prompt is the plainest statement of the task with a foreground cue and no hedging (prompt 5), and it is the prompt used for every LLaVA number elsewhere in this appendix. Since a linear probe on the visual input to the language model reaches 78% (Section B.9), the gap is not a matter of prompt engineering; it arises from delegating a 200-way metric decision to a language model.

B.4 Closed-set compliance: out-of-label-space answers

A closed-set annotator must return an element of $\mathcal{Y}$. We measure the out-of-label-space rate $\mathrm{OOLS} = |\{i : \hat{y}_i \notin \mathcal{Y}\}|/N$ after deterministic normalization (case, whitespace, underscore and trailing-punctuation folding), so that harmless formatting differences are not counted.

The two models behave very differently. The recent Qwen3-VL-8B follows the closed-set instruction almost perfectly: on the TinyImageNet pool its OOLS rate (answers normalized to their first token) is below 1% at every C (0.72, 0.34, 0.07, 0.14 and 0.16% at C=10, 20, 100, 150 and 200), and the remaining OOLS answers are superclasses or near-synonyms missing from the vocabulary (*snake* for *boa constrictor*, *duck* for *goose*). LLaVA-1.6 is far less compliant (Table 5; measured on the full generated string), and its compliance is unrelated to its accuracy: the most compliant prompt (prompt 2, 0.5% OOLS) has the worst accuracy, because it lets the answer prior of the model dominate (three answers cover 27,000 of 100,000 images), whereas prompts that invite hedging ("if uncertain", "even if unclear") raise OOLS to 9–11%, and the request for silent reasoning (prompt 6) produces 21.6% OOLS and 11.3% refusals. In the QA numbers used elsewhere in this appendix, OOLS answers are counted as errors.

B.5 Candidate-order sensitivity

A closed-set classifier should be invariant to the order in which candidates are listed, i.e. $f(\mathbf{x}, \mathcal{Y}) = f(\mathbf{x}, \boldsymbol{\pi}(\mathcal{Y}))$ for every permutation $\boldsymbol{\pi}$; retrieval scoring satisfies this by construction once the embeddings exist. Placing the *true* class name first, in the middle, or last in the 200-name list given to LLaVA-1.6 (prompt 5) on a stratified 2,000-image subset of TinyImageNet, with everything else fixed, yields 43.1, 33.7 and 26.2% with longest-match parsing of the generated string (34.2, 31.1

Table 6: Decision mechanism of LLaVA-1.6-7B on TinyImageNet (200-way). Left: the 2,000-image stratified subset used for the ablation chain; right: the full 100,000-image pool. CLIP ViT-L/14 zero-shot on the same images is the retrieval reference; the vision tower of LLaVA is initialized from this CLIP model.

Decision mechanism	2k subset	100k pool
Generation, 200 names in prompt (exact match)	35.4	37.8
Generation, 200 names, single 336 px tile	37.5	–
Generation, CLIP top-5 names in prompt	60.9	–
Likelihood scoring, 200 names, raw	39.1	37.9
Likelihood scoring, 200 names, blank-image calibrated	45.7	44.5
CLIP ViT-L/14 zero-shot (retrieval)	74.5	74.4
Oracle of {calibrated scoring, CLIP}	78.0	78.0

and 22.5% with exact match; random order on the same images: 38.4 and 35.4%), while OOLS rises from 10.0 to 12.2%. Moving the correct name from the first to the last position thus costs 16.9 points, a primacy effect consistent with the position biases reported for language models choosing among listed options (Zheng et al., 2024; Liu et al., 2024b). The Qwen runs shuffle the list independently per image with a fixed seed, so their accuracies average over positions; we did not run a per-image consistency analysis across repeated shuffles.

B.6 Free generation vs. candidate likelihood scoring

Generation commits greedily to one token sequence and may leave the candidate list. Likelihood scoring instead computes, for each of the 200 class names, the log-probability that the same model assigns to that name given the image and a fixed question, and takes the arg-max. This turns the QA model into a metric read-out over the closed set, with identical weights and image tokens.

Table 6 shows that replacing generation with scoring adds only 0.1–3.7 points with raw scores, so the generative decoding step is not where most of the accuracy is lost. The calibrated scorer agrees with CLIP on 46% of the images and is correct where CLIP is wrong on only 3.5%; the two read-outs are thus not complementary enough for an ensemble to rescue the QA pathway (oracle of the two: 78.0).

B.7 Language-model prior effects

Following contextual calibration (Zhao et al., 2021), we feed blank gray, white, and noise images with the same prompt and subtract the resulting per-class log-likelihoods from the real-image scores. Calibration raises scoring accuracy from 39.1 to 45.7 on the 2k subset and from 37.9 to 44.5 on the full pool (Table 6), which means that about a sixth of the QA gap to CLIP is due to an image-independent preference for certain class strings. The same prior shows up in generation. On blank images, LLaVA answers with a fixed set of words (*stopwatch*, *hourglass*, *drum*) or refuses, and on real images the five most over-predicted classes (*desk*, *pole*, *dragonfly*, *bee*, *drumstick*) receive 1,260–1,907 answers each against 500 true images, at 7–19% precision. These classes are the same under every prompt, so the prior belongs to the model rather than to the wording. Qwen3-VL-8B shows a milder version of the same effect: *cliff* is predicted 1,886 times and *plate* 1,339 times, against 500 true images each. The effect does not depend on image difficulty. Binning images by the margin of CLIP itself, calibrated LLaVA scoring reaches 17.6 / 32.1 / 55.4 / 72.9% from the hardest to the easiest quartile, while CLIP reaches 34.5 / 70.4 / 93.5 / 99.4%; even images that a linear read-out classifies with 99% accuracy are answered correctly only 73% of the time by the language model.

B.8 Candidate-space reduction

If the difficulty lies in resolving one class among many, a shortlist should help. We therefore let the retrieval annotator propose its top-K candidates and the QA model decide among them (Figure 6). List length turns out to be the single most influential factor we can manipulate: with LLaVA on TinyImageNet, generation from the top-5 of CLIP reaches 60.9% against 35.4% from the full list

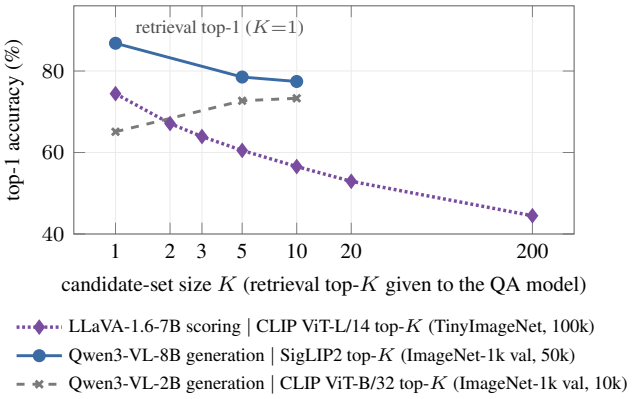

Figure 6: **Letting the QA model choose among the retrieval top-K.** K=1 is the retrieval annotator itself. Shortening the list helps the QA model relative to the full vocabulary (LLaVA: 44.5 at K=200 to 60.5 at K=5), but re-ranking the shortlist of a strong retriever *lowers* accuracy (SigLIP2 86.8 to 78.5 at K=5 with Qwen3-VL-8B) and helps only when the retriever is weak (CLIP ViT-B/32 65.1 to 73.3 at K=10 with Qwen3-VL-2B). Curves follow the color code of Figure 5.

Table 7: Linear-probe accuracy (%, test split, 3-seed mean) on the tensors along the path from the read-out point of CLIP to the input of the language model (TinyImageNet, 100k images, single 336 px tile). Each cell is head alone / head fused with CLIP text scores. The projector output is what Mistral-7B receives.

Tensor	dim	oracle labels	Sinkhorn labels	LLaVA answer
CLIP projected CLS (zero-shot pathway)	768	81.6 / 82.6	80.4 / 81.1	–
final-layer CLS	1024	79.6 / 82.6	77.9 / 80.9	–
layer −2 CLS (discarded by LLaVA)	1024	77.9 / 81.9	76.4 / 80.3	–
layer −2 patch mean (kept by LLaVA)	1024	79.0 / 82.7	77.5 / 81.1	–
projector output, mean-pooled	4096	78.4 / 82.6	76.8 / 81.0	37.8

(+25.5; Table 6), and the top-5 of CLIP contains the true class for 91% of the images. The QA model therefore retains useful visual knowledge, but it adds nothing on top of a strong retriever: on the ImageNet-1k validation set, Qwen3-VL-8B choosing among the top-5 of SigLIP2 (recall 98.0%) scores 78.5%, below the 86.8% of SigLIP2 itself; the 2B model scores 77.5%, and top-10 shortlists are slightly worse. The same re-ranking does help a weak retriever (CLIP ViT-B/32: 65.1 to 73.3 with the top-10), which confirms that the knowledge of the QA model is real but only useful when the retrieval ranking is poor. For the encoders used in the main paper, retrieval is the stronger decision maker at every K, and the unlabeled pool improves it further (SigLIP2 on ImageNet-1k: 85.5 to 87.8, Section 4).

B.9 Is the visual information still there? Linear probes

LLaVA-1.6 reads a CLIP-initialized ViT-L/14 tower: it takes the layer −2 patch tokens (the CLS token is discarded), maps them through a two-layer MLP projector into the language model, and generates. A linear probe on each intermediate tensor (Alain & Bengio, 2017) tells us whether class information is lost on the vision side or in the decision of the language model. The probes are our ridge head under the 50/25/25 fit/val/test protocol of the paper with 3 seeds, fitted either with oracle labels (an information ceiling) or with the label-free Sinkhorn pseudo-labels of the main method.

The tensor handed to the language model supports 78.4% with a linear head alone and 82.6% when fused with the text scores, which is indistinguishable from the pathway of CLIP itself (81.6 / 82.6) and 40 points above what the language model generates from it (Table 7). A substantial amount of class-discriminative information therefore survives the visual pathway but is not translated into the final many-class answer. One side result is relevant to the main paper: the class-mean prototype

collapses to 53–54% on the patch and projector features, whereas the whitened ridge head recovers 78–79%, because these features are far more anisotropic than the projected CLS of CLIP.

B.10 Computational efficiency

The cost panel of Figure 5 separates the two costs (one NVIDIA RTX 6000 Ada; 202 GPU-hours for the QA model against 6 h for SigLIP2 on the 1.28M ImageNet-1k images). For the QA model, every image requires a forward pass through a multimodal language model with the whole vocabulary in context, followed by either generation or C likelihood evaluations; the throughput of 0.3–1.8 images per second cannot be improved by caching, since nothing is reusable across images. For retrieval, the expensive part is one encoder pass per image, after which the similarity matrix and the class statistics of the main method amount to seconds of linear algebra: on the full 1.28M-image pool, the balanced assignment takes about 10 s per prior and the ridge head a single solve of 2.4 s.

B.11 Failure modes of QA-based closed-set annotation

Four kinds of failure recur throughout Sections B.4–B.7, all with the full 200-name list in the prompt: answers outside the vocabulary (superclasses and near-synonyms, e.g. *Bird* for *black stork*, *snake* for *boa constrictor*, *duck* for *goose*); verbose strings that embed an invalid label (*"The main foreground object is 'conversion'."*); refusals that comment on image quality (11.3% of the images under prompt 6); and answer-prior collapse (*bee* given to 10,726 images under prompt 2; *cliff* to 1,886 by Qwen3-VL-8B). Position sensitivity (Section B.5) is a population-level effect and has no single-image example.

B.12 Summary

These experiments do not show that QA-based VLMs are inferior in general. They do show that, for many-class closed-set annotation, the generative interface adds cost and decision complexity without adding accuracy: inference is 34–400 times slower on ImageNet-1k, accuracy falls faster as the vocabulary approaches a thousand classes, the answer depends on list position and on an image-independent prior over class strings, and a linear read-out of the same visual input beats the generated answer by 40 points. A retrieval shortlist helps a QA model, but the QA model does not beat the strong retriever that produced the shortlist. Retrieval-based VLMs are therefore the more direct and more efficient foundation for this setting, and their cached embeddings are what the label-free calibration of the main paper builds on, at negligible extra cost.

C Datasets, Checkpoints, Protocol and Reproducibility

Notation. Table 8 collects the hyperparameters and protocol quantities of the main text, together with the value each takes when it is held fixed. Symbols defined inline in Sections 2 and 3 ($\mathbf{z}_i$, $\mathbf{t}_c$, S, P, $\hat{y}_i$, G, $\mathbf{S}_c$, $\mathbf{w}_c$) are omitted.

Annotation pools. Table 9 lists the pools. Each pool is the official training set; the corresponding validation or test sets are not used anywhere, with the single exception of the ImageNet-1k validation images in the candidate-reduction study (Appendix B.8). Class names are the canonical names shipped with each dataset (for ImageNet, the readable names used in the zero-shot evaluation of CLIP, in which two pairs of classes share the names *missile* and *sunglasses*, so four of the 1,000 anchors coincide). SUN397 names are the full hierarchical labels with underscores and slashes replaced by spaces (*apartment building outdoor*, *bakery shop*). An earlier version of our pipeline kept only the last path component, which collapsed 100 of the 397 classes onto eight anchors (43 onto *outdoor*, 44 onto *indoor*). All SUN397 numbers in this paper use the corrected anchors; zero-shot accuracy with SigLIP2 rose from 57.8% to 75.0% after the fix, and the qualitative findings did not change.

The two long-tailed pools are the standard ImageNet-LT and Places-LT training lists (Liu et al., 2019), taken as row subsets of the ImageNet-1k and Places365 caches. They share the anchors of the parent pool, and both lists are included in the code release (Appendix E).

Table 8: Hyper-parameters and protocol quantities. "Fixed" gives the value used whenever the symbol is not searched; the validated policy searches $\alpha \in \{0.1, 1, 10\}$, $\tau \in \{0, 1\}$ and $\lambda \in \{0, 0.02, \ldots, 1\}$ on the validation split.

Symbol	Meaning	Defined	Fixed
$\mathcal{D} = \{\mathbf{x}_i\}$, N	annotation pool of N unlabeled images	§2	
$\mathcal{Y}$, C	fixed vocabulary of C class names	§2	
D	embedding dimension	§2	1536 / 512 / 768
fit	per-class split of each pool; heads on fit	§4.1	50%, 5 seeds
validation	per-class split of each pool; hyper-parameters on validation	§4.1	25%, 5 seeds
test	per-class split of each pool; accuracy on test	§4.1	25%, 5 seeds
marginal	vector of class sizes of an assignment	§2	
purity	vector of fraction of correct pseudo-labels	§2	
$\hat{\boldsymbol{\pi}}$	estimated class prior: arg-max marginal of the anchor scores	§3	
τ, $\boldsymbol{\pi}_\tau$	prior shrinkage; $\boldsymbol{\pi}_\tau = (1-\tau)\,\text{uniform} + \tau\hat{\boldsymbol{\pi}}$	§3	$\tau = 1$ (label-free)
ε	Sinkhorn temperature	Eq. 1	$\varepsilon^{-1} = 100$, 50 iterations
α	ridge regularizer of the whitened head	Eq. 2	$\alpha = 1$ (label-free)
λ	fusion weight on the anchor score	Eq. 3	0.3 SigLIP2, 0.5 CLIP
ρ	stress test: imbalance ratio	§4.5	$\rho \in \{1, 10, 100\}$
f	stress test: fraction of classes retained	§4.5	$f \in \{.25, .5, 1\}$

Table 9: Annotation pools.

Pool	N	C	imbalance (max/min images per class)
CIFAR-100 (Krizhevsky, 2009)	50 000	100	1×
TinyImageNet (Le & Yang, 2015)	100 000	200	1×
Food-101 (Bossard et al., 2014)	75 750	101	1×
SUN397 (Xiao et al., 2010)	108 754	397	23.6× (100–2 361)
Places365 (Zhou et al., 2018)	1 803 460	365	1.6× (3 068–5 000)
ImageNet-1k (Deng et al., 2009)	1 281 167	1000	1.8× (732–1 300)
ImageNet-LT (Liu et al., 2019)	115 846	1000	256× (5–1 280)
Places-LT (Liu et al., 2019)	62 500	365	996× (5–4 980)

Encoders. We use SigLIP2 ViT-gopt/16 at 384 px (*timm/ViT-gopt-16-SigLIP2-384* via *open_clip*; D=1536), OpenAI CLIP ViT-B/16 (D=512) and ViT-L/14@336 (D=768) via the official *clip* package, and DINOv2-base (*facebook/dinov2-base*, CLS token, D=768). Images use the default preprocessing of each model. Image embeddings are L_2-normalized and cached together with the $N \times C$ similarity matrix. The SigLIP2 TinyImageNet cache is the one exception: it was computed with the Hugging Face implementation of the same checkpoint (*google/siglip2-giant-opt-patch16-384*), whose stored similarity matrix is the calibrated logit of SigLIP, approximately $108\cos - 16$, rather than $100\cos$. All SigLIP2 TinyImageNet results use this cache.

Prompt templates. Every class name is embedded with the single template `a photo of a {label}` and L_2-normalized. We confirmed this by re-encoding the class names with the single template and reproducing the cached similarity matrices to within 0.05 (maximum absolute error on 100×cosine; for the SigLIP2 TinyImageNet cache, after fitting its scale and offset), whereas the 80-template ensemble does not reproduce them.

Splits and seeds. For each pool and seed $s \in \{0, \ldots, 4\}$, the images of every class are permuted with `RandomState`$(1000 + s)$ and divided 50/25/25 into fit, validation and test splits. Balanced assignment (Stage 2a) runs once on the whole pool and uses no labels. Heads (Stage 2b) are built on the fit split. Validated hyper-parameters are chosen on the validation split from $\alpha \in \{0.1, 1, 10\}$, $\tau \in \{0, 1\}$ and $\lambda \in \{0, 0.02, \ldots, 1\}$. The label-free policy uses $\alpha = 1$, $\tau = 1$, and the $\lambda \in \{0.2, 0.3, 0.4, 0.5\}$ with the highest mean validation accuracy on the other five pools; the "transferred" variant of Figure 12 chooses the (τ, λ) pair from $\{0, 0.5, 1\} \times \{0.2, 0.3, 0.4, 0.5\}$ in the same way (Appendix D); the label-free prototype recipe uses the mean validated λ of the other five pools, rounded to the 0.02 grid (five seeds). Accuracy is always measured on the test split of the pool. Sinkhorn uses 50 iterations at temperature $\varepsilon^{-1} = 100$ in log space. The main-text tables report means over seeds and wins over the 24 (or 18 CLIP-family) encoder–dataset pairs. Paired t-tests

over the 18 CLIP-family pairs are reported where stated, but they should be read as indicative only, since pairs share encoders; the win counts are our primary evidence.

Class subsets. Nested subsets are drawn once with `RandomState(42)` from the class indices (TinyImageNet: 10, 20, 50, 100, 150; ImageNet-1k: 200, 400, 600, 800) and keep every training image of the selected classes. The same subsets are used for the QA annotators of Appendix B and for the scaling experiment of Section 4.4.

Hardware and cost. All experiments ran on NVIDIA RTX 6000 Ada Generation GPUs (48 GB). Encoding the ImageNet-1k pool takes six GPU-hours with SigLIP2, two with ViT-L/14 and half an hour with ViT-B/16; every subsequent operation runs in seconds (Figure 5, cost panel).

D Evaluation Settings and Use of Labels

The input to the annotator is the unlabeled pool and the class names. Labels of the pool enter the paper in three roles only: to measure accuracy (every result), to tune hyper-parameters in the rows marked *validated* (a reference), and to build the *oracle* rows and diagnostics. This appendix defines the two protocols and the two hyper-parameter policies, lists the settings of every table and figure (Table 10), and compares our settings with those of the methods we evaluate against (Table 11).

Protocols. In the *transductive* protocol, the annotator labels a set of unlabeled images using only those images: balanced assignment, whitened head, and prediction all run on the same set, with no split. This is the deployment setting and the protocol of the published methods we compare with; we apply it to the test split of each pool so that every method is scored on the same images (Tables 3, 25 and 26). In the *fit-split* protocol, used for controlled comparisons, every class is divided 50/25/25 into fit, validation, and test images (five seeds, Appendix C); balanced assignment runs once on the whole pool, so it sees all images but none of their labels, heads are built on the fit images, and accuracy is measured on the test images. The two protocols give nearly the same accuracy: with the label-free policy, the six-pool means are 79.99 / 67.59 / 74.33 (transductive) and 80.04 / 67.53 / 74.50 (fit split) for SigLIP2 / ViT-B/16 / ViT-L/14 (Table 3).

Hyper-parameter policies. The *label-free* policy uses no label of the target pool. It fixes α=1 and the estimated prior (τ=1) for every pool and transfers λ per encoder: for each target pool, $\lambda \in \{0.2, 0.3, 0.4, 0.5\}$ is the value with the highest mean validation accuracy over the other five pools at α=1, τ=1. This gives 0.5 for CLIP on every pool and 0.3 for SigLIP2 on every pool except TinyImageNet (0.4). ImageNet-LT and Places-LT additionally leave out their parent pools (ImageNet-1k and Places365), whose images they share, and obtain 0.3 (SigLIP2) and 0.5 (CLIP); the stress-test pools, which are subsets of the main pools, use the value of their parent pool. The validation images of the target pool are never used. The label-free prototype recipe takes the mean validated λ of the other five pools, rounded to the 0.02 grid. The only labels behind any label-free number are therefore those of the validation splits of *other* pools. The *validated* policy selects $\alpha \in \{0.1, 1, 10\}$, $\tau \in \{0, 1\}$ and $\lambda \in \{0, 0.02, \ldots, 1\}$ on the labeled validation images of the target pool (for the prototype recipe, λ alone). It is a reference for what target labels would add, not a recommended use of them: a whitened head trained on the same labeled quarter is more accurate still (Figure 10).

Oracles and diagnostics. Rows that use ground-truth labels or class proportions to build the annotator are marked *oracle* and excluded from the bold comparisons: the oracle-label prototype and whitened head (Figure 2, Table 21), OTTER with the true prior (quoted in Appendix I.13, not tabulated), and the true proportions of Figure 16. The selection controls (Figure 9, Table 19) and the fusion diagnostics of Appendix H score each configuration at its accuracy-optimal λ on the evaluated images (an oracle λ). They measure what a mechanism can contribute, not a deployable accuracy.

Relation to the settings of prior work. Table 11 summarizes the main settings of the closest methods as described in their papers. TransCLIP, InMaP, Frolic, IOTA, LaFTer, and TDA, like our label-free annotator, use no label of the target images; OTTER takes the class proportions as given, and Tip-Adapter, the closest to our fusion of anchor and visual evidence in form, builds its cache from a few labeled images per class rather than from an unlabeled pool. The methods also differ in which images they adapt on, in whether they train, and in how they treat the class proportions; ours is the only one that estimates the proportions from the unlabeled pool by default.

Table 10: Protocol and hyper-parameter policy of every result. "validated" names the scalars chosen on the labeled validation split of the target pool; "label-free" rows use no label of the target pool.

Result	Protocol	Policy
Table 2, CLIP-family rows	fit split	label-free and validated rows
Table 2, DINOv2 rows; Tables 22, 23	fit split	validated
Tables 3, 24, 25	transductive and fit split	label-free; last row validated (reference)
Table 26	transductive	label-free
Figure 14	fit split	α, λ validated; τ as labeled
Figure 4	fit split	label-free and validated curves
Tables 4, 20, 21; Figure 11	fit split	α, λ validated; τ validated or fixed as labeled
Figures 15, 17	fit split	α, λ validated; τ as labeled
Figures 12, 13, 16	fit split	label-free configurations; validated as reference
Figure 18	fit split	λ validated for both heads
Figure 10	fit split	head trained on the labeled validation quarter (control)
Figures 2, 7; Table 12	fit split or whole pool	oracle labels (diagnostic)
Figure 9, Table 19, Appendix H	hold-out or in-sample	oracle λ (diagnostic)

Table 11: Settings of the compared methods in their main setting (zero-shot, or few-shot for Tip-Adapter), as described in each paper.

Method	Unlabeled images	Target labels	Class proportions	Adaptation
TransCLIP	test set (transductive)	none (few-shot variant: labeled shots)	uniform	block updates, no gradients
InMaP	test set (transductive)	none	uniform (estimated option)	gradient steps on class proxies
Frolic	test set (transductive)	none	uniform; pre-training bias estimated	closed form and fixed-point iterations
OTTER	test set (transductive)	class proportions given	given	optimal transport
IOTA	test set (transductive)	none (few-shot variant: labeled shots)	uniform or empirical	transport and prototype updates
LaFTer	unlabeled training split	none (LLM-generated descriptions)	not modeled	prompt and normalization-layer training
Tip-Adapter	none (labeled shots only)	few labeled images per class	not modeled	feature cache (fine-tuned keys)
TDA	test stream	none	not modeled	feature caches
ours	pool (transductive/ fit split)	none (label-free policy)	estimated from the pool	closed form

E Source Code

Code. The code is provided in the supplementary material as `beyond-text-anchors.zip`, an anonymized copy of our repository. The `README.md` of the repository describes the environment, the expected layout of the datasets, how to build the embedding caches, and the command behind each table and figure of the paper. The repository has four parts.

Library. Loaders for the eight annotation pools (with the SUN397 class names of Appendix C and the ImageNet-LT and Places-LT training lists), the encoders with the single prompt template, and the three stages of the method: balanced assignment, the whitened head, and fusion with the anchor. The published baselines of Table 3 (TransCLIP, InMaP, Frolic, TDA and OTTER) are implemented on the same caches and splits.

Experiments. One script per experiment, each writing a CSV file: the main comparison and build-up, the label-free policies and the scaling in C, the stress test, the imbalance analyses and prior re-estimation, the selection controls, the closed-form head against SGD, the DINOv2 transfer, and the numerical checks of Appendix F.

QA annotators. The LLaVA-1.6 and Qwen3-VL annotators and the analyses of Appendix B.

Tables and results. Scripts that turn the CSV files into the tables and figures, together with the CSV files of our own runs, so that every number in the paper can be checked without recomputing it.

Cost and determinism. Encoding the pools is the only expensive step (Appendix C); every other script runs from the caches on a single GPU in seconds to minutes. We re-ran each script on at least one pool and compared its output with the released CSV file. The outputs are identical except where non-deterministic GPU kernels are involved: the SGD-trained heads of Appendix I.9 change by up to 0.5 points after one to five epochs and by less than 0.1 points from 20 epochs on, and the K-means prototypes of the selection controls change by less than 0.01 points. The caches themselves (up to 17 GB for one encoder and pool) exceed the size limit of the anonymous copy; the repository regenerates them (all but the SigLIP2 TinyImageNet cache; see Appendix C), and they will be released with the public version.

F THEORETICAL ANALYSIS

Under a minimal spherical class model (Assumptions 1–4) we prove one statement for each design decision of Section 3. Averaging over *all* pseudo-labeled images attenuates label noise, whereas confidence selection biases the class statistic even when the labels are perfect (Lemmas 1–2). Balanced assignment is the information projection of the zero-shot posterior onto the couplings with the prescribed marginal (Proposition 1). The ridge head is a whitened class mean computed from the same statistic $\mathbf{S}_c$ as the prototype; it is consistent for fixed regularization, and it attenuates the shared spectral component that makes class means collinear (Propositions 2–4). Finally, under a mean-squared surrogate of the risk, the optimal fusion weight follows an inverse-sample-size law with a bias floor for every head in this family, and the result is robust to imperfect error orthogonality (Theorem 1, Lemma 3). The practical message is to use all samples without filtering and to read $\mathbf{S}_c$ through the whitened map. Caveats are collected in Section F.6 and proofs in Appendix G.

F.1 SETTING, ESTIMATORS, AND ASSUMPTIONS

The quantities $\mathbf{z}_i$, $\mathbf{t}_c$, $\boldsymbol{\pi}$, G, $\mathbf{S}_c$, n_c and α are as in Sections 2 and 3, with G, $\mathbf{S}_c$ and n_c computed over a single index set $\mathcal{I}$ of N images (the fit split). Embeddings lie on $\mathbb{S}^{D-1}$, and we linearize around the class direction. Since the measured within-class spread is 18°–39° (image-to-prototype cosine 0.78–0.95, Appendix H.1), results marked *leading order* are approximations checked against Table 12, not bounds. The concentration results use $\|\mathbf{z}\| = 1$.

Definition 1 (Heads and population targets). *With $\hat{y}_i \in [C]$ the pseudo-labels on $\mathcal{I}$, $\mathbf{S}_c = \sum_{i\in\mathcal{I}:\,\hat{y}_i=c} \mathbf{z}_i$, n_c the class count, $G = \sum_{i\in\mathcal{I}} \mathbf{z}_i\mathbf{z}_i^\top$ and $\mathrm{norm}(\mathbf{v}) = \mathbf{v}/\|\mathbf{v}\|$, the* prototype head *is $\mathbf{p}_c = \mathrm{norm}(\mathbf{S}_c/n_c)$ and the* ridge head *is $\mathbf{w}_c(\alpha) = \mathrm{norm}\left((G+\alpha I)^{-1}\mathbf{S}_c\right)$. With $M = \mathbb{E}[\mathbf{z}\mathbf{z}^\top]$, $\bar{\alpha} = \alpha/N$, $\boldsymbol{\mu}_c$ the class mean and $\tilde{\boldsymbol{\mu}}_c = \mathbb{E}[\mathbf{S}_c/n_c]$ the contaminated class mean of Lemma 1, their population* limits *are $\bar{\mathbf{p}}_c = \mathrm{norm}(\tilde{\boldsymbol{\mu}}_c)$ and $\bar{\mathbf{w}}_c = \mathrm{norm}\left((M+\bar{\alpha}I)^{-1}\tilde{\boldsymbol{\mu}}_c\right)$ (Proposition 3). The* target $\mathbf{h}_c^{\mathrm{opt}}$*, the population class direction minimizing the surrogate risk of Section F.5, is common to all heads and in general differs from each limit, since neither the class mean nor its whitened version is the optimal linear read-out even with clean labels (this is the gap Finding 2 measures). A generic head is $\hat{\mathbf{h}}_c$, with limit $\bar{\mathbf{h}}_c$.*

Assumption 1 (Spherical class model). *Embeddings of class k are i.i.d. draws $\mathbf{z} = \boldsymbol{\mu}_k + \boldsymbol{\xi}$ with mean-zero noise $\boldsymbol{\xi}$ of covariance Σ_k and $\mathrm{tr}(\Sigma_k) = \sigma_k^2$; since $\|\mathbf{z}\| = 1$, $\boldsymbol{\xi}$ has bounded support and is sub-Gaussian.*

Assumption 2 (Biased, noiseless text reference). *$\mathbf{t}_k = \boldsymbol{\mu}_k + \mathbf{b}_k$ with deterministic bias $\mathbf{b}_k$ (the class-level modality gap); $b_t^2 = \|\mathbf{t}_k - \mathbf{h}_k^{\mathrm{opt}}\|^2$ is the squared anchor bias against the common target. The bias is not small ($\cos(\mathbf{t}_k, \boldsymbol{\mu}_k)$ is 0.11 to 0.16 for SigLIP2, Table 15, so $\|\mathbf{b}_k\| > \|\boldsymbol{\mu}_k\|$); only the value of b_t^2 is used.*

Assumption 3 (Spiked second moment). *$M = \varrho\,\mathbf{u}\mathbf{u}^\top + M_r$ with $M_r\mathbf{u} = 0$ and $\|M_r\|_{\mathrm{op}} < \varrho$, and class directions share the spike: $\boldsymbol{\mu}_k = \beta_k\mathbf{u} + \boldsymbol{\Delta}_k$ with $\beta_k > 0$ and $\langle \mathbf{u}, \boldsymbol{\Delta}_k\rangle = 0$. No separation between β_k^2 and $\|\boldsymbol{\Delta}_k\|^2$ is assumed; the measured ratio is finite (Table 12).*

Assumption 4 (Quantified error obliqueness). *There is $\gamma \in [0, 1)$ such that for every class k and every head $\hat{\mathbf{h}}_k$, $\left|\mathbb{E}\langle \mathbf{t}_k - \mathbf{h}_k^{\mathrm{opt}},\ \hat{\mathbf{h}}_k - \mathbf{h}_k^{\mathrm{opt}}\rangle\right| \le \gamma\,\|\mathbf{t}_k - \mathbf{h}_k^{\mathrm{opt}}\|\left(\mathbb{E}\|\hat{\mathbf{h}}_k - \mathbf{h}_k^{\mathrm{opt}}\|^2\right)^{1/2}$.*

Table 12 (SigLIP2, per pool) reports the measured quantities behind these assumptions. The top eigenvalue carries 48–75% of the spectrum, and the prototype inter-class cosine of 0.71–0.88 corresponds to $\beta^2/\|\boldsymbol{\Delta}\|^2 \approx 2.4$–$7.3$ (first display of Proposition 4). For Assumption 4, we use the proxy correlation $\hat{\gamma}_k = \cos(\mathbf{t}_k, \hat{\mathbf{e}}_k)$ of the contamination displacement $\hat{\mathbf{e}}_k = \mathbf{w}_k^{\mathrm{pseudo}} - \mathbf{w}_k^{\mathrm{oracle}}$; on the four pools where it was computed (not tabulated) it has mean ≤ 0.03, 95% quantile ≤ 0.09 and maximum ≤ 0.15. Because this proxy uses $\mathbf{t}_k$ rather than $\mathbf{t}_k - \mathbf{h}_k^{\mathrm{opt}}$ and captures only the contamination part of the error, $\gamma \le 0.15$ should be read as a working bound rather than a measurement; Lemma 3 quantifies the effect of any γ.

F.2 AGGREGATION DILUTES LABEL NOISE; SELECTION CONCENTRATES BIAS

Let $\mathcal{I}_k$ be the index set assigned to class k, $n = |\mathcal{I}_k|$, q_k (the purity) the fraction of $\mathcal{I}_k$ with true label k, $\boldsymbol{\mu}_k^{\mathrm{cor}} = \mathbb{E}[\mathbf{z} \mid y = k,\ i \in \mathcal{I}_k]$ the mean of the correctly assigned images (a score-truncated

Table 12: Measured quantities behind the assumptions (SigLIP2, whole pool, α=1). "share": fraction of $\mathrm{tr}\, M$ in the top eigenvalue; inter-class cosines: means over class pairs of $\mathbf{p}_k$ and of $\mathbf{w}_k$; $\cos(\mathbf{t}_k, \cdot)$: mean cosine between an anchor and its own class vector; purity: fraction of correct pseudo-labels for the arg-max and for balanced assignment onto the uniform and the estimated priors.

		inter-class cosine		$\cos(\mathbf{t}_k, \cdot)$		purity (%)		
Pool	share (%)	prototypes	whitened	$\mathbf{p}_k$	$\mathbf{w}_k$	arg-max	uniform	estimated
CIFAR-100	75	0.88	-0.0100	+0.141	+0.195	87.1	89.4	87.3
TinyImageNet	68	0.84	-0.0048	+0.149	+0.138	77.1	82.9	77.5
Food-101	65	0.82	-0.0099	+0.160	+0.149	94.8	95.4	95.1
SUN397	50	0.73	-0.0021	+0.165	+0.091	75.0	66.0	77.4
Places365	48	0.78	-0.0023	+0.140	+0.017	44.7	47.3	44.8
ImageNet-1k	51	0.71	-0.0010	+0.162	+0.025	85.5	86.9	85.7

sample, displaced from $\boldsymbol{\mu}_k$ by the selection effect of Lemma 2), and $\boldsymbol{\nu}_k = \boldsymbol{\mu}_k^{\mathrm{cor}} + \boldsymbol{\delta}_k$, σ^2_{int} the mean and spread of the intruders.

Lemma 1 (Contamination dilution)**.** *Under Assumption 1, conditionally on the assignment, the empirical class mean* $\hat{\mathbf{m}}_k = \mathbf{S}_k/n$ *satisfies* $\mathbb{E}[\hat{\mathbf{m}}_k] = \tilde{\boldsymbol{\mu}}_k = \boldsymbol{\mu}_k^{\mathrm{cor}} + (1-q_k)\,\boldsymbol{\delta}_k$ *and* $\|\hat{\mathbf{m}}_k - \tilde{\boldsymbol{\mu}}_k\| = O_P(\tilde{\sigma}_k/\sqrt{n})$*, with* $\tilde{\sigma}_k^2 = q_k\sigma_k^2 + (1-q_k)\sigma^2_{\mathrm{int}} + q_k(1-q_k)\|\boldsymbol{\delta}_k\|^2$ *the spread of the assigned mixture. For any* deterministic *matrix* B*,* $B\hat{\mathbf{m}}_k$ *has contamination error* $(1-q_k)B\boldsymbol{\delta}_k$ *and fluctuation* $O_P(\|B\|_{\mathrm{op}}\,\tilde{\sigma}_k/\sqrt{n})$*; the* data-dependent *whitening map is handled in Proposition 3.*

Because intruders are confusable with the class, $\|\boldsymbol{\delta}_k\|$ is small relative to the inter-class separation, and the label error is attenuated both by the mixture weight *and* by this confusability. A 22.9% label error moves the class statistic by only $1 - \cos = 0.46\%$ ($5.5°$) from its oracle-label counterpart (Table 14: 13–37% error yields $1 - \cos = 0.08$–0.64%, or $2.3°$–$6.5°$; Table 15 shows the separate, slight selection shift toward the anchor). Aggregation thus acts as the noise filter, and no selection step is needed.

Lemma 2 (Selection bias of truncated class means; Gaussian surrogate)**.** *Let* $\mathbf{z} \sim \mathcal{N}(\boldsymbol{\mu}_k, \Sigma_k)$ *(a within-class surrogate that drops the norm constraint) and keep* $\{\mathbf{z} : \mathbf{a}^\top\mathbf{z} \geq t_0\}$ *for a unit selector* $\mathbf{a}$ *(confidence or margin mining:* $\mathbf{a} = \mathbf{t}_k/\|\mathbf{t}_k\|$*). With standardized threshold* $\theta = (t_0 - \mathbf{a}^\top\boldsymbol{\mu}_k)/\sqrt{\mathbf{a}^\top\Sigma_k\mathbf{a}}$*, standard normal density and distribution function* ϕ*,* Φ*, and inverse Mills ratio* $h(\theta) = \phi(\theta)/(1-\Phi(\theta))$*,*

$$\mathbb{E}\big[\mathbf{z} \,\big|\, \mathbf{a}^\top\mathbf{z} \geq t_0\big] \;=\; \boldsymbol{\mu}_k \;+\; \frac{\Sigma_k\,\mathbf{a}}{\sqrt{\mathbf{a}^\top\Sigma_k\mathbf{a}}}\,h(\theta). \tag{4}$$

Corollary 1 (Three testable consequences)**.** *Under the surrogate of Lemma 2: (i) the bias in Equation 4 does not depend on label correctness, so it is present in full with oracle labels; (ii)* h *is strictly increasing in* t_0*, so the bias grows as the keep rate shrinks; (iii) its direction is* $\Sigma_k\mathbf{a}$*: selection displaces the class statistic along the selector, toward the anchor when* $\mathbf{a} \propto \mathbf{t}_k$*, which can only increase the correlation that Assumption 4 bounds.*

All three consequences are observed empirically. With 100%-pure labels, top-margin selection at a matched keep rate loses to a random subset of the same size on every backbone–dataset pair tested ($70/70$ cells; up to -2.1 pp at 5% keep); the loss is monotone in the keep rate; and in the full pipeline the standard top-25% filter scores 0.47 pp below no filtering even though it raises purity. Selection is therefore the wrong tool for mean estimation, regardless of how reliable the labels are.

F.3 Stage 2a: balanced assignment is an information projection

Proposition 1 (Sinkhorn label repair)**.** *No distributional assumption is used. Let* $P_0 \in \mathbb{R}^{N\times C}_{>0}$ *be the entrywise positive row-softmax of the zero-shot logits at temperature* ε*,* $\mathbf{r} = \mathbf{1}/N$*,* $\boldsymbol{\pi}$ *a class prior with* $\pi_k > 0$ *for every* k *and* $\sum_k \pi_k = 1$*, and* $U(\mathbf{r}, \boldsymbol{\pi}) = \{P \geq 0 : P\mathbf{1} = \mathbf{r},\ P^\top\mathbf{1} = \boldsymbol{\pi}\}$ *the transport polytope. The Sinkhorn iteration converges to the information projection* $P^\star = \arg\min_{P\in U(\mathbf{r},\boldsymbol{\pi})} \mathrm{KL}(P \,\|\, P_0)$ *(Cuturi, 2013; Benamou et al., 2015), and Stage 2a outputs* $\hat{y}_i =$

$\arg\max_k P^\star_{ik}$. *The* coupling *satisfies the class-marginal constraint exactly; the hard labels inherit it only approximately, and exactly if every $N\pi_k$ is an integer and $P^\star$ is a vertex of $U(\mathbf{r}, \boldsymbol{\pi})$ (the zero-temperature limit).*

The marginal-bias component of pseudo-label error, which is intrinsic to confidence-ranked pseudo-labels even on balanced data (Wang et al., 2022), is thus removed up to the rounding gap, while the assignment stays as close as possible (in KL) to the evidence of the model. Classes with $\hat{\pi}_k = 0$ are excluded from the estimated prior and receive no images. With $\boldsymbol{\pi}_\tau = (1-\tau)\,\mathrm{uniform} + \tau\,\hat{\boldsymbol{\pi}}$ at $\tau = 1$, the target is the arg-max marginal itself, so the projection only reassigns images within that marginal (purity $77.1 \to 77.5\%$ on TinyImageNet). When the prior is correct, the repair is substantial: on TinyImageNet (true marginal 500 per class), arg-max pseudo-labels span 15–2,657 per class ($177\times$) against 321–717 ($2.2\times$) after repair onto the uniform prior. With the prior matched to each pool, purity rises on 6/6 datasets ($+0.6$ to $+5.8$ pp; Table 12) without any supervision (on SUN397 only with the estimated prior, $75.0 \to 77.4$, since the uniform prior lowers it to 66.0), and under imposed class imbalance validation moves $\tau \to 1$, restoring up to $+6.7$ pp over the uniform prior.

F.4 Stage 2b: the ridge head is the whitened prototype

Proposition 2 (Whitening identity and endpoint)**.** *This is an algebraic identity and uses no assumptions. (i) With X the matrix with rows $\mathbf{z}_i$, $i \in \mathcal{I}$, and Y the one-hot pseudo-label matrix, $\mathbf{w}_c(\alpha)$ is the L_2-normalized c-th column of the ridge solution $W = (X^\top X + \alpha I)^{-1} X^\top Y$: both heads share the sufficient statistic $\mathbf{S}_c = (X^\top Y)_{:,c}$ and differ only in the read-out map, $(G+\alpha I)^{-1}$ versus $n_c^{-1} I$. (ii) $\mathbf{w}_c(\alpha) \to \mathbf{p}_c$ as $\alpha \to \infty$: the prototype is the infinite-regularization endpoint of the same family. (iii) The read-out map uses no labels, so all label information enters through $\mathbf{S}_c$, and Lemma 1 applies to the systematic error of the ridge head unchanged.*

Numerically, the identity holds to machine precision, and at large α the minimum over classes of $\cos(\mathbf{w}_c, \mathbf{p}_c)$ equals 1.000000: the prototype recipe is contained in the ridge family as a limiting case, not merely approximated by it.

Proposition 3 (Concentration and consistency of the whitened head)**.** *Let $\bar{\alpha} = \alpha/N > 0$, $\hat{B} = (G/N + \bar{\alpha} I)^{-1}$, $B = (M + \bar{\alpha} I)^{-1}$, and $\lambda_{\min} \geq 0$ the smallest eigenvalue of M. Under Assumption 1 (using $\|\mathbf{z}_i\| = 1$), for any $\delta \in (0,1)$, with probability at least $1-\delta$,*

$$\left\| G/N - M \right\|_{\mathrm{op}} \;\leq\; \sqrt{\frac{2\|M\|_{\mathrm{op}} \log(2D/\delta)}{N}} + \frac{2\log(2D/\delta)}{3N} \;=:\; \epsilon_N, \tag{5}$$

and on that event, whenever $\epsilon_N < \lambda_{\min} + \bar{\alpha}$,

$$\left\| \hat{B}\mathbf{S}_c/n_c - B\,\tilde{\boldsymbol{\mu}}_c \right\| \;\leq\; \frac{\|\mathbf{S}_c/n_c - \tilde{\boldsymbol{\mu}}_c\|}{\lambda_{\min} + \bar{\alpha} - \epsilon_N} \;+\; \frac{\epsilon_N \,\|\tilde{\boldsymbol{\mu}}_c\|}{(\lambda_{\min} + \bar{\alpha} - \epsilon_N)(\lambda_{\min} + \bar{\alpha})}, \tag{6}$$

where $\|\mathbf{S}_c/n_c - \tilde{\boldsymbol{\mu}}_c\| = O_P(\tilde{\sigma}_c/\sqrt{n_c})$ by Lemma 1. Since $\mathrm{norm}(\cdot)$ is $2/\|B\tilde{\boldsymbol{\mu}}_c\|$-Lipschitz near $B\tilde{\boldsymbol{\mu}}_c$, for fixed $\bar{\alpha} > 0$ the normalized head is consistent for its limit: $\|\mathbf{w}_c(\alpha) - \bar{\mathbf{w}}_c\| = O_P\big(n_c^{-1/2} + \sqrt{\log D/N}\big)$, a label-driven term and a label-free Gram term from all N samples.

This is a fixed-$\bar{\alpha}$ statement. At $\alpha = 1$ of the pipeline ($\bar{\alpha} = 1/N$), Equation 6 is informative only if $\lambda_{\min}(M)$ is bounded away from zero, which we have not measured: our check that the top eigenvector of G computed from half the data matches the full-data one to $\cos \geq 0.99$ on every pool tests the dominant direction, not the bottom of the spectrum that governs $(G+\alpha I)^{-1}$. We therefore claim no rate at the operating point, where the whitened head is justified by Proposition 2(i) and by its measured accuracy.

Proposition 4 (Whitening attenuates the shared component; leading order)**.** *Under Assumption 3, for classes $j \neq k$, at leading order in the within-class spread (so that $\mathbf{S}_c/n_c \approx \boldsymbol{\mu}_c$),*

$$\cos(\mathbf{p}_j, \mathbf{p}_k) \;\approx\; \frac{\beta_j\beta_k + \langle \boldsymbol{\Delta}_j, \boldsymbol{\Delta}_k \rangle}{\sqrt{\beta_j^2 + \|\boldsymbol{\Delta}_j\|^2}\sqrt{\beta_k^2 + \|\boldsymbol{\Delta}_k\|^2}} \tag{7}$$

which tends to 1 as $\beta^2/\|\boldsymbol{\Delta}\|^2$ grows, whereas the whitened directions satisfy

$$\cos\big(\mathbf{w}_j(\alpha), \mathbf{w}_k(\alpha)\big) \;\approx\; \frac{\kappa^2\,\beta_j\beta_k + \langle \tilde{\boldsymbol{\Delta}}_j, \tilde{\boldsymbol{\Delta}}_k \rangle}{\sqrt{\kappa^2\beta_j^2 + \|\tilde{\boldsymbol{\Delta}}_j\|^2}\,\sqrt{\kappa^2\beta_k^2 + \|\tilde{\boldsymbol{\Delta}}_k\|^2}}, \qquad \kappa = \frac{N\lambda_r + \alpha}{N\varrho + \alpha} \;<\; 1, \tag{8}$$

with $\tilde{\mathbf{\Delta}}_k = (N\lambda_r + \alpha)(NM_r + \alpha I)^{-1}\mathbf{\Delta}_k$ and λ_r a reference eigenvalue of M_r (exact when $\mathbf{\Delta}_j, \mathbf{\Delta}_k$ lie in one eigenspace of M_r, approximate otherwise): the shared component is attenuated by the eigenvalue ratio κ. The proposition does not bound $\langle \tilde{\mathbf{\Delta}}_j, \tilde{\mathbf{\Delta}}_k \rangle$: near-orthogonality of the whitened vectors is measured, not proved; the statement says only that prototype coherence comes from the spiked spectrum of the representation, not from the classes.

On all six pools (SigLIP2, Table 12) the mean inter-class cosine is 0.71–0.88 for prototypes and $|\cos| \leq 0.01$ for whitened heads: whitening turns a set of nearly collinear class vectors into a nearly orthogonal one. The ablation that changes only the head, with labels fixed, gives a mean gain of $+0.96$ pp over 18 backbone–dataset pairs (Table 4), which is the accuracy value of that decorrelation.

Remark 1 (Derived, not trained). W is the unique global minimizer of the ridge objective and is obtained by one $D \times D$ solve, so gradient descent on that objective could only return the same point. A linear head trained on a *different* objective (cross-entropy with AdamW) for 300 epochs remains 0.63 pp below the closed form on average (17/18 pairs) at $18\times$ the compute, and its accuracy *decreases* from epoch 50 to 300 on the largest datasets; under label noise, the quadratic objective with explicit ridge control is the better choice. Iterating the pipeline (relabel–refit) loses a further 0.79 pp after five rounds (Figure 18) even though purity rises, which is the same purity–accuracy dissociation as in Lemma 2.

F.5 The optimal fusion weight: an inverse-sample-size law with a bias floor

The calibrated classifier is $\mathbf{g}_k(\lambda) = \lambda\, \mathbf{t}_k + (1-\lambda)\, \hat{\mathbf{h}}_k$, where $\hat{\mathbf{h}}_k$ is any head of Definition 1 built from n images of class k. We analyze the mean-squared surrogate $R(\lambda) = \mathbb{E}\|\mathbf{g}_k(\lambda) - \mathbf{h}_k^{\mathrm{opt}}\|^2$ with minimizer $\lambda^\star$. R is not the 0–1 risk: it penalizes deviations in every direction, whereas the 0–1 risk depends only on score differences between classes. Hence $\lambda^\star$ is surrogate-optimal, and the interior optimum, the functional form in n, and the comparative statics are hypotheses that Corollary 3 tests against measured accuracy.

Theorem 1 (Optimal fusion, estimator-generic)**.** *Under Assumptions 1, 2 and 4 with $\gamma = 0$, assume the head error decomposes to leading order in $1/n$ (delta method for the normalized head) as $\mathbb{E}\|\hat{\mathbf{h}}_k - \mathbf{h}_k^{\mathrm{opt}}\|^2 = b_h^2 + \sigma_h^2/n$, with n-independent floor $b_h^2 = \|\bar{\mathbf{h}}_k - \mathbf{h}_k^{\mathrm{opt}}\|^2$ (misspecification, i.e. the limit is not the target, plus the contamination displacement of Lemma 1 and Proposition 3) and sampling variance σ_h^2/n; b_t^2 is the squared anchor bias of Assumption 2. Then $R(\lambda) = \lambda^2 b_t^2 + (1-\lambda)^2(b_h^2 + \sigma_h^2/n)$ is strictly convex with unique minimizer*

$$\boxed{\mathrm{odds}(\lambda^\star) := \frac{\lambda^\star}{1-\lambda^\star} = A + \frac{B}{n}, \qquad A = \frac{b_h^2}{b_t^2}, \quad B = \frac{\sigma_h^2}{b_t^2}} \tag{9}$$

(the odds law*). The head enters only through (b_h^2, σ_h^2), so switching from the prototype to the ridge head changes the constants, not the form; for one λ shared by all classes the pooled surrogate $\sum_k R_k$ has the same form with b_t^2, b_h^2, σ_h^2 summed over classes; and because b_h^2 contains the misspecification term, the floor A does not vanish with oracle labels (Corollary 3).*

Lemma 3 (Robustness to imperfect obliqueness)**.** *Under Assumption 4 with signed correlation exactly $\gamma \in (-1, 1)$, write $V = b_h^2 + \sigma_h^2/n$ and $\eta = \sqrt{V}/b_t$ (the head-to-anchor error ratio). The minimizer of the perturbed risk $R_\gamma(\lambda) = \lambda^2 b_t^2 + (1-\lambda)^2 V + 2\lambda(1-\lambda)\gamma\, b_t\sqrt{V}$ is*

$$\lambda^\star(\gamma) = \frac{\eta\,(\eta - \gamma)}{(\eta-\gamma)^2 + 1 - \gamma^2}, \tag{10}$$

which is interior ($0 < \lambda^\star < 1$) if and only if $\gamma < \min(\eta,\, 1/\eta)$: the interior optimum survives any γ smaller than both the head-to-text error ratio and its inverse.

At the fitted operating point ($\eta = \sqrt{0.377} = 0.61$), the working bound $|\gamma| \leq 0.15$ moves $\lambda^\star$ from 0.274 to 0.239 ($\gamma = +0.15$) or 0.300 ($\gamma = -0.15$), i.e. by at most 0.04, which lies well inside the measured flat region (± 0.1 in λ costs < 0.2 points). Equation Equation 10 agrees exactly with direct numerical minimization on 2,000 random (b_t, V, γ) instances.

Corollary 2 (Interior optimum)**.** *Under the assumptions of Theorem 1 and Lemma 3, if $b_t^2 > 0$, $b_h^2 + \sigma_h^2/n > 0$ and $\gamma < \min(\eta, 1/\eta)$, then $0 < \lambda^\star < 1$: the blend has strictly lower surrogate risk than both pure classifiers, for every finite n and every purity level.*

Corollary 3 (Empirical check of the law for the ridge head)**.** *In the sample-size sweep of Figure 7 (SigLIP2, TinyImageNet, single split; ridge head from n oracle-labeled images per class with G from all images of the pool; $\lambda^\star$ the accuracy-maximizing weight on a 0.01 grid over a disjoint evaluation split of 250 images per class), regressing $\mathrm{odds}(\lambda^\star)$ on $1/n$ over $n \in \{1, 2, 5, 10, 25, 50, 100, 250\}$ gives $\mathrm{odds}(\lambda^\star) = 0.377 + 1.209/n$ ($A = 0.377$, $B = 1.209$) with $R^2 = 0.966$, against $R^2 = 0.39$ for the no-floor law $\mathrm{odds}(\lambda^\star) = B/n$ on the same response. The fit reproduces the measured $\lambda^\star$ from 0.61 at $n = 1$ to 0.28 at $n = 250$ (measured 0.23); since the $n = 1$ point carries leverage 0.80, the constants are determined mainly by small n, and we attach no standard errors to them. The implied asymptote $\lambda^\star_\infty = A/(1 + A) = b_h^2/(b_h^2 + b_t^2) = 0.27$ matches the 0.29 of the prototype head (same sweep, Appendix H.3). With oracle labels, the floor contains no contamination, so it measures the misspecification term of Theorem 1: the anchor carries class information that the whitened head does not converge to (Finding 2: fusion beats the oracle-label head on every pool). The smaller variance coefficient of the ridge head ($B = 1.21$ versus 4.48 for the prototype) is the quantity behind the head gain in the ablation.*

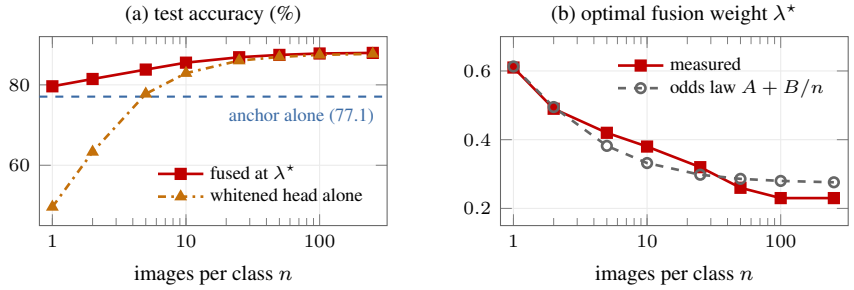

Figure 7: Sample-size sweep for the ridge head (SigLIP2, TinyImageNet; oracle labels, n images per class). (a) The whitened head alone overtakes the anchor between $n = 5$ and $n = 10$; the fusion is above both at every n, from 79.6% at $n = 1$ to 88.0% at $n = 250$. (b) The accuracy-optimal $\lambda^\star$ falls from 0.61 to 0.23 along the odds law $\mathrm{odds}(\lambda^\star) = A + B/n$ of Theorem 1.

Corollary 4 (Monotone comparative statics)**.** *Under the assumptions of Theorem 1, $\lambda^\star$ is (i) increasing in b_h^2: degrading label purity shifts weight to text, observed on every dataset when moving from oracle to arg-max labels; (ii) decreasing in b_t^2: a stronger text classifier receives more weight; (iii) decreasing in n toward the floor $\lambda^\star_\infty = A/(1 + A) > 0$.*

F.6 Assumption audit

Each limitation of scope is stated where it first matters: the leading-order linearization in Section F and Proposition 4; the absence of a rate at $\alpha = 1$ in Proposition 3; the proxy-based working bound $\gamma \le 0.15$; and the distinction between surrogate and 0–1 risk in Section F.5. In addition, all heads are compared against the single target $\mathbf{h}_k^{\mathrm{opt}}$, so the bias floor mixes misspecification and contamination, which the theory does not separate. Claims about the relative accuracy of different heads rest on the measured ablations (Table 4), not on Theorem 1.

G Proofs

Proof of Lemma 1. Conditionally on the assignment, $\hat{\mathbf{m}}_k = q_k \bar{\mathbf{z}}^{\mathrm{cor}} + (1-q_k)\bar{\mathbf{z}}^{\mathrm{int}}$ with $\mathbb{E}\bar{\mathbf{z}}^{\mathrm{cor}} = \boldsymbol{\mu}_k^{\mathrm{cor}}$, $\mathbb{E}\bar{\mathbf{z}}^{\mathrm{int}} = \boldsymbol{\nu}_k$, giving $\tilde{\boldsymbol{\mu}}_k = \boldsymbol{\mu}_k^{\mathrm{cor}} + (1-q_k)\boldsymbol{\delta}_k$. The fluctuation averages n independent bounded (hence sub-Gaussian) vectors of total variance $\tilde{\sigma}_k^2$ (within-component spreads plus the between-component term $q_k(1-q_k)\|\boldsymbol{\delta}_k\|^2$), so vector Bernstein gives $O_P(\tilde{\sigma}_k/\sqrt{n})$; for deterministic B use linearity and $\|B(\hat{\mathbf{m}}_k - \tilde{\boldsymbol{\mu}}_k)\| \le \|B\|_{\mathrm{op}}\|\hat{\mathbf{m}}_k - \tilde{\boldsymbol{\mu}}_k\|$. □

Proof of Lemma 2. Write $u = \mathbf{a}^\top \mathbf{z} \sim \mathcal{N}(\mathbf{a}^\top \boldsymbol{\mu}_k, s^2)$ with $s^2 = \mathbf{a}^\top \Sigma_k \mathbf{a}$ and decompose $\mathbf{z} = \boldsymbol{\mu}_k + \Sigma_k \mathbf{a}\,(u - \mathbf{a}^\top \boldsymbol{\mu}_k)/s^2 + \mathbf{v}$ with $\mathbf{v}$ mean-zero Gaussian and independent of u (regression decomposition; this is where Gaussianity is used). Conditioning on $u \ge t_0$ leaves $\mathbb{E}[\mathbf{v}] = 0$ and $\mathbb{E}[u - \mathbf{a}^\top \boldsymbol{\mu}_k \mid u \ge t_0] = s\,h(\theta)$, which is Equation 4. Monotonicity of h gives Corollary 1(ii), label-agnosticism of the display gives (i), and the direction $\Sigma_k \mathbf{a}$ gives (iii). □

Proof of Proposition 1. For $P_0 > 0$ entrywise and marginals $\mathbf{r}, \boldsymbol{\pi} > 0$ of equal mass, $U(\mathbf{r}, \boldsymbol{\pi})$ is a nonempty compact polytope and $P \mapsto \mathrm{KL}(P\|P_0)$ strictly convex, so the projection exists and is unique, and Sinkhorn–Knopp (alternating Bregman projection onto the two marginal constraints) converges to it (Cuturi, 2013, Lemma 2), (Benamou et al., 2015). Taking $\mathbf{r} = \mathbf{1}/N$ while P_0 has unit row sums shifts the objective by the constant $\log N$, not the minimizer. If every $N\pi_k$ is an integer, the vertices of $U(\mathbf{r}, \boldsymbol{\pi})$ are scaled assignment matrices with exactly $N\pi_k$ ones per column, so when $P^\star$ is a vertex the arg-max labels reproduce the marginal exactly and otherwise only up to the rounding of a fractional coupling. □

Proof of Proposition 2. (i) With one-hot Y, $(X^\top Y)_{:,c} = \mathbf{S}_c$ and $X^\top X = G$ on $\mathcal{I}$, so column c of $(G + \alpha I)^{-1} X^\top Y$ is $(G + \alpha I)^{-1}\mathbf{S}_c$. (ii) $(G + \alpha I)^{-1}\mathbf{S}_c = \alpha^{-1}\big(\mathbf{S}_c - G\mathbf{S}_c/\alpha + O(\alpha^{-2})\big)$, whose direction converges to that of $\mathbf{S}_c$, i.e. $\mathbf{p}_c$. (iii) G contains no labels. □

Proof of Proposition 3. Equation 5 is matrix Bernstein for the sum of the independent, centered, bounded matrices $\mathbf{z}_i\mathbf{z}_i^\top - M$ (Tropp, 2015): $\mathbf{z}_i\mathbf{z}_i^\top$ and M are positive semidefinite with operator norm at most 1, so $\|\mathbf{z}_i\mathbf{z}_i^\top - M\|_{\mathrm{op}} \le 1$, and $\mathbb{E}(\mathbf{z}_i\mathbf{z}_i^\top - M)^2 = \mathbb{E}[\mathbf{z}\mathbf{z}^\top] - M^2 \preceq M$ gives the variance proxy $\|M\|_{\mathrm{op}}$. For Equation 6, write $\hat{B}\,\mathbf{S}_c/n_c - B\tilde{\boldsymbol{\mu}}_c = \hat{B}(\mathbf{S}_c/n_c - \tilde{\boldsymbol{\mu}}_c) + (\hat{B} - B)\,\tilde{\boldsymbol{\mu}}_c$. On the Bernstein event, the Weyl inequality gives $\lambda_{\min}(G/N) \ge \lambda_{\min} - \epsilon_N$, hence $\|\hat{B}\|_{\mathrm{op}} \le (\lambda_{\min} + \bar{\alpha} - \epsilon_N)^{-1}$ when $\epsilon_N < \lambda_{\min} + \bar{\alpha}$, and the resolvent identity $\hat{B} - B = \hat{B}\,(M - G/N)\,B$ gives $\|\hat{B} - B\|_{\mathrm{op}} \le \epsilon_N\,\big[(\lambda_{\min} + \bar{\alpha} - \epsilon_N)(\lambda_{\min} + \bar{\alpha})\big]^{-1}$, which is Equation 6. The class-mean fluctuation is $O_P(\tilde{\sigma}_c/\sqrt{n_c})$ by Lemma 1, and $\|\mathrm{norm}(\mathbf{b}) - \mathrm{norm}(\mathbf{a})\| \le 2\|\mathbf{b} - \mathbf{a}\|/\|\mathbf{a}\|$ for $\mathbf{a} = B\tilde{\boldsymbol{\mu}}_c \ne 0$, $\mathbf{b} \ne 0$ transfers the rate to the normalized head. □

Proof sketch of Proposition 4. By Lemma 1, at leading order $\mathbf{S}_c/n_c \approx \boldsymbol{\mu}_c = \beta_c\mathbf{u} + \boldsymbol{\Delta}_c$; the first display is the cosine of two such vectors, using $\langle \mathbf{u}, \boldsymbol{\Delta} \rangle = 0$. For the second, decompose $(G + \alpha I)^{-1}$ on the eigenbasis of $G \approx N\varrho\,\mathbf{u}\mathbf{u}^\top + NM_r$, an eigendecomposition because $M_r\mathbf{u} = 0$: the $\mathbf{u}$-component of $\mathbf{S}_c$ is scaled by $(N\varrho + \alpha)^{-1}$ and a component in an eigenspace of M_r at eigenvalue λ_r by $(N\lambda_r + \alpha)^{-1}$; factoring out the latter gives the ratio κ and the vectors $\tilde{\boldsymbol{\Delta}}$, and cross terms between $\mathbf{u}$ and the range of M_r vanish by orthogonality. Replacing M by G/N costs $O(\epsilon_N)$ per eigenvalue (Weyl) and $\mathbf{S}_c/n_c \approx \boldsymbol{\mu}_c$ costs the contamination and sampling terms of Lemma 1; neither is tracked, hence leading order. □

Proof of Theorem 1 and Lemma 3. $\mathbf{g}_k(\lambda) - \mathbf{h}_k^{\mathrm{opt}} = \lambda\,(\mathbf{t}_k - \mathbf{h}_k^{\mathrm{opt}}) + (1-\lambda)\,(\hat{\mathbf{h}}_k - \mathbf{h}_k^{\mathrm{opt}})$. The first term contributes $\lambda^2 b_t^2$, the second $(1-\lambda)^2 V$ with $V = b_h^2 + \sigma_h^2/n$, and the cross term $2\lambda(1-\lambda)\gamma' b_t\sqrt{V}$ with $|\gamma'| \le \gamma$ (Assumption 4), giving R_γ. For $|\gamma| < 1$, R_γ is strictly convex since its leading coefficient is $b_t^2 + V - 2\gamma b_t\sqrt{V} = (b_t - \gamma\sqrt{V})^2 + (1 - \gamma^2)V > 0$. Setting $R'_\gamma(\lambda) = 0$ and substituting $\eta = \sqrt{V}/b_t$ yields Equation 10 by the identity $1 + \eta^2 - 2\gamma\eta = (\eta - \gamma)^2 + 1 - \gamma^2$. Positivity of the numerator requires $\eta > \gamma$, and $\lambda^\star < 1$ is equivalent to $1 - \gamma\eta > 0$; together these give interiority if and only if $\gamma < \min(\eta, 1/\eta)$. At $\gamma = 0$, Equation 10 reduces to $\eta^2/(1 + \eta^2)$, which rearranges to Equation 9; the corollaries follow from the closed form. □

H Anchor–Representative Fusion: Full Results

All numbers in this appendix are computed in closed form from the cached SigLIP2 and CLIP ViT-B/16 image embeddings $\mathbf{z}_i$ and image-to-class-text logits of the CIFAR-100, TinyImageNet, SUN397, Places365 and ImageNet-1k pools (Flowers-102 is added for two image-side diagnostics), with class names embedded by the single template `a photo of a {label}.` (Appendix C). Every cache passed our validation check (unit-norm embeddings, no clipped cosines, arg-max of the logits reproduces the cached labels, text-embedding recovery within 0.02 of unit norm). Notation follows Sections 2 and 3: $\mathbf{t}_k$ is the text embedding of class k and $\mathbf{p}_k$ the normalized mean of the images assigned to it, grouped either by ground truth (the *oracle* prototype, an upper bound) or by the arg-max of the model itself (the *pseudo* prototype). Prototypes use the leave-one-out correction unless stated otherwise. Throughout this appendix, $\lambda^\star$ denotes the accuracy-optimal fusion weight on the evaluated images, i.e. an oracle λ. These are diagnostics of what fusion can achieve, not deployable settings; the deployable policies are defined in Section 4.1 and Appendix D.

H.1 Geometry: Text and Prototype References Are Near-Orthogonal

We interpolate the two class references with $\lambda \in [0, 1]$,

$$\tilde{t}_k \;=\; \lambda \mathbf{t}_k + (1-\lambda)\mathbf{p}_k, \qquad s_{i,k} \;=\; \mathbf{z}_i^\top \tilde{t}_k, \tag{11}$$

optionally renormalizing $\tilde{t}_k$ (the *normalized* blend). Since the score is linear in the weights, Equation 11 is a convex combination of two similarity matrices, and the whole λ sweep can be run from the cache without a model forward pass; blending classifier weights is the same as ensembling two classifiers at the score level. **For a given class, the text embedding and the visual prototype are close to perpendicular**: $\cos(\mathbf{t}_k, \mathbf{p}_k)$ averaged over classes (oracle prototypes) is 0.11 to 0.16 for SigLIP2 ($81°$–$83°$) and 0.28 to 0.36 for CLIP ViT-B/16 ($69°$–$74°$), so CLIP is consistently better aligned (Table 15, oracle column). This is the modality gap at the class level, and it is what makes Equation 11 worthwhile: two near-orthogonal references have decorrelated errors, so their ensemble is informative rather than redundant.

Raw cosine distances are not comparable across modalities or datasets. As a consequence of this near-orthogonality, the mean image-to-true-text cosine (0.09 to 0.13 for SigLIP2) and the mean image-to-true-prototype cosine (0.77 to 0.94, i.e. $21°$–$39°$) live on incommensurable scales. The image–image anisotropy floor (mean cosine between random cross-class image pairs) also varies from 0.48 to 0.74 across datasets, so prototype distances cannot be compared across datasets either. Ordering the datasets by text distance does not reproduce their accuracy ordering, and the pooled mean margin (own class minus best other) is negative for the text classifier on Places365 despite its 44.7% accuracy. We therefore report accuracy and rank statistics instead of distances.

Effect of leave-one-out. The leave-one-out correction, which removes each image from the prototype it is scored against, changes prototype accuracy by at most 0.7 points at the class sizes used in this work (-0.73 on SUN397 with 100–2 361 images per class, -0.31 on TinyImageNet, -0.09 on Places365). Its effect is largest relative to the class size on Flowers-102, whose 10 images per class give 100% in-sample and 99.90% leave-one-out accuracy.

H.2 Main Result: the λ Ablation

Figure 8 gives the central result of this appendix: **on all ten backbone–dataset configurations, an interior λ strictly outperforms both endpoints**. Neither the text classifier ($\lambda = 1$) nor the prototype classifier ($\lambda = 0$) is ever optimal.

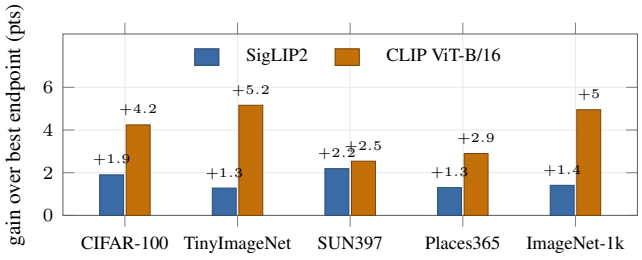

Figure 8: λ ablation with oracle (ground-truth-grouped) LOO prototypes: gain of the better of the raw and normalized blends at its accuracy-optimal $\lambda^\star$ (oracle λ) over the better of its two endpoints, the text classifier (λ=1) and the prototype classifier (λ=0). The blend is ahead of both on every pool and backbone, by 1.3 to 2.2 points for SigLIP2 and 2.5 to 5.2 for CLIP ViT-B/16; the normalized $\lambda^\star$ lies between 0.36 and 0.59 in every case.

The optimum is broad: on TinyImageNet every $\lambda \in [0.33, 0.47]$ lies within 0.1 points of the best, so λ does not need to be tuned precisely, which is what makes a transferred value viable. **The weaker backbone benefits more**: CLIP gains $+3.96$ points on average versus $+1.64$ for SigLIP2, which suggests that the headroom shrinks as encoders improve. Renormalization is not uniformly better (it wins on SUN397 by $+0.55$ and loses slightly on Places365 by -0.19), but it makes $\lambda^\star$ far more consistent across datasets, which is why we prefer it for transfer.

Table 13: Complementarity of the two classifiers (SigLIP2, raw blend at the oracle $\lambda^\star$). "ceiling" is the accuracy of an oracle that picks the correct endpoint per image; the last column is the fraction of that headroom captured by a single global λ.

Dataset	text ok / proto bad	proto ok / text bad	both bad	ceiling	captured
CIFAR-100	2 297	2 743	3 684	92.63%	41%
TinyImageNet	3 975	11 561	11 368	88.63%	32%
SUN397	6 883	14 416	12 762	88.27%	26%

H.3 Why Blending Helps

The gain has more than one source. Varying the number of labeled images per prototype separates a variance effect from a bias effect, and the geometry of the blended weights adds a third.

(i) Variance reduction: text acts as a prior. We vary the number n of oracle-labeled images per prototype on SigLIP2 / TinyImageNet, with a disjoint held-out evaluation split (the sweep of Figure 7 with the prototype head in place of the ridge head). The prototype alone needs about $n = 5$ images per class to overtake the anchor. The fusion is ahead of both at every n: its gain over the prototype falls from $+28.7$ points at $n = 1$ to $+1.4$ at $n = 250$ but never disappears, and $\lambda^\star$ decreases from 0.84 at $n = 1$ to 0.33 at $n = 100$ (0.38 at $n = 250$) as the prototype becomes more reliable. This is the classical shrinkage pattern: a noisy empirical estimate is pulled toward a stable prior, and less so as its variance decreases.

(ii) An irreducible complementary component. If the effect were pure variance reduction, the gain would vanish as $n \to \infty$. It does not: the gain plateaus at roughly 1.5 points from $n = 100$ onward, and Places365 still gains $+1.30$ with about 4941 images per class. Table 13 quantifies the complementarity directly and bounds what a perfect per-image router could achieve.

Between 2 297 and 6 883 images are classified correctly *only* by the text head. A single global λ captures just 26–41% of the attainable headroom, which is why we regard per-image or per-class weighting as a promising direction for future work.

(iii) Improved classifier conditioning. Prototype weight vectors are severely crowded: on SigLIP2 / TinyImageNet their mean inter-class cosine is 0.843, against 0.729 for the text directions. For $\lambda \geq 0.45$ the blend is better separated than *either* component, reaching 0.697 at $\lambda = 0.7$, because the shared cone component partially cancels while the class-discriminative parts add. At the accuracy-optimal $\lambda^\star = 0.40$, however, the blend lies between the two components, so conditioning explains part of the gain over the prototype but not the gain over the anchor. Separation keeps improving past $\lambda^\star$ while accuracy declines, which indicates that $\lambda^\star$ balances conditioning against fidelity to the image distribution.

H.4 From Oracle to Pseudo-Labels: Dilution, Repair, Selection and the Trained Head

Sections H.2–H.3 group prototypes by ground truth and therefore give upper bounds. Replacing the labels with the predictions of the model itself brings us to the setting of the main text, whose systematic ablations are in Table 4. The diagnostics below explain why that table looks the way it does: why noisy labels cost less than one would expect, why filtering them costs more, and why a trained head is the better use of the same labels.

Aggregation dilutes label noise. A pseudo-*label* can never be more accurate than the classifier that produced it, but a pseudo-*prototype* can, because independent errors cancel under averaging (Lemma 1). Table 14 shows 13–37% label error turning into 0.08–0.64% prototype error, so the class statistic is far more reliable than any individual label. Grouping by the decisions of the anchor itself has a different cost: it rotates the prototype toward the text embedding (Table 15), and $\lambda^\star$ rises from 0.26 to 0.74 as the prototypes degrade, which means that the fusion has to lean on the anchor exactly when the visual side is least independent of it.

Table 14: Label noise versus prototype noise. "prototype error" is $1 - \cos(\mathbf{p}_k^{\text{pseudo}}, \mathbf{p}_k^{\text{oracle}})$ averaged over non-empty classes. Aggregation suppresses label noise by 39–158×.

Backbone	Dataset	label error	prototype error	reduction	empty classes
SigLIP2	CIFAR-100	12.85%	0.08%	158×	0 / 100
SigLIP2	TinyImageNet	22.93%	0.46%	50×	0 / 200
CLIP ViT-B/16	TinyImageNet	36.59%	0.44%	84×	0 / 200
SigLIP2	SUN397	24.99%	0.64%	39×	0 / 397

Table 15: $\cos(\mathbf{t}_k, \mathbf{p}_k)$ under different grouping rules. Values increase monotonically toward the text direction as supervision is removed and as filtering becomes more aggressive.

Setting	oracle	pseudo	top-50%	top-25%	top-10%
SigLIP2 / TinyImageNet	+0.1422	+0.1493	+0.1596	+0.1638	+0.1668
SigLIP2 / SUN397	+0.1606	+0.1648	+0.1703	+0.1731	+0.1750
CLIP ViT-B/16 / TinyImageNet	+0.2929	+0.3012	+0.3098	+0.3150	+0.3194

Selection bias, not label noise, is the cost of confidence filtering. With perfect labels, the top-10% margin subset costs 4.9 points relative to a random 10% subset of the same size (Table 16); purity is identical in the two rows, so the loss can only come from which images are kept. The hold-out version of this control on all caches is in Figure 9 (Appendix I.2). Pseudo-label prototypes retain 42% (TinyImageNet, both backbones) and 39% (SUN397) of the oracle gain, but this fraction hides two failure cases: on CLIP / TinyImageNet the pseudo prototype alone (62.69%) falls below the text classifier that produced its labels (63.41%), and on the 24× imbalanced SUN397 the arg-max marginal spans 39 to 1,621 images per class against a true range of 100 to 2,361, so some classes are represented by a handful of images while others absorb their neighbors.

Repair. Balanced assignment raises purity from 77.1% to 82.9% and alignment with the oracle prototype from 0.9954 to 0.9976 without supervision (Table 17); with a prior that matches the pool, it is the one label-free operation in the table that raises purity and alignment at once. The alternatives do not: self-training degrades monotonically (79.75 → 78.60 over five iterations), soft weighting is neutral at $T{=}1$ and collapses at higher temperature, and orthogonalizing $\mathbf{p}_k$ against $\mathbf{t}_k$ is harmful for both backbones, strongly so for CLIP (62.69 → 54.72). The last result is informative in itself: the text-aligned component of the prototype carries signal, not just bias, so removing the overlap between the two references removes evidence.

Selection granularity. Per-class selection beats a global threshold by 11.9 to 15.9 points at a matched keep rate despite lower purity (SUN397 at 10%: 73.11 versus 43.00 for the prototype alone, with 0 versus 142 empty classes; Table 18). Confidence is not comparable across classes, so a global threshold silently deletes the hard classes. Even with the right granularity, no keep rate beats using all pseudo-labels for a trained head (82.76 → 81.92 → 81.13 → 80.25), which is why the method of the main text filters nothing.

Trained head. A closed-form ridge classifier $W = (X^\top X + I)^{-1} X^\top Y$ on the same embeddings and labels is uniformly better than prototype averaging, by +1.7 to +7.1 points (SigLIP2 / TinyImageNet 88.43 versus 84.97 with oracle labels and 82.76 versus 79.75 with pseudo-labels; CLIP ViT-B/16 73.98 versus 66.84 and 67.38 versus 62.69; SUN397 85.90 versus 82.67 and 78.74 versus 77.04). The gain appears with oracle and with pseudo-labels alike, so it comes from the read-out map rather than from the labels. On this cache the trained head largely absorbs the text contribution ($\lambda^\star \to 0$ with pseudo-labels on SigLIP2 / TinyImageNet), whereas over the 18 encoder–pool pairs fusion still adds +0.85 on top of the whitened head (Table 4); the anchor remains useful in general even where one well-behaved cache can do without it. The whitened head of Section 3 resolves these observations, and Table 4 shows that balanced labels and the whitened head compose.

Table 16: Oracle versus pseudo-label prototypes (in-sample prototypes throughout, for a like-for-like comparison). The two control rows use ground-truth labels and differ only in how the 10% subset is chosen (top-10% zero-shot margin within each class); the pseudo-label filtering rows are in Table 18.

Setting	prototype source	purity	λ=0	best acc.	$\lambda^\star$
SigLIP2 TinyImageNet (text: 77.07)	oracle, all	100.0%	84.97	86.12	0.26
	pseudo, all	77.1%	79.75	80.87	0.44
	control: oracle, top-10% margin	100.0%	79.42	83.19	0.52
	control: oracle, random 10%	100.0%	**84.30**	85.81	0.32
SigLIP2 SUN397 (text: 75.01)	oracle, all	100.0%	82.67	84.08	0.46
	pseudo, all	75.0%	77.04	78.55	0.54
	control: oracle, random 10%	100.0%	81.47	83.85	0.52
CLIP ViT-B/16 TinyImageNet (text: 63.41)	oracle, all	100.0%	66.84	71.23	0.48
	pseudo, all	63.4%	62.69	66.61	0.54
	control: oracle, random 10%	100.0%	65.58	70.64	0.54

Table 17: Improving pseudo-prototypes. "align" is $\cos(\mathbf{p}_k, \mathbf{p}_k^{\text{oracle}})$. All rows are label-free except ORACLE. † uniform-prior Sinkhorn fails on SUN397, whose classes are imbalanced by 23.6×; the estimated-prior variant repairs this.

Method	Backbone / Dataset	purity	align	λ=0	best acc.
SigLIP2 / TinyImageNet (text 77.07, ORACLE 86.12)					
M0 pseudo argmax (baseline)		77.1%	0.9954	79.75	80.87
M1 soft-weighted (T=1)		—	0.9950	79.68	80.64
M2 Sinkhorn soft, uniform		—	0.9969	81.52	82.54
M2b Sinkhorn hard, uniform		82.9%	**0.9976**	**81.80**	**82.89**
M2c Sinkhorn hard, estimated		77.5%	0.9949	79.51	80.72
M3 self-training (5 iters)		—	0.9953	78.60	80.98
M4 orthogonalized vs. text		—	—	78.95	80.89
M6 argmax + conf. top-75%		91.6%	0.9919	80.94	81.96
M7 M2b + conf. top-75%		93.4%	0.9946	81.33	82.22
CLIP ViT-B/16 / TinyImageNet (text 63.41, ORACLE 71.23)					
M0 pseudo argmax		63.4%	0.9956	62.69	66.61
M2b Sinkhorn hard, uniform		68.4%	**0.9977**	**64.21**	**68.35**
M3 self-training (5 iters)		—	0.9946	59.26	66.35
M4 orthogonalized vs. text		—	—	54.72	66.59
M6 argmax + conf. top-75%		76.1%	0.9943	63.40	66.83
SigLIP2 / SUN397 (text 75.01, ORACLE 84.08)					
M0 pseudo argmax		75.0%	0.9936	77.04	78.55
M2b Sinkhorn hard, uniform†		66.0%	0.9884	72.75	77.43
M2c Sinkhorn hard, estimated		77.4%	**0.9940**	**77.93**	**79.34**
M8 M2c + conf. top-75%		83.2%	0.9931	77.53	79.15

I Additional Results

I.1 Main table with standard deviations

The sample standard deviation over the five seeds is at most 0.7 in every cell of Table 2 (the largest is 0.69, for CLIP ViT-B/16 on CIFAR-100; "zero-shot" is the test-split accuracy of the anchors, so it also varies with the split). Every gain of ours (validated) over the prototype recipe exceeds the sum of the two standard deviations, with the exception of DINOv2 on Food-101 (+0.08) and SUN397 (+0.02).

Table 18: Per-class versus global confidence selection at matched keep rates (SigLIP2 / TinyImageNet, pseudo-labels). Global thresholds delete entire hard classes; per-class selection guarantees coverage. "+text" is the best λ blend. Reference at 100% keep: the "pseudo, all" row of Table 16 (prototype 79.75 / 80.87) and 82.76 for the linear probe alone and blended.

				prototype		linear probe	
keep	selection	purity	empty	alone	+text	alone	+text
50%	per-class	89.6%	0	78.43	80.21	81.92	82.00
50%	global	98.4%	10	78.59	80.53	80.75	81.54
25%	per-class	92.7%	0	**76.28**	**79.46**	**81.13**	**81.52**
25%	global	99.8%	19	73.74	77.07	76.38	79.05
10%	per-class	94.3%	0	**73.52**	**78.87**	**80.25**	**81.14**
10%	global	99.9%	50	61.63	77.07	64.32	78.02

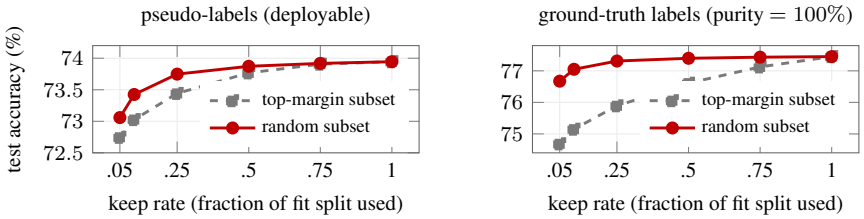

Figure 9: **Confidence filtering lowers accuracy even with perfect labels.** Representatives from a fraction (keep rate) of the fit split chosen by the highest zero-shot margin or at random (10 draws), fused with the anchor; mean over 14 caches. Right: ground-truth labels, so purity is 100% for both rules and the gap is selection bias alone (Appendix I). Each rule is scored at its accuracy-optimal λ on the evaluated half (an oracle λ, for diagnosis only).

I.2 Selection controls

Figure 9 averages the hold-out protocol over 14 caches (SigLIP2 and ViT-B/16 on all six pools; ViT-L/14 on CIFAR-100 and TinyImageNet). Table 19 applies the same protocol to the 10 caches without ImageNet-1k and Places365, which is why its absolute accuracies are higher than those in the figure. At each keep rate, the top-margin, top-confidence, and lowest-entropy rules are compared with ten random subsets of the same per-class size; Table 19 reports the margin rule, and the pattern is the same for the confidence and entropy rules and under the leave-one-out protocol. The table separates the two effects that confidence filtering mixes. With ground-truth labels, purity is perfect for both rules, so the gap between the margin and random columns is selection bias alone; it is present at every keep rate and largest when few images are kept. With pseudo-labels, the gap shrinks, because the filter now also removes wrong labels, but it never reverses: the purity gained by filtering does not pay for the images it discards. The statistical evidence matches this reading. Per-image McNemar tests on TinyImageNet favor the random subset in every cell at keep $\leq 25\%$ with ground-truth labels, and also with pseudo-labels for SigLIP2; with pseudo-labels and the CLIP encoders, between 0 and 10 of the 10 random draws differ significantly, depending on encoder and protocol, so for the weaker encoders the two rules are closer to a tie. Inside the full K-means prototype pipeline (12 caches, 10 seeds) the picture is the same: at keep ≤ 0.25 the random subset wins on 8–11 of the 12 caches (-0.4 to -1.7 points on average for margin selection), while at keep 0.75–0.9 the two rules are within 0.9 points of each other, since a filter that keeps most images can do little harm. The one exception is SUN397, the most imbalanced pool, where keeping the top-margin half beats using all samples by up to 0.7 points; there the filter presumably also thins the over-populated classes of the arg-max marginal, an effect that balanced assignment obtains without discarding any image.

I.3 What the validation labels would buy if used directly

The validated policy uses the labeled validation quarter only to choose (α, τ, λ). Figure 10 instead fits the whitened head on that quarter with its true labels (α=1) and fuses it with the anchor, either

Table 19: Margin-selected versus random subsets at matched keep rate, hold-out protocol, mean over the 10 caches without ImageNet-1k and Places365 (test accuracy %; each cell at its accuracy-optimal λ on the evaluated half, an oracle λ).

	pseudo-labels			ground-truth labels		
keep	margin	random	p	margin	random	p
0.05	77.56	77.73	0.19	79.60	81.31	$6.7 \cdot 10^{-5}$
0.10	77.86	78.19	0.010	80.04	81.77	$3.3 \cdot 10^{-5}$
0.25	78.29	78.61	0.009	80.73	82.08	$2.7 \cdot 10^{-5}$
0.50	78.63	78.77	0.031	81.40	82.19	$2.1 \cdot 10^{-5}$
0.75	78.79	78.83	0.23	81.92	82.23	$7.3 \cdot 10^{-6}$
1.00	78.87	78.87	–	82.26	82.26	–

at λ=0.5 or not at all. The supervised head beats ours (validated) on 18 of 18 pairs at λ=0.5 and on 13 when used alone, by 3.0 points on average for the better of the two, with the largest margins on SUN397 and Places365. The validated numbers in the paper should therefore be read as a reference for what the label-free method loses to hyper-parameter transfer, not as a recommendation to spend labels on validation. If labels are available, they are better spent on training the head (five seeds).

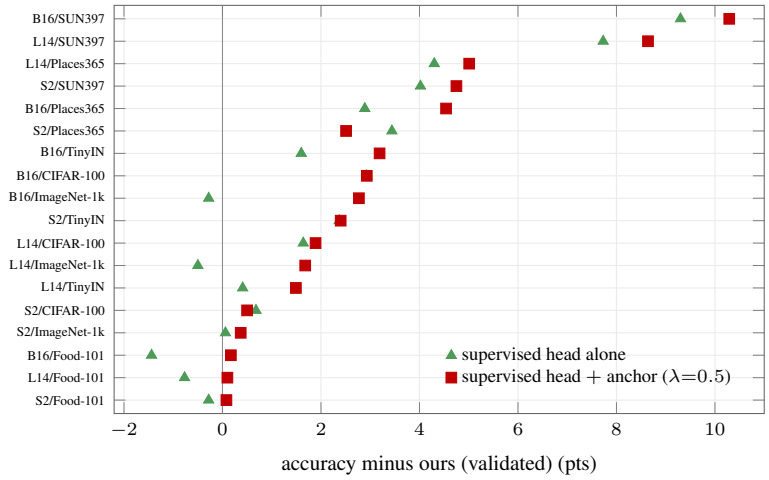

Figure 10: Labeled-quarter control (test accuracy, 5 seeds): a whitened head fitted on the validation quarter with its true labels (α=1), alone and fused with the anchor at λ=0.5, shown as accuracy minus ours (validated) on each of the 18 caches (sorted by the better of the two). The supervised head beats the validated policy on 18 of 18 caches, by 3.0 points on average for the better of the two, with the largest margins on SUN397 and Places365; the head alone is ahead on 13.

I.4 Build-up with the label-free prior throughout

Table 4 selects τ on the validation split, which on five of the six pools picks the uniform prior. Table 20 repeats the build-up with the estimated prior held fixed (τ=1, the label-free choice; α and λ are still validated). On these near-balanced pools, balanced labels then add only +0.08 to the prototype recipe, and the full method adds only +0.03 to the whitened head on arg-max labels; the label-free gain over the prototype recipe (+0.99) comes from the head. The prior contributes where the pool is imbalanced: on SUN397, balanced labels add +0.73 to the prototype and the full method adds +0.89 to the whitened head, and on the long-tailed pools (Table 25) the estimated prior is what keeps every method above zero-shot.

Table 20: Build-up with the estimated prior fixed (τ=1) throughout (test accuracy %, mean over 3 CLIP-family backbones $\times$ 5 seeds; α, λ validated; last columns as in Table 4, against the last row). Rows that do not depend on the prior are identical to Table 4.

Method	TinyIN	CIFAR-100	SUN397	Food-101	Places365	ImageNet-1k	mean	wins	p
zero-shot (anchor only)	71.64	77.75	67.39	90.01	41.93	76.56	70.88	18/18	$1 \cdot 10^{-6}$
prototype recipe	75.14	79.46	71.65	90.53	44.54	78.52	73.31	17/18	$7 \cdot 10^{-5}$
+ balanced labels, estimated prior (2a)	74.92	79.41	72.38	90.56	44.57	78.52	73.39	17/18	$2 \cdot 10^{-5}$
whitened head on arg-max (2b)	76.86	80.95	73.04	90.98	44.81	78.96	74.27	6/18	0.79
whitened head alone (2a + 2b, no fusion)	75.97	80.58	73.40	90.00	43.86	77.33	73.52	16/18	$1 \cdot 10^{-4}$
ours, label-free prior (2a + 2b + fusion)	76.47	80.67	73.93	91.03	44.75	78.92	74.30	–	–

Table 21: Oracle-label ceilings for Table 4 (test accuracy %, mean over 3 CLIP-family backbones $\times$ 5 seeds; "wins": pairs in which the full method with validated λ is ahead of the row). Paired t-tests over the 18 pairs: ours loses to the oracle-label prototype in 11/18 ($p = 9.3 \times 10^{-3}$) and to the oracle-label whitened head in 18/18 ($p = 1.2 \times 10^{-4}$), which bounds the remaining headroom.

Method	TinyIN	CIFAR-100	SUN397	Food-101	Places365	ImageNet-1k	mean	wins
ours (2a + 2b + fusion)	78.99	82.33	74.07	91.21	45.95	80.12	75.45	–
prototype, oracle labels	79.28	82.01	79.05	90.85	49.84	81.05	77.02	7/18
whitened head, oracle labels	81.53	84.51	82.10	91.44	50.83	82.19	78.77	0/18

I.5 Decomposition of the gain and headroom captured

Figure 11 decomposes the gain over the prototype recipe into the contribution of balanced labels alone, the whitened head alone, and both together, per cache (validated hyperparameters, single split). The two components are not alternatives: the mean gain of both over the prototype recipe is +2.16 points (18/18 caches), and the fraction of the headroom to the oracle-label whitened head that is captured averages 63.6% per cache (the ratio of the mean gain over zero-shot to the mean headroom over zero-shot is 57.8%), so most of what perfect labels would add is already obtained without any. In-sample decompositions on SigLIP2 give +2.02 (labels), +1.88 (head) and +4.44 (both) on TinyImageNet, and +0.56, +0.13 and +0.82 on ImageNet-1k, i.e. +0.53 and +0.12 points more than the sum of the parts. The mild super-additivity has a simple reading: cleaner labels give the whitened head a better statistic to read, and the whitened map makes better use of the repaired labels than the class mean does.

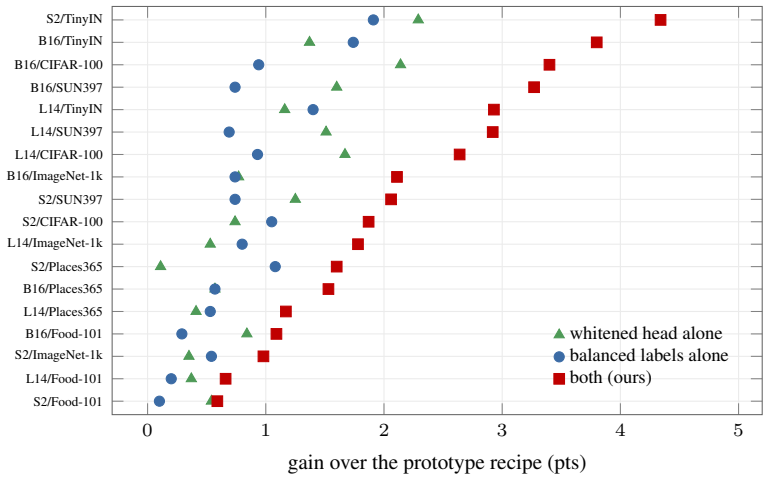

Figure 11: Decomposition of the gain over the prototype recipe on each of the 18 caches (test accuracy, validated λ): the whitened head alone (arg-max labels), balanced labels alone (class mean), and both. The two compose on every cache, with mean gains of +1.02, +0.83 and +2.16 points; both together capture on average 63.6% of the headroom to the oracle-label whitened head.

I.6 HYPER-PARAMETER SENSITIVITY AND THE LABEL-FREE POLICY GRID

The optimum in λ is broad: on TinyImageNet every $\lambda \in [0.33, 0.47]$ is within 0.1 points of the best, so a transferred value loses little. The choice of prior matters more, and it is decided by the balance of the pool. The purity of the balanced labels for $\tau \in \{0, 1\}$ on the six SigLIP2 pools is 82.9/77.5 (TinyIN), 89.4/87.3 (CIFAR-100), 66.0/77.4 (SUN397), 95.4/95.1 (Food-101), 47.3/44.8 (Places365) and 86.9/85.7 (ImageNet-1k), against 77.1/87.1/75.0/94.8/44.7/85.5 for the arg-max: the uniform prior repairs the labels on every balanced pool, and only the estimated prior repairs them on the imbalanced one, where the uniform prior makes them worse than the arg-max. Figure 13 shows the test accuracy of every fixed (τ, λ) configuration with $\alpha = 1$, averaged over the 18 CLIP-family caches and 5 seeds; the validated configuration averages 75.45. The uniform prior (τ=0) is best on average because five of the six pools are balanced, so the average says more about the benchmark than about the annotator. The per-pool transferred choices are derived from the same grid by leaving the target pool out and ranking the configurations by validation accuracy on the other five pools (Figure 12, Appendix D). For SigLIP2, the SUN397 loss under τ=0 (77.3 against 80.3 at τ=1, λ=0.3) is large enough to tip the transferred choice for the balanced pools to τ=0.5, which costs them 0.3 to 0.9 points; the CLIP encoders keep τ=0 throughout. A single imbalanced pool thus changes the transferred policy for all the others, which is the argument for fixing the estimated prior rather than transferring τ.

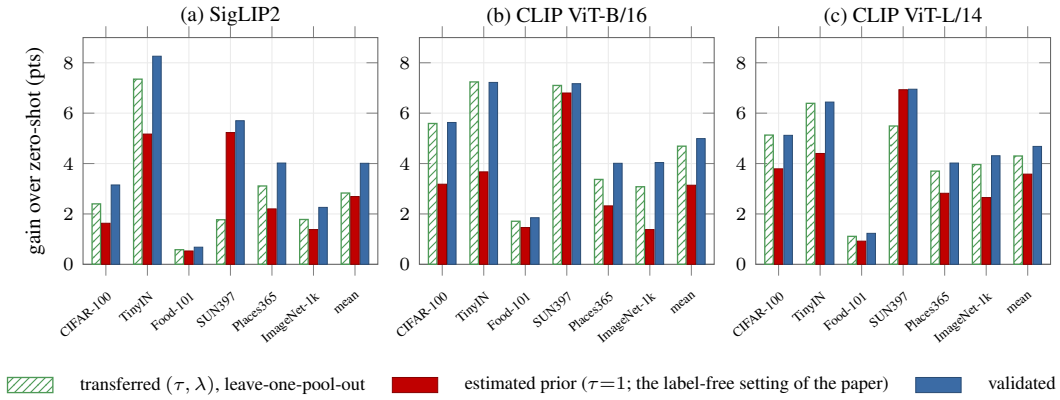

Figure 12: Label-free policies versus validated hyper-parameters: gain over zero-shot per pool (test accuracy, 5 seeds, α=1; absolute values in the released result files, Appendix E). "transferred" takes the (τ, λ) pair with the best validation accuracy on the other five pools; it prefers the uniform prior because five pools are balanced, which costs it 3.5 points on SUN397 with SigLIP2 and makes it collapse on long-tailed pools (Table 25). The label-free setting of the paper uses the estimated prior with the transferred λ; means over the six pools: SigLIP2 80.18 / 80.04 / 81.36, ViT-B/16 69.08 / 67.53 / 69.38, ViT-L/14 75.22 / 74.50 / 75.60 (transferred / estimated / validated).

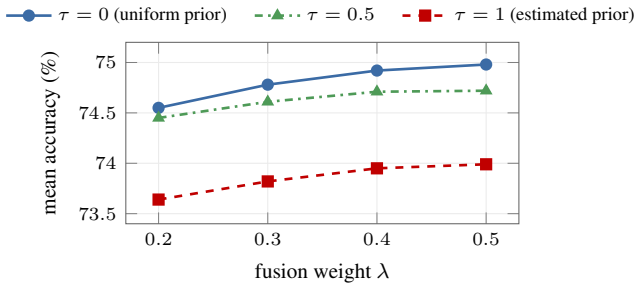

Figure 13: Label-free hyper-parameter grid: mean test accuracy over 18 caches $\times$ 5 seeds for fixed (τ, λ) with $\alpha = 1$. On these near-balanced pools, the uniform prior is best at every λ, and $\lambda = 0.5$ is best for every prior.

I.7 STRESS TEST

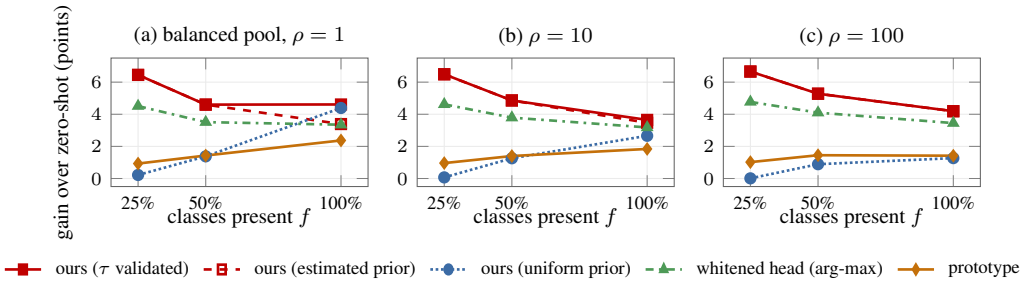

Figure 14: **Stress test: the estimated prior is what makes the annotator robust.** Gain over zero-shot (points, mean over 18 caches $\times$ 3 seeds) when the largest class has ρ times as many images as the smallest (panels) and only a fraction f of the classes is present (x-axis); the annotator scores all C classes and is never told which are absent. The uniform prior loses its gain as classes go missing and as ρ grows; the estimated prior gains most exactly there. Details in Appendix I.7.

Figure 14 averages the stress test over 18 caches $\times$ 3 seeds; the per-cache results (18 caches $\times$ 9 conditions $\times$ 3 seeds, 2 430 rows) are in the released result files (Appendix E). The averages hide no failure: the worst per-condition mean of the validated method is $+3.65$ points over zero-shot, and its worst single cache–condition cell is $+0.74$, so the annotator never falls below the anchors it starts from, whatever the imbalance and however many classes are missing. The two label-free alternatives do fall below them: the uniform-prior variant reaches -0.84 (SUN397 at ρ=100) and the prototype recipe -0.33. The difference is the prior; with a uniform prior on a pool that is not uniform, balanced assignment forces images onto absent classes and undoes the gain of the head.

I.8 CLASS IMBALANCE: PER-CLASS VIEW AND LABEL-FREE PRIORS

Figure 15 extends the per-class view of Section 4.5 to all three encoders and to the mildly imbalanced Places365 (1.6$\times$) and ImageNet-1k (1.8$\times$) pools. On these two pools, the uniform prior is the right one, and the arg-max estimate of the prior is far off: its class sizes span 17–33$\times$ on Places365 and leave classes empty on ImageNet-1k, against a true spread of 1.6–1.8$\times$. The classes of each pool are split into thirds by training-set size, and "tail" denotes the third with the fewest images.

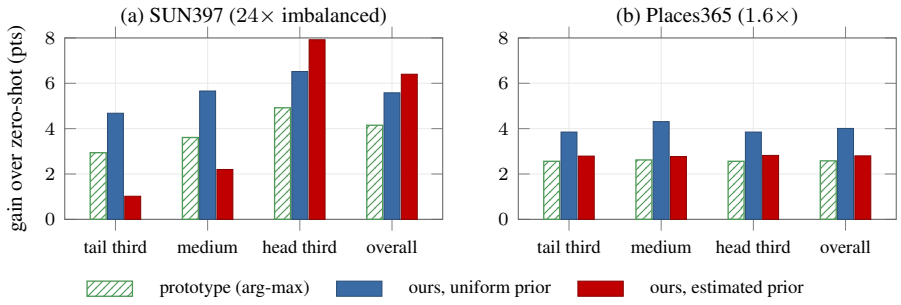

Figure 15: Per-class view of the class prior on the two imbalanced pools: gain over zero-shot by class-size third (test split, 3 seeds; α and λ validated, τ fixed as indicated), mean over SigLIP2, CLIP ViT-B/16 and ViT-L/14; "tail" is the third with the fewest training images. On SUN397, the estimated prior gains on the head third and loses on the tail, and overall accuracy follows the head; on the mildly imbalanced Places365, the uniform prior is right and is better on every third.

A label-free prior by re-estimation. The estimated prior of Stage 2a is the arg-max marginal, which is itself distorted by the anchors (177$\times$ spread on the balanced TinyImageNet). A better label-free estimate becomes available after one pass of the annotator: run Stage 2a with the uniform prior, fuse at the transferred λ, predict the whole pool, and use the predicted marginal as the prior

for a second pass. Figure 16 compares this estimate with the uniform prior, the arg-max estimate, and the true class proportions, all with α=1 and the transferred λ. One round recovers most of the SUN397 gap (+2.7 to +3.2 points over the uniform prior) but costs up to 1.2 points on the balanced pools, 0.6 to 1.0 on Places365 and 0.4 to 1.4 on ImageNet-1k, because the predicted marginal of an imperfect annotator is itself skewed. A second round drifts further, and over the 18 caches one round ties the uniform prior on average (75.0 against 75.0). Under the imposed imbalance of the stress test, however, re-estimation is far safer than the uniform prior, though it remains below the arg-max estimate: on the twelve CIFAR-100, TinyImageNet, Food-101 and SUN397 caches its gain over zero-shot is +3.7 to +4.9 points in every condition (mean +4.4), against −0.1 to +4.7 for the uniform prior (mean +1.4) and +3.8 to +8.0 for the arg-max estimate (mean +6.0). Estimating the prior without labels is thus possible but not free, and the label-free policy of the paper keeps the estimated prior (τ=1).

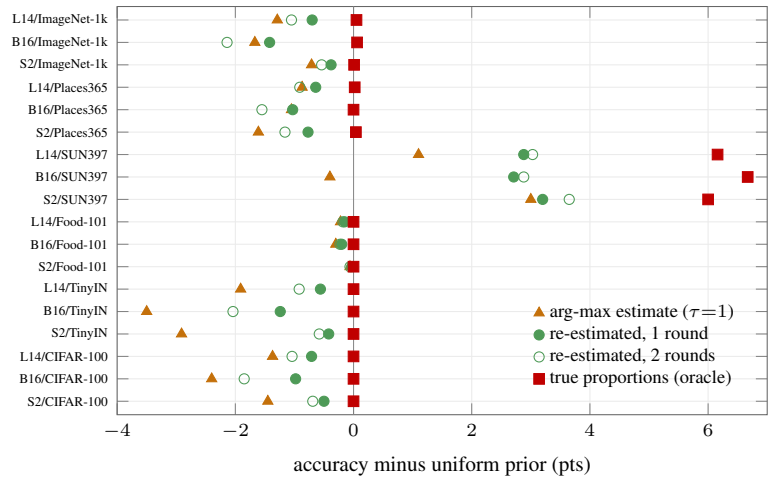

Figure 16: Label-free class priors (test accuracy, 3 seeds, α=1, transferred λ), shown per cache as the difference from the uniform prior: the arg-max estimate (τ=1), the prior re-estimated from the predictions of the annotator itself after one and two rounds, and, for reference, the true class proportions. One round of re-estimation recovers most of the SUN397 gap (+2.7 to +3.2 over uniform) but costs up to 1.2 points on the balanced pools; only the true proportions are never worse than uniform.

Macro and per-tertile accuracy on the long-tailed benchmarks. Figure 17 repeats the tertile analysis of Figure 15 on ImageNet-LT and Places-LT (3 seeds; classes split into thirds by training count). The estimated prior raises overall accuracy by 0.9 to 4.5 points over the prototype recipe, but overall accuracy is dominated by the head third: macro accuracy changes by −1.4 to +0.2 relative to the prototype recipe, and the tail third is 1.7 to 6.3 points below our own uniform-prior variant, which is the best choice for the tail on every cache. The prior thus trades tail recall for head accuracy, and which side of this trade an annotator should take depends on what the labels are for.

I.9 Closed form versus trained heads

Figure 18 compares the closed-form whitened head with a linear head trained by cross-entropy SGD for 1, 50, and 300 epochs on the same pseudo-labels and split, and with five rounds of self-training (re-labeling the pool with the current head and refitting).

I.10 Transfer to DINOv2 and choice of annotator

Table 22 reports DINOv2-base features annotated with the labels of SigLIP2 on all six pools, and Table 23 compares different label sources on TinyImageNet. The transfer shows that the method is not tied to the encoder that produced the anchors: balanced labels help the prototype, the whitened head on its own matches it, and the two together with fusion are best on five of the six pools, while the oracle-label row shows that most of the remaining gap is label quality rather than the head.

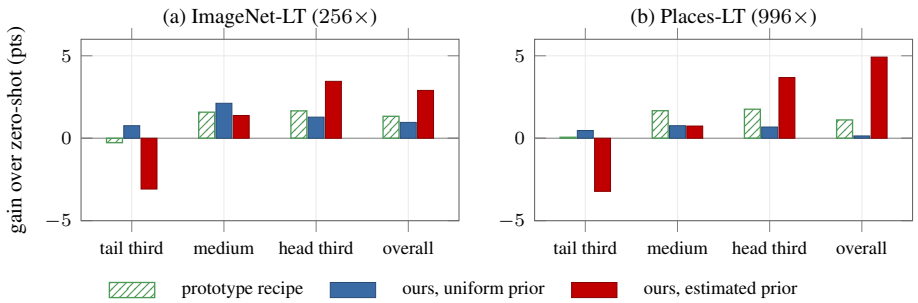

Figure 17: Per-tertile view on the long-tailed benchmarks (test split, 3 seeds; α, λ validated, τ fixed as indicated): gain over zero-shot by class-size third, mean over SigLIP2, CLIP ViT-B/16 and ViT-L/14. The estimated prior wins overall by lifting the head third, which dominates the class-frequency-weighted accuracy, while it loses on the tail third, where the uniform prior is best; macro accuracy stays within 1.4 points of the prototype recipe.

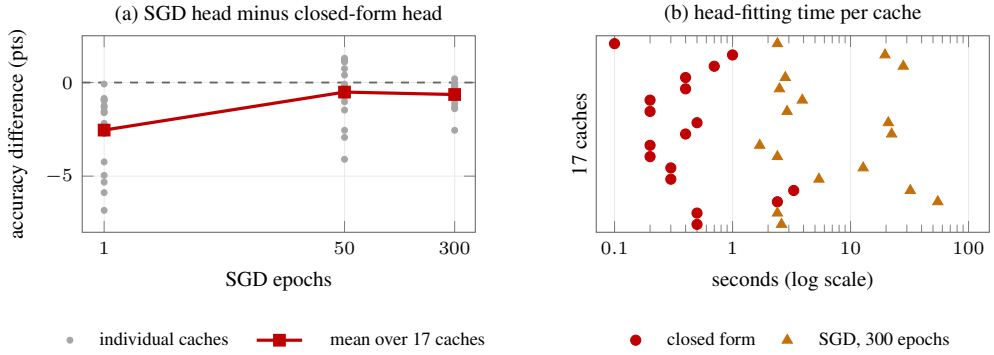

Figure 18: Closed-form whitened head versus linear heads trained by cross-entropy SGD on identical pseudo-labels and split (18 caches, single split). (a) Accuracy of the SGD head minus the closed-form head after 1, 50 and 300 epochs; gray: individual caches, red: mean. At 300 epochs SGD trails the closed form on 17 of 18 caches (-0.63 on average); five self-training rounds do not close the gap (-0.79, worse on 16 of 18). (b) Head-fitting time per cache: the closed form takes 0.1–3.3 s where 300 epochs take 1.7–55 s.

The second table makes the same point from the other side. What DINOv2 contributes is a better representation, not better labels: grouping its features by its own clusters gives the lowest purity, and the higher purity obtained by refining our labels with DINOv2 does not translate into higher fused accuracy.

Table 22: DINOv2-base features with SigLIP2 pseudo-labels (test accuracy %, 5 seeds, validated α and λ; scores fused with SigLIP2 anchors).

method	TinyIN	CIFAR-100	SUN397	Food-101	Places365	ImageNet-1k	mean
SigLIP2 zero-shot	77.07	87.15	75.01	94.77	44.73	85.45	77.36
DINOv2 + prototype	84.38	90.42	78.35	94.68	47.65	86.30	80.30
DINOv2 + prototype + balanced labels	85.64	91.14	77.92	94.73	48.36	86.56	80.72
DINOv2 + whitened head	84.32	90.29	**78.80**	94.72	47.30	86.29	80.29
DINOv2 + ours	**86.09**	**91.26**	78.37	**94.76**	**48.43**	**86.63**	**80.92**
DINOv2 + whitened head, oracle labels	88.50	92.01	83.02	94.79	51.68	87.06	82.84

I.11 Scaling with the vocabulary: full results

Figure 4 reports the mean of 3 seeds (the std of ours is ≤ 0.45 everywhere); the subsets keep all training images of the selected classes (Appendix C). The point of the experiment is that the gain

Table 23: Label source for a DINOv2 whitened head on TinyImageNet (5 seeds). "dino refine" re-annotates with DINOv2 seeded by our labels.

label source	grouped by	purity	head alone	fused with anchors
SigLIP2 arg-max	anchors	77.1	81.74	84.32
SigLIP2 balanced (ours)	anchors	82.9	83.18	**86.09**
DINOv2 k-means	DINOv2 clusters	75.4	75.38	83.98
DINOv2 refine	seeded by ours	83.9	82.39	86.05

does not shrink as the vocabulary grows, which would be the case if the method mainly resolved easy confusions among a few classes. On every backbone and both pools, the gain over zero-shot is larger at the largest C than at the smallest (ImageNet-1k: $+1.5 \rightarrow +2.3$ for SigLIP2, $+3.2 \rightarrow +4.1$ for ViT-B/16, $+2.4 \rightarrow +4.3$ for ViT-L/14), although it is not monotone at every step (TinyImageNet, SigLIP2: $+6.1$ at C=50, $+5.7$ at C=100). On TinyImageNet (not shown in the figure), CLIP ViT-L/14 goes from $+1.5$ at C=10 to $+6.6$ at C=200, and ViT-B/16 from $+1.8$ to $+7.3$; the weaker the encoder, the more the pool adds as the vocabulary grows. The label-free curve uses the label-free policy of the paper (τ=1, α=1, transferred λ).

I.12 Published baselines: reproduction details

Table 24 gives the per-pool numbers behind Table 3. Implementation details (Appendix E) are as follows. TransCLIP is ported from the official solver in zero-shot mode (λ=1, 10 outer and 5 inner iterations, 3-nearest-neighbor Laplacian, GMM with a shared diagonal variance, means initialized from the 8 most confident samples per class); the per-sample variance and likelihood sums are computed in closed form, which lets the solver run on hundreds of thousands of images. InMaP uses the official Sinkhorn (τ_t=0.01, 20 iterations) and proxy optimization (lr 10, 2,000 iterations, τ_i=0.04, confidence threshold 0.6) and predicts with the learned vision proxies. We verified the port in three ways: its two functions return Sinkhorn couplings and proxies that are bit-identical to those of the official `main.py` on our caches; the official script, run unchanged on ImageNet val with CLIP ViT-B/16 and its 7-template ensemble, gives 68.81% zero-shot and 72.53% InMaP against the published 68.75% and 72.55%; and our port applied to the features and text classifier dumped by that run gives 72.52%. The differences from the published setting are shared by every method in Table 3: the single-template anchors of our caches instead of the 7-template ensemble of InMaP, and our imbalanced training pools instead of the balanced test splits of its paper, on which it reports $+5.2$ points for SUN397 with ViT-B/16. Because the Sinkhorn step of InMaP assumes a uniform class prior, we also run the estimated-prior option of the official code (γ=1, the softmax marginal as prior), which is the counterpart of our τ=1 row; ours uses the arg-max marginal and OTTER its own estimate. The released Frolic code contains the Gaussian (text-mean, diagonal-covariance) logits, two InMaP runs from template and Gaussian pseudo-labels with γ=1 Sinkhorn, their logit ensemble, and the eigenvector label-bias correction (11 rounds), but not the LLM-generated prompt distribution used in the paper; we run the released components with the template prompt in place of the LLM prompts. TDA uses the official positive and negative cache hyperparameters of the matching config (`sun397.yaml`, `food101.yaml`, or `imagenet.yaml` for pools without their own config) and streams the test split in a seed-dependent random order. All methods use the same seeds and test splits as Table 2.

I.13 Long-tailed pools and published baselines under stress

Table 25 repeats the comparison of Table 3 on the two standard long-tailed benchmarks, ImageNet-LT (256×) and Places-LT (996×), built as subsets of our ImageNet-1k and Places365 caches. Table 26 runs the transductive published methods under the stress protocol of Figure 14 (the test split of the stressed pool serves as the unlabeled set; TDA is omitted). The long-tailed pools make the role of the prior visible in a way the balanced pools cannot. Every method that assumes a uniform class prior collapses on them: InMaP and OTTER with the uniform prior fall 10 to 20 points below zero-shot, and our own uniform-prior variant 3 to 11, while TransCLIP stays within 1.8 points of zero-shot and Frolic within 0.6 points on ImageNet-LT but 3.4 to 5.3 points below it on Places-LT. Balancing toward a prior that the data do not follow is worse than not adapting at all. When each method uses

Table 24: Published training-free methods per pool (test accuracy %, mean of 3 seeds). Best per column within each encoder block in bold.

Encoder	method	CIFAR-100	TinyIN	Food-101	SUN397	Places365	ImageNet-1k	mean
SigLIP2	zero-shot retrieval	86.97	77.09	94.72	75.33	44.71	85.50	77.39
	TDA (Karmanov et al., 2024), streamed	87.95	80.61	94.81	76.77	46.09	86.01	78.71
	TransCLIP (Zanella et al., 2024)	86.53	80.18	94.57	78.44	47.30	86.64	78.94
	Frolic (Zhu et al., 2024), released core‡	88.40	81.71	95.17	76.35	47.80	86.76	79.36
	InMaP (Qian et al., 2023), uniform prior	**90.32**	84.64	**95.53**	69.53	**49.15**	**88.13**	79.55
	InMaP, estimated prior (γ=1)	87.81	78.56	95.17	77.35	45.92	86.56	78.56
	OTTER (Shin et al., 2024), estimated prior	87.04	77.47	95.00	77.54	44.79	85.75	77.93
	ours, transductive, uniform prior	90.19	**85.43**	95.46	76.91	48.68	87.69	80.73
	ours, transductive, estimated prior	88.57	81.89	95.35	80.28	46.98	86.88	79.99
	ours, fit split, label-free	88.66	82.24	95.25	80.28	46.94	86.88	80.04
	ours, fit split, validated	90.18	85.33	95.40	**80.75**	48.76	87.76	**81.36**
ViT-B/16	zero-shot retrieval	69.03	63.50	84.59	60.96	40.21	68.10	64.40
	TDA (Karmanov et al., 2024), streamed	70.09	66.88	84.97	64.78	41.67	69.41	66.30
	TransCLIP (Zanella et al., 2024)	71.11	66.95	85.99	66.75	43.30	70.92	67.51
	Frolic (Zhu et al., 2024), released core‡	72.46	68.42	85.78	63.23	43.21	71.39	67.41
	InMaP (Qian et al., 2023), uniform prior	74.18	70.65	**86.78**	61.89	44.09	**73.09**	68.44
	InMaP, estimated prior (γ=1)	70.19	64.84	85.64	62.60	41.62	69.51	65.73
	OTTER (Shin et al., 2024), estimated prior	68.58	62.80	84.97	63.21	40.25	68.20	64.67
	ours, transductive, uniform prior	74.48	**70.95**	86.47	67.10	43.80	71.76	69.10
	ours, transductive, estimated prior	71.90	67.29	86.10	67.59	42.69	69.93	67.59
	ours, fit split, label-free	72.18	67.17	86.15	67.68	42.52	69.48	67.53
	ours, fit split, validated	**74.63**	70.72	86.54	**68.05**	**44.21**	72.14	**69.38**
ViT-L/14	zero-shot retrieval	77.05	74.64	90.41	66.51	40.85	76.15	70.94
	TDA (Karmanov et al., 2024), streamed	78.72	76.83	90.53	68.68	42.01	77.27	72.34
	TransCLIP (Zanella et al., 2024)	79.36	78.42	91.25	71.89	44.00	78.98	73.98
	Frolic (Zhu et al., 2024), released core‡	80.54	78.76	91.35	67.88	44.08	79.26	73.65
	InMaP (Qian et al., 2023), uniform prior	81.82	80.80	**92.07**	65.11	**44.93**	**81.01**	74.29
	InMaP, estimated prior (γ=1)	78.39	76.99	91.30	68.75	42.90	78.23	72.76
	OTTER (Shin et al., 2024), estimated prior	77.38	74.64	90.76	68.97	41.07	76.45	71.54
	ours, transductive, uniform prior	**82.18**	**81.07**	91.70	71.71	44.69	80.26	75.27
	ours, transductive, estimated prior	80.60	78.64	91.41	72.78	43.74	78.80	74.33
	ours, fit split, label-free	80.85	78.90	91.39	73.40	43.69	78.80	74.50
	ours, fit split, validated	**82.18**	80.94	91.70	**73.42**	44.89	80.46	**75.60**

an estimated prior (ours: the arg-max marginal), ours in the transductive protocol is ahead of InMaP and OTTER on all six caches (ImageNet-LT $88.0/72.4/80.2$ against $87.4/71.2/79.1$ for the best published method; Places-LT $53.1/48.7/48.9$ against $52.2/47.1/47.4$), and the fit-split label-free policy is within 0.4 points of the validated one, so the transferred scalars hold up on pools unlike those they were transferred from. Given the *true* test class proportions (an oracle, as in the evaluation of OTTER itself), OTTER reaches 90.5 / 77.3 / 83.5 on ImageNet-LT and 60.4 / 56.6 / 57.0 on Places-LT, which is 2 to 8 points above ours. On long-tailed pools, the prior is therefore worth more than any head, and estimating it without labels (Figure 16) is the largest remaining headroom. Under the stress protocol the same picture holds on all 18 caches (486 condition–seed cells): the uniform-prior methods lose up to 31 points on average at f=0.25, InMaP with the estimated prior gains $+2.2$ on average, and ours with the estimated prior, label-free, gains $+3.9$, beats zero-shot in all 162 cache–condition cells (worst $+0.57$), and is ahead of every other method in 89 to 96% of the condition–seed cells. What separates the methods here is robustness to a wrong assumption about the class distribution, not peak accuracy on a balanced benchmark.

Table 25: Long-tailed pools, full version of the right half of Table 3 (test accuracy %, 3 seeds; fit-split rows 5 seeds). "estimated prior" is the arg-max marginal for ours and the label-free estimate native to each method for InMaP (softmax marginal) and OTTER. Best per column in bold.

	ImageNet-LT (256×, C=1000)			Places-LT (996×, C=365)		
Method	SigLIP2	ViT-B/16	ViT-L/14	SigLIP2	ViT-B/16	ViT-L/14
zero-shot retrieval	86.65	69.05	77.36	50.09	43.28	43.55
TDA, streamed	86.89	69.62	77.87	49.86	43.13	42.80
TransCLIP	85.77	70.34	78.31	48.29	43.62	43.13
Frolic, released core	86.79	68.90	77.95	46.67	38.01	38.91
InMaP, uniform prior	70.96	59.03	64.70	32.64	31.00	30.82
InMaP, estimated prior	87.37	71.19	79.14	51.93	45.30	46.35
OTTER, uniform prior	67.96	55.55	60.49	30.39	28.68	28.31
OTTER, estimated prior	87.38	70.91	78.62	52.16	47.08	47.36
ours, transductive, uniform prior	80.93	66.17	74.21	38.71	37.01	36.81
ours, transductive, estimated prior	87.96	72.35	80.19	53.06	48.70	48.87
prototype recipe (fit split, validated)	87.50	70.56	79.24	51.05	44.35	44.95
prototype recipe (fit split, label-free)	87.41	70.25	79.00	50.74	43.81	43.99
ours, fit split, label-free	88.25	72.42	80.73	53.19	**48.83**	**49.33**
ours, fit split, validated	**88.41**	**72.78**	**80.78**	**53.36**	48.73	**49.33**

Table 26: Published transductive methods under the stress protocol: gain over zero-shot (points) per imbalance ratio ρ and fraction of classes present f, mean over the 18 CLIP-family caches, and 3 seeds; "ours" rows are label-free; transductive protocol on the stressed test split, so the "ours" rows are not cell-by-cell comparable with Figure 14 (fit-split protocol, 18 caches). Best per column in bold.

	$\rho = 1$			$\rho = 10$			$\rho = 100$			
Method	f=.25	.5	1	.25	.5	1	.25	.5	1	mean
TransCLIP	-4.08	-0.57	+2.62	-4.08	-0.87	+1.56	-4.40	-1.52	-0.42	-1.31
InMaP, uniform prior	-29.50	-11.34	+3.21	-29.85	-13.64	-3.68	-30.78	-18.50	-13.91	-16.44
InMaP, estimated prior	+2.98	+2.54	+1.51	+2.76	+2.28	+1.54	+2.58	+2.12	+1.78	+2.23
Frolic, released core	-6.79	-2.22	+2.63	-6.16	-1.86	+1.49	-6.29	-2.47	-0.33	-2.45
ours, uniform prior	-13.96	-3.25	**+4.10**	-14.78	-4.95	+0.94	-16.38	-8.17	-5.00	-6.83
ours, estimated prior	**+4.79**	**+3.89**	+3.09	**+4.71**	**+3.90**	**+2.97**	**+4.64**	**+3.89**	**+3.25**	**+3.90**