CAN CLASS NAMES FORECAST PER-CLASS ACCURACY FOR DATA-FREE CLIP CUSTOMIZATION?

Anonymous authors
Paper under double-blind review

ABSTRACT

Data-free customization trains a compact student from class names alone, but gives no per-class report when training finishes. We study a cheap way to estimate that gap without any real image of the target vocabulary. Our model forecast from name and inversion-based statistics targets *student* rather than teacher per-class accuracy on a data free customization system. We evaluate it under two protocols within and between classes. Scoring takes seconds once a synthetic pool exists; the estimate is coarse, Spearman r_s trailing a labeled teacher oracle on global rank for three of four vocabularies under both protocols. But on compositional and fine-grained name lists it is even better than estimation from teacher with real labeled data: class-split worst-30% precision beats the oracle on Flowers-102 and FGVC-Aircraft, and CUB-200 reaches 0.46 leave-one-out global rank without target photographs. However, Our model cannot replace labeled evaluation when real images are provided, but it is a cheap per-class estimate that requires no real images.

1 INTRODUCTION

Open-vocabulary CLIP customization lets a practitioner define a new classifier from class names alone (Wei et al., 2025; Radford et al., 2021). A frozen text–image teacher embeds the names, inverts them into synthetic training images, and distills a compact student that can deploy without collecting any real photograph of the target task. That workflow removes the usual first step of curating labeled data for every new vocabulary someone wants to support.

The system leaves a blind spot at deployment time. Averaged over eleven customized tasks, the teacher reaches 85.65% and the text-only student 64.81%, yet the headline numbers report no per-class breakdown and there is no held-out test set when customization finishes. The average can hide severe local failure: on Flowers-102 the student reaches only 15.83%, and (Wei et al., 2025) trace part of the collapse to names CLIP misreads—"globe-flower" read as a globe and a flower, "balloon flower" confused with a balloon. Until someone collects labeled images and evaluates class by class, a user who typed those names learns nothing about *which* classes drive the average down or deserve the first round of inspection.

If real images of the task were already available, one would simply evaluate the student on them. The open question is how well class names, together with synthetic inversions from the same customization pipeline, can forecast per-class *student* accuracy without any real images.

We answer with a deliberately simple instrument: a linear forecast ϕ of six statistics, fit on calibration classes and scored on held-out classes under leave-one-vocabulary-out and class-split protocols. It is *cheap* to score, *partially effective* on some name lists, but not enough to replace labeled evaluation when real images exist.

We study a student-specific estimate that uses no real image of the held-out classes under either protocol. The teacher reference requires labeled test images of every class and is therefore a higher-information oracle rather than a same-budget competitor. The contribution is a cheap indication of which classes to inspect first when only names and synthetic inversions are available.

Against this teacher reference, ϕ gives a partial and vocabulary-dependent answer. Global rank trails the oracle on Caltech-101, Flowers-102, and CUB-200 under both protocols, and is near zero

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.

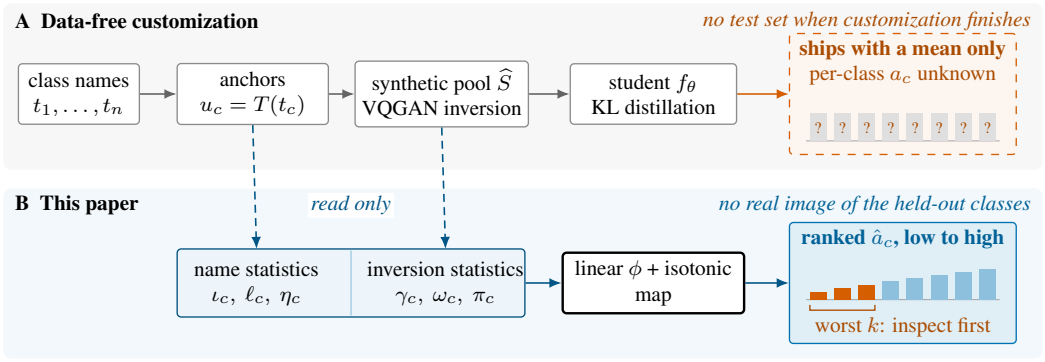


Figure 1: **Setting and approach. (A)** Data-free customization inverts class names through a frozen CLIP teacher, distills a compact student from the synthetic pool, and ships it with a mean accuracy but no per-class a_c and no test set. **(B)** We read six statistics from two by-products the pipeline already produces, the anchors u_c and the first inversions (dashed taps), fit a linear map ϕ with isotonic recalibration on calibration classes only, and return classes ranked by forecast accuracy $\hat{a}_c$; the bottom k form the shortlist to inspect first. No real image or label of the held-out classes is read anywhere in (B).

on Caltech-101, while FGVC-Aircraft overlaps the oracle. The more useful result is at the bottom of the ranking: the worst-class shortlist is competitive on Flowers-102 and FGVC-Aircraft under class split, and on FGVC-Aircraft under leave-one-vocabulary-out.

- **A cheap, no-real-image estimate of per-class student accuracy.** ϕ scores six name and inversion-based statistics in seconds with no held-out real image. Rank is low leave-one-out and under a $70/30$ class split, but worst 30% beats a labeled oracle on two of four vocabularies.
- **Student-specific, not teacher-proxied.** Residual variance after regressing student on teacher per-class accuracy is large; ϕ targets the student directly rather than assuming the teacher rank suffices.
- **Honest scope: partial effectiveness and structural limits.** Observation 1 bounds label-free prediction. Text-only statistics suffice on Caltech; compositional names need inversion features. ϕ is pre-class triage, not a substitute for labeled evaluation.

2 Related Work

Data-free distillation and customization. Inversion-based methods synthesize a surrogate set from a frozen teacher and distill a student from it (Yin et al., 2020; Fang et al., 2021; 2022; Wei et al., 2025; Xuan et al., 2023). Because no validation data exist at customization time, practitioners lack a built-in per-class report when training completes. (Wei et al., 2025) improve open-vocabulary customization with image-text matching and meta distillation, but their weakest results still trace to misread class names.

Predicting performance without target labels. Language-only model selection predicts how a vision-language model will perform on a task described in text (Zohar et al., 2023; Yi et al., 2024; Robbins et al., 2025), which is the *teacher's* whole-task performance, not the distilled student's accuracy on each class. Accuracy estimators and failure predictors need images of the target task (Garg et al., 2022; Lafon et al., 2025); frequency accounts of weak zero-shot classes need the pretraining corpus (Udandarao et al., 2024; Parashar et al., 2024). (Yi et al., 2024) bridge text and image statistics using open-source datasets, much as our map is fit on labeled calibration classes—from other vocabularies or other classes within the same vocabulary—before it is applied to held-out classes without target real images.

Per-class failure in compressed and customized models. (Hooker et al., 2019) showed that networks with similar top-line accuracy can disagree on a narrow subset of inputs concentrated on underrepresented tails. Data-free customization inherits that structure at the class level: failures are uneven across the vocabulary even when the mean is acceptable. Prior work improves the synthe-

sizer or reports aggregate gains (Wei et al., 2025; Frank & Davis, 2025); we ask instead how well class names and pipeline by-products can forecast which classes the *student* will miss before any real image of the target task exists.

3 WHAT CAN BE ESTIMATED WITHOUT REAL IMAGES

Frozen encoders map images and text to a shared unit sphere. A vocabulary $V = (t_1, \ldots, t_n)$ fixes one anchor per class, $u_c = T(\text{prompt}(t_c))$, stacked as rows of U. Inversion produces synthetic images $\hat{x}_j$ and a pool $\widehat{S}$; a student f_θ is trained on $\widehat{S}$ to match the teacher's soft labels. Real images of class c follow P_c, which customization never observes; per-class accuracy is $a_c = \Pr_{x \sim P_c}[\arg\max f_\theta(x) = c]$.

3.1 IDENTIFIABILITY CEILING

Any statistic computed without real images of the target task is a function of the anchors, the first inversions, and the random seed—never of P_c directly.

Observation 1. A statistic that uses no real image of the task is a function of (U, seed), so it takes the same value on two tasks with the same class names and different real distributions, whose per-class accuracies a_c and a'_c can differ. Any label-free forecast therefore succeeds only to the extent that the effect of P_c on a_c is predictable from the anchors.

This is a statement about identifiability, not about ϕ: no scorer of any capacity recovers a_c from names alone, so the ceiling is fixed before an estimator is chosen. In a controlled instance, holding the vocabulary fixed and moving the real distribution changes how often the teacher predicts a class from 0.080 to 0.024, with every decision rule unchanged.

3.2 WHAT CUSTOMIZATION EXPOSES

What customization *does* expose, before any photograph of the target vocabulary exists, is the data the student is about to learn from. The pool $\widehat{S}$ and the teacher's reading of it are by-products of a step the pipeline already pays for, and together with the seed they determine the student. P_c is out of reach, but the surrogate distribution the student actually consumes is fully measurable at customization time. ϕ is built on that side of the split—three statistics read the anchors, three read the first inversions see Section 4.

3.3 STUDENT-SPECIFIC DISAGREEMENT

Surrogate labels are teacher labels, so a student can differ from the teacher on class c only by disagreeing with it there:

$$\big|\text{acc}_c(f_\theta) - \text{acc}_c(g)\big| \;\le\; \text{dis}_c := \Pr_{x \sim P_c}\big[\arg\max f_\theta(x) \neq h(x)\big]. \tag{1}$$

A condition that degrades teacher and student alike moves both terms on the left together and leaves the gap unchanged; only a condition that reaches the student without reaching the teacher enlarges dis_c. Synthesis is such a condition. The anchor u_c is not only the teacher's classifier at test time, it is also the target of the inversion objective, so a misread name corrupts the pool the student trains on while the teacher still reads real images through its image encoder—and may recognize the class despite the same faulty anchor. Pool statistics therefore proxy dis_c, whereas anchor statistics sit upstream of both channels and proxy a_c more than the gap. The ablation follows this ordering: on Flowers-102 removing π_c costs the most, removing ℓ_c or η_c almost nothing see Table 7.

3.4 READING THE POOL VERSUS RESHAPING IT

Pool statistics proxy student–teacher disagreement equation (1), and the ablation shows they carry signal on compositional vocabularies; that is reading what inversion already produced, not changing how synthesis runs. Pool statistics read the synthesis channel in equation (1); anchor statistics read the text channel upstream of both. We do not treat this split as a formal prediction about which statistic will dominate on every vocabulary—the main audit reports rank and triage quality directly (Tables 2, 3, 6).

3.5 Implication for synthesis knobs

Across five seeds, deliberate pool reshaping shifts mean teacher–student disagreement by at most 1.2 points and mean accuracy by less (Table 5; Figure 2a). We treat this as a targeted null on synthesis *intervention*, not as proof that no synthesis change could ever matter, and not as a contradiction of using inversion-side *statistics* in ϕ. Early triage should therefore score the anchors and first inversions that customization already exposes—via ϕ—rather than retune synthesis hoping to re-rank classes.

4 Forecast ϕ and evaluation

4.1 Audit setup

We audit the released text-only pipeline of (Wei et al., 2025) with CLIP ViT-B/32 teacher, VQGAN inversion (Esser et al., 2021), and a ResNet-18 student (He et al., 2016) trained by two-stage KL distillation on the pooled synthetic images. We use four vocabularies—Caltech-101, Flowers-102, the first 150 classes of CUB-200 (Wah et al., 2011), and FGVC-Aircraft (Maji et al., 2013), with five seeds and 64 synthetic images per class. Labels enter only to fit ϕ on calibration data and to score predictions against realized a_c on held-out vocabularies or classes; ϕ itself never sees a real image of the held-out target.

4.2 Statistics

For each class we compute six quantities that use no real image and no label formulas in Appendix C. Three probe how the text encoder reads the name. *Anchor isolation* $\iota_c = 1 - \max_{k \neq c} \cos(u_c, u_k)$ is high when u_c is far from every rival anchor, typical for generic Caltech names. *Compositional leakage* ℓ_c is large when a hyphenated or multi-token name such as "globe-flower" is read as a composition of parts rather than as a single concept. *Paraphrase rank* η_c is large when paraphrases of the name embed closer to other classes than to u_c, signaling ambiguity in text space.

Three statistics probe how the teacher reads the class's synthetic inversions, the images ϕ may inspect before distillation finishes. *Cycle consistency* γ_c is the fraction of those inversions whose nearest anchor is u_c. *Margin dispersion* ω_c measures how spread out the teacher's confidence gaps are across inversions. *Prototype alignment* π_c measures whether inversion embeddings drift away from u_c. Together they summarize whether the teacher "believes" the synthetic pool matches the name.

4.3 Scoring ϕ

Algorithm 1 Forecast ϕ

Require: frozen encoders E, T; vocabulary V; reference noun list R; linear coefficients and isotonic map fit on calibration classes only; the first $b_0 \geq 2$ inversions per class, already produced by the pipeline

$u_c \leftarrow T(\text{prompt}(t_c))$ for all c ▷ class names only; no target image

$\iota_c \leftarrow 1 - \max_{k \neq c} \cos(u_c, u_k)$

$\ell_c \leftarrow u_c^\top P_c u_c$, or 0 with a missing indicator if t_c has no components

$\eta_c \leftarrow$ paraphrase rank of t_c against R

$\gamma_c, \omega_c, \pi_c \leftarrow$ cycle consistency, margin dispersion, and prototype alignment of the first b_0 inversions of class c

$s_c \leftarrow$ least-squares linear index of $(\iota_c, \ell_c, \eta_c, \gamma_c, \omega_c, \pi_c)$

$\hat{a}_c \leftarrow$ isotonic map of s_c, fit on the calibration classes

return classes ranked by increasing $\hat{a}_c$ ▷ a shortlist, not a certified accuracy

The forecast ϕ is their least-squares linear index with isotonic recalibration on calibration classes; an indicator handles missing ℓ_c for single-token names. Fitting uses realized per-class labels from calibration classes only—other public vocabularies under leave-one-vocabulary-out, or the training

Table 1: **Information budget.** What each scorer may use for a new vocabulary. ϕ uses no target photographs; while the oracle needs.

	Label-free ϕ	Teacher (labeled oracle)
Real images of target vocabulary	none	all test images
Labels of target vocabulary	none	test-image class labels
Other inputs	names, inversions; calibration labels (other vocabs or classes)	same anchors
Requires target photographs?	no	yes

70% under class split; scoring applies the fixed coefficients to held-out classes from names and first inversions alone.

4.4 Held-out protocols

Under leave-one-vocabulary-out, each held-out vocabulary is scored by a map fit on the remaining three datasets: for example, when Flowers-102 is held out, ϕ is fit on Caltech-101, CUB-200, and FGVC-Aircraft and then applied to every flower class using only its name and inversions. Isotonic recalibration maps the linear index to a calibrated accuracy scale on the calibration classes; it is refit for each split but uses no image or label from the held-out vocabulary. Under class split, the same procedure holds out 30% of classes within one vocabulary and fits on the other 70%. We repeat every run with five synthesis seeds $(0, \ldots, 4)$ and 64 synthetic images per class. Real images enter *only* when we measure realized a_c on official test splits and when we score the teacher oracle; ϕ never reads them.

Leave-one-vocabulary-out fits on all classes of three vocabularies and scores on the fourth, faithful to deployment, because a data-free user has no labels for their own vocabulary. *Class split* fits on 70% of the classes of one vocabulary and scores on the other 30%. We report Spearman r_s between predicted and realized a_c, because the user's decision is a ranking of which classes to check first. We also report *worst-k precision* with $k = \lfloor 0.30\, n_{\text{eval}} \rfloor$: the fraction of the true lowest-accuracy classes recovered among the k classes ϕ ranks lowest. The teacher reference is per-class zero-shot accuracy on labeled test images with the same anchors; we treat it as an oracle upper bound see Table 1, not a same-budget competitor.

Table 1 lists what each scorer may use at deployment time. ϕ may read class names, the first synthetic inversions of the vocabulary, and realized per-class labels from calibration data used only to fit the linear coefficients, other public vocabularies under leave-one-vocabulary-out, or other classes within the same vocabulary under class split. It may not read any real image or label of the held-out classes. The teacher oracle may read all labeled test images of those classes and compute per-class zero-shot accuracy with the same anchors. Trailing the oracle on Spearman r_s therefore does not mean ϕ failed; it means labeled images carry information ϕ deliberately forgoes. The fair question is whether ϕ recovers enough of the student-specific rank, especially among the weakest classes, to matter before target photographs exist.

4.5 Information budget

Customization already pays for VQGAN inversion and distillation; synthesis dominates total time, while scoring ϕ is a negligible add-on once a pool exists. ϕ does not shortcut that step for the full six-feature map, inversion-side statistics read pipeline by-products. The point is not to replace labeled evaluation when photographs exist, but to offer a per-class estimation to the users without any real images of the held-out classes. A text-only variant using just $(\iota_c, \ell_c, \eta_c)$ needs no inversions; on Caltech-101 it matches the full map, on Flowers-102 it loses most triage signal see Appendix Table 7. Workflow: after inversion, score with ϕ; if there are labeled images, validate with real images.

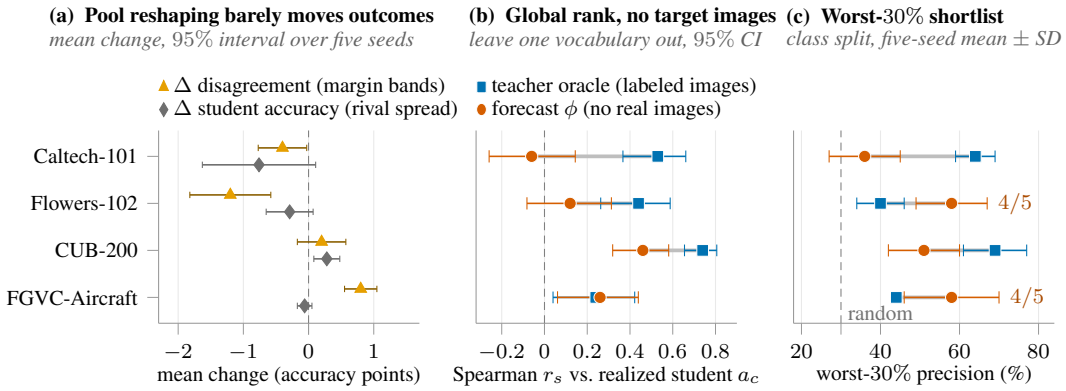


Figure 2: **Audit summary.** Rows are vocabularies; all panels share them. **(a)** Two deliberate pool-reshaping arms: margin-band reshaping shifts mean teacher–student disagreement by at most 1.2 points (triangles) and rival-directed spread shifts mean student accuracy by less (diamonds); bars are 95% intervals over five seeds (Tables 5, 4). **(b)** Leave-one-vocabulary-out Spearman r_s between predicted and realized per-class *student* accuracy for the label-free forecast ϕ (orange) and the labeled teacher oracle (blue); bars are approximate class-level 95% intervals (Table 4). Grey connectors show the gap; only FGVC-Aircraft overlaps. **(c)** Worst-30% precision under class split (Table 3); markers as in (b); the dashed line is the random base rate and the numbers count seeds on which ϕ beats the oracle. Class-split r_s is in Table 2.

4.6 Global rank of per-class accuracy

Figure 2b and Table 2 report Spearman r_s between predicted and realized a_c. We read the table under two protocols. *Leave-one-vocabulary-out* holds out an entire vocabulary the map never saw during fitting—the faithful test for a data-free user who customizes a new list of class names. *Class split* holds out 30% of classes within one vocabulary and is an easier upper bound: some classes share a domain with the calibration set.

Under leave-one-vocabulary-out, ϕ reaches r_s between -0.06 and 0.46; under class split, between 0.07 and 0.52. Both are low-precision estimates that trail the labeled oracle on three of four vocabularies, as expected when the oracle sees every test image. Class split is an easier upper bound because some held-out classes share a domain with the calibration classes in the same vocabulary. The informative question is whether this cheap read carries *enough* signal to be partially useful before photographs exist.

Caltech-101: Names are short and visually grounded ("accordion", "airplane"). ϕ is near zero under both protocols ($r_s = -0.06 \pm 0.04$ leave-one-out; 0.07 ± 0.08 class split), while the oracle remains strong (0.53 ± 0.02). Text-side statistics add little when anchors are already well separated see Appendix Table 7.

Flowers-102: Names are compositional and easily confused ("globe-flower", "balloon flower"). Leave-one-out rank is modest (0.12 ± 0.03) but class split reaches 0.42 ± 0.13, and ϕ wins worst-30% precision on four of five seeds see Section 4.7. Compositional leakage ℓ_c and paraphrase rank η_c carry the most signal here.

CUB-200. Fine-grained bird names are crowded in embedding space. ϕ reaches 0.46 ± 0.01 leave-one-out and 0.52 ± 0.05 class split, among the strongest global ranks, though the oracle stays ahead (0.74 ± 0.01 and 0.79 ± 0.04). Inversion-side statistics (γ_c, ω_c, π_c) contribute alongside text features.

FGVC-Aircraft. Leave-one-out intervals for ϕ (0.26 ± 0.05) and the oracle (0.24 ± 0.03) overlap; ϕ wins three of five seeds on global rank and four of five on worst-30% precision. Here the oracle is itself weak, so a name-and-inversion read can be competitive before target photographs exist.

Under class split, approximate class-level 95% intervals are about ± 0.35 see Table 4; under leave-one-out they are about ± 0.2. Seeds share classes, so seed SD understates class-level uncertainty; we report both.

Table 2: **Global rank: forecasting per-class student accuracy.** Spearman r_s between predicted and realized per-class accuracy under leave-one-vocabulary-out and class split. The label-free forecast ϕ uses no real image of the held-out classes; the teacher reference uses labeled test images. p is a two-sided Fisher z test of the reported $r_s(\phi)$ against zero, with variance $1.06/(n-3)$ as in Table 4. Δr_s is the five-seed mean of $r_s(\phi) - r_s(\text{teacher})$, with a paired bootstrap 95% interval from 700 resamples of the held-out classes. Seeds won counts synthesis seeds on which the teacher has the higher r_s.

		Spearman r_s, mean ± SD			Seeds won
Protocol	Held-out	Label-free ϕ (p vs. 0)	Teacher (labeled)	Δr_s [95% CI]	by teacher
Leave one vocabulary out	Caltech-101	-0.06 ± 0.04 (0.57)	0.53 ± 0.02	-0.59 $[-0.83, -0.32]$	5/5
	Flowers-102	0.12 ± 0.03 (0.24)	0.44 ± 0.04	-0.32 $[-0.55, -0.13]$	5/5
	CUB-200	0.46 ± 0.01 (< 0.001)	0.74 ± 0.01	-0.28 $[-0.43, -0.17]$	5/5
	FGVC-Aircraft	0.26 ± 0.05 (0.011)	0.24 ± 0.03	$+0.02$ $[-0.16, +0.23]$	2/5
Class split (30% held out)	Caltech-101	0.07 ± 0.08 (0.73)	0.51 ± 0.03	-0.44 $[-0.84, -0.04]$	5/5
	Flowers-102	0.42 ± 0.13 (0.022)	0.62 ± 0.05	-0.20 $[-0.55, +0.05]$	5/5
	CUB-200	0.52 ± 0.05 (< 0.001)	0.79 ± 0.04	-0.27 $[-0.45, +0.00]$	5/5
	FGVC-Aircraft	0.44 ± 0.12 (0.016)	0.46 ± 0.11	-0.02 $[-0.43, +0.30]$	4/5

Two patterns stand out. First, ϕ is strongest on fine-grained vocabularies (CUB-200, FGVC-Aircraft) and on compositional flower names, where text-side ambiguity is high. Second, even when global Spearman trails the oracle, as on CUB-200 class split (0.52 vs. 0.79), the deployment question is whether the *bottom* of the list is right; Section 4.7 shows that global rank and triage quality can diverge.

4.7 Which classes to check first

Deployment rarely requires a perfect global rank; it requires knowing which classes to validate first. Suppose a user will eventually label k classes for a spot check. Random selection recovers a k/n fraction of the true worst classes by base rate alone (≈30% here). A useful triage scorer should do substantially better.

Table 3 reports worst-30% precision under the class split: among the $k = \lfloor 0.30\, n_{\text{eval}} \rfloor$ classes ϕ ranks lowest, what fraction are truly among the k lowest-accuracy classes? On Flowers-102, ϕ recovers $58\% \pm 9\%$ of the true worst classes against $40\% \pm 6\%$ for the oracle, winning on four of five seeds; on FGVC-Aircraft the margin is $58\% \pm 12\%$ against $44\% \pm 0\%$, again with four of five wins. On CUB-200 ($51\% \pm 9\%$ vs. $69\% \pm 8\%$) and Caltech-101 ($36\% \pm 9\%$ vs. $64\% \pm 5\%$) the oracle remains ahead. The asymmetry is informative: ϕ helps most where names are ambiguous or fine-grained, not where generic object names already separate cleanly in text space.

4.8 Limits and ablations

ϕ's usefulness depends on the validation budget. Sweeping k from 10% to 50% of the held-out classes (Table 6), ϕ beats random inspection on CUB-200 and FGVC-Aircraft under both protocols and on Flowers-102 under class split, while on Caltech-101 it tracks random. It beats the labeled oracle only on Flowers-102 and FGVC-Aircraft, and only at intermediate budgets: on FGVC-Aircraft under leave-one-vocabulary-out it wins on every seed from 20% to 40%, but the margin shrinks as the budget grows and reverses by 50% on Flowers-102. On Caltech-101 and CUB-200 the oracle leads at every budget, and at the tightest budget ϕ is weak except on Flowers-102—it locates a weak region of the vocabulary rather than the single worst class.

These budget sweeps clarify the role of ϕ: it is a pre-image estimate for the interval before target labeled data exist, not a replacement for labeled evaluation afterward.

We do not claim that ϕ removes the need for customization or labeled validation, or that it matches oracle accuracy globally. We claim only that it is a cheap, low-precision ranked shortlist for the regime where the user has names and synthetic inversions but not target real images—and that on two of four vocabularies this shortlist beats a labeled oracle on the deployment-critical worst-30% metric under class split.

Table 3: **Triage quality: worst-30% precision (class split, five seeds).** For $k = \lfloor 0.30\, n_{\text{eval}} \rfloor$, precision is $|\widehat{B}_k \cap B_k|/k$, where B_k and $\widehat{B}_k$ are the true and predicted bottom-k class sets. "ϕ wins" counts seeds on which ϕ exceeds the teacher oracle.

Vocabulary	n_{eval}	k	ϕ (%)	Teacher (%)	ϕ wins
Caltech-101	30	9	36 ± 9	64 ± 5	$0/5$
Flowers-102	31	9	58 ± 9	40 ± 6	$4/5$
CUB-200	45	13	51 ± 9	69 ± 8	$0/5$
FGVC-Aircraft	30	9	58 ± 12	44 ± 0	$4/5$

4.9 Reading the audit for practitioners

Table 3 can be read as a concrete validation budget. On Flowers-102 class split, $k = 9$ classes are flagged; ϕ places about 58% of the true nine weakest classes in that shortlist on average—roughly five of nine—whereas the labeled oracle recovers about four of nine. Random inspection at the same budget would recover about three. The gain is modest in absolute terms but obtained without any real image of the vocabulary. On Caltech-101 the oracle recovers about six of nine true worst classes and ϕ about three; there a practitioner should not rely on ϕ alone for triage. A text-only variant see Appendix Table 7 matches the full map on Caltech but loses most signal on Flowers, clarifying when class names alone suffice.

Leave-one-out is the conservative read for a new user vocabulary. Global Spearman is lower there than under class split, yet CUB-200 still reaches 0.46 ± 0.01 and FGVC-Aircraft overlaps the oracle. We therefore report both protocols: leave-one-out answers "will this work on a vocabulary unlike my calibration sets?"; class split answers "how much signal is present when some same-domain classes were seen during fitting?" A deployment recommendation that follows from both is domain-dependent. Favor ϕ for compositional or fine-grained name lists; treat it as a weak prior on generic object vocabularies and confirm with labeled images as soon as they exist.

5 Discussion

Implications. Data-free customization removes the test set along with the data requirement. Our work asks how well class names, with optional inversion-side statistics, forecast per-class student accuracy in that regime. We answer under leave-one-vocabulary-out and class split. The honest synthesis defines ϕ as a *cheap, low-precision, partially effective* student audit: global rank is modest and usually below a labeled oracle under both protocols, yet on compositional and fine-grained vocabularies the worst-class shortlist can beat that oracle under class split, and scoring costs seconds rather than a full labeled evaluation pass.

The recommended workflow is staged. After inversion, score ϕ and flag the bottom-ranked classes; treat the ranking as a weak prior, not a calibrated accuracy report, and send the labeling budget there first. Once labeled images of those classes exist, ϕ has done its job: measure a_c directly and let the measurement replace the forecast. The teacher then serves diagnosis rather than confirmation, a class it also misses has a name CLIP cannot ground, which argues for better anchors or exclusion; a class it handles but the student misses was lost in synthesis, the regime of equation (1), which argues for example images.

What we do not claim. ϕ does not avoid synthesis for the full six-feature map. Once target photographs exist, labeled evaluation is inexpensive relative to VQGAN inversion and should replace ϕ for accuracy. We do not claim high global precision: on Caltech-101 ϕ is near useless under both protocols, and even on CUB-200 it trails the oracle by 0.28 Spearman points leave-one-out and 0.27 under class split. The method is worth using only when photographs are absent and the vocabulary is compositional or fine-grained.

When to trust the forecast. The empirical pattern supports narrower conditions than a vocabulary type alone. First, the task must be to shortlist a weak region of the vocabulary: ϕ is competitive at moderate budgets, k from roughly 20% to 40% of classes, its margin disappears by 50%, and it is never a calibrated accuracy report. Second, the names must be compositional or fine-grained, as in Flowers-102 and FGVC-Aircraft; on the one split we ablate, the signal there comes from the inversion-side statistics and from π_c above all see Table 7, so a text-only variant does not suffice, on

Caltech-101 it merely ties a full map that is itself near chance. Third, cross-vocabulary calibration transfers only in part: under leave-one-vocabulary-out, triage survives on FGVC-Aircraft and global rank on CUB-200, while Flowers-102 falls to chance. Even on vocabularies where the teacher explains only part of the student rank, ϕ can still fail, as on Caltech-101.

Future work. The audit measures a forecast; it does not act on one. Does repair pay—if flagged classes receive better anchors, example images, or exclusion, does per-class accuracy rise more than under the same effort spent uniformly? And is the shortlist a good rule for spending a scarce budget, such as the few real images per class that (Wei et al., 2025) use for recovery? Both ask whether a coarse, label-free ranking is good enough to drive a decision, which is a different question from how well it ranks.

Limitations. The audit on Leave-one-vocabulary-out is the faithful deployment test, calibration labels come from other public vocabularies; class split is optimistic because calibration labels come from sibling classes in the same vocabulary. Perfect per-class prediction remains impossible without real images according to Observation 1. The teacher reference shares test-set noise with the evaluation target, and CLIP pretraining may overlap the evaluation sets. Calibration vocabularies are public benchmarks, not user-uploaded private lists.

6 Conclusion

Data-free CLIP customization ships no per-class report and provides no test set at customization time. We ask how well class names, together with synthetic inversions from the same pipeline, can forecast per-class student accuracy before any real image of held-out classes exists. A linear map ϕ of six statistics, audited leave-one-vocabulary-out and under a $70/30$ class split, is our instrument: cheap to score, effective on some name lists, but can't replace real image estimation when have real images.

Global rank trails a labeled oracle on Caltech-101, Flowers-102, and CUB-200 under both protocols; on Caltech-101 ϕ is near zero. FGVC-Aircraft is the exception: its rank overlaps the oracle's under both protocols (0.26 against 0.24 leave-one-out, 0.44 against 0.46 under class split). The strongest rank is CUB-200's, 0.46 leave-one-out and 0.52 under class split, but there the oracle still leads, including on worst-30% precision. The shortlist beats the oracle on Flowers-102 and FGVC-Aircraft under class split see Table 3; on Flowers-102 that advantage is confined to class split, since leave-one-out rank is 0.12, while on FGVC-Aircraft the same worst-30% comparison also wins leave-one-out. Observation 1 explains why perfection is impossible without real images. The honest answer to our title is *partially, for some vocabularies under some protocols*: a cheap shortlist before photographs exist, not a substitute for labeled student evaluation afterward.

AI use statement

We used generative AI for drafting and editing, figure layout, and audit code development. We did not use generative AI for synthetic data generation, proof writing, or interpreting results without author review. All claims were verified by the authors, who take full responsibility for the final manuscript.

Ethics statement

The audit uses public datasets and involves no human subjects. Data-free customization inherits the teacher's view of which concepts are legible; we recommend scoring ϕ before deployment and validating flagged classes once labeled images exist.

Reproducibility statement

The audit protocol, splits, and Algorithm 1 are in Section 4; feature definitions and experiment tables are in Appendices C and B. Code to reproduce the reported audit is provided in the supplementary material.

REFERENCES

Patrick Esser, Robin Rombach, and Bjorn Ommer. Taming transformers for high-resolution image synthesis. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021.

Gongfan Fang, Jie Song, Xinchao Wang, Chengchao Shen, Xingen Wang, and Mingli Song. Contrastive model inversion for data-free knowledge distillation. In *International Joint Conference on Artificial Intelligence*, 2021.

Gongfan Fang, Kanya Mo, Xinchao Wang, Jie Song, Shitao Bei, Haofei Zhang, and Mingli Song. Up to 100x faster data-free knowledge distillation. In *AAAI Conference on Artificial Intelligence*, 2022.

Logan Frank and Jim Davis. What makes a good dataset for knowledge distillation? In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 23755–23764, 2025.

Saurabh Garg, Sivaraman Balakrishnan, Zachary C. Lipton, Behnam Neyshabur, and Hanie Sedghi. Leveraging unlabeled data to predict out-of-distribution performance. In *International Conference on Learning Representations (ICLR)*, 2022.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2016.

Sara Hooker, Aaron Courville, Gregory Clark, Yann Dauphin, and Andrea Frome. What do compressed deep neural networks forget? *arXiv preprint arXiv:1911.05248*, 2019.

Marc Lafon, Yannis Karmim, Julio Silva-Rodríguez, Paul Couairon, Clément Rambour, Raphaël Fournier-Sniehotta, Ismail Ben Ayed, Jose Dolz, and Nicolas Thome. ViLU: Learning vision-language uncertainties for failure prediction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 17807–17817, 2025.

Subhransu Maji, Esa Rahtu, Juho Kannala, Matthew Blaschko, and Andrea Vedaldi. Fine-grained visual classification of aircraft. *arXiv preprint arXiv:1306.5151*, 2013.

Shubham Parashar, Zhiqiu Lin, Tian Liu, Xiangjue Dong, Yanan Li, Deva Ramanan, James Caverlee, and Shu Kong. The neglected tails in vision-language models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 12988–12997, 2024.

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*, 2021.

Kevin Robbins, Xiaotong Liu, Yu Wu, Le Sun, Grady McPeak, Abby Stylianou, and Robert Pless. Will it zero-shot? predicting zero-shot classification performance for arbitrary queries. In *IEEE International Conference on Data Mining Workshops (ICDMW)*, 2025.

Vishaal Udandarao, Ameya Prabhu, Adhiraj Ghosh, Yash Sharma, Philip Torr, Adel Bibi, Samuel Albanie, and Matthias Bethge. No "zero-shot" without exponential data: Pretraining concept frequency determines multimodal model performance. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024.

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

Yongxian Wei, Zixuan Hu, Li Shen, Zhenyi Wang, Chun Yuan, and Dacheng Tao. Open-vocabulary customization from CLIP via data-free knowledge distillation. In *International Conference on Learning Representations*, 2025.

Yunyi Xuan, Weijie Chen, Shicai Yang, Di Xie, Luojun Lin, and Yueting Zhuang. Distilling vision-language foundation models: A data-free approach via prompt diversification. In *ACM International Conference on Multimedia*, 2023.

Chao Yi, Yu-Hang He, De-Chuan Zhan, and Han-Jia Ye. Bridge the modality and capability gaps in vision-language model selection. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024.

Hongxu Yin, Pavlo Molchanov, Jose M Alvarez, Zhizhong Li, Arun Mallya, Derek Hoiem, Niraj K Jha, and Jan Kautz. Dreaming to distill: Data-free knowledge transfer via DeepInversion. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020.

Orr Zohar, Shih-Cheng Huang, Kuan-Chieh Wang, and Serena Yeung. LOVM: Language-only vision model selection. In *Advances in Neural Information Processing Systems (Datasets and Benchmarks Track)*, 2023.

A FORMAL RESULTS

Frozen encoders map images and text to a shared sphere; vocabulary V fixes anchors u_c and centered anchors $\bar{u}_c = u_c - \bar{u}$. The six audit statistics used by ϕ are defined in Appendix C.

Assumption 1 (Softmax-mediated objectives)**.** The anchors enter synthesis and distillation only through the teacher's softmax over V, plus anchor-free terms; the student's initialization does not depend on the anchors.

Theorem 1 (Vocabulary sufficiency)**.** *Under Assumption 1, the trained student depends on the vocabulary only through $\bar{U}$: with encoders, generator, and hyperparameters fixed, it is a measurable function of $(\bar{U}, seed)$. For any fixed real distribution, per-class accuracy and per-class student-teacher disagreement are determined by $\bar{U}$ up to the seed.*

B ADDITIONAL EXPERIMENT TABLES

Table 4: **Approximate uncertainty derived from the reported summaries.** Top: approximate 95% intervals for the held-out Spearman correlations of Table 2, from Fisher's z with variance $1.06/(n-3)$ for n held-out classes; seeds share classes, so class sampling dominates. "Separate" means the two intervals do not overlap. Bottom: 95% intervals over five seeds (Student t) for the reshaping effects of Table 5. Computed by *triage_approx_uncertainty.py*. The paired bootstrap over classes in Table 2 supersedes the top block for Δr_s.

Protocol	Vocabulary	n	Label-free ϕ	Teacher (labeled)	Intervals
Leave one out	Caltech-101	100	$-0.06\ [-0.26, +0.14]$	$0.53\ [+0.37, +0.66]$	separate
Leave one out	Flowers-102	102	$+0.12\ [-0.08, +0.31]$	$0.44\ [+0.26, +0.59]$	overlap
Leave one out	CUB-200	150	$+0.46\ [+0.32, +0.58]$	$0.74\ [+0.66, +0.81]$	separate
Leave one out	FGVC-Aircraft	100	$+0.26\ [+0.06, +0.44]$	$0.24\ [+0.04, +0.42]$	overlap
Class split	Caltech-101	30	$+0.07\ [-0.31, +0.43]$	$0.51\ [+0.17, +0.74]$	overlap
Class split	Flowers-102	31	$+0.42\ [+0.07, +0.68]$	$0.62\ [+0.33, +0.80]$	overlap
Class split	CUB-200	45	$+0.52\ [+0.25, +0.71]$	$0.79\ [+0.63, +0.88]$	overlap
Class split	FGVC-Aircraft	30	$+0.44\ [+0.08, +0.70]$	$0.46\ [+0.10, +0.72]$	overlap

Arm	Vocabulary	Mean (points)	95% interval	Nonzero	Within ± 1.75
Margin bands	Caltech-101	-0.4	$[-0.8, 0.0]$	borderline	yes
Margin bands	Flowers-102	-1.2	$[-1.8, -0.6]$	yes	borderline
Margin bands	CUB-200	$+0.2$	$[-0.2, +0.6]$	no	yes
Margin bands	FGVC-Aircraft	$+0.8$	$[+0.6, +1.0]$	yes	yes
Rival spread	Caltech-101	-0.76	$[-1.63, +0.11]$	no	yes
Rival spread	Flowers-102	-0.29	$[-0.65, +0.07]$	no	yes
Rival spread	CUB-200	$+0.28$	$[+0.08, +0.48]$	yes	yes
Rival spread	FGVC-Aircraft	-0.06	$[-0.17, +0.05]$	no	yes

Table 5: **Pool reshaping on the real pipeline (Margin-band reshaping, Rival-spread reshaping; five seeds per vocabulary).** All changes are in points, as in Figure 2a. Margin bands: prototype fraction swept from 1.0 to 0.5 at fixed budget; "monotone" means disagreement is non-decreasing as budget leaves the prototypes. Rival spread: rival-directed spread minus style-prompt variation on mean per-class accuracy.

	Margin bands			Rival spread	
Vocabulary	Monotone	Δ disagreement	Sweep range	Δ accuracy	Repair wins
Caltech-101	$0/5$	-0.4 ± 0.3	1.3 ± 0.2	-0.76 ± 0.70	$1/5$
Flowers-102	$0/5$	-1.2 ± 0.5	1.7 ± 0.5	-0.29 ± 0.29	$1/5$
CUB-200 (150 classes)	$0/5$	$+0.2 \pm 0.3$	0.9 ± 0.2	$+0.28 \pm 0.16$	$5/5$
FGVC-Aircraft	$0/5$	$+0.8 \pm 0.2$	1.1 ± 0.2	-0.06 ± 0.09	$2/5$
All 20 seeds	$0/20$	max |mean| 1.2	mean 1.25	-0.21 ± 0.55	$9/20$

The last row pools all 20 vocabulary–seed runs. "max |mean| 1.2" is the largest absolute vocabulary-level mean Δ teacher–student disagreement under margin-band reshaping (Flowers-102). "mean 1.25" is the mean sweep range of Δ disagreement across vocabularies; rival-spread mean Δ accuracy is -0.21 ± 0.55 points.

Table 6: **Triage quality across validation budgets (Budget sweep, recomputed; five seeds).** For a budget of $k = \lfloor \beta n_{\text{eval}} \rfloor$ classes, precision is $|\widehat{B}_k \cap B_k|/k$, where B_k and $\widehat{B}_k$ are the true and predicted lowest-accuracy class sets, as in Table 3. Random inspection recovers the base rate k/n_{eval}, listed once per row because flooring makes it differ across vocabularies by at most 1.5 points. Bold marks the cells on which the label-free ϕ exceeds the labeled teacher oracle. The $\beta = 30\%$ class-split row reproduces Table 3. Held-out sizes are 30, 31, 45, 30 under class split and 100, 102, 150, 100 under leave one vocabulary out.

		Caltech-101		Flowers-102		CUB-200		FGVC-Aircraft	
Budget β	Random	ϕ	Teacher	ϕ	Teacher	ϕ	Teacher	ϕ	Teacher
Class split (30% of classes held out)									
10%	10	13 ± 16	33 ± 0	$\mathbf{67 \pm 0}$	33 ± 0	20 ± 10	60 ± 12	20 ± 27	80 ± 16
20%	20	20 ± 12	53 ± 12	$\mathbf{47 \pm 7}$	33 ± 0	29 ± 5	64 ± 11	$\mathbf{53 \pm 16}$	40 ± 8
30%	29	36 ± 8	64 ± 4	$\mathbf{58 \pm 8}$	40 ± 5	51 ± 8	69 ± 7	$\mathbf{58 \pm 11}$	44 ± 0
40%	40	40 ± 6	67 ± 5	$\mathbf{63 \pm 4}$	53 ± 8	66 ± 4	81 ± 3	50 ± 5	52 ± 8
50%	49	47 ± 8	75 ± 3	60 ± 4	77 ± 3	73 ± 3	85 ± 2	67 ± 7	69 ± 3
Leave one vocabulary out (fit on the other three)									
10%	10	0 ± 0	26 ± 8	$\mathbf{14 \pm 5}$	12 ± 4	12 ± 5	48 ± 3	14 ± 10	50 ± 6
20%	20	25 ± 0	47 ± 7	21 ± 5	38 ± 2	29 ± 2	57 ± 4	$\mathbf{56 \pm 5}$	33 ± 2
30%	30	35 ± 2	61 ± 3	33 ± 4	45 ± 2	53 ± 3	75 ± 5	$\mathbf{63 \pm 4}$	44 ± 2
40%	40	36 ± 5	66 ± 2	42 ± 3	56 ± 2	58 ± 1	77 ± 2	$\mathbf{65 \pm 2}$	53 ± 3
50%	50	47 ± 3	68 ± 1	50 ± 3	65 ± 3	64 ± 2	81 ± 1	$\mathbf{71 \pm 2}$	69 ± 2

Seed counts on which ϕ exceeds the oracle, for $\beta = 10, 20, 30, 40, 50\%$: Flowers-102 under class split $5, 4, 4, 3, 0$; FGVC-Aircraft under class split $0, 3, 4, 1, 0$ and under leave one vocabulary out $0, 5, 5, 5, 4$; Caltech-101 and CUB-200 never.

Table 7: **Which statistics carry signal (Feature ablation; one split holding out 30% of classes, seed 0).** Left: Spearman r_s of each scorer, with the teacher as the labeled reference. Right: r_s of ϕ with one statistic removed; with none removed it equals the "All six" row. A single seed and split, so differences of a few hundredths are within noise.

	r_s			r_s, one removed	
Scorer	Caltech	Flowers	Removed	Caltech	Flowers
Teacher (labeled reference)	0.57	0.68	ι	0.24	0.37
Text statistics (ι, ℓ, η)	0.17	0.20	ℓ	0.15	0.30
Inversion statistics (γ, ω, π)	0.37	0.32	η	0.25	0.31
All six, ϕ	0.16	0.31	γ	0.14	0.29
Teacher plus statistics	0.49	0.57	π	0.22	0.14

C AUDIT COMPONENTS AND REPRODUCTION DETAILS

Statistics. All cosines use unit-norm embeddings, and every statistic below uses no real image and no label.

- Anchor isolation: $\iota_c = 1 - \max_{k \neq c} \cos(u_c, u_k)$.
- Compositional leakage: $\ell_c = u_c^\top P_c u_c$, the squared norm of the projection of u_c onto the span of the normalized embeddings of the tokens of t_c and, when the user supplies one, a task-level superordinate term such as "flower". The projection is approximated by ridge regression with ridge 10^{-6} and clipped to $[0, 1]$. It is undefined for a name with no components, in which case ϕ receives 0 and a missing indicator.
- Paraphrase rank: η_c measures how often CLIP-style paraphrases of t_c embed closer to other reference nouns than to u_c; templates and M are fixed in released code.
- Cycle consistency: γ_c is the fraction of the first b_0 inversions of class c whose most similar anchor, over the vocabulary extended by R, is u_c.
- Margin dispersion: ω_c is the sample standard deviation, over the same inversions, of the gap between their largest and second-largest cosine to the anchors of V; it needs $b_0 \geq 2$.
- Prototype alignment: $\pi_c = \cos(\bar{e}_c, u_c)$, where $\bar{e}_c$ is the mean of the normalized embeddings of those inversions; higher values mean less drift.

The released code names these rho, lambda, eta (*retrieval_hubness*), gamma, sigma, pi (*prototype_drift*). In this paper they are ι, ℓ, η, γ, ω, and π, renamed to avoid clashes with the Spearman correlation, the ridge parameter, and the isotropic variance, and to describe what η and π measure.