ALIGNLEDGER-CLIP: GROUP-WISE AUDITING AND GUARDED ADAPTATION FOR FROZEN VISION-LANGUAGE EMBEDDINGS

Anonymous authors
Paper under double-blind review

ABSTRACT

Aggregate retrieval scores can conceal regressions within semantic groups when adapting frozen vision–language embeddings. We introduce ALIGNLEDGER-CLIP, a framework that connects group-wise auditing to adaptation and update selection. A ledger records group geometry, directional retrieval, and representation changes; its statistics weight gap-closing and preservation objectives for lightweight cached-feature adapters. A held-out guard selects an interpolation scale subject to recall and group-degradation constraints, with the frozen representation as a fallback. The same ledger then measures the deployed update on evaluation records. We study 12 evaluation slices from 10 source datasets with CLIP and SigLIP, separating transductive audits from disjoint fit/guard/evaluation experiments. In the latter, guarded ALIGNLEDGER-CLIP achieves 38.27 average Recall@1 with 20.01% degraded groups, while identically guarded CLIP-Refine achieves 38.53 with 21.65%. These retrieval–degradation profiles expose the local cost of adaptation alongside its aggregate benefit.

1 INTRODUCTION

Dual-encoder vision–language models support zero-shot recognition and cross-modal retrieval. CLIP introduced large-scale contrastive image–text pretraining (Radford et al., 2021), ALIGN scaled this approach to noisy web data (Jia et al., 2021), and SigLIP replaced the batch-wise softmax objective with pairwise sigmoid losses (Zhai et al., 2023). Their frozen representations support efficient adaptation through learned prompts (Zhou et al., 2022b;a), residual feature adapters (Gao et al., 2024), cache-based classifiers (Zhang et al., 2022), and coupled multimodal prompts (Khattak et al., 2023).

Post-hoc adaptation can improve retrieval by changing embedding geometry. CyCLIP regularizes cross-modal and intra-modal consistency (Goel et al., 2022), AlignCLIP improves representation alignment (Eslami & de Melo, 2024), and CLIP-Refine performs lightweight post-pretraining alignment through contrastive distillation (Yamaguchi et al., 2025). The modality gap can affect instance retrieval and group-level semantics differently (Grassucci et al., 2026). Average performance can also hide failures on minority or shifted groups (Sagawa et al., 2020; Koh et al., 2021). A global retrieval gain therefore leaves open which classes, attributes, relations, or compositional groups improved or degraded.

This issue spans several evaluation regimes. ImageNet-A, ImageNet-R, and ObjectNet expose structured recognition failures (Hendrycks et al., 2021b;a; Barbu et al., 2019). COCO and Visual Genome provide object, attribute, and relation groups (Lin et al., 2014; Krishna et al., 2017). DocVQA and ChartQA test document and chart semantics (Mathew et al., 2021; Masry et al., 2022), while Winoground and SugarCrepe evaluate compositional image–text matching (Thrush et al., 2022; Hsieh et al., 2023). These settings call for an adaptation interface that records group-local alignment, representation drift, and retrieval damage alongside the aggregate score.

We propose ALIGNLEDGER-CLIP, a group-wise alignment ledger for frozen vision–language embeddings. Its central interface is a *ledger–adapt–guard–re-ledger* loop: record each group's geometry and retrieval, fit a lightweight update, calibrate its scale against both recall and group degradation, and record the resulting group-level changes. Group gaps weight correction and preservation during

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.

fitting; the guard can also be applied to existing cached-feature adapters. A deterministic metadata map selects one objective branch per dataset–backbone slice.

We evaluate 12 benchmark slices from 10 source datasets with two frozen backbones. Disjoint fit/guard/evaluation experiments reveal different retrieval–degradation tradeoffs even after applying the same guard: CLIP-Refine attains higher aggregate retrieval, while ALIGNLEDGER-CLIP has a lower fraction of degraded groups. Transductive audits and component ablations complement this comparison by measuring adaptation on the cached benchmark and identifying which objectives drive its gains.

Contributions. The main contributions of this paper are:

- We introduce a group-wise ledger that connects local geometry to correction and preservation, and records directional retrieval, drift, and degradation before and after adaptation.
- We instantiate a cached-feature adaptation loop with identity-initialized maps and a reusable calibration guard that selects update scales using recall and group-degradation constraints.
- We evaluate retrieval and group degradation jointly under disjoint data roles and transductive audits, including equally guarded baselines and ablations of the adaptation objectives.

2 RELATED WORK

Vision-language pretraining. Large-scale vision–language pretraining aligns images and descriptions to learn transferable visual representations. CLIP popularized contrastive dual encoders for zero-shot transfer through text prompts (Radford et al., 2021). ALIGN scaled this recipe to billions of noisy image–alt-text pairs (Jia et al., 2021). SigLIP introduced a pairwise sigmoid loss that improves scaling behavior and batch-size flexibility (Zhai et al., 2023). These models provide the interface used here: image and text encoders produce normalized embeddings, and nearest-neighbor or class-prompt similarity gives the prediction. ALIGNLEDGER-CLIP operates independently of the pretraining objective. It takes a frozen backbone and measures and adapts group-local alignment in its cached embeddings.

Prompt and adapter-based CLIP adaptation. CLIP adaptation often learns a small set of parameters around a frozen backbone. CoOp learns context vectors in text prompts; CoCoOp conditions prompts on the input image to improve transfer to unseen classes (Zhou et al., 2022b;a). CLIP-Adapter and Tip-Adapter tune residual feature transformations or cache-based classifiers (Gao et al., 2024; Zhang et al., 2022). MaPLe couples prompts in the visual and textual branches to improve cross-modal adaptation (Khattak et al., 2023). These methods are effective when labeled examples or few-shot class prompts represent the target task well. Our setting also requires identifying and controlling which semantic groups move during adaptation. ALIGNLEDGER-CLIP therefore adapts cached features and makes group-level diagnostics part of both training and reporting.

Group-robust adaptation and alignment. Group-robust CLIP adaptation already extends beyond global accuracy. You et al. (2024) construct a calibration set and use contrastive representation calibration without group annotations. Project-Probe-Aggregate infers groups using biased classifiers and aggregates group weights to improve robust classification (Zhu et al., 2025). Our manifest-defined groups instead provide an explicit unit for measuring retrieval changes and constraining an update. This interface complements post-pretraining alignment methods such as CLIP-Refine (Yamaguchi et al., 2025) and group-wise modality-gap analysis (Grassucci et al., 2026): their updates can be assessed with the same before/after ledger and calibration guard.

3 METHOD

ALIGNLEDGER-CLIP connects four operations on cached image and text embeddings: build a group-wise ledger, fit an adapter, calibrate the update on held-out records, and rebuild the ledger for evaluation. The backbone remains frozen. The ledger and guard accept any adapter that returns embeddings in the same space; we instantiate this interface with identity-initialized linear maps and fixed objective branches (Figure 1).

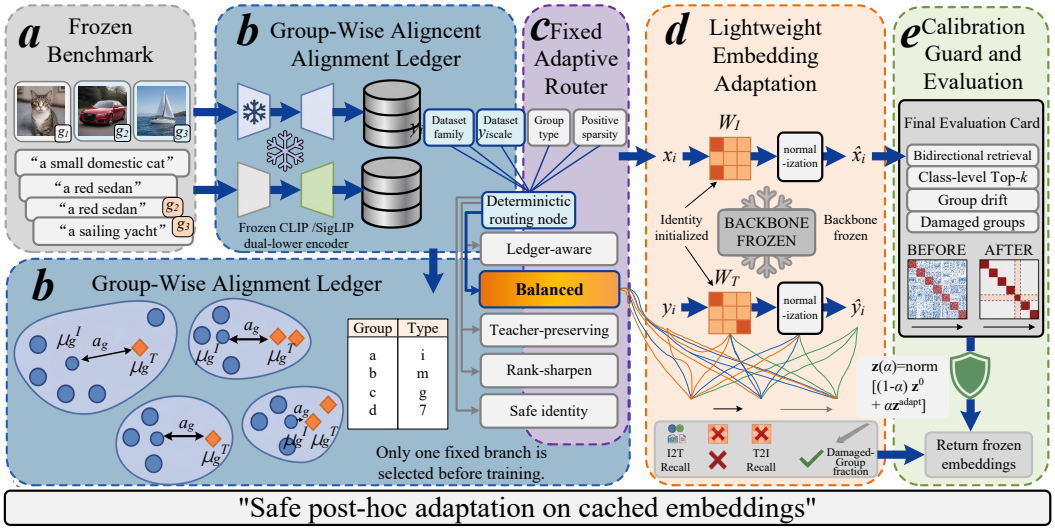


Figure 1: Overview of ALIGNLEDGER-CLIP. Frozen CLIP or SigLIP embeddings are summarized by a group-wise ledger. A metadata-only router selects one fixed branch for the dataset–backbone slice. Identity-initialized maps adapt the cached embeddings, after which a held-out guard backtracks or rejects the update. Re-ledgering reports retrieval, compactness, drift, and damaged groups.

3.1 PROBLEM SETTING AND GROUP MANIFEST

For a dataset–backbone slice, let

$$\mathcal{D} = \{(x_i, y_i, g_i)\}_{i=1}^{N}, \qquad x_i, y_i \in \mathbb{S}^{d-1}, \tag{1}$$

where x_i and y_i are frozen normalized embeddings of a paired image and text, and $g_i \in \mathcal{G}$ is the group identifier supplied by the manifest. Each materialized row has one group identifier; an underlying image that participates in several semantic slices is represented by a separate row in each slice. Groups denote classes, attributes, relations, question types, or compositional perturbations depending on the benchmark. We write $s_{ij}^0 = \tau^{-1} x_i^\top y_j$ for frozen cross-modal similarity. For a new dataset, the manifest specifies the semantic categories to audit using task annotations or fixed metadata rules; the benchmark constructions are listed in Table 8.

The method accepts three record-level data roles: $\mathcal{D}_{\text{fit}}$, $\mathcal{D}_{\text{guard}}$, and $\mathcal{D}_{\text{eval}}$. In the strict protocol they are disjoint:

$$\mathcal{D} = \mathcal{D}_{\text{fit}} \dot{\cup} \mathcal{D}_{\text{guard}} \dot{\cup} \mathcal{D}_{\text{eval}}. \tag{2}$$

The maps see only $\mathcal{D}_{\text{fit}}$, the guard sees only $\mathcal{D}_{\text{guard}}$, and reported metrics use only $\mathcal{D}_{\text{eval}}$. Groups too small to populate all three roles remain evaluation-only. We also report a transductive cached-benchmark audit, where $\mathcal{D}_{\text{fit}} = \mathcal{D}_{\text{eval}}$ and no held-out claim is made. Group labels supervise adaptation and define the audit. The learned maps are shared across groups, so they also apply to evaluation-only groups without requiring a group-specific parameter.

3.2 GROUP-WISE ALIGNMENT LEDGER

For group g, let $\mathcal{I}_g = \{i : g_i = g\}$, and define frozen centroids, gap, and within-modality compactness as

$$\begin{aligned} \mu_g^M &= \text{norm}\left(\frac{1}{|\mathcal{I}_g|}\sum_{i\in\mathcal{I}_g} z_i^M\right), \quad a_g = \|\mu_g^I - \mu_g^T\|_2, \\ c_g^M &= 1 - \frac{1}{|\mathcal{I}_g|}\sum_{i\in\mathcal{I}_g} z_i^{M\top}\mu_g^M. \end{aligned} \tag{3}$$

where $M \in \{I, T\}$, $z_i^I = x_i$, $z_i^T = y_i$, and $\text{norm}(z) = z/(\|z\|_2 + \epsilon)$.

For each direction, the ledger records retrieval $R_g^{I\to T}, R_g^{T\to I}$ and group-hit consistency $K_g^{I\to T}, K_g^{T\to I}$. A group-hit occurs when a top-k cross-modal neighbor belongs to the query group; the reported damage rates and guard use $k = 1$. Retrieval uses paired positives for image–text benchmarks and group positives for classification. The frozen ledger entry is

$$\mathfrak{L}_g^0 = \begin{pmatrix} n_g, m_g, \mu_g^I, \mu_g^T, a_g, c_g^I, c_g^T, \\ K_g^{I\to T}, K_g^{T\to I}, R_g^{I\to T}, R_g^{T\to I} \end{pmatrix}, \tag{4}$$

where $n_g = |\mathcal{I}_g|$ and $m_g \geq 0$ is a prescribed manifest weight, equal to one by default.

The frozen gap determines how strongly a group is corrected or preserved:

$$\begin{aligned} \bar{a} &= \max\left(\max_{h\in\mathcal{G}} a_h, \epsilon_a\right), \qquad & w_g^{\text{close}} &= m_g \frac{a_g}{\bar{a}}, \\ w_g^{\text{keep}} &= \max\Big(0.05,\, m_g\left(1 - \frac{a_g}{\bar{a}}\right)\Big). \end{aligned} \tag{5}$$

The positive floor retains preservation pressure even for the largest-gap group, while ϵ_a makes the all-zero-gap case well-defined. All weights are computed once from the frozen fit ledger. The full-scale classification SHARPEN profile sets both weights to one, retaining the same gap and preservation terms without group reweighting.

3.3 LEDGER-AWARE ADAPTATION OBJECTIVES

We learn two bias-free linear maps initialized to identity:

$$\begin{aligned} \widehat{x}_i &= \text{norm}(W_I x_i), \\ \widehat{y}_i &= \text{norm}(W_T y_i), \\ W_I &= W_T = I_d \quad \text{at initialization.} \end{aligned} \tag{6}$$

At full scale, losses are evaluated on shuffled batches $\mathcal{B} \subset \mathcal{D}_{\text{fit}}$; for smaller cached suites, the same equations use $\mathcal{B} = \mathcal{D}_{\text{fit}}$. Let $\widehat{S}_{ij}^{\mathcal{B}} = \tau^{-1}\widehat{x}_i^\top \widehat{y}_j$ for $i, j \in \mathcal{B}$. The bidirectional paired contrastive loss is

$$\mathcal{L}_{\text{con}} = \frac{1}{2}\left[\text{CE}(\widehat{S}^{\mathcal{B}}, \text{diag}) + \text{CE}(\widehat{S}^{\mathcal{B}\top}, \text{diag})\right]. \tag{7}$$

Using batch-local centroids $\widehat{\mu}_{g,\mathcal{B}}^I$ and $\widehat{\mu}_{g,\mathcal{B}}^T$, the ledger-gap and pointwise preservation terms are

$$\begin{aligned} \mathcal{L}_{\text{gap}} &= \frac{1}{|\mathcal{G}_{\mathcal{B}}|}\sum_{g\in\mathcal{G}_{\mathcal{B}}} w_g^{\text{close}} \|\widehat{\mu}_{g,\mathcal{B}}^I - \widehat{\mu}_{g,\mathcal{B}}^T\|_2^2, \\ \mathcal{L}_{\text{keep}} &= \frac{1}{|\mathcal{B}|}\sum_{i\in\mathcal{B}} w_{g_i}^{\text{keep}} \left(\|\widehat{x}_i - x_i\|_2^2 + \|\widehat{y}_i - y_i\|_2^2\right). \end{aligned} \tag{8}$$

The first term closes large frozen group gaps; the second limits the movement of individual embeddings from the frozen reference.

For classification, $\mathcal{P}_i = \{j \in \mathcal{B} : g_j = g_i\}$ replaces the single paired positive. With $r_i^+ = \max_{j\in\mathcal{P}_i} \widehat{S}_{ij}^{\mathcal{B}}$ and $r_i^- = \max_{j\notin\mathcal{P}_i} \widehat{S}_{ij}^{\mathcal{B}}$, the image-to-text term is

$$\ell_i^I = -\log \frac{\sum_{j\in\mathcal{P}_i}\exp(\widehat{S}_{ij}^{\mathcal{B}})}{\sum_{k\in\mathcal{B}}\exp(\widehat{S}_{ik}^{\mathcal{B}})} + \max(0, \gamma + r_i^- - r_i^+), \tag{9}$$

and ℓ_i^T is symmetric. We set $\mathcal{L}_{\text{rank}} = \frac{1}{2|\mathcal{B}|}\sum_i(\ell_i^I + \ell_i^T)$.

The router chooses from one unified branch set:

$$\mathcal{B}_{\text{route}} = \left\{\begin{matrix} \text{AWARE, BALANCED, REFINE,} \\ \text{SHARPEN, IDENTITY} \end{matrix}\right\}. \tag{10}$$

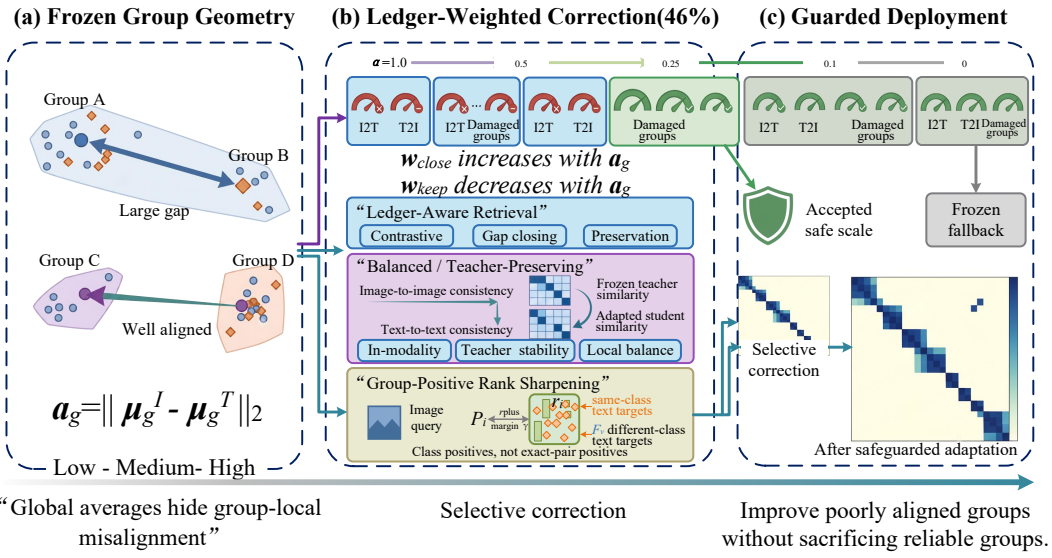


Figure 2: Group-wise mechanism. Frozen group geometry determines gap-closing and preservation weights. The selected branch performs paired alignment, balanced stabilization, frozen-reference refinement, group-positive rank sharpening, or identity abstention. Guarded deployment then accepts the largest passing interpolation scale or returns the frozen representation. Full-scale classification uses uniform weights in the SHARPEN profile. The diagram shows representative checkpoints along the backtracking path; the complete candidate grid is specified in Section 3.5.

Their objectives are

$$\begin{aligned}
\mathcal{L}_{\text{AWARE}} &= \mathcal{L}_{\text{con}} + \lambda_{\text{gap}}\mathcal{L}_{\text{gap}} + \lambda_{\text{keep}}\mathcal{L}_{\text{keep}}, \\
\mathcal{L}_{\text{BALANCED}} &= \mathcal{L}_{\text{AWARE}} + \lambda_{\text{in}}\mathcal{L}_{\text{in}} + \lambda_{\text{teach}}\mathcal{L}_{\text{teach}}, \\
\mathcal{L}_{\text{REFINE}} &= \lambda_{\text{ref}}\mathcal{L}_{\text{teach}} + \lambda_{\text{rand}}\mathcal{L}_{\text{rand}}, \\
\mathcal{L}_{\text{SHARPEN}} &= \mathcal{L}_{\text{rank}} + \lambda_{\text{gap}}\mathcal{L}_{\text{gap}} \\
&\quad + \lambda_{\text{keep}}\mathcal{L}_{\text{keep}} + \lambda_{\text{teach}}\mathcal{L}_{\text{teach}}.
\end{aligned} \tag{11}$$

Here $\mathcal{L}_{\text{in}}$ maintains within-modality discrimination, $\mathcal{L}_{\text{teach}}$ distills frozen cross-modal similarities, and $\mathcal{L}_{\text{rand}}$ regularizes toward fixed random anchors. IDENTITY performs no optimization. The auxiliary losses are defined in Section A.7.

Figure 2 summarizes how group geometry, branch objectives, and guarded deployment interact.

3.4 DETERMINISTIC BRANCH ROUTING

Routing assigns one branch to each dataset–backbone slice; it is fixed across its evaluation samples and semantic groups. Let

$$\phi(\mathcal{D}) = \begin{pmatrix} \text{dataset family, backbone, group type,} \\ \text{scale, positive mode} \end{pmatrix} \tag{12}$$

contain metadata available before optimization. A fixed ordered map $b^\star = \mathcal{R}(\phi(\mathcal{D})) \in \mathcal{B}_{\text{route}}$ selects the branch and its fixed coefficient profile. The selected branch does not change during training and $\mathcal{R}$ never reads calibration or evaluation scores. Broadly, dense full-scale classification uses SHARPEN; sparse or unsupported full-scale slices may use IDENTITY; large pair-retrieval slices use REFINE; attribute/relation slices with stronger within-modality constraints use BALANCED; and all remaining slices use AWARE. The complete scale- and backbone-specific map is given in Section A.9.

3.5 CALIBRATION GUARD AND FROZEN FALLBACK

After fitting W_I, W_T, the guard evaluates a descending set of positive update scales $\mathcal{A} = \{1, 0.5, 0.25, 0.1, 0.05, 0.01\}$. For either modality, it forms

$$z_i(\alpha) = \text{norm}\Big((1-\alpha)z_i^0 + \alpha z_i^{\text{adapt}}\Big). \tag{13}$$

Let $\Delta R^{I\to T}_{\mathrm{guard}}(\alpha)$ and $\Delta R^{T\to I}_{\mathrm{guard}}(\alpha)$ be guard-set Recall@1 changes relative to the frozen representation, and let $D_{\mathrm{guard}}(\alpha)$ be the damaged-group fraction defined below. Guard recalls are macro-averaged over groups and expressed on the $[0,1]$ scale, so a tolerance of 0.05 means five percentage points. The selected scale is

$$\alpha^\star = \max\left\{ \alpha \in \mathcal{A} : \begin{array}{l} \Delta R^{I\to T}_{\mathrm{guard}}(\alpha) \geq -0.05, \\ \Delta R^{T\to I}_{\mathrm{guard}}(\alpha) \geq -0.05, \\ D_{\mathrm{guard}}(\alpha) \leq 0.20 \end{array} \right\}. \tag{14}$$

If the feasible set is empty, $\alpha^\star = 0$ and the method returns the frozen embeddings. The constraints apply to the observed guard records; evaluation-group degradation is measured separately by re-ledgering.

The complete procedure is given in Algorithm 1.

3.6 Re-Ledgering, Damage Accounting, and Complexity

For the deployed scale, we rebuild every entry as $\widehat{\mathfrak{L}}_g$. Define the average bidirectional group consistency and its change by

$$\bar{K}_g = \frac{1}{2}\left(K_g^{I\to T} + K_g^{T\to I}\right), \qquad \Delta\bar{K}_g = \widehat{\bar{K}}_g - \bar{K}_g^0. \tag{15}$$

A group is damaged exactly when its group-hit consistency decreases:

$$h_g = \mathbf{1}[\Delta\bar{K}_g < 0], \qquad D_{\mathrm{group}} = \frac{1}{|\mathcal{G}|}\sum_{g\in\mathcal{G}} h_g. \tag{16}$$

We separately report $\Delta R_g^{I\to T}$, $\Delta R_g^{T\to I}$, and Δa_g, so a smaller centroid gap cannot conceal worse retrieval. D_{group} measures the prevalence of negative changes, giving each group equal weight. The ledger retains group sizes and directional changes to distinguish their magnitude and support; the sign-based rate alone is not a measure of severity or statistical significance. The summary drift score is

$$A_{\mathrm{drift}} = \frac{1}{|\mathcal{G}|}\sum_g |\Delta a_g| + \frac{1}{|\mathcal{G}|}\sum_g \max(0, -\Delta\bar{K}_g). \tag{17}$$

For batch size B, dense projection and batch-local similarities require $O(Bd^2 + B^2 d)$ time and $O(d^2 + Bd + B^2)$ training memory, including the two maps and their optimizer states. Ledger centroids and compactness statistics stream in $O(Nd)$ time, with $O(|\mathcal{G}|d)$ ledger storage. Retrieval uses query-blocked exact top-k search without materializing an $N \times N$ matrix; the largest retrieval slice uses the stated FAISS streaming backend. The full ledger is retained, while optimization and retrieval memory are bounded by the batch or search-block size.

4 Experiments

4.1 Experimental Setup

We evaluate AlignLedger-CLIP on 12 benchmark slices from 10 source datasets, spanning robustness classification, object–attribute–relation retrieval, document and chart retrieval, and compositional image–text matching. The frozen backbones are CLIP ViT-B/32 (Radford et al., 2021) and SigLIP ViT-B/16 (Zhai et al., 2023). We construct three evaluation suites of increasing scale. The multi-seed suite evaluates three seeds over 720 dataset–backbone–method–seed runs. The paper-scale suite contains 42,678 image–text pairs at seed 42, whereas the full-scale suite removes the sample cap and contains 2,514,471 pairs at seed 42. The datasets include ImageNet-1K (Deng et al., 2009); ImageNet-A, ImageNet-R, and ObjectNet (Hendrycks et al., 2021b;a; Barbu et al., 2019); COCO object, attribute, and relation slices, and Visual Genome relations (Lin et al., 2014; Krishna et al., 2017); DocVQA and ChartQA (Mathew et al., 2021; Masry et al., 2022); and Winoground and SugarCrepe (Thrush et al., 2022; Hsieh et al., 2023). The three COCO slices are counted separately throughout the tables and suite averages. Group compositions appear in Table 8.

Method	R@1	Δ	Degraded	Adapt./Frozen
Frozen reference	36.03	0.00	0.00	–
CR-off	35.37	-0.66	43.44	–
CR-off + calibrated scaling	**38.53**	+2.50	21.65	88.9/11.1
CR-off + reference reversion	36.09	+0.06	**4.75**	22.2/77.8
ALIGNLEDGER-CLIP uncalibrated	34.71	-1.32	43.40	–
ALIGNLEDGER-CLIP + calibrated scaling	38.27	+2.24	20.01	87.5/12.5
AL-fixed + calibrated scaling	37.59	+1.55	15.47	77.8/22.2
AL-balanced + calibrated scaling	37.37	+1.34	16.57	76.4/23.6

Table 1: Disjoint-partition results on the paper-scale suite. R@1 and degraded-group rates are percentages averaged over 12 benchmark slices, two backbones, and three partition seeds. Δ is measured relative to the frozen representation on the same evaluation subset. Adapt./Frozen reports the percentages of runs using the adapted or frozen representation.

We compare with the frozen backbone, ModGap (Grassucci et al., 2026), CLIP-Refine (Yamaguchi et al., 2025), AlignCLIP (Eslami & de Melo, 2024), CyCLIP (Goel et al., 2022), and fixed branches of our adapter. CR denotes CLIP-Refine; "-off" identifies an objective implemented using official source code within our cached-feature pipeline. AL-fixed and AL-balanced denote AWARE and BALANCED; teacher-refined denotes REFINE. These are cached-feature comparisons using common embeddings, manifests, data loaders, and retrieval evaluation.

We distinguish two protocols. The disjoint-partition experiment uses separate fit, guard, and evaluation records and compares adapters under identical guard settings. The multi-seed, paper-scale, and full-scale suite comparisons and classification ablations are *transductive cached-benchmark audits*, with fitting and reporting on the same records.

The primary metric is averaged Recall@1:

$$\mathrm{AvgR@1} = \frac{1}{2}\left(\mathrm{R@1}_{\mathrm{i2t}} + \mathrm{R@1}_{\mathrm{t2i}}\right), \tag{18}$$

Here, $\mathrm{R@1}_{\mathrm{i2t}}$ is image-to-text Recall@1 and $\mathrm{R@1}_{\mathrm{t2i}}$ is text-to-image Recall@1. ImageNet/ObjectNet use class-level positives after repeated prompts are collapsed; the other slices use manifest-paired image–text retrieval. For DocVQA and ChartQA this measures retrieval of the paired text, rather than answer accuracy; Winoground and SugarCrepe likewise use retrieval over the manifest candidate pool rather than their original coupled-matching or distractor tests. Degraded-group rates report $100D_{\mathrm{group}}$ from Equation (16).

4.2 DISJOINT-PARTITION COMPARISON WITH A SHARED GUARD

We split image–text pairs within each semantic group into $\mathcal{D}_{\mathrm{fit}}$, $\mathcal{D}_{\mathrm{guard}}$, and $\mathcal{D}_{\mathrm{eval}}$, keeping each pair's two modalities together. The target proportions are 20%/20%/60%, with at least one pair per role; groups with fewer than three pairs remain evaluation-only. This experiment covers all 12 benchmark slices, both backbones, and partition seeds 42, 43, and 44. All calibrated-scaling variants use the same fit/guard/evaluation records, scale grid, and acceptance thresholds. Reference reversion uses the same constraints but tests only $\alpha = 1$, otherwise returning frozen.

Calibration-based update selection. Calibration improves both post-hoc alignment methods (Table 1). For ALIGNLEDGER-CLIP, scaling raises R@1 from 34.71 to 38.27 and reduces degraded groups from 43.40% to 20.01%. CR-off likewise improves from 35.37 to 38.53 R@1, with degraded groups falling from 43.44% to 21.65%. Calibration therefore contributes substantially to both methods. Reference reversion reduces degraded groups to 4.75% by rejecting 77.8% of updates, leaving R@1 near frozen; interpolation retains more of the adaptation gain.

Calibrated CR-off reaches 38.53 R@1, exceeding calibrated ALIGNLEDGER-CLIP by 0.26 points. For ALIGNLEDGER-CLIP minus CR-off, the dataset- and slice-level bootstrap intervals are $[-0.51, -0.07]$ and $[-0.52, -0.04]$, with W/T/L $4/13/7$ across 24 dataset–backbone combinations. ALIGNLEDGER-CLIP has a 1.64-point lower degraded-group rate. The fixed AWARE and

Method	Multi	Paper	Full	Class.
ALIGNLEDGER-CLIP	**49.41**	**45.81**	**36.81**	**76.50**
CR-off	48.63	45.73	36.24	75.24
CR	32.17	31.46	36.24	75.24
AlignCLIP-off	46.20	42.16	34.01	71.55
AL-balanced	45.40	41.59	33.80	71.33
AL-fixed	45.45	41.51	33.52	71.03
ModGap	45.13	40.89	32.93	70.30
ModGap-off	45.13	40.89	32.93	70.30
Frozen	33.50	32.62	29.72	63.42
CyCLIP-off	30.13	26.67	22.71	47.51

Table 2: Transductive cached-benchmark R@1 (%). Multi, Paper, and Full average bidirectional retrieval; Class. averages class-level R@1 on the four full-scale robustness datasets. CR denotes CLIP-Refine; AL denotes ledger-aware; "-off" denotes the official-cached variant.

Variant	R@1	R@5	R@10	Degr.	Δ
ALIGNLEDGER-CLIP	**76.50**	**90.62**	**94.03**	18.42	0.00
w/o rank sharpening	69.60	88.30	92.68	31.74	-6.89
w/o preservation	71.49	86.92	90.93	32.32	-5.01
w/o teacher stability	76.29	90.42	93.85	19.36	-0.21
w/o ledger gap	76.32	90.49	94.00	18.62	-0.18
Frozen	63.42	84.13	88.93	0.00	-13.08

Table 3: Transductive component ablation on the four full-scale classification slices. Scores and degraded-group rates (Degr.) are percentages. Δ is the R@1 difference from the complete method.

BALANCED branches reduce this rate further, at lower R@1. Thus routing changes the measured retrieval–degradation tradeoff, while fixed REFINE matches the strongest retrieval baseline. Additional group-supervised baselines and per-dataset results appear in Section A.6.

4.3 TRANSDUCTIVE CACHED-BENCHMARK AUDIT

Suite-level comparison. Table 2 reports differences from CR-off of 0.78, 0.08, and 0.57 R@1 points for the multi-seed, paper-scale, and full-scale audits. On the four robustness-classification datasets, adaptive group-positive ranking improves class-level R@1 by 1.26 points. The aggregate ordering also holds at R@5 and R@10, although CLIP-Refine performs better on several datasets and retrieval cutoffs. ModGap and ModGap-off implement the same cached objective and agree at the reported precision. At full scale, CR and CR-off also use the same distillation and random-feature objective with zero paired-contrastive weight.

Statistical comparison. The paired bootstrap in Table 6 gives dataset-level differences from CR-off. Mean differences are positive on the multi-seed and full-scale suites, but every confidence interval includes zero. The aggregate advantage is therefore not statistically significant across datasets. Competing methods perform better on ObjectNet, ChartQA, Winoground, and several higher-k classification slices.

4.4 ABLATION STUDIES

Component ablation. Rank sharpening and preservation account for most of the classification gain (Table 3). Removing rank sharpening reduces R@1 by 6.89 points. Removing preservation reduces it by 5.01 points and raises the degraded-group rate from 18.42% to 32.32%. Removing teacher stability or the centroid-gap term changes R@1 by at most 0.21 points on this subset. Full-scale SHARPEN uses uniform group weights, so these results characterize its ranking and regularization terms; they do not isolate the benefit of gap-derived weighting.

Multi-case qualitative analysis. Figure 3 extends the single-case mechanism check to stingray, Persian-cat, and lynx examples. Frozen CLIP and both adapted baselines retain plausible semantic or

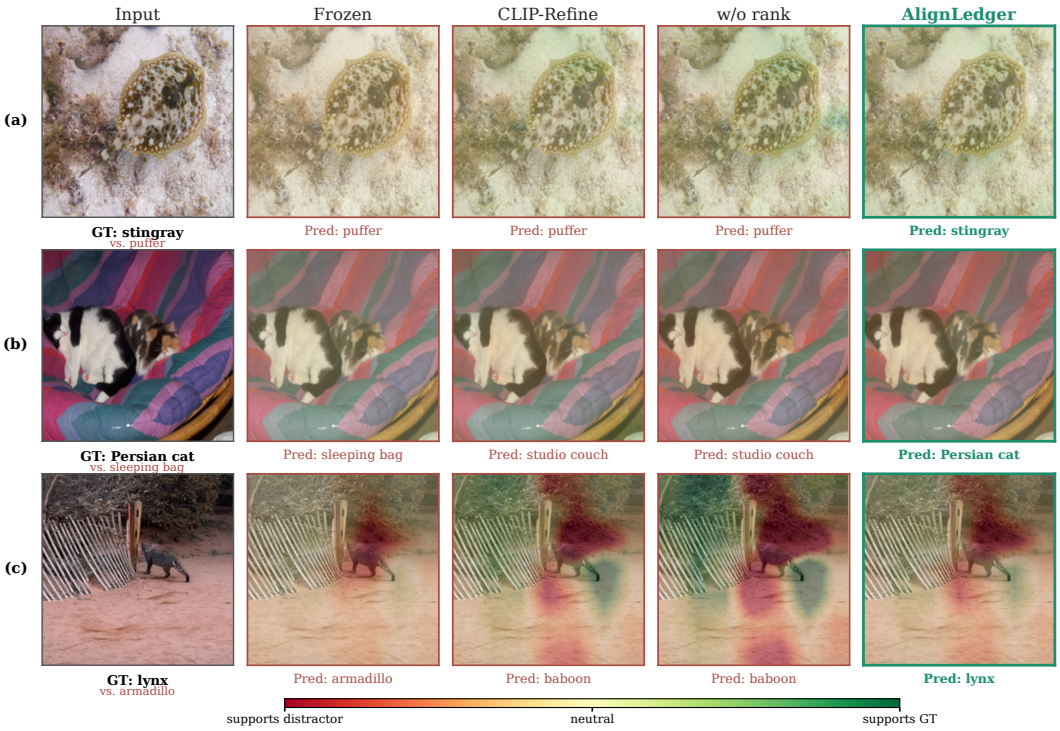


Figure 3: Multi-case qualitative comparison on ImageNet-A. Each row shows an input followed by patch-occlusion maps for Frozen CLIP, CLIP-Refine, the adaptive variant without rank sharpening, and ALIGNLEDGER-CLIP. Green and red borders indicate correct and incorrect top-1 predictions. The shared color scale measures the change in the ground-truth-minus-Frozen-distractor margin when masking one ViT patch: green supports the ground-truth class and red supports the distractor. For diagnostic clarity, the three distinct-class examples are selected by the fixed rule that ALIGNLEDGER-CLIP is correct while all three compared variants are incorrect; they are not used for model selection or aggregate claims.

contextual confounders; ALIGNLEDGER-CLIP restores the ground-truth class. The occlusion maps show spatially distributed decision changes, with no single common patch explaining them across methods. The appendix gives the detailed stingray–puffer analysis and its $+0.144$ corrected margin.

4.5 SCALE AND SEED COVERAGE

Every aggregate averages CLIP ViT-B/32 and SigLIP ViT-B/16. The paper-scale suite caps DocVQA and ChartQA at 1,024 examples each; the full-scale suite uses the complete datasets and streaming retrieval for Visual Genome. Only the multi-seed suite uses three seeds, so the paper- and full-scale audits serve as scale comparisons. The disjoint-partition experiment separately measures variation across three partition seeds on the paper-scale suite.

5 CONCLUSION

ALIGNLEDGER-CLIP connects group-wise measurement, cached-feature adaptation, and calibrated update selection for frozen vision–language embeddings. Evaluating retrieval together with group degradation exposes tradeoffs that an aggregate score alone misses. The disjoint-partition results show that calibration benefits both ALIGNLEDGER-CLIP and CLIP-Refine, while the classification ablations identify ranking and preservation as the largest contributors to adaptation gains. The framework provides a common way to inspect and control these updates; its applicability depends on meaningful group manifests and calibration records representative of the intended evaluation setting.

AI USE STATEMENT

Generative AI tools were used solely to assist with and polish the writing of this manuscript. The authors reviewed and revised all AI-assisted content and take full responsibility for the final manuscript.

REPRODUCIBILITY STATEMENT

The paper specifies the cached image–text feature protocol, semantic-group ledger, identity-initialized adapters, deterministic routing, calibration guard, datasets, backbones, and evaluation metrics. The appendix gives the method specification, complete Top-K matrices, strict-guard tables, and candidate settings used in the reported experiments. This repository contains the manuscript source, figures, tables, and references; it does not include the complete feature-extraction and training code, cached feature exports, checkpoints, or external model assets. Full replication requires those artifacts together with the public datasets and model versions named in the paper.

REFERENCES

Andrei Barbu, David Mayo, Julian Alverio, William Luo, Christopher Wang, Dan Gutfreund, Josh Tenenbaum, and Boris Katz. Objectnet: A large-scale bias-controlled dataset for pushing the limits of object recognition models. In *NeurIPS*, pp. 9448–9458, 2019.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pp. 248–255. Ieee, 2009.

Sedigheh Eslami and Gerard de Melo. Mitigate the gap: Investigating approaches for improving cross-modal alignment in CLIP. *CoRR*, abs/2406.17639, 2024.

Peng Gao, Shijie Geng, Renrui Zhang, Teli Ma, Rongyao Fang, Yongfeng Zhang, Hongsheng Li, and Yu Qiao. Clip-adapter: Better vision-language models with feature adapters. *Int. J. Comput. Vis.*, 132(2):581–595, 2024.

Shashank Goel, Hritik Bansal, Sumit Bhatia, Ryan Rossi, Vishwa Vinay, and Aditya Grover. Cyclip: Cyclic contrastive language-image pretraining. *Advances in Neural Information Processing Systems*, 35:6704–6719, 2022.

Eleonora Grassucci, Giordano Cicchetti, Emanuele Frasca, Aurelio Uncini, and Danilo Comminiello. Closing the modality gap aligns group-wise semantics. *arXiv preprint arXiv:2601.18525*, 2026.

Dan Hendrycks, Steven Basart, Norman Mu, Saurav Kadavath, Frank Wang, Evan Dorundo, Rahul Desai, Tyler Zhu, Samyak Parajuli, Mike Guo, et al. The many faces of robustness: A critical analysis of out-of-distribution generalization. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 8340–8349, 2021a.

Dan Hendrycks, Kevin Zhao, Steven Basart, Jacob Steinhardt, and Dawn Song. Natural adversarial examples. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 15262–15271, 2021b.

Cheng-Yu Hsieh, Jieyu Zhang, Zixian Ma, Aniruddha Kembhavi, and Ranjay Krishna. Sugarcrepe: Fixing hackable benchmarks for vision-language compositionality. *Advances in neural information processing systems*, 36:31096–31116, 2023.

Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In *International conference on machine learning*, pp. 4904–4916. PMLR, 2021.

Muhammad Uzair Khattak, Hanoona Abdul Rasheed, Muhammad Maaz, Salman H. Khan, and Fahad Shahbaz Khan. Maple: Multi-modal prompt learning. In *CVPR*, pp. 19113–19122. IEEE, 2023.

Pang Wei Koh, Shiori Sagawa, Henrik Marklund, Sang Michael Xie, Marvin Zhang, Akshay Balsubramani, Weihua Hu, Michihiro Yasunaga, Richard Lanas Phillips, Irena Gao, et al. Wilds: A benchmark of in-the-wild distribution shifts. In *International conference on machine learning*, pp. 5637–5664. PMLR, 2021.

Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, et al. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International journal of computer vision*, 123(1):32–73, 2017.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pp. 740–755. Springer, 2014.

Ahmed Masry, Jia Qing Tan, Shafiq Joty, Enamul Hoque, et al. Chartqa: A benchmark for question answering about charts with visual and logical reasoning. In *Findings of the association for computational linguistics: ACL 2022*, pp. 2263–2279, 2022.

Minesh Mathew, Dimosthenis Karatzas, and CV Jawahar. Docvqa: A dataset for vqa on document images. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pp. 2200–2209, 2021.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pp. 8748–8763. PmLR, 2021.

Shiori Sagawa, Aditi Raghunathan, Pang Wei Koh, and Percy Liang. An investigation of why overparameterization exacerbates spurious correlations. In *International Conference on Machine Learning*, pp. 8346–8356. PMLR, 2020.

Tristan Thrush, Ryan Jiang, Max Bartolo, Amanpreet Singh, Adina Williams, Douwe Kiela, and Candace Ross. Winoground: Probing vision and language models for visio-linguistic compositionality. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5238–5248, 2022.

Shin'ya Yamaguchi, Dewei Feng, Sekitoshi Kanai, Kazuki Adachi, and Daiki Chijiwa. Post-pre-training for modality alignment in vision-language foundation models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 4256–4266, 2025.

Chenyu You, Yifei Min, Weicheng Dai, Jasjeet S. Sekhon, Lawrence Staib, and James S. Duncan. Calibrating multi-modal representations: A pursuit of group robustness without annotations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 26140–26150, 2024.

Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid loss for language image pre-training. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 11975–11986, 2023.

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. Springer, 2022.

Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Conditional prompt learning for vision-language models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 16816–16825, 2022a.

Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Learning to prompt for vision-language models. *International journal of computer vision*, 130(9):2337–2348, 2022b.

Beier Zhu, Jiequan Cui, Hanwang Zhang, and Chi Zhang. Project-probe-aggregate: Efficient fine-tuning for group robustness. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 25487–25496, 2025.

A Appendix

A.1 R@1 Dataset Matrix and Aggregate Ranking

Table 4 expands the compact main-paper aggregate table into a dataset-level all-method R@1 matrix. Table 5 summarizes the same results as suite-level method rankings.

Multi-seed suite													
Method	Avg	IN-1K	IN-A	IN-R	ObjNet	COCO-O	COCO-A	COCO-R	VG-R	ChartQA	DocVQA	Wino	Sugar
AlignLedger-CLIP	**49.41**	69.35	**56.46**	85.43	**53.95**	**40.85**	**50.72**	**36.47**	**5.78**	**44.24**	**28.83**	**71.52**	**49.35**
CR-off	48.63	**72.72**	54.64	**85.79**	**53.95**	**40.85**	50.62	**36.47**	3.32	38.93	25.44	**71.52**	**49.35**
CR	32.17	**72.72**	54.64	**85.79**	**53.95**	16.68	28.17	20.02	0.61	3.63	2.68	20.32	26.80
AlignCLIP-off	46.20	68.13	54.20	84.75	53.02	34.70	47.85	33.33	3.27	43.12	25.52	61.61	44.93
AL-balanced	45.40	67.69	52.93	84.98	51.58	32.79	46.96	32.37	**5.78**	40.82	23.89	62.08	43.00
AL-fixed	45.45	67.63	53.64	84.64	49.64	33.11	47.42	32.47	4.83	40.67	25.13	63.72	42.55
ModGap	45.13	67.59	53.24	84.68	49.91	33.20	46.96	32.86	2.25	40.38	25.02	62.46	43.01
ModGap-off	45.13	67.59	53.24	84.68	49.91	33.20	46.96	32.86	2.25	40.38	25.02	62.46	43.01
Frozen	33.50	64.65	36.40	79.25	44.20	27.68	39.26	26.42	1.17	12.79	6.16	28.44	35.62
CyCLIP-off	30.13	46.17	30.89	79.49	40.15	16.46	36.25	24.09	1.27	5.71	10.64	39.95	30.44
Paper-scale suite													
Method	Avg	IN-1K	IN-A	IN-R	ObjNet	COCO-O	COCO-A	COCO-R	VG-R	ChartQA	DocVQA	Wino	Sugar
AlignLedger-CLIP	**45.81**	75.63	59.44	86.47	**58.93**	**33.88**	**46.70**	**33.29**	**5.41**	**23.23**	**23.03**	**71.52**	**32.18**
CR-off	45.73	**79.22**	**62.91**	**88.29**	**58.93**	**33.88**	**46.70**	**33.29**	5.27	13.52	**23.03**	**71.52**	**32.18**
CR	31.46	**79.22**	**62.91**	**88.29**	**58.93**	13.74	20.10	15.19	1.22	0.91	2.05	18.85	16.16
AlignCLIP-off	42.16	74.50	58.29	85.46	52.89	28.34	42.19	29.22	4.76	21.78	18.60	61.61	28.31
AL-balanced	41.59	73.68	57.70	85.28	52.49	26.72	41.59	27.57	5.16	21.71	18.21	62.08	26.86
AL-fixed	41.51	73.63	58.44	85.65	52.34	26.79	41.54	27.51	4.80	17.81	19.25	63.72	26.66
ModGap	40.89	73.09	57.63	85.08	51.93	26.92	41.57	27.47	3.74	15.48	18.04	62.46	27.27
ModGap-off	40.89	73.09	57.63	85.08	51.93	26.92	41.57	27.47	3.74	15.48	18.04	62.46	27.27
Frozen	32.62	69.30	41.34	81.33	47.40	24.56	36.33	24.23	3.15	5.22	5.92	28.44	24.21
CyCLIP-off	26.67	49.89	24.85	75.88	39.32	13.16	30.62	20.06	2.39	0.78	3.09	39.95	20.06
Full-scale suite													
Method	Avg	IN-1K	IN-A	IN-R	ObjNet	COCO-O	COCO-A	COCO-R	VG-R	ChartQA	DocVQA	Wino	Sugar
AlignLedger-CLIP	**36.81**	**85.25**	**64.96**	**90.48**	65.30	**22.53**	**31.08**	**11.67**	**0.48**	4.65	**6.37**	38.47	**20.46**
CR-off	36.24	82.85	59.04	90.18	**68.90**	**22.53**	**31.08**	**11.67**	0.38	3.48	5.80	38.47	**20.46**
CR	36.24	82.85	59.04	90.18	**68.90**	**22.53**	**31.08**	**11.67**	0.38	3.48	5.80	38.47	**20.46**
AlignCLIP-off	34.01	79.51	55.78	88.03	62.86	18.10	28.33	10.23	0.28	**5.04**	5.07	36.76	18.12
AL-balanced	33.80	78.32	56.69	87.68	62.62	17.19	27.25	10.14	0.35	3.87	4.54	40.23	16.69
AL-fixed	33.52	78.20	55.95	87.73	62.25	16.76	26.79	9.76	0.26	3.12	4.33	**40.40**	16.66
ModGap	32.93	77.92	53.85	87.44	62.01	16.35	26.71	9.21	0.15	3.26	2.97	38.96	16.38
ModGap-off	32.93	77.92	53.85	87.44	62.01	16.35	26.71	9.21	0.15	3.26	2.97	38.96	16.38
Frozen	29.72	71.92	44.18	84.36	53.21	14.60	26.57	8.69	**0.48**	4.06	2.78	28.44	17.39
CyCLIP-off	22.71	48.84	26.54	73.67	41.00	12.12	21.01	7.35	0.20	0.64	0.99	25.71	14.44

Table 4: Dataset-level all-method matrix under the transductive audit protocol. Values are average bidirectional R@1 in percentage units. The **Avg** column averages all dataset–backbone slices; the remaining columns average both backbones within each dataset. The multi-seed suite averages seeds 42, 43, and 44; the paper-scale and full-scale suites use a single seed. Bold marks the best method within each suite column. Ours denotes AlignLedger-CLIP.

A.2 Transductive Paired Bootstrap

Table 6 supplements the aggregate comparison in Table 2. All three confidence intervals include zero.

A.3 Classification Top-K Audit

Table 7 reports class-level R@1/R@5/R@10 for the four robustness-classification datasets.

Method	Med. R@1	Paper R@1	Full R@1
ALIGNLEDGER-CLIP	**49.41**	**45.81**	**36.81**
CLIP-Refine (official cached)	48.63	45.73	36.24
CLIP-Refine	32.17	31.46	36.24
AlignCLIP (official cached)	46.20	42.16	34.01
Balanced ledger-aware	45.40	41.59	33.80
Ledger-aware	45.45	41.51	33.52
ModGap	45.13	40.89	32.93
ModGap (official cached)	45.13	40.89	32.93
Frozen	33.50	32.62	29.72
CyCLIP (official cached)	30.13	26.67	22.71

Table 5: Method ranking under the transductive audit protocol. Values are average bidirectional R@1 percentages over 24 dataset–backbone slices.

Suite	Δ R@1	95% CI	W/T/L
Multi-seed	+0.78	$[-0.41, 2.00]$	4/6/2
Paper-scale	+0.08	$[-1.46, 2.14]$	2/7/3
Full-scale	+0.57	$[-0.53, 1.79]$	5/6/1

Table 6: Paired bootstrap for the transductive audit against CR-off. Benchmark slices are resampled after averaging the two backbones. W/T/L denotes wins, ties, and losses using a 0.10-point tie threshold.

Method	Med. R@1	Med. R@5	Med. R@10	Paper R@1	Paper R@5	Paper R@10	Full R@1	Full R@5	Full R@10
ALIGNLEDGER-CLIP	66.30	87.10	91.93	70.12	89.43	93.37	**76.50**	90.62	94.03
CR-off	**66.77**	**87.41**	**92.19**	**72.34**	**90.57**	**94.36**	75.24	**91.41**	**94.81**
CR	**66.77**	**87.41**	**92.19**	**72.34**	**90.57**	**94.36**	75.24	**91.41**	**94.81**
AlignCLIP-off	65.03	86.12	91.38	67.78	87.62	92.17	71.55	89.40	93.42
AL-balanced	64.29	85.94	91.26	67.29	87.57	92.09	71.33	88.98	93.16
AL-fixed	63.89	85.83	91.24	67.52	87.58	92.19	71.03	89.19	93.13
ModGap	63.86	85.62	91.14	66.94	87.19	92.02	70.30	88.92	92.77
ModGap-off	63.86	85.62	91.14	66.94	87.19	92.02	70.30	88.92	92.77
Frozen	56.13	79.78	86.67	59.84	82.25	88.10	63.42	84.13	88.93
CyCLIP-off	49.18	73.65	81.23	47.49	72.80	79.61	47.51	74.17	80.74

Table 7: Transductive ImageNet/ObjectNet class-level top-k audit. Values average ImageNet-1K, ImageNet-A, ImageNet-R, and ObjectNet over both backbones; all entries are percentages.

Family	Benchmark	Samples	Groups	Sampling rule
COCO grounding	COCO Objects	5,942	80	object categories, max 80 per category
COCO grounding	COCO Attributes	2,352	11	color keywords, max 250 per group
COCO grounding	COCO Relations	3,000	12	relation keywords, max 250 per group
Robustness	ImageNet-A	3,996	200	max 25 per class
Robustness	ImageNet-R	5,000	200	max 25 per class
Robustness	ObjectNet	4,695	313	max 15 per class
Relations	Visual Genome Relations	6,000	1,469	max 30 per predicate
Charts	ChartQA validation	1,024	2	validation parquet, max 1,024 rows
Documents	DocVQA validation	1,024	9	validation parquet, max 1,024 rows
Robustness	ImageNet-1K validation	5,000	1,000	labeled validation classes, 5 images per class
Compositional	Winoground	800	70	full correct-pair manifest
Compositional	SugarCrepe	3,845	7	data split, max 600 per file
Total		42,678		

Table 8: Composition of the paper-scale validation suite.

A.4 Paper-Scale Suite Composition

A.5 Qualitative Diagnostics

The case study and patch-occlusion figures diagnose individual predictions. They replay the cached-feature adapters from the full-scale audit and are excluded from model selection.

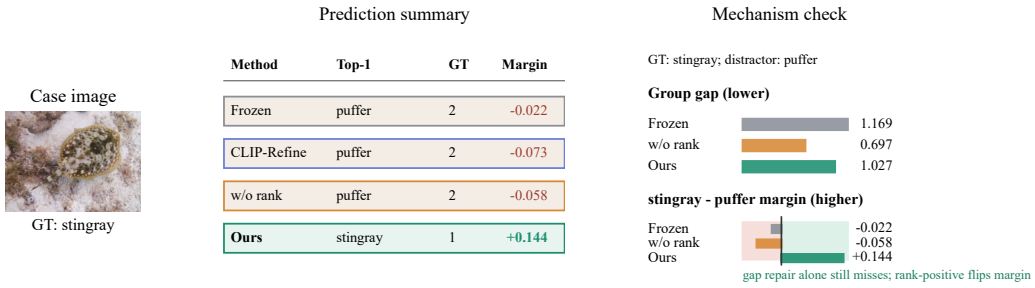
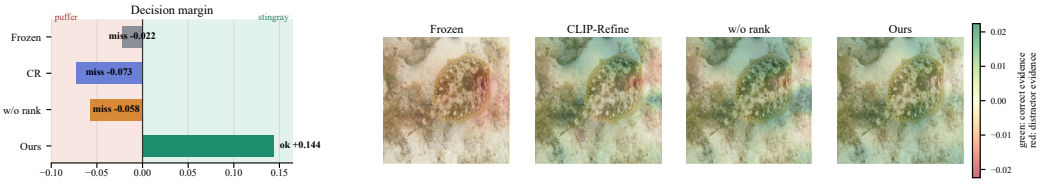


Figure 4: ImageNet-A full-scale case-study visualization. The compact prediction table reports each method's top-1 class, the ground-truth rank among the shown candidates, and the local *stingray*-versus-*puffer* margin. The frozen backbone, CLIP-Refine, and the adaptive variant without rank-sharpening all rank *puffer* first, whereas AlignLedger ranks the correct class first. The mechanism panel compares the image–text ledger gap for the ground-truth group and the same local decision margin. The no-rank variant can reduce the group gap while still leaving the local class margin negative; AlignLedger's rank-positive branch flips that margin.



(a) Decision margin after adaptation.

(b) Patch-level occlusion sensitivity maps.

Figure 5: Patch-occlusion sensitivity visualization for the same ImageNet-A example as fig. 4. Figure 5a defines the reading rule: negative margin means the distractor *puffer* scores higher, while positive margin means the correct *stingray* scores higher. Only AlignLedger crosses the zero line into the correct region. Figure 5b masks one ViT patch at a time; green patches support the correct margin, and red patches support the distractor margin. The diagnostic uses the frozen CLIP image encoder with method-specific embedding adapters to inspect local decisions. It is excluded from the training objective.

A.6 Strict Held-Out Guard-Fair Audit

The strict held-out audit uses the same adaptation, calibration, and evaluation partitions across 12 paper-scale benchmark slices, two backbones, and split seeds 42, 43, and 44. The tables report method, family, or dataset averages; the committed exporter also retains per-seed and per-backbone rows.

Method	R@1	Δ	Damaged	Pass	Fallback
CR-off	35.37	-0.66	43.44	–	–
CR-off + calibrated scaling	38.53	+2.50	21.65	88.89	11.11
CR-off + reference reversion	36.09	+0.06	4.75	22.22	77.78
Prototype	25.69	-10.34	25.79	–	–
ALIGNLEDGER-CLIP uncalibrated	34.71	-1.32	43.40	–	–
ALIGNLEDGER-CLIP + calibrated scaling	38.27	+2.24	20.01	87.50	12.50
Balanced + calibrated scaling	37.37	+1.34	16.57	76.39	23.61
Ledger-aware + calibrated scaling	37.59	+1.55	15.47	77.78	22.22
Teacher-refined + calibrated scaling	38.53	+2.50	21.65	88.89	11.11
Frozen	36.03	0.00	0.00	–	–
Rank + preservation	30.60	-5.43	52.05	–	–
Rank-only	30.34	-5.69	53.25	–	–
Linear probe	28.60	-7.43	60.79	–	–

Table 9: Strict held-out fair-method summary. R@1, damaged groups, guard pass rate, and fallback rate are percentages after averaging over datasets, both backbones, and three split seeds. Δ is measured relative to the frozen representation on the same evaluation partition.

Comparison	Baseline	Δ	Dataset CI	Slice CI	W/T/L	Worst LOO
Ours vs CR guard	CR-off + calibrated scaling	-0.26	[-0.51, -0.07]	[-0.52, -0.04]	4/13/7	-0.29 w/o COCO-Attr
Ours vs CR fallback	CR-off + reference reversion	+2.17	[0.80, 4.03]	[1.03, 3.53]	17/3/4	+1.42 w/o IN-A
Ours vs no guard	ALIGNLEDGER-CLIP uncalibrated	+3.55	[1.87, 5.39]	[2.24, 4.92]	17/4/3	+2.86 w/o Wino
Router vs fixed ledger	Ledger-aware + calibrated scaling	+0.68	[0.19, 1.20]	[0.23, 1.14]	17/2/5	+0.54 w/o COCO-Obj
Router vs fixed balanced	Balanced + calibrated scaling	+0.89	[0.37, 1.43]	[0.38, 1.40]	17/2/5	+0.75 w/o IN-A
Router vs fixed refine	Teacher-refined + calibrated scaling	-0.26	[-0.51, -0.07]	[-0.52, -0.04]	4/13/7	-0.29 w/o COCO-Attr

Table 10: Paired strict held-out statistics for guarded ledger comparisons. Deltas are R@1 percentage points for ALIGNLEDGER-CLIP + calibrated scaling minus the listed baseline, computed over 24 dataset–backbone slices after averaging split seeds. Bootstrap intervals resample datasets or slices.

Method	Class.	COCO	VG-R	Doc/Chart	Wino/Sugar	Avg	Damaged
CR-off	61.25	34.12	4.00	9.45	27.05	35.37	43.44
CR-off + calibrated scaling	63.71	38.06	3.99	9.91	34.77	38.53	21.65
CR-off + reference reversion	58.57	37.05	3.99	7.39	34.47	36.09	4.75
Prototype	66.33	1.27	5.23	3.20	13.77	25.69	25.79
ALIGNLEDGER-CLIP uncalibrated	59.27	34.15	3.72	9.58	27.08	34.71	43.40
ALIGNLEDGER-CLIP + calibrated scaling	63.35	38.09	3.67	9.18	34.75	38.27	20.01
Balanced + calibrated scaling	61.78	37.11	3.37	8.84	34.49	37.37	16.57
Ledger-aware + calibrated scaling	62.13	37.16	3.39	9.39	34.42	37.59	15.47
Teacher-refined + calibrated scaling	63.71	38.06	3.99	9.91	34.77	38.53	21.65
Frozen	58.38	37.05	3.81	7.48	34.46	36.03	0.00
Rank + preservation	58.88	25.80	3.56	1.91	23.48	30.60	52.05
Rank-only	58.71	25.55	3.50	1.32	23.25	30.34	53.25
Linear probe	54.35	23.30	3.39	2.63	23.61	28.60	60.79

Table 11: Cross-family strict held-out audit. Values are average bidirectional R@1 percentages; damaged is the average fraction of groups whose bidirectional top-1 group-hit consistency drops relative to frozen.

Dataset	Frozen	CR-off	CR-G	AL	AL-G	Fix-L	Fix-B	Fix-R
ChartQA	7.36	10.34	11.88	10.78	10.47	9.89	9.04	11.88
COCO-Attr	46.34	40.08	46.72	40.07	46.82	46.74	46.56	46.72
COCO-Obj	31.58	32.28	33.86	32.25	33.87	31.60	31.60	33.86
COCO-Rel	33.24	30.01	33.59	30.12	33.57	33.14	33.16	33.59
DocVQA	7.60	8.55	7.93	8.37	7.90	8.90	8.63	7.93
IN-1K	67.08	66.94	68.45	63.59	67.98	66.94	66.29	68.45
IN-A	40.61	48.94	51.50	46.26	51.01	48.77	48.56	51.50
IN-R	80.52	83.04	84.94	81.15	84.49	83.67	83.62	84.94
ObjectNet	45.33	46.10	49.94	46.10	49.94	49.14	48.66	49.94
SugarCrepe	33.04	29.98	34.26	30.04	34.22	33.23	33.23	34.26
VG-R	3.81	4.00	3.99	3.72	3.67	3.39	3.37	3.99
Winoground	35.88	24.13	35.27	24.13	35.27	35.61	35.75	35.27

Table 12: Per-dataset strict held-out R@1 averages for selected methods. Each row averages six runs: two backbones times three split seeds.

CR-G = CR-off + guard; AL = unguarded adaptive ledger; AL-G = guarded adaptive ledger; Fix-L/Fix-B/Fix-R are fixed ledger-aware, balanced, and teacher-refine guarded branches.

A.7 Auxiliary Adaptation Objectives

The branch objectives in Equation (11) use the following stability terms, with batch notation as in Section 3.3.

The balanced branch also uses within-modality discrimination:

$$\begin{aligned} \ell^M_{\mathrm{in},i} &= -\log \frac{\exp(\widehat{z}_i^{M\top}\widehat{z}_i^M/\tau)}{\sum_{j\in\mathcal{B}}\exp(\widehat{z}_i^{M\top}\widehat{z}_j^M/\tau)}, \\ \mathcal{L}_{\mathrm{in}} &= \frac{1}{2|\mathcal{B}|}\sum_{i\in\mathcal{B}} w_{g_i}^{\mathrm{close}}\left(\ell^I_{\mathrm{in},i}+\ell^T_{\mathrm{in},i}\right), \end{aligned} \tag{19}$$

where $\widehat{z}_i^I = \widehat{x}_i$ and $\widehat{z}_i^T = \widehat{y}_i$.

To preserve cross-modal rankings, we distill the frozen similarities without a separate teacher model. For image query i,

$$\begin{aligned} q_i^I &= \beta e_i + (1-\beta)\,\mathrm{softmax}(s^0_{i:}), \quad p_i^I = \mathrm{softmax}(\widehat{S}^{\mathcal{B}}_{i:}), \\ q_i^T &= \beta e_i + (1-\beta)\,\mathrm{softmax}(s^0_{:i}), \quad p_i^T = \mathrm{softmax}(\widehat{S}^{\mathcal{B}}_{:i}), \end{aligned} \tag{20}$$

where e_i is the paired one-hot target and s^0 is computed on the same batch. The frozen-reference stability loss is

$$\mathcal{L}_{\mathrm{teach}} = \frac{1}{2|\mathcal{B}|}\sum_{i\in\mathcal{B}}\left[\mathrm{KL}(q_i^I\|p_i^I)+\mathrm{KL}(q_i^T\|p_i^T)\right]. \tag{21}$$

Here $\mathcal{L}_{\mathrm{rand}}$ is the random-feature regularizer used by the refinement branch:

$$\mathcal{L}_{\mathrm{rand}} = \frac{1}{|\mathcal{B}|d}\sum_{i\in\mathcal{B}}\left(\|\widehat{x}_i - r_i^I\|_2^2 + \|\widehat{y}_i - r_i^T\|_2^2\right), \tag{22}$$

where r_i^I, r_i^T are normalized random anchors generated under the fixed run seed and $\lambda_{\mathrm{ref}} = \lambda_{\mathrm{keep}}$ in the implementation. IDENTITY performs no optimization. The three stability mechanisms are distinct: $\mathcal{L}_{\mathrm{keep}}$ limits pointwise displacement, $\mathcal{L}_{\mathrm{in}}$ maintains within-modality discrimination, and $\mathcal{L}_{\mathrm{teach}}$ preserves the frozen cross-modal score distribution.

A.8 Complete Adaptation Procedure

Algorithm 1 summarizes the fit, calibration, and evaluation steps. The routing map and coefficient profiles follow below.

Algorithm 1: Guarded ledger-aware adaptation

Input: Fit, guard, and evaluation roles; manifest metadata ϕ
Output: Guarded embeddings and before/after ledgers
Build $\{\mathfrak{L}_g^0\}$ and frozen weights on the fit role
$b^\star \leftarrow \mathcal{R}(\phi)$ before optimization
if $b^\star =$ IDENTITY **then**
  set $W_I = W_T = I_d$
else
  optimize $\mathcal{L}_{b^\star}$ on fit batches
$\alpha^\star \leftarrow 0$
for $\alpha \in \mathcal{A}$ in descending order **do**
  re-ledger $z(\alpha)$ on the guard role
  if both recall and group-damage constraints pass **then**
    $\alpha^\star \leftarrow \alpha$; **break**
Apply $z(\alpha^\star)$ to the evaluation role and re-ledger

A.9 Deterministic Router and Guard Specification

Table 13 expands the metadata map used by the final adaptive method. Routing is applied once to a complete dataset–backbone slice. Each cell lists the CLIP/SigLIP branch, and the first applicable dataset row is used. The abbreviations are Aware (A), Balanced (B), frozen-reference Refine (R), group-positive Sharpen (S), and Identity (I). "IN" denotes ImageNet.

Dataset family	Small	Medium	Paper	Full
IN-1K / IN-R	A/A	A/A	A/A	S/S
IN-A	A/B	A/B	A/B	S/S
ObjectNet	A/A	R/R	R/R	R/R
COCO object	B/A	B/R	R/R	R/R
COCO attribute	B/B	R/B	R/R	R/R
COCO relation	B/B	R/R	R/R	R/R
Visual Genome relation	B/B	B/B	R/B	I/I
ChartQA	A/A	A/A	A/B	R/I
DocVQA	A/A	B/A	R/R	R/A
Winoground / SugarCrepe	A/A	R/R	R/R	R/R

Table 13: Complete branch map for the evaluated dataset families. Each entry is CLIP/SigLIP. Scale is read from the manifest name and otherwise inferred from pair count (≤ 256: small; ≤ 4096: medium; larger: full). The map is fixed before optimization and contains no calibration or evaluation metric.

The router starts from

$$(\lambda_{\text{gap}}, \lambda_{\text{keep}}, \lambda_{\text{in}}, \lambda_{\text{teach}}) = (0.1, 0.1, 0.25, 0.1).$$

Table 14 lists the nondefault coefficient profiles. Coefficients for terms absent from the selected branch are inactive. The metadata lookup is fixed during each run; it does not search over branches using guard or evaluation scores.

Slice / backbone / scale	$(\lambda_{\text{gap}}, \lambda_{\text{keep}}, \lambda_{\text{in}}, \lambda_{\text{teach}})$	Other overrides
Classification except ObjectNet / both / full	$(0.05, 0.2, 0.25, 0.2)$	2 epochs; unit group weights
COCO attribute / SigLIP / small, medium	$(0.1, 0.1, 0.25, 0.03)$	–
ChartQA / CLIP / medium, paper	$(0.1, 0.2, 0.25, 0.1)$	–
ChartQA / SigLIP / medium	$(0.05, 0.2, 0.25, 0.1)$	–
ChartQA / SigLIP / paper	$(1.0, 0.1, 0.5, 0.1)$	–
DocVQA / CLIP / small	$(0.3, 0.1, 0.25, 0.1)$	–
DocVQA / SigLIP / small	$(1.0, 0.1, 0.25, 0.1)$	LR 10^{-2}; 5 epochs
DocVQA / CLIP / medium	$(0.1, 0.1, 0.1, 0.03)$	–
DocVQA / SigLIP / medium, full	$(0.1, 0.2, 0.25, 0.1)$	–
COCO relation / SigLIP / small	$(0.1, 0.1, 0.1, 0.1)$	–
COCO object / CLIP / small	$(0.03, 1.0, 0.05, 0.3)$	LR 5×10^{-4}; 1 epoch
COCO object / CLIP / medium	$(0.01, 1.0, 0.02, 0.3)$	LR 10^{-6}; 1 epoch; $\mathcal{A} = \{0.05\}$
VG relation / CLIP / small	$(0.3, 0.1, 0.25, 0.1)$	–
VG relation / both / medium; SigLIP / paper	$(1.0, 0.1, 0.5, 0.1)$	–

Table 14: Coefficient overrides for the cached-feature router. Unlisted entries use the default profile above. The strict paper-scale runs use 10 epochs and learning rate 0.003, subject to these overrides. The medium COCO-object profile is outside the strict paper-scale comparison.

For the strict audit, the guard tests

$$\mathcal{A} = \{1, 0.5, 0.25, 0.1, 0.05, 0.01\}$$

in descending order. A scale passes only if both guard-set Recall@1 changes are at least -0.05 on the $[0, 1]$ scale (five percentage points) and the guard-set damaged-group fraction is at most 0.20. Guard recall is macro-averaged over groups; group-hit consistency uses $k = 1$. The first passing scale is used on the evaluation role. When no positive scale passes, the adapter is discarded and the frozen representation is returned. The guard thresholds and candidate order are shared by the guarded ALIGNLEDGER-CLIP and guarded-baseline comparisons. Reference reversion uses the same records and thresholds with $\mathcal{A} = \{1\}$, returning frozen when the full update fails.

A.10 ALL-METHOD TOP-K DATASET MATRICES

Tables 15 to 20 report the full all-method by all-dataset Top-5 and Top-10 matrices. Values are percentages and use the same final evaluation protocol as the main transductive audit tables: class-level ImageNet/ObjectNet evaluation, stratified ImageNet-1K capped sampling, full-scale retrieval rows, and the unweighted rank-sharpen full-classification branch. Each table includes an equal-weight **Avg** column over the 12 benchmark-slice columns.

Method	Avg	IN-1K	IN-A	IN-R	ObjNet	COCO-O	COCO-A	COCO-R	VG-R	ChartQA	DocVQA	Wino	Sugar
ALIGNLEDGER-CLIP	**83.87**	90.09	**82.67**	95.69	**79.97**	**85.54**	95.92	**87.93**	**19.84**	**91.36**	82.87	**99.38**	**95.21**
CR-off	83.17	**92.52**	81.28	**95.88**	**79.97**	**85.54**	**96.02**	**87.93**	13.27	85.56	**85.51**	**99.38**	**95.21**
CR	54.42	**92.52**	81.28	**95.88**	**79.97**	41.86	65.66	55.43	2.60	9.49	10.16	51.38	66.86
AlignCLIP-off	80.86	89.35	81.18	95.38	78.58	76.79	93.69	83.64	13.00	90.33	81.13	96.31	90.97
AL-balanced	80.42	88.95	81.55	95.40	77.86	74.37	92.64	81.59	**19.84**	86.33	80.63	96.46	89.40
AL-fixed	80.14	89.05	82.01	95.55	76.69	74.17	93.08	82.06	16.38	85.55	80.16	97.75	89.22
ModGap	79.42	88.92	81.42	95.50	76.66	74.47	92.78	81.76	9.13	86.82	79.24	96.86	89.45
ModGap-off	79.42	88.92	81.42	95.50	76.66	74.47	92.78	81.76	9.13	86.82	79.24	96.86	89.45
Frozen	60.49	87.90	67.86	92.76	70.61	60.28	81.69	69.04	3.95	27.64	21.48	64.00	78.68
CyCLIP-off	57.00	74.14	60.23	93.66	66.55	39.49	77.93	62.66	4.35	15.62	35.91	81.59	71.93

Table 15: Top-5 all-method dataset matrix for the multi-seed suite. Values are average bidirectional R@5 percentages over the same backbone/seed aggregation used in the main table. The Avg column is the equal-weight mean over the 12 dataset columns. Bold marks the best method within each column.

Method	Avg	IN-1K	IN-A	IN-R	ObjNet	COCO-O	COCO-A	COCO-R	VG-R	ChartQA	DocVQA	Wino	Sugar
ALIGNLEDGER-CLIP	**89.84**	93.91	**89.78**	97.06	**86.98**	**93.15**	99.60	**96.32**	**32.68**	**95.31**	94.62	**99.91**	**98.72**
CR-off	89.17	**95.73**	88.66	**97.39**	**86.98**	**93.15**	**99.66**	**96.32**	23.76	91.75	**98.06**	**99.91**	**98.72**
CR	62.40	**95.73**	88.66	**97.39**	**86.98**	54.74	78.63	70.76	4.13	12.73	16.27	62.42	80.35
AlignCLIP-off	87.55	93.40	89.16	96.92	86.04	86.89	98.54	94.10	22.65	95.02	93.55	97.93	96.35
AL-balanced	87.70	93.35	89.18	97.05	85.45	84.87	98.31	92.83	**32.68**	92.14	92.82	98.12	95.56
AL-fixed	87.11	93.25	89.38	97.12	85.20	84.93	98.20	92.78	26.47	90.97	92.86	98.73	95.43
ModGap	86.34	93.15	89.19	97.08	85.14	85.11	98.15	92.56	16.01	92.19	93.84	98.19	95.50
ModGap-off	86.34	93.15	89.19	97.08	85.14	85.11	98.15	92.56	16.01	92.19	93.84	98.19	95.50
Frozen	68.60	92.41	79.50	95.55	79.20	72.11	91.64	82.76	7.24	31.98	28.54	73.89	88.42
CyCLIP-off	66.09	82.21	71.34	95.74	75.64	49.71	89.25	77.39	7.70	20.80	49.75	89.31	84.18

Table 16: Top-10 all-method dataset matrix for the multi-seed suite. Values are average bidirectional R@10 percentages over the same backbone/seed aggregation used in the main table. The Avg column is the equal-weight mean over the 12 dataset columns. Bold marks the best method within each column.

Method	Avg	IN-1K	IN-A	IN-R	ObjNet	COCO-O	COCO-A	COCO-R	VG-R	ChartQA	DocVQA	Wino	Sugar
ALIGNLEDGER-CLIP	**79.03**	93.36	85.61	95.95	**82.80**	**76.78**	**94.26**	**81.88**	**18.15**	**65.83**	**71.89**	**99.38**	**82.41**
CR-off	77.18	**95.31**	**87.35**	**96.81**	**82.80**	**76.78**	**94.26**	**81.88**	17.27	40.08	**71.89**	**99.38**	**82.41**
CR	50.39	**95.31**	**87.35**	**96.81**	**82.80**	35.03	51.62	45.10	3.19	3.24	7.48	47.90	48.81
AlignCLIP-off	74.78	92.83	84.02	95.60	78.04	66.68	89.67	74.94	16.15	63.53	64.57	96.31	75.03
AL-balanced	73.90	92.40	84.07	95.49	78.32	63.72	88.57	73.03	17.24	63.23	61.54	96.46	72.76
AL-fixed	73.06	92.24	84.16	95.62	78.32	63.92	88.42	72.85	16.10	49.99	64.04	97.75	73.28
ModGap	71.73	91.94	83.14	95.52	78.16	64.12	88.07	72.96	13.20	46.76	56.78	96.86	73.25
ModGap-off	71.73	91.94	83.14	95.52	78.16	64.12	88.07	72.96	13.20	46.76	56.78	96.86	73.25
Frozen	58.13	90.25	72.28	93.76	72.71	55.31	78.38	64.30	9.51	14.43	17.44	64.00	65.15
CyCLIP-off	50.77	81.14	52.87	91.11	66.09	32.67	70.60	54.77	7.50	3.41	12.40	81.59	55.04

Table 17: Top-5 all-method dataset matrix for the paper-scale suite. Values are average bidirectional R@5 percentages over the same backbone/seed aggregation used in the main table. The Avg column is the equal-weight mean over the 12 dataset columns. Bold marks the best method within each column.

Method	Avg	IN-1K	IN-A	IN-R	ObjNet	COCO-O	COCO-A	COCO-R	VG-R	ChartQA	DocVQA	Wino	Sugar
ALIGNLEDGER-CLIP	**86.43**	96.12	91.38	97.39	**88.60**	**86.77**	**98.65**	**92.07**	**27.83**	**77.73**	**86.87**	**99.91**	**93.84**
CR-off	84.53	**97.58**	**93.15**	**98.11**	**88.60**	**86.77**	**98.65**	**92.07**	26.39	52.46	**86.87**	**99.91**	**93.84**
CR	57.87	**97.58**	**93.15**	**98.11**	**88.60**	47.05	65.53	60.60	4.50	4.28	11.73	57.40	65.88
AlignCLIP-off	83.11	95.76	90.54	97.23	85.13	78.05	96.34	87.33	24.99	76.48	78.70	97.93	88.86
AL-balanced	82.41	95.62	90.51	97.18	85.04	75.48	95.80	85.83	26.55	74.18	77.67	98.12	86.91
AL-fixed	81.48	95.31	90.48	97.20	85.77	75.62	95.70	85.88	24.79	62.20	79.09	98.73	86.94
ModGap	80.31	95.23	90.03	97.12	85.69	75.74	95.46	85.86	21.03	58.52	73.78	98.19	87.07
ModGap-off	80.31	95.23	90.03	97.12	85.69	75.74	95.46	85.86	21.03	58.52	73.78	98.19	87.07
Frozen	66.36	93.99	81.85	95.93	80.62	66.67	89.05	78.13	14.46	18.59	23.27	73.89	79.84
CyCLIP-off	58.87	87.91	62.16	93.53	74.86	41.48	83.49	68.35	11.00	5.66	17.65	89.31	71.07

Table 18: Top-10 all-method dataset matrix for the paper-scale suite. Values are average bidirectional R@10 percentages over the same backbone/seed aggregation used in the main table. The Avg column is the equal-weight mean over the 12 dataset columns. Bold marks the best method within each column.

Method	Avg	IN-1K	IN-A	IN-R	ObjNet	COCO-O	COCO-A	COCO-R	VG-R	ChartQA	DocVQA	Wino	Sugar
ALIGNLEDGER-CLIP	**61.09**	**96.65**	84.12	97.07	84.63	**70.04**	**78.66**	**49.05**	**1.23**	11.61	**20.55**	**81.62**	**57.78**
CR-off	61.00	96.60	**84.71**	**97.42**	**86.89**	**70.04**	**78.66**	**49.05**	1.07	9.02	19.14	**81.62**	**57.78**
CR	61.00	96.60	**84.71**	**97.42**	**86.89**	**70.04**	**78.66**	**49.05**	1.07	9.02	19.14	**81.62**	**57.78**
AlignCLIP-off	57.87	95.44	80.78	96.36	85.03	59.56	72.87	42.34	0.84	**13.18**	17.71	78.04	52.29
AL-balanced	57.11	95.09	80.67	96.22	83.93	56.24	71.09	41.55	0.97	11.46	16.72	80.79	50.65
AL-fixed	56.50	94.92	81.60	96.15	84.07	54.60	69.99	40.06	0.78	9.28	15.59	81.18	49.75
ModGap	55.34	94.75	81.06	95.99	83.90	52.85	69.41	37.45	0.49	9.80	10.38	78.80	49.15
ModGap-off	55.34	94.75	81.06	95.99	83.90	52.85	69.41	37.45	0.49	9.80	10.38	78.80	49.15
Frozen	51.10	91.48	74.88	94.46	75.68	42.90	67.72	34.71	**1.23**	10.18	9.07	64.00	46.93
CyCLIP-off	44.66	82.66	49.48	92.32	72.21	38.00	56.94	30.16	0.57	1.93	4.28	65.77	41.58

Table 19: Top-5 all-method dataset matrix for the full-scale suite. Values are average bidirectional R@5 percentages over the same backbone/seed aggregation used in the main table. The Avg column is the equal-weight mean over the 12 dataset columns. Bold marks the best method within each column.

Method	Avg	IN-1K	IN-A	IN-R	ObjNet	COCO-O	COCO-A	COCO-R	VG-R	ChartQA	DocVQA	Wino	Sugar
ALIGNLEDGER-CLIP	**69.64**	98.15	90.27	98.28	89.43	**84.74**	**90.26**	**68.24**	**1.81**	15.21	**28.75**	**88.50**	**82.08**
CR-off	69.45	**98.23**	**91.37**	**98.49**	**91.16**	**84.74**	**90.26**	**68.24**	1.73	12.10	26.54	**88.50**	**82.08**
CR	69.45	**98.23**	**91.37**	**98.49**	**91.16**	**84.74**	**90.26**	**68.24**	1.73	12.10	26.54	**88.50**	**82.08**
AlignCLIP-off	66.56	97.49	88.45	97.85	89.88	74.72	85.14	60.38	1.35	**17.97**	24.53	84.89	76.06
AL-balanced	65.84	97.21	88.52	97.78	89.14	71.20	84.18	59.71	1.48	16.24	23.85	87.08	73.67
AL-fixed	64.92	97.07	88.62	97.70	89.13	68.96	83.18	57.55	1.20	13.20	22.20	87.15	73.14
ModGap	63.50	96.96	87.64	97.57	88.90	67.62	82.27	54.37	0.78	13.98	14.85	84.60	72.48
ModGap-off	63.50	96.96	87.64	97.57	88.90	67.62	82.27	54.37	0.78	13.98	14.85	84.60	72.48
Frozen	59.01	94.52	82.44	96.47	82.27	54.62	80.80	48.86	**1.81**	12.79	12.77	73.89	66.83
CyCLIP-off	52.89	89.09	59.55	95.33	79.00	49.81	71.16	44.08	0.88	3.24	6.19	75.40	60.93

Table 20: Top-10 all-method dataset matrix for the full-scale suite. Values are average bidirectional R@10 percentages over the same backbone/seed aggregation used in the main table. The Avg column is the equal-weight mean over the 12 dataset columns. Bold marks the best method within each column.