CONDITION ON THE GALLERY, NOT JUST THE QUERY: QUERY EQUALIZATION AND SPARSE TRANSPORT FOR DEBIASING VISION-LANGUAGE RETRIEVAL

Anonymous authors
Paper under double-blind review

ABSTRACT

Text-to-image retrieval with a vision-language model such as CLIP ranks the images of a gallery against a text query, and for a stereotyped query the retrieved set can have a demographic composition far from that of the gallery. A retrieval score has two arguments, the query and the image, yet many post-hoc corrections derive their edit from text prompts alone. We propose BALLAST (Barycentric Level-Allocated Latent Sparse Transport), a debiasing method for frozen CLIP models that conditions the correction on unlabeled images from the gallery's population. From prompts naming the protected attribute, BALLAST gives each reference image a soft membership over the attribute groups, and this one estimate drives two corrections: a repair of the gallery, which moves the groups' activations in the code of a frozen sparse autoencoder part of the way toward a shared barycenter on the coordinates where the groups differ, and a closed-form equalization of each query against the repaired gallery's group centroids. No per-image annotation is read and no gradient-based training takes place. On the SEM benchmark, with three retrieval datasets and four CLIP backbones, BALLAST lowers both KL divergence and MaxSkew below every SEM variant in all 24 comparisons on the two ViT backbones where it cuts KL on the stereotype queries by a factor of 2.3 to 5.0 relative to the best SEM variant per cell. The query equalization supplies most of the improvement, and its subspace has to come from the gallery: the same projection built from the attribute prompts alone raises KL above base CLIP on all 24 retrieval cells of the benchmark.

1 INTRODUCTION

Contrastive vision-language models such as CLIP (Radford et al., 2021) learn from large, uncurated collections of image-text pairs, and their embeddings carry the associations of that data (Birhane et al., 2021; Agarwal et al., 2021; Kay et al., 2015). In text-to-image retrieval those associations decide which images come back: a query such as "a photo of a criminal" over a gallery of face images can return a set whose demographic composition departs sharply from that of the gallery.

Retraining is rarely an option, so a correction has to act on the embeddings the frozen model produces. What such a correction needs is an estimate of how the images of each demographic group score against a query in the gallery being searched, because that is what retrieval fairness measures: the composition of the retrieved set against the composition of the gallery.

Existing corrections take that estimate from elsewhere. Most derive their edit from text prompts (Guimard et al., 2026; Chuang et al., 2023; Adila et al., 2024; Molahasani et al., 2025; Jang et al., 2025): prompts reveal what the model associates with a group, but they do not measure how the images of that group score in the gallery at hand. The methods that do look at images of the target population either require per-image group labels (Gerych et al., 2024; Wang et al., 2026) or use the images only for a mean shift or a rescaling of scores (Lu et al., 2025; Hasanebrahimi et al., 2026). The gallery itself, the one object every fairness metric is defined against, has not been the source.

We make it the source. Unlabeled images from the gallery's population already contain what the prompts lack, and in a deployed system the gallery being searched can supply them itself. Scoring each reference image against prompts that name the protected attribute gives it a soft membership over the groups, and from these memberships alone we read off how the groups differ in this gallery, both on average and in their full distributions.

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.

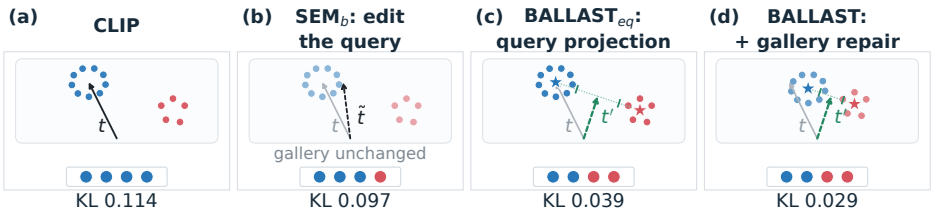


Figure 1: Gallery-conditioned debiasing. Colors mark pseudo-groups, arrows denote queries, stars mark soft group centroids, and trays show the top of the retrieved set. SEM$_b$ edits the query, BALLAST$_{eq}$ equalizes the query against gallery-derived centroids, and BALLAST also repairs the gallery. Panels are schematic; KL values average 25 query families (Table 1); more in App. E.

BALLAST (Barycentric Level-Allocated Latent Sparse Transport) applies this one estimate on both sides of the score. It repairs the gallery, moving the groups' image representations part of the way toward each other in the sparse code of a frozen autoencoder, where the groups' differences are concentrated, and it equalizes each query so that no group is favored on average by the repaired gallery. Neither step reads a group label or trains anything; the repair is calibrated without labels on the reference gallery. Figure 1 illustrates the difference from editing the query alone.

On the SEM benchmark, BALLAST improves on every variant of SEM, the closest prior method, on both fairness metrics in every retrieval setting of the ViT backbones, and the lead survives other retrieval depths and a single fixed configuration. Ablations show that the query equalization carries most of the gain, and that its power comes from the gallery: the same projection built from the prompts alone is worse than no correction at all. A fit made with a single query transfers to most queries it never saw. Our contributions are fourfold.

- **A gallery-conditioned formulation of post-hoc debiasing.** The estimate a correction needs can be read from unlabeled images of the gallery itself, without per-image labels, and one such estimate can serve both sides of the retrieval score.
- **A label-free gallery repair.** A sparse barycentric transport that moves the groups' activations toward each other in the code of a frozen sparse autoencoder, with no training and a contraction guarantee (App. P, Thm. 1).
- **A gallery-conditioned query equalization.** A closed-form projection with nothing to tune that removes a query's average preference between the groups of the repaired gallery.
- **Evidence that the gallery is what matters.** BALLAST beats every SEM variant across the benchmark's ViT settings, the same projection fails when its subspace comes from prompts instead of the gallery, and a fit made with one query transfers to unseen queries.

2 Related Work

Prompt-derived and reference-based debiasing. Many methods derive their corrections from text prompts. Orth-Proj and its calibrated form Orth-Cali (Chuang et al., 2023) project prompted bias directions out of the query. RoboShot (Adila et al., 2024) projects LLM-derived harmful directions out of the image embeddings and amplifies helpful ones, and TIE (Lu et al., 2025) shifts each group's image embeddings along its own prompt direction by the group's mean projection, with groups annotated or, in TIE*, inferred from prompts. PRISM (Molahasani et al., 2025) learns a projection from LLM-generated descriptions, and ZSDebias (Jang et al., 2025) learns debiased image embeddings from an LLM bias corpus and generic unlabeled images. Concurrent work treats bias as a subspace (Zhao et al., 2026) or gives a closed-form both-modality correction (Lian et al., 2026). Reference-based methods use images of the target population. BendVLM (Gerych et al., 2024) refines each query at test time against labeled reference neighbors. W4D (Wang et al., 2026) aligns prompt-induced query distributions with group-labeled reference distributions. DAT (Hasanebrahimi et al., 2026), evaluated on classification only, rescales each image's group-prompt similarities by its local density against held-out references labeled by class and, except in DAT*, by attribute. Each of these consumes per-image labels or leaves the gallery embeddings untouched.

Sparse-feature interventions and SEM. SAE atoms have been used as an intervention space for fairness: existing work selects atoms and suppresses or projects out their directions (Bărbălau et al., 2025; Yalavarthi et al., 2026; An et al., 2026), S&P (Bărbălau et al., 2025) with a labeled selector in its main configuration. SEM (Guimard et al., 2026), the method BALLAST builds on, encodes each

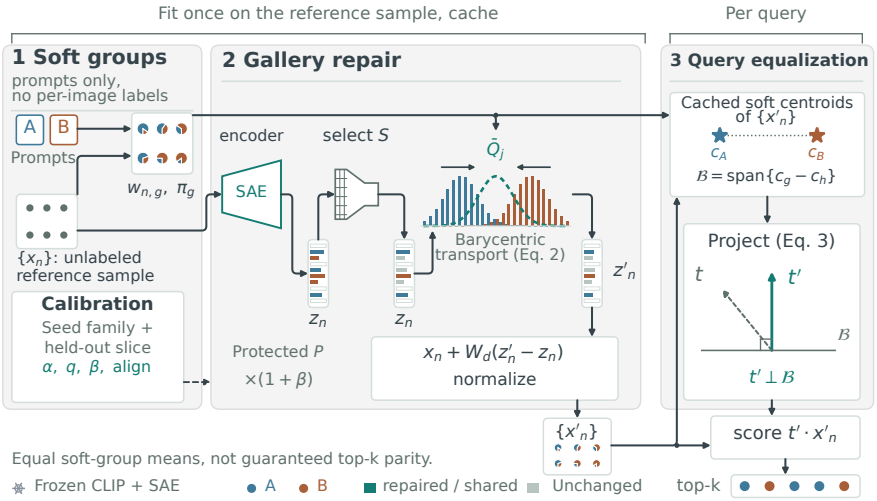


Figure 2: BALLAST overview. (1) Attribute prompts and an unlabeled reference sample yield soft group weights. (2) Selected SAE activations are partially transported toward shared per-atom barycenters (Eq. 2); only the decoded changes are added to the image embeddings. (3) Each query is equalized against the repaired reference centroids (Eq. 3). The gallery repair is calibrated on one seed query family, and the fitted repair and projection are cached across queries.

query with a frozen Matryoshka SAE, attenuates the query's bias-sensitive atoms while seeking to retain its content atoms, and decodes; the gallery is unchanged.

Distributional repair and gallery-aware retrieval. Barycentric repair of feature or prediction distributions under demographic parity is classical in tabular fairness (Feldman et al., 2015; Gordaliza et al., 2019; Chzhen et al., 2020). FairCLIP (Luo et al., 2024) trains an optimal-transport alignment of labeled groups for CLIP, and Linear-AcT (Rodriguez et al., 2025) maps dense activations from a labeled source to a labeled target distribution. BALLAST instead transports selected sparse activations of unlabeled pseudo-groups partway toward their barycenter without training (Sec. 3.3); no prior work, to our knowledge, transports group-conditioned SAE activations.

Kong et al. (2023) balance each retrieved set with CLIP-predicted groups at test time (further precedents in App. N); BALLAST repairs the gallery embeddings once, with no per-query reranking.

What is new. What distinguishes BALLAST is one estimate, fitted from an unlabeled reference gallery and attribute prompts and cached across queries, that serves both a sparse barycentric repair of that gallery and a gallery-derived query equalization. The closest label-free relatives, TIE* and DIAL (Yalavarthi et al., 2026), stop at one fixed edit, a per-pseudo-group mean shift along a prompt direction or a projection out of a spurious SAE subspace; DAT* reads class labels and rescales scores. None aligns the groups' distributions or conditions the query on the gallery. The ingredients, a projection along difference vectors and soft groups from CLIP scores, are standard tools; new is the gallery-conditioned estimate they share and the two operators built on it.

3 Method

BALLAST works from a reference sample of images drawn from the same population as the gallery to be searched, estimates how the groups of a protected attribute differ in embedding space, and uses that estimate to correct both the gallery's image embeddings and the queries, in three stages (Fig. 2). Soft group weights are inferred from attribute prompts (Sec. 3.2); selected activation distributions of the gallery are transported toward their barycenter in the code of a frozen sparse autoencoder (Sec. 3.3); and each query is equalized against the repaired gallery's group centroids (Sec. 3.4). The first two stages are fitted once per reference sample and cached; the third runs per query. Sec. 3.5 describes the label-free calibration of the gallery repair and the classification extension.

3.1 Problem setup and information regime

CLIP (Radford et al., 2021) maps text and images into a shared space $\mathbb{R}^d$ with frozen encoders E_{txt} and E_{vis}. A retrieval query with embedding t is scored against a gallery of image embeddings $\{x_n\}_{n=1}^N$ by the cosine $s(t, x_n)$, and the k highest-scoring images are returned; zero-shot classification labels an image with the class whose prompt scores highest. We measure retrieval fairness

by comparing the group composition of the retrieved set with that of the gallery it was drawn from, which is why the correction should be conditioned on that gallery.

BALLAST therefore fits on an unlabeled *reference gallery*, where it estimates how the groups differ, and applies the fitted correction when answering queries against the *evaluation gallery*, which is drawn from the same population. In a deployed system the gallery being searched can itself serve as the reference gallery. We fix four terms. A *group* is a value of the protected attribute, such as one gender or one race category of the benchmark. A *pseudo-group* is the prompt-based counterpart of a group: each image belongs to it with a soft membership weight inferred from the attribute prompts (Sec. 3.2). An *atom* is one feature learned by the sparse autoencoder. An *activation* is how strongly that feature is present in an image, zero when the feature is absent. *Label-free* means that no per-image group or class annotation is read, while prompts naming the attribute are required.

Inputs and outputs. BALLAST receives the frozen CLIP encoders, SEM's frozen sparse autoencoder (Guimard et al., 2026), the benchmark's attribute prompts, the *task prompts* (the seed query and its five paraphrases in retrieval, the class prompts and their paraphrases in classification), SEM's diverse prompts, and the unlabeled reference gallery. It returns the repaired gallery embeddings and one closed-form correction rule, computed once per reference gallery and applied to every retrieval query; what is fitted where and what each method consumes are listed in App. B (Table 5).

3.2 Soft groups and sparse codes

Soft groups from prompts. We embed each group's attribute prompts with E_{txt} and average them into an anchor $\bar{t}_g$. Every reference image then receives a *soft membership weight* over the G pseudo-groups, $w_{n,g} \propto \exp\big(\cos(x_n, \bar{t}_g)/\tau\big)$ with $\tau = 0.03$ and $\sum_g w_{n,g} = 1$, computed from dense CLIP cosines. The *prior* $\pi_g = \frac{1}{N}\sum_n w_{n,g}$ is the share of pseudo-group g in the reference gallery; when we compare it with the annotated group shares, we call it the pseudo-prior.

The two quantities play different roles. The weights $w_{n,g}$ enter every group-conditioned statistic below, so no hard assignment is ever committed; the prior π_g weights the groups against each other when a common target is formed. The pseudo-prior tracks the annotated group shares on the gender cells but not on UTKFace race, where it stays close to uniform (App. C, Table 12). A soft weight limits the cost of a wrong assignment but does not rule one out (Sec. 4.4).

Sparse codes. A sparse autoencoder represents an embedding $x \in \mathbb{R}^d$ by a nonnegative code $z = \text{ReLU}(W_e(x - b_{\text{pre}}) + b_e) \in \mathbb{R}^s$, $s = 16{,}384 \gg d$, and a linear decoder reconstructs $\hat{x} = W_d z + b_{\text{pre}}$. Column j of W_d is a direction in embedding space, the atom j, and z_j says how strongly that direction is present in x. We use the Matryoshka SAE checkpoints (Zaigrajew et al., 2025) released with SEM, trained once on generic caption embeddings and kept frozen. The code is a useful place to intervene because a group difference that is spread over every coordinate of the dense embedding falls disproportionately on a small set of strong atoms of the sparse code: on the benchmark's reference sets the strongest 6% of atoms carry 38% to 85% of the between-group discrepancy (App. C), so an intervention can be confined to selected coordinates and decoded back. SEM established this premise for text embeddings; App. C measures it on the image side.

Applying a caption-trained SAE to images. The SAE checkpoints never saw images. On the ViT backbones, text embeddings are reconstructed at a mean cosine of 0.99 but image embeddings only at 0.78 to 0.85. On the ResNets, image embeddings are reconstructed at 0.58 to 0.77, and there the codes are also not sparse in practice (App. C). We respond in two ways. BALLAST never replaces an image by its reconstruction; it decodes only the change of the code on the selected atoms (Sec. 3.3). And a binary alignment switch, set by the calibration of Sec. 3.5, decides whether image embeddings are shifted toward the mean embedding of SEM's diverse prompts before encoding (App. B). The codes still carry the information the method needs. The *strength* of an atom, how often it appears among a sample's largest activations (App. B), concentrates the between-group discrepancy.

3.3 Gallery repair by sparse barycentric transport

Consider one selected atom j and two pseudo-groups A and B (Fig. 3, App. B). Suppose the atom is active more strongly, or more often, in the image embeddings of group A. A text query whose embedding has a component along the direction of that atom then scores the two groups differently on average, unless other atoms cancel the difference. The repair removes this difference atom by atom. For each selected atom it pairs equal percentiles of the two groups' activations and moves every activation part of the way toward the average of the two percentile curves, so that the two groups' activation distributions come closer. We use optimal transport for this because it aligns

the whole activation distribution of each group rather than only its mean, moves the activations by the least total amount that this alignment allows, and, in one dimension, has a closed form through quantiles. The steps below form the *gallery repair*, the gallery edit of the tables.

Group distributions and their barycenter. For atom j and group g, the *weighted group distribution* is the distribution of the activations $z_{n,j}$ over the reference images, each image counted with its membership weight $w_{n,g}$. This distribution has the cumulative distribution function $F_{g,j}$ and the quantile function $Q_{g,j}(u)$, $u \in (0,1)$, the activation below which a fraction u of group g's weighted mass lies. The *barycenter* of the G group distributions has the quantile function

$$\bar{Q}_j(u) \;=\; \textstyle\sum_g \pi_g \, Q_{g,j}(u), \tag{1}$$

the prior-weighted average of the group quantile curves. This is the W_2 Wasserstein barycenter of the one-dimensional distributions, the common target that minimizes the π-weighted sum of squared transport costs (Agueh & Carlier, 2011; Chzhen et al., 2020; Gordaliza et al., 2019). The *transport map* of group g is the equal-quantile map $T_{g,j} = \bar{Q}_j \circ F_{g,j}$, which sends the activation at quantile position u to $\bar{Q}_j(u)$; pairing equal quantile positions is the optimal coupling in one dimension. The *interpolated displacement* with strength $a \in [0,1]$ moves an activation a fraction a of the way to its target, $(1-a)\, z + a\, T_{g,j}(z)$, which limits how far any image moves. Throughout the paper, "partial" means this interpolation, never partial optimal transport, and no mass is discarded.

Selection, protection, and the update. After selection every atom is of one of three kinds: *transported* atoms, which the maps move; *protected* atoms, which are exempt from the transport and may be amplified instead; and unchanged atoms, the large majority. Transport is confined to atoms that pass a heuristic screen, with a cap and a fallback (App. B): an atom is selected when its between-group discrepancy, defined as the prior-weighted pairwise W_1 distance between the $Q_{g,j}$, exceeds the largest value observed under random shuffles of the group weights. A *protected* set of task-active, low-discrepancy atoms is exempted from the transport and amplified by a factor $1+\beta$ instead, following SEM's content amplification (App. B). On the transported atoms the displacement strength is atom-specific, $a_j = \alpha\, m_j$: a global α times a soft gate $m_j \in [0,1]$ that favors stronger atoms, with a smooth transition around a strength tier q, a threshold on the rank of the atom strength of Sec. 3.2 (Fig. 4a, App. C). Because an image belongs to every group with weight $w_{n,g}$, its target is the weighted blend of the group maps, which gives the update

$$z'_{n,j} \;=\; (1-a_j)\, z_{n,j} + a_j \textstyle\sum_g w_{n,g}\, T_{g,j}(z_{n,j}), \qquad a_j = \alpha\, m_j. \tag{2}$$

The transport also moves exact zeros: a zero activation can become positive and a positive one can become zero. An activation far above the fitted range is left unchanged (App. B).

Decoding the change. We do not replace an embedding by the decoded repaired code: the autoencoder reconstructs an embedding only approximately, so that would also discard what the code fails to capture. Instead the repaired embedding adds only the decoded *change* of the code, including the amplification of the protected atoms, to the original embedding, $x'_n = \mathrm{normalize}\big(x_n + W_d(z'_n - z_n)\big)$; up to the final rescaling, the unedited atoms and the reconstruction residual stay as they were, as in other SAE interventions (Marks et al., 2025).

Population property and scope. For the transport operator alone, with $a_j = 1$ and hard groups, the group distributions of atom j coincide after the update; with $a_j < 1$ their W_2 distances shrink by the factor $1 - a_j$ while the barycenter itself stays fixed (App. P, Thm. 1). The statement concerns the idealized population operator; App. P gives the slack from soft weights and the departures of finite quantile grids and boundary conventions, and Sec. 4.4 states what it leaves open.

3.4 Gallery-conditioned query equalization

After the transport a query can still score one group higher on average than another; equalization removes exactly that difference and uses the gallery to define it. Once the repair is calibrated we compute the soft centroids $c_g = \sum_n w_{n,g}\, x'_n / \sum_n w_{n,g}$ of the repaired reference gallery and the subspace $\mathcal{B} = \mathrm{span}\{c_g - c_h\}$ of their pairwise differences, of dimension at most $G-1$. Writing $\Pi_{\mathcal{B}}$ for the orthogonal projector onto $\mathcal{B}$ and $\hat{t} = t/\|t\|$, each retrieval query becomes

$$t' \;=\; \|t\|\, \frac{\hat{t} - \Pi_{\mathcal{B}}\hat{t}}{\|\hat{t} - \Pi_{\mathcal{B}}\hat{t}\|}. \tag{3}$$

BALLAST$_{eq}$ applies only this step, with the centroids of the normalized unrepaired gallery (Sec. 4.3). The corrected query has the same inner product with every centroid, $\langle t', c_g\rangle = \langle t', c_h\rangle$,

so its mean score is the same in every pseudo-group; the rescaling only restores the norm, which cosine scoring ignores. The projection pulls no query toward a shared point and displaces a query by at most $\sqrt{2}$ times the norm of its component in $\mathcal{B}$ (App. P, Thm. 2).

The correction is *gallery-conditioned*: $\mathcal{B}$ is estimated from soft centroids of the realized reference gallery, whereas Orth-Proj (Chuang et al., 2023) projects out a span of prompt embeddings. App. B (Table 8) shows that the centroid subspace is far better aligned with the annotated image group direction and removes far less of each query.

Equalization guarantees only that the groups have the same *average* score. Their scores can still spread differently, so a top-k set can still lean toward one group; that residual difference is what the distributional repair of Sec. 3.3 targets. A query that lies entirely inside $\mathcal{B}$ is a degenerate case, which the implementation does not guard (App. B).

3.5 Label-free calibration and classification

Four settings of the gallery repair are chosen per cell and split (Sec. 4.1): the transport strength α, the strength tier q, the amplification β, and the alignment switch; the query equalization has nothing to tune. Each setting is picked from a fixed grid on a held-out slice of the unlabeled reference gallery. Without labels, a candidate is scored by a proxy of the fairness metric: the two-sided MaxSkew of Sec. 4.1 over the top-k held-out images ranked by the task prompts, with the soft pseudo-labels and the pseudo-prior π standing in for annotations. Three guardrails keep the repair faithful: a candidate is admitted only if the repaired embeddings stay close to the originals, the top prompt of the most confident held-out images rarely changes, and the top-k ranking largely overlaps with the original one (thresholds in App. B). Leaving the gallery unrepaired passes all three guardrails trivially and is always admissible; Table 7 reports how often it is chosen. The reported gains thus include the effect of tuning the repair against a proxy of one evaluation metric; with every setting pinned to its most frequently selected value, BALLAST stays below the best SEM variant per cell and metric (Best SEM) on all 24 ViT comparisons, at an aggregate KL within 0.001 of the calibrated method's (App. B, Table 10). Grids, slice sizes, selected atom counts and the calibration prompts are in App. B.

Retrieval focus and zero-shot classification. Retrieval is the setting BALLAST is built for: the composition of a retrieved set is a property of the gallery being searched, and both operators are fitted on unlabeled images from that gallery's population. Zero-shot classification differs in two ways. Each image is scored against a fixed set of class prompts that must stay distinct from one another, so the query equalization is not applied to them, and its fairness metric, worst-group accuracy, needs class labels as well as group labels, neither of which the method reads. We therefore treat classification as a secondary extension, reported in App. F to H. BALLAST-C, the classification variant, scores the repaired gallery against SEM's text edit of the class prompts and then removes a shared per-group shift of the evaluation-gallery margins, a step we call TRIM (Translation Removal in Margins). TRIM reads unlabeled statistics of the gallery being evaluated, so BALLAST-C is *transductive*; the classification tables mark it with a dagger (†) for that reason. The TRIM margin model and its estimator, and plain BALLAST's treatment of class prompts, are in App. G.

4 Experiments

4.1 Experimental setup

Tasks, data, and backbones. We evaluate stereotype-query retrieval on FairFace (Karkkainen & Joo, 2021), UTKFace (Zhang et al., 2017), and CelebA (Liu et al., 2015), and hair-color retrieval on CelebA, following the SEM benchmark (Guimard et al., 2026). We use its four CLIP backbones: ViT-B/16, ViT-L/14@336px, RN50, and RN101. A *cell* is one combination of task, dataset, protected attribute, and backbone; the two ViT backbones give ten stereotype cells and two hair cells. A *query family* is one query concept, such as "addict", with its stored paraphrases. The classification extension uses CelebA and Waterbirds (Sagawa et al., 2020) on the same backbones (App. F).

Metrics. Retrieval reports the KL divergence of the top-500 group composition from the evaluation gallery's prior, and MaxSkew (Geyik et al., 2019), the largest positive log ratio of a group's retrieved share to its gallery share. Hair queries carry relevance labels and add precision@500.

Protocol. Results aggregate ten seed-fixed 50/50 splits per dataset into reference set and evaluation gallery. Per-split ranges describe variability, not confidence intervals (App. M).

Table 1: Stereotype retrieval fairness on the ViT backbones. KL and MaxSkew (MS) are evaluated at $k = 500$; lower is better. Row groups distinguish access to reference images; gray rows use group-labeled reference images and are excluded from the ranking. Bold and underline mark the best and second-best results among methods without reference group labels, with ties at displayed precision. BALLAST$_{eq}$ applies query equalization without gallery repair. SEM$_b$, SEM$_i$ and SEM$_{bi}$ differ only in which prompt resources score the atoms (Table 5).

	FairFace Race		FairFace Gender		UTKFace Race		UTKFace Gender		CelebA Stereo.	
Method	KL↓	MS↓	KL↓	MS↓	KL↓	MS↓	KL↓	MS↓	KL↓	MS↓
ViT-L/14@336px										
No task-specific reference images										
Base CLIP	0.244	0.798	0.114	0.326	0.124	0.461	0.040	0.186	0.237	0.536
RoboShot	0.303	0.925	0.325	0.519	0.236	0.741	0.269	0.466	0.194	0.393
Orth-Proj	0.213	0.783	0.034	**0.164**	0.201	0.493	0.050	0.221	0.099	0.356
Orth-Cali	0.169	0.657	0.052	0.207	0.180	0.528	0.040	0.201	0.054	0.264
PRISM	0.147	0.614	0.051	0.230	0.159	0.543	0.038	0.198	0.061	0.245
ZSDebias	0.178	0.692	0.112	0.321	0.164	0.474	0.112	0.333	0.258	0.557
SEM$_i$	0.147	0.625	0.123	0.328	0.058	0.451	0.033	0.186	0.153	0.413
SEM$_b$	0.194	0.706	0.097	0.298	0.137	0.446	0.047	0.201	0.198	0.480
SEM$_{bi}$	0.155	0.623	0.110	0.299	0.118	0.420	0.054	0.217	0.209	0.489
Unlabeled reference images										
BALLAST$_{eq}$	0.036	0.357	0.039	0.212	0.017	0.213	**0.007**	**0.088**	0.034	0.229
BALLAST	**0.034**	**0.354**	**0.029**	0.182	**0.014**	**0.205**	0.008	**0.088**	**0.033**	**0.228**
KL ratio*	4.4		3.4		4.0		4.2		4.6	
Group-labeled reference images										
BendVLM	0.107	0.580	0.005	0.080	0.086	0.478	0.009	0.106	0.030	0.217
BendSEM$_{bi}$	0.063	0.435	0.007	0.088	0.045	0.329	0.006	0.081	0.042	0.261
ViT-B/16										
No task-specific reference images										
Base CLIP	0.237	0.795	0.139	0.346	0.125	0.476	0.134	0.321	0.314	0.555
RoboShot	0.327	0.893	0.349	0.507	0.220	0.682	0.248	0.397	0.190	0.357
Orth-Proj	0.313	0.818	0.336	0.521	0.282	0.542	0.196	0.388	0.188	0.383
Orth-Cali	0.267	0.789	0.416	0.596	0.242	0.518	0.266	0.457	0.235	0.407
PRISM	0.152	0.643	0.085	0.284	0.142	0.508	0.093	0.293	0.143	0.377
ZSDebias	0.197	0.782	0.122	0.320	0.128	0.624	0.070	0.248	0.195	0.439
SEM$_i$	0.170	0.690	0.087	0.269	0.096	0.407	0.064	0.241	0.173	0.443
SEM$_b$	0.232	0.750	0.098	0.278	0.145	0.502	0.124	0.320	0.240	0.495
SEM$_{bi}$	0.217	0.750	0.088	0.255	0.137	0.499	0.119	0.317	0.224	0.489
Unlabeled reference images										
BALLAST$_{eq}$	0.038	0.352	0.048	0.235	0.035	0.235	**0.011**	**0.109**	0.050	0.227
BALLAST	**0.034**	**0.350**	**0.038**	**0.205**	**0.032**	**0.225**	0.013	0.115	**0.049**	**0.224**
KL ratio*	5.0		2.3		3.0		4.9		3.5	
Group-labeled reference images										
BendVLM	0.098	0.496	0.009	0.105	0.099	0.409	0.009	0.102	0.035	0.237
BendSEM$_{bi}$	0.055	0.434	0.007	0.093	0.055	0.419	0.005	0.078	0.030	0.224

*KL ratio = best SEM KL / BALLAST KL, computed from unrounded means; higher is better.

Baselines and information access. We reproduce the released baseline configurations under the same evaluation harness (App. A). Methods are grouped by their access to task-specific reference images at test time: none, unlabeled reference images, or group-labeled reference images (Table 5, App. B). PRISM and ZSDebias are averaged over their three released seeds (App. D, which also notes ZSDebias's training overlap with UTKFace). W4D (Wang et al., 2026) uses a different five-fold protocol and is not in the tables.

4.2 Retrieval fairness

On the ViT backbones, BALLAST lowers both KL and MaxSkew below all three SEM variants in all 24 comparisons (one per cell and metric), 20 on the stereotype cells (Table 1) and 4 on the hair cells (Table 3). On the stereotype cells, BALLAST reduces KL by a factor of 2.3 to 5.0 relative to the best per-cell SEM variant (ratio rows).

BALLAST and BALLAST$_{eq}$ hold the two best label-free values in 18 of the 20 stereotype comparisons. Two of the 24 comparisons favor a label-free method outside the BALLAST family: Orth-Proj has a lower MaxSkew on FairFace-gender ViT-L/14@336px, where it also beats BALLAST$_{eq}$ on both metrics, and seed-mean PRISM has a lower MaxSkew on hair ViT-B/16 (Table 29). The lead over SEM holds on every ViT cell and metric at $k = 100$ and $k = 1{,}000$ as well (App. L).

On the ResNet backbones (Table 2), BALLAST beats the Best SEM point estimate in 16 of the 20 stereotype comparisons. The four others are the CelebA-stereotype comparisons, where SEM$_i$ leads. Seed-mean ZSDebias is lower than BALLAST on six of the 20 ResNet comparisons; four of the six are the CelebA-stereotype comparisons, both metrics on both ResNets. PRISM and ZSDebias are

Table 2: Stereotype retrieval fairness on the ResNet backbones, KL and MaxSkew (MS) at $k = 500$; conventions as in Table 1, Best SEM being the lowest value per metric and setting. On RN50 the calibration kept the gallery unrepaired in 38 of the 60 splits, so those results are largely the equalizations; full baseline results in App. D.

	FairFace Race		FairFace Gender		UTKFace Race		UTKFace Gender		CelebA Stereo.	
Method	KL↓	MS↓	KL↓	MS↓	KL↓	MS↓	KL↓	MS↓	KL↓	MS↓
RN50										
Best SEM	0.126	0.562	0.031	0.190	0.039	0.265	0.075	0.280	**0.050**	**0.193**
BALLAST$_{eq}$	**0.044**	**0.390**	**0.011**	**0.107**	**0.035**	**0.243**	0.012	**0.109**	0.118	0.388
BALLAST	0.045	0.393	0.012	0.113	**0.035**	**0.243**	**0.011**	**0.109**	0.105	0.357
KL ratio*	2.8		2.7		1.1		6.6		0.5	
RN101										
Best SEM	0.111	0.566	0.037	0.202	0.084	0.341	0.041	0.215	**0.041**	**0.186**
BALLAST$_{eq}$	**0.060**	0.466	0.011	0.107	0.044	**0.260**	**0.012**	**0.114**	0.099	0.354
BALLAST	0.061	**0.464**	**0.010**	**0.104**	**0.041**	0.268	**0.012**	0.118	0.099	0.346
KL ratio*	1.8		3.6		2.1		3.3		0.4	

*KL ratio = best SEM KL / BALLAST KL, computed from unrounded means; higher is better.

Table 3: Hair-color retrieval on CelebA with gender as the protected attribute: KL and MaxSkew (MS) as in Table 1, plus precision@500 (higher is better). Best SEM selects each metric independently across its three variants; full baseline results in App. J.

	ViT-L/14@336px			ViT-B/16		
Method	KL↓	MS↓	Prec.↑	KL↓	MS↓	Prec.↑
Base CLIP	0.148	0.359	0.622	0.179	0.409	0.630
Best SEM	0.135	0.365	**0.732**	0.172	0.344	**0.734**
BALLAST$_{eq}$	0.097	0.323	0.608	0.043	0.231	0.592
BALLAST	**0.070**	**0.294**	0.572	**0.042**	**0.208**	0.585
KL ratio*	1.9			4.1		

*KL ratio = best SEM KL / BALLAST KL, computed from unrounded means; higher is better.

the two baselines that train a component, a projection and an image-embedding debiaser (Sec. 2); PRISM leads BALLAST on none of the 20 ResNet comparisons. On RN50 the calibration selected the identity, that is, no gallery repair (transport strength and amplification both zero), in 38 of the 60 retrieval splits, so the RN50 results are largely those of the equalization alone (App. D). Two comparisons are close at the split level: on FairFace-gender ViT-L/14@336px Orth-Proj's MaxSkew is below BALLAST's on nine of the ten splits, and on UTKFace-race RN50 BALLAST's KL is below SEM$_i$'s on all ten splits, by a margin that vanishes at three decimals in its worst split (App. I).

4.3 Analysis

Fairness and relevance. On the hair-color queries, where the protected attribute is correlated with the relevance labels, BALLAST's lower demographic imbalance comes with lower precision (Table 3). Part of the reason is in the gallery. On the blond family, which seeds the calibration, 94% of the blond images in CelebA are women (28,234 of 29,983), so a top-500 set that matches the gallery's gender shares must draw its men from a far thinner pool of blond images, and precision falls. SEM$_{bi}$ balances less and retains more precision on both backbones (Table 29), and BALLAST$_{eq}$ retains more than BALLAST (precision column of Table 4). No tested equalization or transport strength combines SEM-level precision with BALLAST-level fairness (App. J).

Component ablations. The query equalization supplies most of the improvement, and its gallery-derived subspace is what matters: the same projection with the prompt-anchor difference subspace, which needs no reference images, raises KL above base CLIP on all 24 retrieval cells, whereas BALLAST$_{eq}$ is below this control on every cell and both metrics (48 comparisons) (App. B, Table 9). Over the twelve ViT cells the mean KL is 0.038 for BALLAST$_{eq}$, 0.033 for full BALLAST, and 0.135 for the gallery repair alone, which stays above Best SEM on nine of the twelve cells (Table 22, App. H). The repair alone still lowers KL below base CLIP on all twelve cells; it never touches the query, which every SEM variant edits, and its job is the distributional part of the correction and the centroids the equalization projects against. Table 4 shows the same division on the four settings where the repair contributes most, with every variant recalibrated without labels.

Of the remaining improvement, the transport contributes a share that varies with the setting, and the amplification of the protected atoms, inherited from SEM, contributes almost nothing. With the equalization kept active, adding the transport lowers the twelve-cell mean KL from 0.038 to 0.033, a 13% reduction, and the twelve-cell mean is lower with the transport in every one of the

Table 4: Component ablation on four retrieval settings where gallery repair contributes most; Eq., Tr., and Amp. denote query equalization, transport, and protected-atom amplification, each variant recalibrated without labels. Base CLIP and the per-metric best SEM values are references; bold and underline as in Table 1; Prec. is precision@500 on the hair setting. App. I covers all twelve ViT settings with equalization active; design-choice ablations are in App. H.

	Component			FairFace Race		FairFace Gender		UTKFace Race		CelebA Hair		
Variant	Eq.	Tr.	Amp.	KL↓	MS↓	KL↓	MS↓	KL↓	MS↓	KL↓	MS↓	Prec.↑
ViT-L/14@336px												
Base CLIP				0.244	0.798	0.114	0.326	0.124	0.461	0.148	0.359	0.622
Best SEM				0.147	0.623	0.097	0.298	0.058	0.420	0.135	0.365	0.732
Gallery repair only		✓	✓	0.179	0.696	0.065	0.255	0.082	0.393	0.075	**0.266**	0.572
BALLAST$_{eq}$	✓			0.036	0.357	0.039	0.212	0.017	0.213	0.097	0.323	**0.608**
Eq. + Amp.	✓		✓	0.036	0.357	0.040	0.215	0.017	0.214	0.097	0.322	**0.608**
Eq. + Tr.	✓	✓		**0.034**	**0.354**	**0.028**	**0.182**	**0.014**	**0.204**	**0.070**	0.294	0.572
BALLAST	✓	✓	✓	**0.034**	**0.354**	0.029	**0.182**	**0.014**	0.205	**0.070**	0.294	0.572
ViT-B/16												
Base CLIP				0.237	0.795	0.139	0.346	0.125	0.476	0.179	0.409	0.630
Best SEM				0.170	0.690	0.087	0.255	0.096	0.407	0.172	0.344	0.734
Gallery repair only		✓	✓	0.204	0.767	0.118	0.317	0.102	0.443	0.122	0.327	**0.608**
BALLAST$_{eq}$	✓			0.038	0.352	0.048	0.235	0.035	0.235	0.043	0.231	0.592
Eq. + Amp.	✓		✓	0.038	0.354	0.049	0.235	0.035	0.236	0.043	0.232	0.593
Eq. + Tr.	✓	✓		**0.034**	**0.350**	**0.037**	**0.204**	**0.032**	0.226	**0.042**	**0.208**	0.585
BALLAST	✓	✓	✓	**0.034**	**0.350**	0.038	0.205	**0.032**	**0.225**	**0.042**	**0.208**	0.585

ten splits (per-cell values in App. I). Adding the amplification to BALLAST$_{eq}$ leaves its aggregates unchanged at the displayed precision, so the fairness gains come from the two gallery-conditioned operators (App. I, Table 22). The design choices behind the soft weights, the strength tier and the barycenter are ablated in App. H.

Transfer to unseen queries. The fit is calibrated on the first query family of each cell and cached, so the remaining 24 of the 25 stereotype families and three of the four hair families are evaluated against a repair that never saw their queries (Table 31, App. K). BALLAST lowers KL below base CLIP on 214 of the 240 unseen stereotype family evaluations across the ten ViT cells. On the hair cells, all three unseen families improve on ViT-B/16 and one of three on ViT-L/14@336px.

Computational cost. BALLAST does its work once per gallery: in the instrumented run (App. O, medians over the retrieval cells) a split is fitted in 11.7 s, after which each further query family costs 14 ms against 35 ms for SEM$_{bi}$; under the benchmark protocol, which fits every split afresh, a split costs 12.8 s against 0.9 s, so at the benchmark's query counts SEM remains an order of magnitude cheaper.

4.4 Limitations

The evidence is scoped to the benchmark protocol: an unlabeled half of each evaluation dataset as the reference set, one fixed calibration family, a transductive classification extension (App. G), and demographic parity with proxy annotations as every method's target. The pseudo-groups are estimates rather than calibrated probabilities (App. C), so the effect of a wrong assignment is limited but not removed. Neither operator guarantees parity of a retrieved set: the transport aligns selected code marginals rather than the joint distribution, and the equalization matches group mean scores; fidelity after decoding is checked empirically by the calibration floors rather than guaranteed by a bound. On the hair queries balance comes with lower precision, an observed operating point rather than a necessary cost (Sec. 4.3, App. J). The method treats one protected attribute at a time; intersectional repair, reference-to-deployment mismatch, sensitivity to the calibration family and a protocol-matched comparison with W4D are left to future work.

5 Conclusion

BALLAST conditions post-hoc debiasing on the gallery it is deployed on: one estimate of the gallery's group geometry, from unlabeled reference images and attribute prompts, drives a sparse barycentric repair of the gallery and a closed-form equalization of each query. On the SEM benchmark the projection carries most of the retrieval gain and the transport a cell-dependent share; both pay for fairness with precision on group-correlated queries. Whether the estimate survives a gap between the reference sample and the deployed gallery is the question the method now poses.

AI USE STATEMENT

We used large language model assistants, under the authors' direction and with every output reviewed by the authors, to polish and draft text from the authors' outlines, to search for related work the authors had missed in their manual literature survey, to discuss design alternatives and help implement and check the experiment code, and to help check the derivations in the appendix. The method, the experiments and the claims are the authors' own; every number in the paper traces to records produced by the released harness and was verified by the authors, who take full responsibility for this work.

ETHICS STATEMENT

This work aims to reduce demographic bias in the retrieval and zero-shot classification behavior of deployed vision-language models. It uses only standard public benchmark datasets (FairFace, UTKFace, CelebA, Waterbirds) and their existing group annotations, demographic on the three face datasets and background on Waterbirds. These annotations serve as evaluation proxies and are imperfect: the benchmark's group taxonomies are coarse and its fairness target encodes demographic parity, and both caveats apply to every method compared.

The method is label-free in the sense that it reads no per-image annotation, but it is not attribute-agnostic: like the baselines, it consumes a small set of prompts naming the protected attribute. The per-image soft group weights it computes are an intermediate used only for group statistics and are not surfaced as predictions about individuals. Limitations, including the fairness and precision trade-off and the treatment of one bias attribute at a time, are discussed in Sec. 4.4.

REPRODUCIBILITY STATEMENT

Code, configurations, and per-configuration records are released. Every record carries the per-split values of every metric and checksums of every input, and rerunning the released configurations reproduces every reported value; the runs and the checks that accompany them are described in App. M. All 512 baseline configurations of the benchmark were regenerated under the same harness: 481 deviate from the released values by less than 0.005 and the remaining 31 by less than 0.02 (App. A). Every table value is produced from the records. Hyperparameters and calibration grids are listed in App. B, the calibration principle in Sec. 3.5, and the computational cost in Sec. 4.3 and App. O.

REFERENCES

Dyah Adila, Changho Shin, Linrong Cai, and Frederic Sala. Zero-shot robustification of zero-shot models. In *International Conference on Learning Representations*, volume 2024, pp. 34402–34434, 2024.

Sandhini Agarwal, Gretchen Krueger, Jack Clark, Alec Radford, Jong Wook Kim, and Miles Brundage. Evaluating clip: towards characterization of broader capabilities and downstream implications. *arXiv preprint arXiv:2108.02818*, 2021.

Martial Agueh and Guillaume Carlier. Barycenters in the wasserstein space. *SIAM Journal on Mathematical Analysis*, 43(2):904–924, 2011. doi: 10.1137/100805741. URL https://doi.org/10.1137/100805741.

Na Min An, Yoonna Jang, Yusuke Hirota, Ryo Hachiuma, Isabelle Augenstein, and Hyunjung Shim. Interpretable debiasing of vision-language models for social fairness. *arXiv preprint arXiv:2602.24014*, 2026.

Antonio Bărbălau, Cristian Daniel Păduraru, Teodor Poncu, Alexandru Tifrea, and Elena Burceanu. Rethinking sparse autoencoders: Select-and-project for fairness and control from encoder features alone. *arXiv preprint arXiv:2509.10809*, 2025.

Hugo Berg, Siobhan Hall, Yash Bhalgat, Hannah Kirk, Aleksandar Shtedritski, and Max Bain. A prompt array keeps the bias away: Debiasing vision-language models with adversarial learning. In *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pp. 806–822, 2022.

Abeba Birhane, Vinay Uday Prabhu, and Emmanuel Kahembwe. Multimodal datasets: misogyny, pornography, and malignant stereotypes. *arXiv preprint arXiv:2110.01963*, 2021.

Tolga Bolukbasi, Kai-Wei Chang, James Y Zou, Venkatesh Saligrama, and Adam T Kalai. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. *Advances in neural information processing systems*, 29, 2016.

Neil Chowdhury, Franklin Wang, Sumedh Shenoy, Douwe Kiela, Sarah Schwettmann, and Tristan Thrush. Nearest neighbor normalization improves multimodal retrieval. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 22571–22582, 2024.

Ching-Yao Chuang, Varun Jampani, Yuanzhen Li, Antonio Torralba, and Stefanie Jegelka. Debiasing vision-language models via biased prompts, 2023. URL https://arxiv.org/abs/2302.00070.

Evgenii Chzhen, Christophe Denis, Mohamed Hebiri, Luca Oneto, and Massimiliano Pontil. Fair regression with wasserstein barycenters. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS '20, Red Hook, NY, USA, 2020. Curran Associates Inc. ISBN 9781713829546.

Sepehr Dehdashtian, Lan Wang, and Vishnu Boddeti. Fairerclip: Debiasing clip's zero-shot predictions using functions in rkhss. In *International Conference on Learning Representations*, volume 2024, pp. 26185–26202, 2024.

Michael Feldman, Sorelle A Friedler, John Moeller, Carlos Scheidegger, and Suresh Venkatasubramanian. Certifying and removing disparate impact. In *proceedings of the 21th ACM SIGKDD international conference on knowledge discovery and data mining*, pp. 259–268, 2015.

Walter Gerych, Haoran Zhang, Kimia Hamidieh, Eileen Pan, Maanas Sharma, Thomas Hartvigsen, and Marzyeh Ghassemi. Bendvlm: Test-time debiasing of vision-language embeddings. *Advances in Neural Information Processing Systems*, 37:62480–62502, 2024.

Walter Gerych, Cassandra Parent, Quinn Perian, Rafiya Javed, Justin Solomon, and Marzyeh Ghassemi. WRING out the bias: A rotation-based alternative to projection debiasing. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=tkE29O0jzF.

Sahin Cem Geyik, Stuart Ambler, and Krishnaram Kenthapadi. Fairness-aware ranking in search & recommendation systems with application to linkedin talent search. In *Proceedings of the 25th acm sigkdd international conference on knowledge discovery & data mining*, pp. 2221–2231, 2019.

Soumya Suvra Ghosal and Yixuan Li. Distributionally robust optimization with probabilistic group. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pp. 11809–11817, 2023.

Paula Gordaliza, Eustasio Del Barrio, Gamboa Fabrice, and Jean-Michel Loubes. Obtaining fairness using optimal transport theory. In Kamalika Chaudhuri and Ruslan Salakhutdinov (eds.), *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 2357–2365. PMLR, 09–15 Jun 2019. URL https://proceedings.mlr.press/v97/gordaliza19a.html.

Quentin Guimard, Federico Bartsch, Simone Caldarella, Rahaf Aljundi, Elisa Ricci, and Massimiliano Mancini. SEM: Sparse Embedding Modulation for Post-Hoc Debiasing of Vision-Language Models. In *Findings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2026.

Afsaneh Hasanebrahimi, Hanxun Huang, Christopher Leckie, and Sarah Erfani. Density-aware translation of spurious correlations in zero-shot vlms. *arXiv preprint arXiv:2606.01710*, 2026.

Taeuk Jang, Hoin Jung, and Xiaoqian Wang. Target bias is all you need: Zero-shot debiasing of vision-language models with bias corpus. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 01–12. IEEE, 2025.

Hoin Jung, Taeuk Jang, and Xiaoqian Wang. A unified debiasing approach for vision-language models across modalities and tasks. *Advances in Neural Information Processing Systems*, 37: 21034–21058, 2024.

Kimmo Karkkainen and Jungseock Joo. Fairface: Face attribute dataset for balanced race, gender, and age for bias measurement and mitigation. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pp. 1548–1558, 2021.

Matthew Kay, Cynthia Matuszek, and Sean A Munson. Unequal representation and gender stereotypes in image search results for occupations. In *Proceedings of the 33rd annual acm conference on human factors in computing systems*, pp. 3819–3828, 2015.

Fanjie Kong, Shuai Yuan, Weituo Hao, and Ricardo Henao. Mitigating test-time bias for fair image retrieval. *Advances in neural information processing systems*, 36:71520–71539, 2023.

Tangzheng Lian, Guanyu Hu, Yijing Ren, Dimitrios Kollias, and Oya Celiktutan. A closed-form solution for debiasing vision-language models with utility guarantees across modalities and tasks. *arXiv preprint arXiv:2603.12998*, 2026.

Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *Proceedings of the IEEE international conference on computer vision*, pp. 3730–3738, 2015.

Shenyu Lu, Junyi Chai, and Xiaoqian Wang. Mitigating spurious correlations in zero-shot multimodal models. In *International Conference on Learning Representations*, volume 2025, pp. 14727–14753, 2025.

Yan Luo, Min Shi, Muhammad Osama Khan, Muhammad Muneeb Afzal, Hao Huang, Shuaihang Yuan, Yu Tian, Luo Song, Ava Kouhana, Tobias Elze, et al. Fairclip: Harnessing fairness in vision-language learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 12289–12301, 2024.

Samuel Marks, Can Rager, Eric J Michaud, Yonatan Belinkov, David Bau, and Aaron Mueller. Sparse feature circuits: Discovering and editing interpretable causal graphs in language models. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=I4e82CIDxv.

Mahdiyar Molahasani, Azadeh Motamedi, Michael Greenspan, Il-Min Kim, and Ali Etemad. Prism: Reducing spurious implicit biases in vision-language models with llm-guided embedding projection. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 688–697. IEEE, 2025.

Alex Oesterling, Claudio M Verdun, Carol X Long, Alexander Glynn, Lucas M Paes, Sajani Vithana, Martina Cardone, and Flavio P Calmon. Multi-group proportional representation in retrieval. *Advances in Neural Information Processing Systems*, 37:114601–114655, 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 Marina Meila and Tong Zhang (eds.), *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 8748–8763. PMLR, 18–24 Jul 2021. URL https://proceedings.mlr.press/v139/radford21a.html.

Pau Rodriguez, Arno Blaas, Michal Klein, Luca Zappella, Nicholas Apostoloff, Xavier Suau, et al. Controlling language and diffusion models by transporting activations. In *International Conference on Learning Representations*, volume 2025, pp. 89812–89855, 2025.

Shiori Sagawa, Pang Wei Koh, Tatsunori B. Hashimoto, and Percy Liang. Distributionally robust neural networks. In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=ryxGuJrFvS.

Filippo Santambrogio. Optimal transport for applied mathematicians: Calculus of variations, pdes, and modeling. 2015. URL https://api.semanticscholar.org/CorpusID:124181096.

Ashish Seth, Mayur Hemani, and Chirag Agarwal. Dear: Debiasing vision-language models with additive residuals. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6820–6829. IEEE, 2023.

Chengye Wang, Yuyuan Li, XiaoHua Feng, Xiaolin Zheng, and Chaochao Chen. Test-time debiasing with probabilistic prompts via wasserstein distance in vision-language models. In *Forty-third International Conference on Machine Learning*, 2026. URL https://openreview.net/forum?id=awggg9j40l.

Jialu Wang, Yang Liu, and Xin Wang. Are gender-neutral queries really gender-neutral? mitigating gender bias in image search. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pp. 1995–2008, 2021.

Bharat Chandra Yalavarthi, Nalini Ratha, and Venu Govindaraju. Label-free mitigation of spurious correlations in vlms using sparse autoencoders. In *International Conference on Learning Representations*, volume 2026, pp. 31820–31838, 2026.

Vladimir Zaigrajew, Hubert Baniecki, and Przemyslaw Biecek. Interpreting CLIP with hierarchical sparse autoencoders. In *Forty-second International Conference on Machine Learning*, 2025. URL https://openreview.net/forum?id=5MQQsenQBm.

Zhifei Zhang, Yang Song, and Hairong Qi. Age progression/regression by conditional adversarial autoencoder. In *2017 IEEE conference on computer vision and pattern recognition (CVPR)*, pp. 4352–4360. IEEE, 2017.

Dachuan Zhao, Weiyue Li, Zhenda Shen, Yushu Qiu, Bowen Xu, Haoyu Chen, and Yongchao Chen. Bias is a subspace, not a coordinate: A geometric rethinking of post-hoc debiasing in vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10230–10240, 2026.

CONTENTS OF THE APPENDIX

A Reproduction of the SEM benchmark **14**

B Implementation details **14**

B.1 Implementation constants 21

C Premise evidence and reconstruction fidelity of the SAE **22**

D Retrieval tables for the ResNet backbones **24**

E Qualitative retrieval examples **26**

F Zero-shot classification tables **26**

G Classification extension: method **26**

H Component attribution in BALLAST-C **28**

I Per-split ranges and the factorial retrieval ablation **31**

J Hair-color retrieval and the strength sweeps **33**

K Query-family transfer **34**

L Evaluation depth **36**

M Provenance **36**

N Additional related work **38**

O Computational and memory cost **38**

P Theoretical properties of the repair **39**

A REPRODUCTION OF THE SEM BENCHMARK

This section audits our regeneration of the baselines against the released values, configuration by configuration. We regenerated all 512 released baseline configurations under the same harness. In the per-configuration audit, 481 configurations deviate from the released values by less than 0.005, and the remaining 31 deviate by less than 0.02. These 31 are twenty-six ZSDebias and four Bend-family runs, whose training or refinement is stochastic, plus a single Orth-Cali configuration at 0.0054. The released code includes this audit: every configuration with its metric-wise deviations, its largest deviation and its band.

B IMPLEMENTATION DETAILS

This section collects the implementation choices behind Sec. 3: what BALLAST receives, returns, and fits, and what each compared method consumes (Table 5); then the gallery repair in the order of Sec. 3.3, the equalization subspace and a control that replaces it with the prompt-anchor subspace, the calibration, what it selects and a fixed configuration that bypasses it, the alignment refit, the centroids and degenerate queries, and the classification variants. Table 11 closes the section.

Table 5: Test-time information of the listed methods, per reference/evaluation split. All methods receive the task prompts and the frozen CLIP encoders; the SEM and BALLAST families also use the frozen SAE. Blocks are the information-access groups of Sec. 4.1; only the last block reads a per-image annotation. BALLAST$_{eq}$ is BALLAST with transport and amplification off; † marks the transductive variant, which also reads unlabeled margins of the evaluation gallery.

Method	Text-side inputs	Image-side inputs and what is estimated
No task-specific reference images		
Orth-Proj, Orth-Cali, RoboShot	attribute or concept prompts	none; a projection from prompts
SEM$_i$, SEM$_b$, SEM$_{bi}$	query and 10 stored paraphrases (SEM$_i$); query and 20 attribute prompts per group (SEM$_b$); both (SEM$_{bi}$); 328 diverse prompts	none; per-query atom scores
Unlabeled reference images		
BALLAST, BALLAST$_{eq}$	20 attribute prompts per group; first query family and 5 stored paraphrases; diverse prompts	unlabeled reference fold, gallery repaired once; soft group weights, atom selection, quantile tables, four calibrated choices, group centroids, one projection
BALLAST-C†	as BALLAST, plus the SEM$_b$ or SEM$_{bi}$ prompt sets for the class prompts	as BALLAST, plus unlabeled margins of the evaluation gallery (TRIM translations)
Group-labeled reference images		
BendVLM, BendSEM$_{bi}$	query (and the SEM inputs)	group-labeled reference images; test-time query refinement

Algorithm 1 BALLAST: fit once per reference gallery, reuse across query families

Require: reference embeddings $\{x_n\}$, attribute prompts, task prompts, diverse prompts (alignment shift)

1: $w_{n,g} \leftarrow$ soft membership weights from the attribute prompts; $\pi_g \leftarrow \frac{1}{N}\sum_n w_{n,g}$
2: $z_n \leftarrow \mathrm{ReLU}(W_e(x_n - b_{\mathrm{pre}}) + b_e)$
3: select atoms: between-group W_1 above a max-of-8 shuffle screen; if < 8 pass, augment with the 64 highest-cost eligible atoms
4: **for** each selected atom j and group g **do**
5: build $Q_{g,j}$ (zero-inflated, weighted); $\bar{Q}_j \leftarrow \sum_g \pi_g Q_{g,j}$
6: **end for**
7: calibrate $(\alpha, q, \beta, \mathrm{align})$ label-free on a held-out slice; refit the tables on the full reference fold
8: $c_g \leftarrow$ soft centroids of the repaired reference gallery; $\mathcal{B} \leftarrow \mathrm{span}\{c_g - c_h\}$
apply to gallery: $z'_{n,j} \leftarrow (1-a_j)z_{n,j} + a_j \sum_g w_{n,g} T_{g,j}(z_{n,j})$ on transported atoms; $z'_{n,j} \leftarrow (1+\beta)z_{n,j}$ on protected atoms
$x'_n \leftarrow \mathrm{normalize}(x_n + W_d(z'_n - z_n))$
apply to query: normalize t, remove its $\mathcal{B}$-component, rescale to the original norm if nonzero

Inputs, outputs, and what is fitted where. Table 6 lists every prompt set BALLAST reads, with its count, its role, and verbatim examples, and Table 5 lists what each compared method consumes. Soft group weights, atom selection, transport tables, and centroids are estimated on the reference gallery; the four image-side choices are calibrated on a held-out slice of it with one seed query family (Sec. 3.5), and the fitted state is then held fixed for every later query family. Evaluation annotations are used only to compute metrics.

Algorithm 1 summarizes the fit and the two application steps with the symbols of Sec. 3, and Figure 3 illustrates the gallery repair on one atom with a toy example: the transport of each group's activations toward the barycenter, the placement of exact zeros, and the soft blend of the group targets that is decoded back. The paragraphs that follow fix the conventions behind each of these steps.

Group distributions and quantile tables. For the atoms that the selection below retains, the group quantile functions are estimated on the full reference gallery with the soft weights. The estimate uses a 64-knot grid on $u \in [0.002, 0.998]$ under the midpoint convention, which takes the cumulative weight up to the middle of each sample. It interpolates linearly between order statistics and is clamped outside the knot range.

Strength and atom selection. The released checkpoints realize the Matryoshka hierarchy as nested top-k sparsity levels used during training, not as nested dictionary prefixes. At inference the code is a plain ReLU, so no static coarse dictionary exists. We therefore quantify how coarse an atom is by its strength, the fraction of a 4,096-image reference subsample (seed 1) in which the atom is among the sample's top-k_s activations. Here k_s is the coarsest training level of the checkpoint, which is 256, 256, 512, and 128 for ViT-B/16, ViT-L/14@336px, RN50, and RN101.

Table 6: The prompt resources of the benchmark and how BALLAST reads them. Counts are read from the released template files; examples are verbatim. No prompt names an image; the only per-image information BALLAST touches is the unlabeled reference gallery.

Prompt set	Role	Count	Examples
Stereotype query families	The task prompt of every stereotype retrieval cell; the first family, *addict*, seeds the label-free calibration, and the other 24 are evaluated against the cached fit	25 families, one seed query each shared by the FairFace, UTKFace and CelebA cells	*"A photo of an addict."*, *"A photo of a criminal."*, *"A photo of a hacker."*, *"A photo of a thug."*
Hair-color query families	The task prompts of the two hair cells, the one retrieval task with relevance labels; *blond* seeds the calibration	4 families	*"A photo of a celebrity with blond hair."*, *"A photo of a celebrity with gray hair."*
Stored paraphrases	Each query family carries ten stored entries, the seed and nine paraphrases. SEM_i and SEM_{bi} read all ten; BALLAST reads the seed and the first five stored entries of the calibration family as its task prompts	10 per family 250 stereotype, 40 hair	for *addict*: *"An image of an addict."*, *"A person struggling with addiction."*, *"A person with a substance abuse problem."*
Attribute prompts	Name the values of the protected attribute. Their mean text embedding per group is the anchor from which BALLAST computes the soft membership weights (Sec. 3.2); SEM_b and SEM_{bi} score atoms with the same sentences	20 per group gender 40 FairFace race 140 (7 groups) UTKFace race 100 (5 groups) Waterbirds background 40	*"A portrait of a man."*, *"A portrait of a woman."*; *"A close-up of a Black person's face, smiling."*; *"A portrait of an Asian man."*; *"A photo featuring a grassy field as the background."*, *"The ocean as the dominant background."*
Class prompts	The task prompts of the classification cells, each with ten stored entries; BALLAST-C applies SEM's text edit to them and BALLAST reads the seed and the first five entries	2 classes per cell 10 entries each	*"A photo of a celebrity with dark hair."* / *"A photo of a celebrity with blond hair."*; *"This is a picture of a landbird."* / *"This is a picture of a waterbird."*; paraphrases such as *"A famous person with dark hair."*, *"A photo of a waterfowl."*
Diverse prompt bank	Task-neutral sentences released with SEM. Their mean text embedding is the shift of the alignment switch (Sec. 3.2); SEM uses their activations as the reference against which it scores atoms	328 sentences	*"A firefighter in full gear holding a water hose."*, *"A chef meticulously plating a gourmet dish in a kitchen."*, *"A Stone Age person painting on a cave wall."*
Baseline-specific sets	Stored per family for the prompt-derived baselines and never read by BALLAST: spurious prompts (Orth-Proj, Orth-Cali, PRISM-mini, BendVLM), candidate prompts (Orth-Cali, BendVLM), and harmful and helpful prompt pairs (RoboShot)	one spurious prompt per group 4 to 6 harmful and 3 to 5 helpful pairs per family	*"A photo of a male person."* / *"A photo of a female person."*; *"A photo of a male addict."*; RoboShot: *"a photo of a poor addict"* against *"a photo of a wealthy addict"*

The strength enters the transport through the gate of Eq. 2. The strength-rank percentile r_j of atom j is the rank of its strength in descending order, divided by s, and the gate is $m_j = \sigma\big((q - r_j)/0.02\big)$. The sigmoid uses a rank-percentile scale of 0.02. Atoms with r_j well below q receive nearly the full strength α, those with r_j well above q receive almost none, and at $r_j = q$ the strength is $\alpha/2$. The "No strength tier" ablation sets $m_j = 1$.

Selection compares the between-group cost of each atom with a shuffle null. The selection statistics are computed on the same subsample with a coarser 32-knot quantile table. The cost of atom j is the prior-weighted pairwise W_1 distance $b_j = \sum_{g<h} \pi_g \pi_h \, W_1(Q_{g,j}, Q_{h,j})$, with W_1 evaluated as the mean absolute difference of the two quantile tables. The null is the maximum of b_j over eight random shuffles of the group weights across images (seeds 100 to 107). An atom is eligible if its gate at the largest strength tier of the calibration grid ($q = 0.25$, below) exceeds 0.05, and it is selected if it is eligible and b_j exceeds its null. At most $4{,}096$ atoms are kept, ranked by cost times strength, and if fewer than eight atoms pass, the 64 highest-cost eligible atoms are forced in. The screen uses W_1, which weights activation gaps linearly and is less dominated by the tail of a zero-inflated distribution than the quadratic W_2 cost; the target is defined by W_2 because the barycenter is the W_2-optimal common distribution. A selected atom whose pseudo-groups do not differ is left unchanged by the transport, so the screen decides which atoms the repair may move, not the correctness of the repair (App. P, Cor. 1).

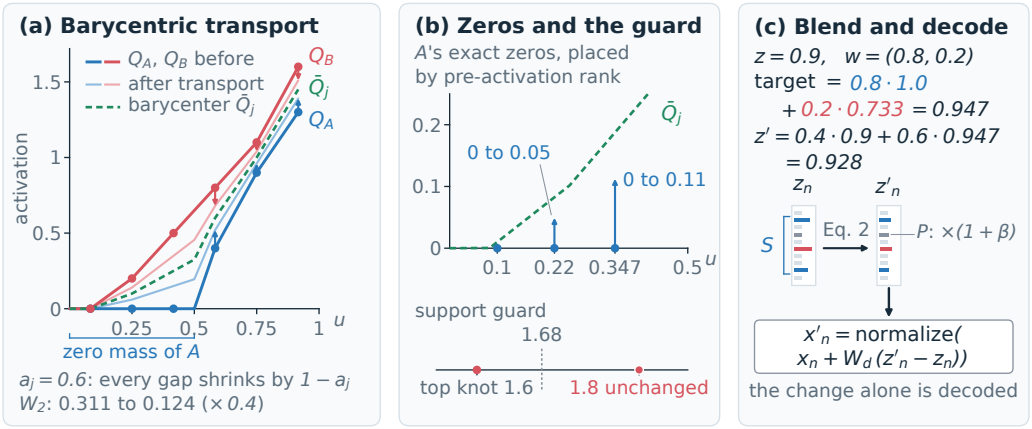


Figure 3: Partial barycentric transport on one SAE atom (toy example). (a) Transport toward the prior-weighted barycenter contracts the between-group W_2 distance by the factor $1 - a_j$ under the idealized hard-group coupling (App. P). (b) The rank of the dense pre-activation places each exact zero within its group's zero mass, which allows new activations; values beyond the support guard remain unchanged. (c) Soft memberships blend the group targets (Eq. 2), and only the decoded changes are added to the original embedding before normalization. The exact contraction in (a) is not guaranteed for this soft blend.

Protected atoms. The protected set is built from the task prompts, that is, the query and its stored paraphrases or the class prompts. We take the 50 atoms with the largest activation over these prompts, and those with cost below the median cost of the selected atoms form the protected set. For classification, the 64 atoms with the largest spread of the class-prompt codes, which we call the class-contrast atoms, are added. Protected atoms are removed from the transported set and amplified instead (below). In retrieval the set is small (zero atoms in 23 of the 120 ViT splits, at most thirteen), and the factorial ablation finds almost no aggregate effect (App. I); in classification it consists mainly of the 64 class-contrast atoms, and it matters on one cell, Waterbirds RN50 (App. H).

Transport application. Two conventions complete the application of Eq. 2. Exact zeros are handled separately: an exact zero is given a quantile position inside the zero mass of its group by the rank of its dense pre-activation among the zero samples of that group, taken from a 16-knot table and scaled by the plateau edge of the group's quantile table. Zero samples therefore never rank above barely positive ones, and the per-group map is monotone at zero. At application time, an activation above 1.05 times the top fitted knot over groups lies outside the estimated support and keeps its value.

Reconstruction and amplification. The repaired embedding adds the decoded change $W_d(z'_n - z_n)$ to the original embedding and renormalizes (Algorithm 1). Protected atoms are amplified on the image side by the factor $1 + \beta$ through the same decoder difference. The amplification follows SEM's content-amplification practice: task atoms that the discrepancy screen does not flag are scaled up so that the task direction dominates the residual bias.

Where the equalization subspace comes from. To see why the source of the equalization subspace matters (Sec. 3.4), Table 8 compares three subspaces with the image group direction on the fold-0 reference half of every retrieval cell. The image group direction is the direction spanned by the annotated group-mean differences of the L2-normalized reference embeddings. The annotations enter this diagnostic only. The first subspace is the soft-centroid subspace $\mathcal{B}$ of BALLAST$_{eq}$. The second is the span of the differences of the prompt anchors $\bar{t}_g$ that the method already computes, which gives a projection with no images and no fit. The third is the span of the raw spurious-prompt embeddings that the benchmark's Orth-Proj removes.

The centroid subspace is aligned with the image group direction at a cosine of 0.92 to 0.999 (median 0.997) and removes a median 2% of a unit query's norm. The prompt-anchor difference is aligned at 0.44 to 0.77 and removes 12% to 46%. Orth-Proj's prompt span is aligned at 0.33 to 0.68 and removes 82% to 90% of every query, since the raw prompt embeddings share the direction of the query itself. Text-derived difference directions thus sit only partly along the image group direction

Table 7: Knobs selected by the calibration and sizes of the fitted sets for the 32 BALLAST configurations of the main run, read from the run logs. Left: on how many of the ten splits the calibration selected $\alpha = 0$ (no transport), the identity ($\alpha = 0$ and $\beta = 0$), a positive boost $\beta > 0$, and the alignment step. Right: size of the fitted support set $|S|$ and of its boosted (protected) subset, minimum to maximum over the ten splits. Classification cells are marked (cls.).

	Splits selecting (of 10)				Size, min to max	
Cell	$\alpha = 0$	Identity	$\beta > 0$	Align.	$\|S\|$	Protected
ViT-L/14@336px						
FairFace race	0	0	0	9	1430 to 4096	0 to 0
FairFace gender	0	0	4	8	1028 to 3592	1 to 4
UTKFace race	0	0	6	9	2477 to 4096	1 to 2
UTKFace gender	0	0	7	10	4096 to 4096	2 to 5
CelebA stereotype	8	5	4	0	1970 to 2119	0 to 1
CelebA hair	0	0	1	1	1970 to 4017	0 to 1
CelebA blond hair (cls.)	0	0	0	10	3889 to 4096	64 to 65
Waterbirds (cls.)	0	0	7	1	1498 to 2999	65 to 66
ViT-B/16						
FairFace race	1	0	5	0	2353 to 2500	2 to 3
FairFace gender	3	1	7	4	1750 to 4058	8 to 12
UTKFace race	2	0	7	5	3970 to 4096	2 to 4
UTKFace gender	0	0	6	1	3110 to 3816	5 to 9
CelebA stereotype	10	2	8	4	3558 to 3945	4 to 13
CelebA hair	0	0	3	8	3588 to 3951	3 to 10
CelebA blond hair (cls.)	0	0	0	7	3516 to 3951	67 to 70
Waterbirds (cls.)	1	1	0	9	1793 to 3278	64 to 72
RN50						
FairFace race	9	9	0	0	4096 to 4096	0 to 0
FairFace gender	0	0	0	4	3903 to 4008	0 to 0
UTKFace race	10	10	0	0	4096 to 4096	0 to 0
UTKFace gender	9	9	0	1	3968 to 4096	0 to 0
CelebA stereotype	0	0	0	10	4096 to 4096	0 to 1
CelebA hair	10	10	0	0	4096 to 4096	0 to 1
CelebA blond hair (cls.)	10	10	0	0	4096 to 4096	64 to 65
Waterbirds (cls.)	0	0	6	8	3976 to 4068	64 to 64
RN101						
FairFace race	2	2	1	1	3908 to 4096	1 to 3
FairFace gender	0	0	1	6	3666 to 4096	1 to 4
UTKFace race	1	1	0	0	4096 to 4096	3 to 3
UTKFace gender	7	3	5	0	4096 to 4096	3 to 4
CelebA stereotype	2	2	1	0	4096 to 4096	2 to 4
CelebA hair	0	0	0	0	4096 to 4096	0 to 0
CelebA blond hair (cls.)	0	0	0	0	4096 to 4096	66 to 66
Waterbirds (cls.)	0	0	0	10	3618 to 3662	64 to 64

and carry a large query-aligned component, which is what an image-derived subspace avoids. The table is a geometric diagnostic and not a retrieval comparison.

The retained rank of the centroid subspace is capped at the exact rank $G-1$, so the table shows $G-1$ in every cell. Without the cap, the numerical rank at the SVD cutoff was one higher in two cells of this diagnostic, UTKFace race ViT-B/16 with five groups and FairFace gender RN101 with two, where the smallest singular value sat at a relative 1.07×10^{-6}, just above the cutoff; the paragraph on the centroids below gives the effect.

Anchor-difference projection as a control. The diagnostic above compares geometries; Table 9 compares outcomes. It applies the query projection of Sec. 3.4 with the prompt-anchor difference subspace in place of the gallery centroid subspace, at the same strength and with no reference images, on all 24 retrieval cells. The anchor projection raises KL above base CLIP on all 24 cells and MaxSkew on 20 of them; its mean KL over the cells is 0.554 against 0.184 for base CLIP. BALLAST$_{eq}$ is lower than the anchor projection on all 48 comparisons, one per cell and metric. The diagnostic above shows why: the prompt-derived subspace is aligned with the image group direction at only 0.44 to 0.77 and removes 12% to 46% of the unit query, so the projection moves the query along directions that the gallery does not share.

Calibration. The calibration searches a fixed grid. The candidate sets are $\alpha \in \{0, 0.5, 0.75, 1\}$, $q \in \{0.06, 0.12, 0.25\}$, $\beta \in \{0, 0.1, 0.2, 0.3\}$, and the alignment switch $\{\text{off}, \text{on}\}$, which gives 96 candidates in all. The held-out slice is 20% of the reference gallery, at most 3,000 images (seed 0).

Table 8: Alignment of three query-side subspaces with the annotated image group direction, per retrieval cell, on the fold 0 reference half (text encoder in fp32 on CPU, image embeddings from disk; centroid subspace of BALLAST$_{eq}$ from at most 20000 reference images, seed 2). The subspaces and the image direction are defined in the paragraph above; only the image direction uses the annotated labels. Left: absolute cosine between each subspace and the image direction (for the race cells the mean over the $G-1$ principal angle cosines), with the rank retained at the 10^{-6} cutoff in parentheses. Right: mean over the cell's queries of the norm of the unit query's component in the subspace, which is what the equalization removes.

	Cosine with image direction			$\|Pq\|$, norm removed from unit query		
Cell	Centroid	Anchor	Orth-Proj	Centroid	Anchor	Orth-Proj
ViT-L/14@336px						
FairFace race	0.962 (6)	0.579	0.456	0.055	0.368	0.846
FairFace gender	0.994 (1)	0.726	0.647	0.021	0.140	0.821
UTKFace race	0.942 (4)	0.569	0.445	0.051	0.206	0.839
UTKFace gender	0.999 (1)	0.730	0.651	0.010	0.140	0.821
CelebA stereotype	0.999 (1)	0.698	0.612	0.021	0.140	0.821
CelebA hair	0.999 (1)	0.698	0.640	0.013	0.122	0.836
ViT-B/16						
FairFace race	0.960 (6)	0.568	0.445	0.042	0.457	0.890
FairFace gender	0.995 (1)	0.772	0.681	0.022	0.175	0.893
UTKFace race	0.948 (4)	0.566	0.453	0.029	0.247	0.885
UTKFace gender	0.998 (1)	0.755	0.663	0.022	0.175	0.893
CelebA stereotype	0.999 (1)	0.731	0.656	0.026	0.175	0.893
CelebA hair	0.999 (1)	0.731	0.671	0.014	0.115	0.890
RN50						
FairFace race	0.935 (6)	0.443	0.334	0.039	0.445	0.876
FairFace gender	0.994 (1)	0.628	0.513	0.021	0.208	0.860
UTKFace race	0.930 (4)	0.441	0.342	0.029	0.274	0.873
UTKFace gender	0.998 (1)	0.624	0.518	0.016	0.208	0.860
CelebA stereotype	0.999 (1)	0.611	0.513	0.017	0.208	0.860
CelebA hair	0.999 (1)	0.611	0.552	0.006	0.150	0.901
RN101						
FairFace race	0.934 (6)	0.547	0.458	0.051	0.337	0.856
FairFace gender	0.994 (1)	0.754	0.494	0.025	0.145	0.836
UTKFace race	0.925 (4)	0.530	0.451	0.038	0.196	0.855
UTKFace gender	0.998 (1)	0.753	0.518	0.019	0.145	0.836
CelebA stereotype	0.999 (1)	0.711	0.492	0.017	0.145	0.836
CelebA hair	0.999 (1)	0.711	0.665	0.013	0.135	0.874

The objective is a pseudo-imbalance computed on this slice. Each held-out image is scored by its maximum similarity over the task prompts, and the top $k = \max(10, \min(500, \lfloor N_{\text{held}}/4 \rfloor))$ images form one set. The mean soft weights $\hat{s}_g$ of this set give the pseudo-imbalance $\max_g |\log(\hat{s}_g/\pi_g)|$. The objective is thus the two-sided form of the MaxSkew functional of Sec. 4.1, the largest absolute log ratio, with soft pseudo-labels at the depth of the evaluation ($k = 500$ on the three face datasets, 294 on Waterbirds).

Feasibility imposes three requirements. The mean fidelity must satisfy $\cos(x, x') \geq 0.985$. The agreement with the unrepaired predictions must be at least 0.98 on the images whose prediction margin is at or above the 60th percentile. The mean top-k overlap between the rankings before and after repair must be at least 0.6. The agreement floor compares the argmax over the task prompts before and after repair, which in retrieval means the seed query and its paraphrases. Among feasible candidates the lowest pseudo-imbalance wins, with fidelity as the tie-breaker.

The classification cells use the same objective, with the two class prompts and their five stored paraphrases as the prompt set. No separate classification objective was used in any reported run.

What the calibration selects. Table 7 lists per cell how often each candidate was selected over the ten splits of the main run: the calibration chose the identity in 38 of the 60 RN50 retrieval splits and $\alpha = 0$ in all ten CelebA-stereotype ViT-B/16 splits, and the protected set holds between zero and thirteen atoms in the retrieval cells and 64 to 72 in classification.

A fixed configuration. Table 10 runs BALLAST with every calibrated knob pinned to its most frequently selected value over the 120 ViT retrieval splits of the main run ($\alpha = 1$, $q = 0.06$, $\beta = 0$, alignment off), so that no split is tuned against the proxy. On the twelve ViT retrieval cells the fixed configuration stays below Best SEM on all 24 comparisons. It tracks the calibrated method in aggregate: its mean KL is 0.034 against the calibrated method's 0.033, and its mean MaxSkew is 0.224 against 0.223. It is lower than the calibrated method on six of the twelve cells for KL and on

Table 9: Anchor-difference projection control on the 24 retrieval cells (KL and MaxSkew at $k = 500$, lower is better). Anchor: the query projection of BALLAST$_{eq}$ applied with the prompt-anchor difference subspace (the anchor subspace of Table 8) in place of the gallery centroid subspace, at the same strength and with no reference images. BALLAST$_{eq}$, BALLAST and Best SEM (the best SEM variant per cell and metric) are the values of the main tables. No ranking marks.

	Base CLIP		Anchor		BALLAST$_{eq}$		BALLAST		Best SEM	
Cell	KL	MS	KL	MS	KL	MS	KL	MS	KL	MS
ViT-L/14@336px										
FairFace Race	0.244	0.798	0.622	1.274	0.036	0.357	0.034	0.354	0.147	0.623
FairFace Gender	0.114	0.326	0.666	0.736	0.039	0.212	0.029	0.182	0.097	0.298
UTKFace Race	0.124	0.461	0.364	0.990	0.017	0.213	0.014	0.205	0.058	0.420
UTKFace Gender	0.040	0.186	0.583	0.701	0.007	0.088	0.008	0.088	0.033	0.186
CelebA Stereo.	0.237	0.536	0.379	0.495	0.034	0.229	0.033	0.228	0.153	0.413
CelebA Hair	0.148	0.359	0.458	0.519	0.097	0.323	0.070	0.294	0.135	0.365
ViT-B/16										
FairFace Race	0.237	0.795	0.722	1.483	0.038	0.352	0.034	0.350	0.170	0.690
FairFace Gender	0.139	0.346	0.712	0.747	0.048	0.235	0.038	0.205	0.087	0.255
UTKFace Race	0.125	0.476	0.276	0.857	0.035	0.235	0.032	0.225	0.096	0.407
UTKFace Gender	0.134	0.321	0.586	0.699	0.011	0.109	0.013	0.115	0.064	0.241
CelebA Stereo.	0.314	0.555	0.426	0.512	0.050	0.227	0.049	0.224	0.173	0.443
CelebA Hair	0.179	0.409	0.516	0.536	0.043	0.231	0.042	0.208	0.172	0.344
RN50										
FairFace Race	0.215	0.734	0.712	1.485	0.044	0.390	0.045	0.393	0.126	0.562
FairFace Gender	0.170	0.351	0.718	0.749	0.011	0.107	0.012	0.113	0.031	0.190
UTKFace Race	0.127	0.477	0.505	1.064	0.035	0.243	0.035	0.243	0.039	0.265
UTKFace Gender	0.153	0.340	0.656	0.723	0.012	0.109	0.011	0.109	0.075	0.280
CelebA Stereo.	0.390	0.623	0.455	0.521	0.118	0.388	0.105	0.357	0.050	0.193
CelebA Hair	0.186	0.367	0.515	0.535	0.101	0.240	0.101	0.240	0.246	0.369
RN101										
FairFace Race	0.203	0.744	0.756	1.388	0.060	0.466	0.061	0.464	0.111	0.566
FairFace Gender	0.144	0.335	0.623	0.719	0.011	0.107	0.010	0.104	0.037	0.202
UTKFace Race	0.153	0.497	0.570	1.187	0.044	0.260	0.041	0.268	0.084	0.341
UTKFace Gender	0.136	0.333	0.568	0.693	0.012	0.114	0.012	0.118	0.041	0.215
CelebA Stereo.	0.300	0.559	0.416	0.507	0.099	0.354	0.099	0.346	0.041	0.186
CelebA Hair	0.204	0.413	0.482	0.526	0.074	0.268	0.062	0.247	0.232	0.445

Table 10: BALLAST with every calibrated knob pinned to one setting on the twelve ViT retrieval cells (KL and MaxSkew at $k = 500$, lower is better): transport strength $\alpha = 1$, tier $q = 0.06$, amplification $\beta = 0$, alignment off, the most frequently selected value of each knob over the 120 ViT retrieval splits of the main run. The calibrated BALLAST and BALLAST$_{eq}$ and Best SEM are the values of the main tables. No ranking marks.

	Fixed BALLAST		BALLAST		BALLAST$_{eq}$		Best SEM	
Cell	KL	MS	KL	MS	KL	MS	KL	MS
ViT-L/14@336px								
FairFace Race	0.035	0.355	0.034	0.354	0.036	0.357	0.147	0.623
FairFace Gender	0.033	0.197	0.029	0.182	0.039	0.212	0.097	0.298
UTKFace Race	0.016	0.222	0.014	0.205	0.017	0.213	0.058	0.420
UTKFace Gender	0.006	0.082	0.008	0.088	0.007	0.088	0.033	0.186
CelebA Stereo.	0.028	0.208	0.033	0.228	0.034	0.229	0.153	0.413
CelebA Hair	0.073	0.287	0.070	0.294	0.097	0.323	0.135	0.365
ViT-B/16								
FairFace Race	0.033	0.352	0.034	0.350	0.038	0.352	0.170	0.690
FairFace Gender	0.022	0.155	0.038	0.205	0.048	0.235	0.087	0.255
UTKFace Race	0.030	0.221	0.032	0.225	0.035	0.235	0.096	0.407
UTKFace Gender	0.012	0.113	0.013	0.115	0.011	0.109	0.064	0.241
CelebA Stereo.	0.069	0.299	0.049	0.224	0.050	0.227	0.173	0.443
CelebA Hair	0.046	0.198	0.042	0.208	0.043	0.231	0.172	0.344

seven for MaxSkew. The calibration therefore adjusts the operating point per split; the gains do not come from it.

Alignment shift and refit. When the alignment switch is on, image embeddings are shifted before encoding by the difference between the mean embedding of the 328 diverse prompts of SEM and the reference-fold image mean. The shift enters only the computation of the code. After calibration, the transport tables are refitted on the full reference gallery with the selected alignment.

Centroids and the retained rank. The soft centroids are computed on at most 20,000 reference images (seed 2). The subspace $\mathcal{B}$ is orthonormalized with a singular-value cutoff of 10^{-6} times the largest singular value and its rank is capped at $G - 1$, the exact rank of the centroid-difference matrix; the equalization strength is 1.

Table 11: Fixed scientific hyperparameters of the reported BALLAST and BALLAST-C runs, grouped by mechanism in pipeline order; an ablation overrides only its named mechanism.

Constant	Value
Soft groups, screen and repair	
Membership weights; tier gate	$\tau = 0.03$; width 0.02
Quantile grids (transport / screen / zero-rank)	$64/32/16$ knots on $u \in [0.002, 0.998]$
Atom screen	$N_s = \min(N, 4096)$ (seed 1); $\lvert S\rvert \le 4096$; 8 shuffles (seeds 100 to 107); eligibility $m_j(q_{\max}{=}0.25) > 0.05$
Fallback and support	if $\lvert S\rvert < 8$, add top-64 eligible; skip if $z_j > 1.05 \max_g Q_{g,j}(0.998)$
Protected atoms	≤ 50 task atoms; 64 class-contrast atoms for classification
Calibration	
Calibration sample and k	$N_h = \min(\lfloor 0.2N \rfloor, 3000)$ (seed 0); $k = \max(10, \min(500, \lfloor N_h/4 \rfloor))$
Calibration rails (fidelity / agreement / overlap)	$0.985/0.98/0.60$; agreement at margin $\ge$ 60th percentile
Alignment bank	mean of 328 fixed diverse prompts
Task prompts at fit	retrieval: first query family, query + 5 stored paraphrases (6 entries, 5 distinct); classification: both class prompts + 5 paraphrases each
Query equalization	
Centroids and equalization	$\le$ 20,000 images (seed 2); strength 1; SVD cutoff $10^{-6}\sigma_{\max}$
Sparse autoencoder	
SAE width; coarsest k_s	16,384; $256/256/512/128$ (ViT-B/16, ViT-L/14@336px, RN50, RN101)
Classification and BALLAST-C	
Classification text edit	class-pull scale 1 (grid disabled); text-side class boost 0 (grid empty)
TRIM	25 iterations; $\sigma_y \ge 10^{-3}$; $\pi_{y\lvert g} \ge 10^{-4}$

The cap matters because the cutoff sits at the level of float32 rounding. In exact arithmetic the smallest singular value of the centroid-difference matrix is zero; in float32 it is rounding noise of relative size about 10^{-6}, the order of the cutoff, so without the cap the retained rank could be $G-1$ or G depending on the device's SVD. A retained extra direction is not innocuous, because it is also removed from the query. In the diagnostic of Table 8, the two cells that retained the direction lost 0.355 and 0.378 of the unit query's norm before the cap, against 0.006 to 0.055 in the other 22 cells. With the cap, the same two cells lose 0.025 and 0.029. In the per-split fits of the main run, 24 cells times ten splits for each variant, the uncapped cutoff had kept the extra direction in 39 of the 240 BALLAST fits, spread over 15 of the 24 cells, and in 26 of the 240 BALLAST$_{eq}$ fits, so every retrieval configuration that uses the equalization was rerun under the cap (App. M). The run logs record the retained rank, and the cap removes the only source of device dependence we found in the equalization path.

Degenerate queries. The degenerate case of Sec. 3.4 is $t \in \mathcal{B}$, a query lying entirely in a subspace of dimension at most $G-1$. In this case the projected direction is zero. The implementation does not guard against it: the "if nonzero" of Algorithm 1 names this hypothesis rather than a guard. A safe fallback would return the original query.

Classification variants. Plain BALLAST pulls the class prompts toward the barycenter means on the transported atoms, with the calibrated (α, q) gates at a fixed multiplier of 1.0. The class-contrast atoms are protected from this pull, and no text-side amplification is applied.

BALLAST-C uses the SEM text edit, with SEM$_b$ on CelebA and SEM$_{bi}$ on Waterbirds, and the classifier direction $\delta = t_A - t_B$ of the normalized edited prompts. Its TRIM step runs 25 alternating updates of class responsibilities, class parameters, translations, and prevalences from a median-split initialization, with the lower bounds $\sigma_y \ge 10^{-3}$ and $\pi_{y|g} \ge 10^{-4}$.

The attribution variants of App. H are two versions of BALLAST-C with pinned candidate sets. The first, the "gallery edit off" row of the tables, pins the candidate sets of both α and β to $\{0\}$, so the text edit and TRIM operate on exactly the normalized unrepaired gallery. The second, the "transport off, amp. kept" row, pins α to $\{0\}$ and calibrates β as usual. Both are recalibrated by the same procedure.

B.1 Implementation constants

Table 11 lists the fixed hyperparameters that bear on the reported results. All backbones share the same grids and defaults apart from the backbone-specific SAE checkpoint. Numerical-stability epsilons and batching constants follow the release.

Table 12: Confidence of the soft membership weights and agreement of the pseudo-prior with the annotated prior, on the fold 0 reference half (up to 20,000 images per setting, the same subsample for every backbone) at $\tau = 0.03$. Acc.: agreement of the argmax weight with the annotated label. Mean max: mean of the largest weight per image. Below 0.6: fraction of images whose largest weight is under 0.6. Prior gap: largest absolute difference, in percentage points, between the pseudo-prior (mean weight per group) and the annotated group frequency of the same images. Annotated labels serve this evaluation only; BALLAST never reads them.

Backbone	Acc.↑	Mean max↑	Below 0.6↓	Prior gap (pt)↓
FairFace race (7 groups)				
ViT-L/14@336px	0.694	0.386	0.972	5.2
ViT-B/16	0.671	0.334	0.998	6.3
RN50	0.601	0.297	1.000	5.9
RN101	0.627	0.313	0.997	6.9
FairFace gender (2 groups)				
ViT-L/14@336px	0.963	0.763	0.047	3.5
ViT-B/16	0.951	0.747	0.082	5.5
RN50	0.907	0.703	0.182	4.7
RN101	0.923	0.712	0.152	6.9
UTKFace race (5 groups)				
ViT-L/14@336px	0.853	0.458	0.889	18.7
ViT-B/16	0.820	0.409	0.935	20.1
RN50	0.779	0.371	0.976	18.9
RN101	0.793	0.391	0.976	19.8
UTKFace gender (2 groups)				
ViT-L/14@336px	0.966	0.781	0.044	3.5
ViT-B/16	0.961	0.752	0.066	4.8
RN50	0.936	0.730	0.118	2.3
RN101	0.946	0.734	0.101	6.6
CelebA gender (2 groups)				
ViT-L/14@336px	0.992	0.836	0.005	1.1
ViT-B/16	0.991	0.821	0.006	0.6
RN50	0.988	0.797	0.013	0.1
RN101	0.989	0.789	0.011	1.3

C Premise evidence and reconstruction fidelity of the SAE

This section first reports the premise study (Figure 4), then measures how far the soft weights are from hard assignments (Table 12), and finally measures how well the frozen SAE reconstructs image embeddings (Table 13) and states what this fidelity means for the method.

Premise study. Figure 4 reports the premise study discussed in Sec. 3.2. Its first measurement is the share of the between-group W_1 mass that falls on the strongest atoms, computed on ViT-B/16 with the fold-0 reference set and annotated groups. Across the five settings, the $1{,}024$ strongest of $16{,}384$ atoms carry 38% to 85% of this mass: 38% on FairFace gender, 41% on UTKFace race, 43% on UTKFace gender, 84% on Waterbirds, and 85% on FairFace race. On the gender settings, these strongest 6% of atoms therefore carry a disproportionate share of the discrepancy, not the bulk of it. The second measurement concerns the pseudo-groups: prompt-derived soft groups computed from SAE codes are within 0.5 to 5.5 accuracy points of dense zero-shot grouping, measured against the annotated groups.

Confidence of the soft weights. Table 12 measures how far the soft weights of Sec. 3.2 are from hard assignments on all four backbones. On the gender settings the mean maximum weight is about 0.75 (0.82 on CelebA), with almost no image above 0.9. On the race settings it is 0.33 on FairFace, with seven groups, and 0.41 on UTKFace, with five groups. Three consequences follow. First, on the race settings the weighted group distributions that the transport aligns are mild re-weightings of the pooled distribution. Second, the slack of Thm. 1(iii) is not small. Third, the pseudo-prior π stays close to uniform on the race settings. This matches FairFace's near-uniform annotated composition but not UTKFace's. For UTKFace's largest group the pseudo-share is 0.22 to 0.24 against an annotated 0.42 on every backbone, whereas on the gender settings the pseudo-prior is within 7 points of the annotated prior. That the pseudo-prior stays close to uniform also explains why the "Uniform barycenter" ablation equals the prior-weighted barycenter in retrieval (Table 22).

Reconstruction of image embeddings. The frozen MSAE checkpoints were trained on caption embeddings, and Table 13 measures their standalone reconstruction of CLIP image embeddings. For every image x of the fold-0 reference set of each dataset we compute the plain ReLU code and its decoding $\hat{x} = W_d z + b_{\text{pre}}$, with the encoder and decoder exactly as the method uses them,

Table 13: Reconstruction fidelity of the frozen caption-trained MSAE on CLIP image embeddings (fold-0 reference set of each dataset, plain ReLU code; computation as in the paragraph above). Cosine between an embedding and its reconstruction, mean and 5th percentile, without (raw) and with (aligned) the alignment shift of the calibration, applied before encoding and to the compared embedding. Active atoms: mean number of nonzero code coordinates per input. The last row of each block gives the same statistics on the 328 diverse text prompts, the modality the SAE was trained on (n/a: the shift is not applied to text). No label is read and nothing is fitted.

		Cosine, raw		Cosine, aligned		Active atoms	
Input	N	Mean	5th pct	Mean	5th pct	Raw	Aligned
ViT-L/14@336px							
FairFace	48,849	0.781	0.736	0.903	0.525	565	1,398
UTKFace	11,852	0.801	0.736	0.873	0.429	608	1,743
CelebA	101,299	0.844	0.778	0.902	0.597	615	1,356
Waterbirds	5,894	0.847	0.802	0.935	0.656	632	1,084
Diverse text prompts	328	0.998	0.996	n/a	n/a	586	n/a
ViT-B/16							
FairFace	48,849	0.797	0.752	0.872	0.581	872	1,003
UTKFace	11,852	0.784	0.696	0.866	0.554	1,218	1,050
CelebA	101,299	0.791	0.690	0.872	0.640	1,151	953
Waterbirds	5,894	0.826	0.773	0.921	0.698	907	740
Diverse text prompts	328	0.996	0.994	n/a	n/a	455	n/a
RN50							
FairFace	48,849	0.583	0.536	0.908	0.902	11,055	12,762
UTKFace	11,852	0.582	0.537	0.908	0.901	10,923	12,760
CelebA	101,299	0.621	0.582	0.908	0.902	10,681	12,761
Waterbirds	5,894	0.672	0.629	0.909	0.904	9,893	12,762
Diverse text prompts	328	0.909	0.868	n/a	n/a	12,617	n/a
RN101							
FairFace	48,849	0.675	0.648	0.856	0.848	11,527	11,240
UTKFace	11,852	0.668	0.642	0.855	0.847	11,534	11,241
CelebA	101,299	0.678	0.652	0.852	0.844	11,547	11,244
Waterbirds	5,894	0.766	0.741	0.858	0.852	11,469	11,242
Diverse text prompts	328	0.893	0.873	n/a	n/a	11,183	n/a

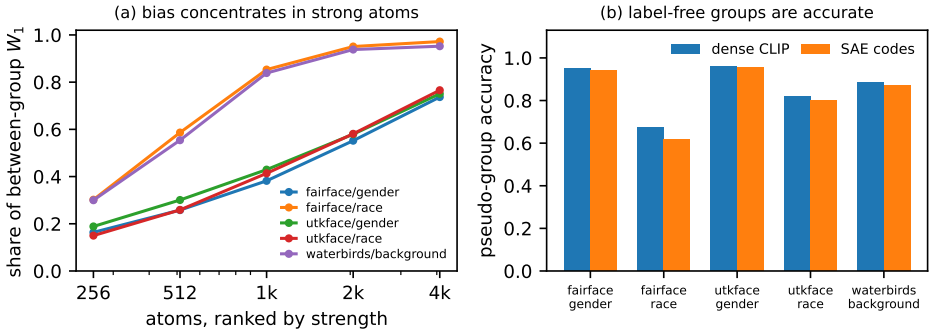


Figure 4: Premise study on reference embeddings (annotations used for analysis only, never by the method). (a) Between-group transport cost concentrates in high-strength atoms, motivating tiered budgets. (b) Soft pseudo-groups obtained from attribute prompts track dense zero-shot grouping accuracy, whether computed from dense embeddings or from SAE codes.

and we report the cosine between x and $\hat{x}$. This is the reconstruction itself, not the identity of the decoder-difference update, which returns x whenever no atom is edited.

Text embeddings are reconstructed almost exactly, with mean cosines of 0.996 and 0.998 on the ViT backbones and of 0.909 and 0.893 on RN50 and RN101. Image embeddings are reconstructed only approximately. Their mean cosine is 0.78 to 0.85 on the ViT backbones and 0.58 to 0.77 on the ResNets, with 5th percentiles as low as 0.54. The modality-gap alignment shift of the calibration raises the means to 0.87 to 0.94 and 0.85 to 0.91, respectively, while widening the lower tail on the ViT backbones.

Sparsity and the inference-time code. The active-atom counts show a further backbone limitation. On the ResNet checkpoints about 10 to 12 thousand atoms are active per image, so the code is not sparse there, whereas the ViT checkpoints activate roughly 560 to $1{,}220$. Applying the coarsest

Table 14: Stereotype retrieval fairness on RN50. KL and MaxSkew (MS) are evaluated at $k = 500$; lower is better. Row groups distinguish access to reference images; gray rows use group-labeled reference images and are excluded from the ranking. Bold and underline mark the best and second-best results among methods without reference group labels, with ties at displayed precision. PRISM and ZSDebias are means over three seeds (Table 16); conventions in App. D.

	FairFace Race		FairFace Gender		UTKFace Race		UTKFace Gender		CelebA Stereo.	
Method	KL↓	MS↓	KL↓	MS↓	KL↓	MS↓	KL↓	MS↓	KL↓	MS↓
No task-specific reference images										
Base CLIP	0.215	0.734	0.170	0.351	0.127	0.477	0.153	0.340	0.390	0.623
RoboShot	0.214	0.706	0.298	0.443	0.152	0.588	0.258	0.415	0.189	0.333
Orth-Proj	0.465	0.997	0.111	0.288	0.340	0.609	0.117	0.312	0.145	0.384
Orth-Cali	0.412	0.911	0.141	0.357	0.307	0.581	0.086	0.257	0.068	0.238
PRISM	0.157	0.632	0.069	0.245	0.134	0.523	0.088	0.265	0.169	0.397
PRISM-mini	0.455	0.985	0.113	0.291	0.336	0.608	0.117	0.311	0.143	0.379
ZSDebias	0.045	**0.380**	0.049	0.218	**0.027**	0.335	0.035	0.183	0.057	0.236
SEM_i	0.126	0.562	0.031	0.206	0.039	0.265	0.112	0.383	**0.050**	**0.193**
SEM_b	0.171	0.652	0.040	0.194	0.107	0.410	0.075	0.282	0.111	0.311
SEM_{bi}	0.153	0.625	0.043	0.190	0.107	0.404	0.079	0.280	0.110	0.307
Unlabeled reference images										
BALLAST_{eq}	**0.044**	0.390	**0.011**	**0.107**	0.035	**0.243**	0.012	**0.109**	0.118	0.388
BALLAST	0.045	0.393	0.012	0.113	0.035	**0.243**	**0.011**	**0.109**	0.105	0.357
Group-labeled reference images										
BendVLM	0.151	0.575	0.006	0.082	0.102	0.444	0.008	0.093	0.029	0.218
BendSEM_{bi}	0.067	0.454	0.005	0.079	0.042	0.365	0.009	0.102	0.010	0.119

Table 15: Stereotype retrieval fairness on RN101. KL and MaxSkew (MS) are evaluated at $k = 500$; lower is better. Row groups distinguish access to reference images; gray rows use group-labeled reference images and are excluded from the ranking. Bold and underline mark the best and second-best results among methods without reference group labels, with ties at displayed precision. PRISM and ZSDebias are means over three seeds (Table 16); conventions in App. D.

	FairFace Race		FairFace Gender		UTKFace Race		UTKFace Gender		CelebA Stereo.	
Method	KL↓	MS↓	KL↓	MS↓	KL↓	MS↓	KL↓	MS↓	KL↓	MS↓
No task-specific reference images										
Base CLIP	0.203	0.744	0.144	0.335	0.153	0.497	0.136	0.333	0.300	0.559
RoboShot	0.222	0.799	0.338	0.494	0.206	0.651	0.322	0.491	0.294	0.454
Orth-Proj	0.321	0.843	0.212	0.408	0.321	0.583	0.163	0.360	0.171	0.372
Orth-Cali	0.297	0.840	0.275	0.469	0.301	0.574	0.202	0.395	0.188	0.349
PRISM	0.152	0.594	0.107	0.282	0.133	0.531	0.127	0.315	0.162	0.361
PRISM-mini	0.312	0.837	0.214	0.410	0.314	0.582	0.167	0.362	0.173	0.374
ZSDebias	0.083	0.593	0.031	0.189	0.091	0.564	0.022	0.165	**0.017**	**0.123**
SEM_i	0.111	0.566	0.037	0.202	0.111	0.401	0.041	0.215	0.041	0.186
SEM_b	0.152	0.637	0.079	0.258	0.084	0.341	0.070	0.245	0.195	0.448
SEM_{bi}	0.140	0.624	0.079	0.258	0.086	0.349	0.069	0.245	0.165	0.394
Unlabeled reference images										
BALLAST_{eq}	**0.060**	0.466	0.011	0.107	0.044	**0.260**	**0.012**	**0.114**	0.099	0.354
BALLAST	0.061	**0.464**	**0.010**	**0.104**	**0.041**	0.268	**0.012**	0.118	0.099	0.346
Group-labeled reference images										
BendVLM	0.122	0.587	0.010	0.106	0.122	0.538	0.013	0.123	0.019	0.173
BendSEM_{bi}	0.059	0.427	0.006	0.087	0.035	0.368	0.012	0.127	0.021	0.183

training-time top-k mask to the code instead of the plain ReLU lowers the image-side reconstruction on twelve of the sixteen dataset and backbone pairs and raises its mean cosine by at most 0.033 on the four RN101 pairs (released with the code), so the gap is one of modality, not of the inference-time code.

Consequences for the method. This gap is why the method decodes only the change of the code on the selected atoms (Sec. 3.3) and why its fidelity floor concerns the repair $\cos(x, x')$ rather than the reconstruction. The codes still carry the group information the method needs (Fig. 4). Their fidelity as a representation of the whole image embedding is a limitation that we state rather than assume away.

D Retrieval tables for the ResNet backbones

This section gives the full retrieval tables for the two ResNet backbones. Table 1 in the main text reports the two ViT backbones with every baseline. Tables 14 and 15 report RN50 and RN101 in the same format, including the BALLAST_{eq} rows, and Table 2 in the main text summarizes them. We first fix the conventions of the full tables, then read the ResNet comparisons against SEM, note how PRISM and ZSDebias are reported, and close with how often the calibration selected the identity on RN50 and how often the transport was active on RN101.

Table 16: Per-seed values of PRISM and ZSDebias on the twelve ViT retrieval cells: the three released seeds, their mean (the value entered in Tables 1 and 29), and the released PRISM-mini variant. KL and MaxSkew (MS) of the top-500 composition, lower is better; precision@500 (Prec.) on the labeled hair queries, higher is better. No ranking marks are applied.

		PRISM					ZSDebias			
Cell	Metric	Seed 0	Seed 1	Seed 2	Mean	PRISM-mini	Seed 0	Seed 1	Seed 2	Mean
ViT-L/14@336px										
FairFace Race	KL↓	0.139	0.153	0.150	0.147	0.209	0.190	0.162	0.181	0.178
FairFace Race	MS↓	0.645	0.585	0.611	0.614	0.780	0.725	0.646	0.704	0.692
FairFace Gender	KL↓	0.069	0.064	0.021	0.051	0.035	0.077	0.175	0.085	0.112
FairFace Gender	MS↓	0.263	0.275	0.152	0.230	0.165	0.271	0.403	0.289	0.321
UTKFace Race	KL↓	0.160	0.159	0.157	0.159	0.197	0.167	0.146	0.179	0.164
UTKFace Race	MS↓	0.504	0.567	0.558	0.543	0.492	0.441	0.455	0.527	0.474
UTKFace Gender	KL↓	0.032	0.033	0.048	0.038	0.051	0.095	0.139	0.103	0.112
UTKFace Gender	MS↓	0.177	0.189	0.227	0.198	0.223	0.304	0.368	0.326	0.333
CelebA Stereo.	KL↓	0.044	0.061	0.078	0.061	0.100	0.266	0.217	0.291	0.258
CelebA Stereo.	MS↓	0.209	0.248	0.278	0.245	0.358	0.570	0.512	0.591	0.557
CelebA Hair	KL↓	0.193	0.148	0.172	0.171	0.143	0.146	0.092	0.111	0.116
CelebA Hair	MS↓	0.266	0.278	0.353	0.299	0.366	0.435	0.271	0.353	0.353
CelebA Hair	Prec.↑	0.660	0.646	0.671	0.659	0.696	0.512	0.486	0.529	0.509
ViT-B/16										
FairFace Race	KL↓	0.135	0.170	0.151	0.152	0.301	0.158	0.231	0.202	0.197
FairFace Race	MS↓	0.609	0.687	0.633	0.643	0.805	0.667	0.864	0.814	0.782
FairFace Gender	KL↓	0.077	0.096	0.082	0.085	0.340	0.120	0.108	0.137	0.122
FairFace Gender	MS↓	0.272	0.294	0.286	0.284	0.523	0.318	0.316	0.326	0.320
UTKFace Race	KL↓	0.134	0.140	0.153	0.142	0.277	0.095	0.201	0.088	0.128
UTKFace Race	MS↓	0.489	0.497	0.537	0.508	0.539	0.526	0.839	0.507	0.624
UTKFace Gender	KL↓	0.078	0.087	0.114	0.093	0.198	0.075	0.055	0.080	0.070
UTKFace Gender	MS↓	0.275	0.291	0.313	0.293	0.389	0.248	0.240	0.257	0.248
CelebA Stereo.	KL↓	0.118	0.129	0.182	0.143	0.190	0.219	0.117	0.248	0.195
CelebA Stereo.	MS↓	0.349	0.341	0.440	0.377	0.384	0.469	0.359	0.489	0.439
CelebA Hair	KL↓	0.075	0.065	0.039	0.060	0.189	0.164	0.296	0.109	0.190
CelebA Hair	MS↓	0.184	0.210	0.163	0.186	0.379	0.408	0.380	0.337	0.375
CelebA Hair	Prec.↑	0.677	0.679	0.654	0.670	0.658	0.516	0.527	0.515	0.519

Table conventions. Rows of every full table form three blocks, the three information-access groups of Table 5: methods with no task-specific reference images, methods using unlabeled reference images, and methods using group-labeled reference images (BendVLM and BendSEM$_{bi}$). The third block is set in gray and excluded from the ranking. Bold marks the best and underline the second-best value at the displayed precision among the first two blocks. BALLAST-C† additionally fits TRIM on unlabeled evaluation-gallery statistics.

Comparison with SEM. The KL ratio rows of Tables 1 and 2 divide unrounded means, so a ratio of displayed values can differ in the last digit: on FairFace-race ViT-L/14@336px, one of the 24 ViT comparisons of Sec. 4.2, the displayed values give 4.3 against the 4.4 of the ratio row. Across the two ResNets, BALLAST beats the best SEM point estimate in 16 of the 20 stereotype comparisons. The exceptions are the four CelebA-stereotype comparisons, which we report as losses. The closest win is UTKFace-race RN50 KL, where BALLAST's 0.035 stands against 0.039 for the best SEM point estimate and BALLAST's per-split KL ranges from 0.030 to 0.039. Elsewhere the ViT-scale improvements recur. On FairFace race with RN50, for example, KL falls from 0.126 under SEM$_i$ to 0.045 under BALLAST.

PRISM and ZSDebias. PRISM and ZSDebias appear as means over their three released seeds, and Table 16 gives the per-seed values. One caveat concerns ZSDebias. Its released adaptor for face images was trained on all of UTKFace, LFW, and FACET, so its UTKFace rows have seen the evaluation galleries.

Equalization and transport on the two ResNets. On RN50 the calibration selected the identity in 38 of the 60 retrieval splits: all ten on hair and on UTKFace race, and nine of ten on UTKFace gender and on FairFace race. BALLAST therefore coincides with BALLAST$_{eq}$ in those splits, and the RN50 comparisons are largely wins of the equalization alone. On RN101, by contrast, the transport was active in 48 of the 60 splits. Both counts are taken from Table 7 (App. B).

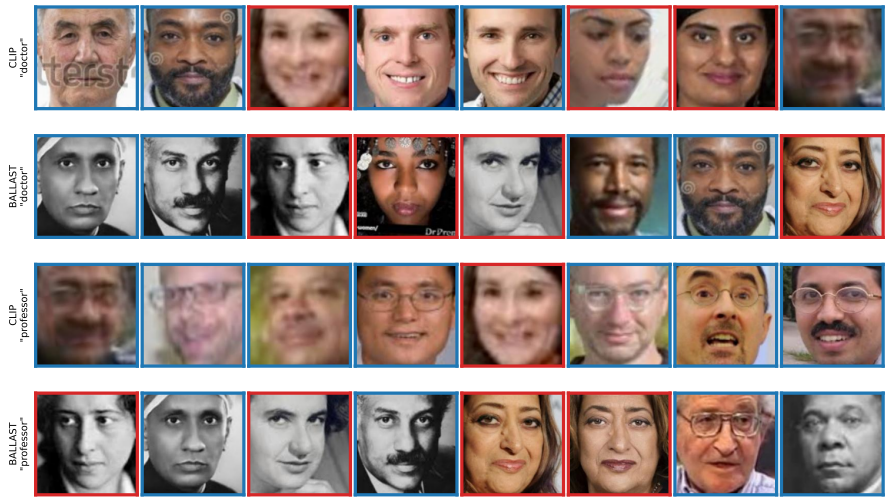


Figure 5: Top-8 retrieved UTKFace images for the queries "a photo of a doctor" (first two rows) and "a photo of a professor" (last two rows), base CLIP above BALLAST for each query (ViT-L/14@336px, one split; border color marks annotated gender, blue male, red female). Base CLIP retrieves predominantly one gender for each occupation; after gallery repair and query equalization the retrieved sets are close to the population balance. The example shows the full method; it does not separate the two operators.

E Qualitative retrieval examples

This section gives one qualitative example of what BALLAST changes in a retrieved set. Figure 5 shows the top-8 retrieved UTKFace images for two occupation queries on one split, before and after BALLAST. For each query, base CLIP retrieves predominantly one gender, and the set retrieved after BALLAST is close to the population balance. The example shows the full method on one split; it does not separate the two operators.

F Zero-shot classification tables

This section gives the zero-shot classification tables for the four backbones. Tables 17 and 18 report the ViT classification cells of the extension of Sec. 3.5, and Tables 19 and 20 report ResNet classification. We first fix the cells and the way the two metrics are computed, and then read the ResNet comparison with SEM.

Cells and metrics. The classification cells are CelebA, with blond hair as the class and gender as the groups, and Waterbirds, with bird type as the class and background as the groups, each on the four backbones. The harness computes accuracy and worst-group accuracy on the predictions pooled over the ten splits. Worst-group accuracy is the minimum over the four (class, group) combinations of a cell.

Comparison with SEM on the ResNets. BALLAST-C exceeds the best per-cell SEM worst-group accuracy on Waterbirds RN50 and on CelebA RN101, and trails it on CelebA RN50 and on Waterbirds RN101. The deficit on Waterbirds RN101, at 0.039, is the largest deficit among the eight (dataset, backbone) cells.

G Classification extension: method

This section gives the full statement of the extension summarized in Sec. 3.5. It states the motivation and the scope of the extension, describes how plain BALLAST and BALLAST-C treat the class prompts, and then gives BALLAST-C's margin model, the estimator of its translation-removal step, and its data-access regime.

Motivation and scope. Used as a zero-shot classifier, CLIP can follow background correlations instead of the object of interest, as on Waterbirds (Sagawa et al., 2020). Worst-group accuracy depends on where each (class, group) combination of the gallery falls relative to the class prompts, so the gallery repair can move it. The extension is secondary and transductive. Classification faces a different obstacle and a different data-access regime from retrieval, and the paragraphs below take the two in turn.

Table 17: Zero-shot classification on ViT-L/14@336px: accuracy (Acc.) and worst-group accuracy (WG), higher is better. Row groups distinguish access to reference images (App. D, Table conventions); gray rows use group-labeled reference images and are excluded from the ranking. Bold and underline mark the best and second-best values among the other rows, ties sharing the mark. † BALLAST-C additionally fits TRIM on unlabeled evaluation-gallery statistics.

	CelebA (gender)		Waterbirds (background)	
Method	Acc.↑	WG↑	Acc.↑	WG↑
No task-specific reference images				
Base CLIP	**0.870**	0.779	**0.862**	0.396
RoboShot	0.848	0.811	**0.862**	0.485
Orth-Proj	0.861	0.785	0.858	0.477
Orth-Cali	0.852	0.815	0.844	0.481
SEM_i	0.791	0.744	0.832	0.523
SEM_b	0.856	**0.824**	0.855	0.628
SEM_{bi}	0.851	0.820	0.835	**0.682**
Unlabeled reference images				
BALLAST	0.862	0.815	0.823	0.546
BALLAST-C†	0.864	0.806	0.831	0.678
Group-labeled reference images				
BendVLM	0.837	0.763	0.820	0.422
BendSEM_{bi}	0.846	0.827	0.808	0.742

Table 18: Zero-shot classification on ViT-B/16: accuracy (Acc.) and worst-group accuracy (WG), higher is better. Row groups distinguish access to reference images (App. D, Table conventions); gray rows use group-labeled reference images and are excluded from the ranking. Bold and underline mark the best and second-best values among the other rows, ties sharing the mark. † BALLAST-C additionally fits TRIM on unlabeled evaluation-gallery statistics.

	CelebA (gender)		Waterbirds (background)	
Method	Acc.↑	WG↑	Acc.↑	WG↑
No task-specific reference images				
Base CLIP	0.748	0.612	**0.829**	0.250
RoboShot	0.788	0.692	0.806	0.263
Orth-Proj	0.744	0.611	0.817	0.290
Orth-Cali	0.747	0.622	0.826	0.373
SEM_i	0.737	0.613	0.801	0.496
SEM_b	0.796	0.711	0.825	0.428
SEM_{bi}	0.795	0.719	0.808	0.547
Unlabeled reference images				
BALLAST	0.770	0.678	0.797	0.376
BALLAST-C†	**0.811**	**0.757**	0.806	**0.550**
Group-labeled reference images				
BendVLM	0.750	0.686	0.816	0.241
BendSEM_{bi}	0.796	0.748	0.780	0.640

Table 19: Zero-shot classification on RN50: accuracy (Acc.) and worst-group accuracy (WG), higher is better. Row groups distinguish access to reference images (App. D, Table conventions); gray rows use group-labeled reference images and are excluded from the ranking. Bold and underline mark the best and second-best values among the other rows, ties sharing the mark. † BALLAST-C additionally fits TRIM on unlabeled evaluation-gallery statistics.

	CelebA (gender)		Waterbirds (background)	
Method	Acc.↑	WG↑	Acc.↑	WG↑
No task-specific reference images				
Base CLIP	0.821	0.768	0.837	0.392
RoboShot	0.842	**0.807**	0.763	0.460
Orth-Proj	0.795	0.723	0.859	0.456
Orth-Cali	0.830	0.801	0.807	**0.709**
SEM_i	0.835	0.799	0.851	0.558
SEM_b	0.848	0.799	0.844	0.577
SEM_{bi}	**0.852**	0.804	**0.862**	0.529
Unlabeled reference images				
BALLAST	0.821	0.768	0.767	0.453
BALLAST-C†	0.849	0.796	0.850	0.620
Group-labeled reference images				
BendVLM	0.809	0.715	0.826	0.608
BendSEM_{bi}	0.848	0.815	0.854	0.653

Table 20: Zero-shot classification on RN101: accuracy (Acc.) and worst-group accuracy (WG), higher is better. Row groups distinguish access to reference images (App. D, Table conventions); gray rows use group-labeled reference images and are excluded from the ranking. Bold and underline mark the best and second-best values among the other rows, ties sharing the mark. † BALLAST-C additionally fits TRIM on unlabeled evaluation-gallery statistics.

	CelebA (gender)		Waterbirds (background)	
Method	Acc.↑	WG↑	Acc.↑	WG↑
No task-specific reference images				
Base CLIP	0.690	0.504	0.800	0.497
RoboShot	0.737	0.598	0.760	0.450
Orth-Proj	0.676	0.488	0.857	0.401
Orth-Cali	0.680	0.507	0.823	0.553
SEM_i	**0.811**	0.759	0.844	0.583
SEM_b	0.795	0.750	0.846	**0.588**
SEM_{bi}	0.791	0.743	0.871	0.544
Unlabeled reference images				
BALLAST	0.664	0.471	0.750	0.363
BALLAST-C†	0.799	**0.765**	**0.874**	0.549
Group-labeled reference images				
BendVLM	0.702	0.490	0.812	0.194
BendSEM_{bi}	0.785	0.701	0.880	0.640

Table 21: Zero-shot classification, worst-group accuracy (WG, higher is better) per (dataset, backbone) cell: Best SEM (per cell), BALLAST, the two BALLAST-C variants without the transport (defined in the text of App. H), and BALLAST-C. Bold and underline mark the best and second-best values per column, including ties. † fits TRIM on unlabeled evaluation-gallery statistics. ViT-L is ViT-L/14@336px, ViT-B is ViT-B/16. Accuracy and every baseline: Tables 17, 18, 19, 20 (App. F); per-split attribution: App. H.

	CelebA (gender groups)				Waterbirds (background groups)			
WG↑	ViT-L	ViT-B	RN50	RN101	ViT-L	ViT-B	RN50	RN101
Best SEM	**0.824**	0.719	**0.804**	0.759	**0.682**	0.547	0.577	**0.588**
BALLAST	0.815	0.678	0.768	0.471	0.546	0.376	0.453	0.363
BALLAST-C†, gallery edit off	0.815	0.720	0.796	0.759	0.668	0.549	0.527	0.547
BALLAST-C†, transport off, amp. kept	0.815	0.720	0.796	0.746	0.663	0.549	0.618	0.547
BALLAST-C†	0.806	**0.757**	0.796	**0.765**	0.678	**0.550**	**0.620**	0.549

Class prompts in the two variants. Class prompts must remain mutually distinct, so the projection of Eq. 3, applied to each class prompt independently, can erase the differences that define the classes. The retrieval correction is not used for them. Plain BALLAST scores the repaired gallery against class prompts that are pulled, on the transported atoms only, toward the barycenter means, with the class-contrast atoms protected. BALLAST-C instead composes the repaired gallery with the fixed text edit of SEM, which is SEM_b on CelebA and SEM_{bi} on Waterbirds (App. B). This composition follows the practice SEM introduced with BendSEM. BALLAST-C then applies TRIM (Translation Removal in Margins), which the remaining paragraphs define.

Margin model. For a binary classifier direction $\delta = t_A - t_B$, the margin of image n is $s_n = x'_n \cdot \delta$. Class prevalence may legitimately differ between groups, but a spurious group effect appears as a shared additive translation of *both* class components of a group's margin distribution. We model the margins as the group-shifted mixture

$$s \mid y, g \;\sim\; \mathcal{N}(\mu_y + \Delta_g,\, \sigma_y^2), \qquad y \mid g \sim \pi_{y|g}, \tag{4}$$

with class means and variances shared across groups, group-specific class prevalences $\pi_{y|g}$, and translations Δ_g. The translations are centered so that $\sum_g \tilde{\pi}_g \Delta_g = 0$ over the evaluation-gallery pseudo-group proportions $\tilde{\pi}_g$.

Removal and estimation. Only the translation is removed. Image n is shifted by $d_n\, \delta/\|\delta\|^2$ with $d_n = \sum_g w_{n,g} \Delta_g$ and then renormalized, so d_n is the removed translation before the final renormalization. The parameters are estimated by a fixed number of alternating updates with the soft weights held fixed (App. B). This procedure is not exact EM for Eq. 4, and we claim neither consistency nor a downstream fairness guarantee.

Data access and the restriction to translations. Because the margins are those of the evaluation gallery, BALLAST-C is *transductive*: it reads unlabeled image statistics of the gallery it is evaluated on, never annotations. Every table marks it with a dagger. Restricting TRIM to translations is deliberate, since richer per-group transforms could absorb legitimate correlation between class and group into the repair.

H Component attribution in BALLAST-C

This section shows that most of BALLAST-C's worst-group gain is associated with SEM's text edit and TRIM rather than with the gallery transport, and that the transport contributes on some cells only. It first compares BALLAST-C with Best SEM per cell, then introduces the two BALLAST-C variants that disable the transport and the two kinds of value in which the comparison is reported, draws three readings from that comparison, and closes with the design-choice ablations that the main text cites for the soft weights, the strength tier, and the barycenter.

BALLAST-C against the best SEM variant. Against Best SEM, the best SEM variant per (dataset, backbone) cell, BALLAST-C's worst-group accuracy is higher in four of the eight cells by point estimate (Table 21; full tables in Tables 17, 18, 19, 20). Three deficits are at most 0.018, as on Waterbirds ViT-L/14@336px with 0.678 against 0.682, and the largest deficit is 0.039 on Waterbirds RN101. The largest gain is 0.043, on Waterbirds RN50 (0.620 against 0.577); CelebA ViT-B/16 also gains (0.757 against 0.719).

Variants without the transport. BALLAST-C is a composite of the repaired gallery, SEM's fixed text edit of the class prompts, and TRIM, so its results cannot be attributed to the gallery transport

Table 22: Every ablation, aggregated over the twelve ViT retrieval cells (KL, MS) and the four ViT classification cells (Acc., WG); n/r: the variant is not run on that task. Base CLIP and Best SEM (per cell and per metric) are references. Component rows switch query equalization (Eq.), transport (Tr.) and amplification (Amp.) as in Table 4, per cell and per split in App. I. Design-choice rows change one choice of BALLAST. Composite rows are BALLAST-C and its two no-transport variants, per cell in Table 23. Every variant is recalibrated without labels. † fits TRIM on unlabeled evaluation-gallery statistics. ViT-L is ViT-L/14@336px, ViT-B is ViT-B/16.

Variant	KL↓	MS↓	Acc.↑	WG↑
References				
Base CLIP	0.169	0.464	0.827	0.509
Best SEM	0.115	0.391	0.833	0.693
Components (Eq., Tr., Amp.)				
Gallery repair only	0.135	0.419	0.813	0.604
BALLAST$_{eq}$	0.038	0.234	n/r	n/r
Eq. + Amp.	0.038	0.234	n/r	n/r
Eq. + Tr.	0.033	0.223	n/r	n/r
BALLAST	0.033	0.223	0.813	0.604
Design choices (BALLAST with one change)				
SAE-code pseudo-grouping	0.048	0.260	0.808	0.623
No strength tier	0.034	0.228	0.815	0.588
Hard assignments	0.032	0.222	0.827	0.520
Uniform barycenter	0.033	0.224	0.811	0.588
No alignment	0.034	0.227	0.807	0.622
Composites (classification only)				
BALLAST-C†, gallery edit off	n/r	n/r	0.825	0.688
BALLAST-C†, transport off, amp. kept	n/r	n/r	0.824	0.687
BALLAST-C†	n/r	n/r	0.828	0.698

Table 23: Component attribution in BALLAST-C: worst-group accuracy (WG, higher is better) per (dataset, backbone) cell. Upper block: Best SEM (per cell), BALLAST, the two BALLAST-C variants without the transport (gallery edit off: SEM text edit and TRIM on the normalized unrepaired gallery; transport off, amp. kept: the protected-atom amplification kept), and BALLAST-C, as pooled-prediction values as in every table. Lower block: per-split differences Δ_{amp} (transport off, amp. kept, minus gallery edit off) and Δ_{tr} (BALLAST-C minus transport off, amp. kept) as median [min, max] over the ten splits, whose center differs from the pooled value, and Splits, the number of splits in which the first-named variant is higher; descriptive ranges only. Every variant is recalibrated without labels; the text edit is SEM$_b$ on CelebA and SEM$_{bi}$ on Waterbirds. On CelebA RN50 the calibration selected the identity on every split, so the three composite values coincide. † fits TRIM on unlabeled evaluation-gallery statistics. ViT-L is ViT-L/14@336px, ViT-B is ViT-B/16.

Cell	Best SEM (per cell)	BALLAST	BALLAST-C† gallery edit off	BALLAST-C† transport off, amp. kept	BALLAST-C† (full)
CelebA ViT-L	0.824	0.815	0.815	0.815	0.806
CelebA ViT-B	0.719	0.678	0.720	0.720	0.757
CelebA RN50	0.804	0.768	0.796	0.796	0.796
CelebA RN101	0.759	0.471	0.759	0.746	0.765
Waterbirds ViT-L	0.682	0.546	0.668	0.663	0.678
Waterbirds ViT-B	0.547	0.376	0.549	0.549	0.550
Waterbirds RN50	0.577	0.453	0.527	0.618	0.620
Waterbirds RN101	0.588	0.363	0.547	0.547	0.549

	Amplification Δ_{amp}		Transport Δ_{tr}	
Cell	Median [min, max]	Splits	Median [min, max]	Splits
CelebA ViT-L	+0.000 [+0.000, +0.000]	0/10	-0.007 [-0.011, -0.006]	0/10
CelebA ViT-B	+0.000 [+0.000, +0.000]	0/10	+0.038 [+0.032, +0.040]	10/10
CelebA RN50	+0.000 [+0.000, +0.000]	0/10	+0.000 [+0.000, +0.000]	0/10
CelebA RN101	+0.000 [-0.062, +0.000]	0/10	+0.007 [+0.005, +0.067]	10/10
Waterbirds ViT-L	-0.011 [-0.041, +0.015]	3/10	+0.016 [-0.005, +0.027]	8/10
Waterbirds ViT-B	+0.000 [+0.000, +0.000]	0/10	+0.001 [-0.002, +0.005]	5/10
Waterbirds RN50	+0.057 [+0.000, +0.116]	7/10	+0.013 [-0.062, +0.117]	8/10
Waterbirds RN101	+0.000 [+0.000, +0.000]	0/10	+0.002 [+0.000, +0.005]	6/10

Table 24: Accuracy (Acc., higher is better) of the variants of Table 23, per cell, from the same runs and under the same conventions. † fits TRIM on unlabeled evaluation-gallery statistics. ViT-L is ViT-L/14@336px, ViT-B is ViT-B/16.

Cell	BALLAST	BALLAST-C† gallery edit off	BALLAST-C† transport off, amp. kept	BALLAST-C† (full)
CelebA ViT-L	0.862	0.860	0.860	0.864
CelebA ViT-B	0.770	0.800	0.800	0.811
CelebA RN50	0.821	0.849	0.849	0.849
CelebA RN101	0.664	0.798	0.790	0.799
Waterbirds ViT-L	0.823	0.834	0.827	0.831
Waterbirds ViT-B	0.797	0.808	0.808	0.806
Waterbirds RN50	0.767	0.865	0.846	0.850
Waterbirds RN101	0.750	0.873	0.873	0.874

alone. Tables 23 and 24 therefore compare it, per cell, with two variants that keep the text edit and TRIM but disable the transport. The first, BALLAST-C† with the gallery edit off, disables the whole gallery edit, so that the text edit and TRIM act on the normalized unrepaired gallery; we call it the no-edit variant below. The second, BALLAST-C† with the transport off and the amplification kept, disables the transport but keeps the protected-atom amplification and calibrates it. Every variant is recalibrated by the label-free procedure.

Pooled and per-split values. The comparison is reported in two kinds of values. The worst-group accuracy that the tables report as "WG" is a pooled-prediction value. The per-split differences instead use per-split worst-group accuracies, whose mean differs from the pooled value. On Waterbirds RN50, for instance, the amplification raises the pooled value from 0.527 to 0.618 while the per-split median difference is 0.057. Three readings follow from the comparison, and the next two paragraphs take them in turn.

Text edit and TRIM. The first two readings concern the text edit and TRIM. First, most of BALLAST-C's worst-group gain over plain BALLAST comes from replacing the class-prompt pull with the text edit: on the four ViT cells the aggregate moves from 0.604 under plain BALLAST to 0.688 under the no-edit variant, which leaves the gallery untouched. Second, TRIM alone changes little. The no-edit variant differs from the SEM variant it wraps by between -0.014 and $+0.009$ worst-group points per cell, which is consistent with its small fitted translations: those translations satisfy $|\Delta_g| \leq 0.0049$. Across the eight classification cells, BALLAST-C's own fitted shifts satisfy $|\Delta_g| \leq 0.006$, with the maximum of 0.0053 on CelebA ViT-L/14@336px, an order of magnitude below the CelebA class separations of 0.039 on ViT-B/16 and 0.055 on ViT-L/14@336px.

Transport and amplification. Third, the transport contributes on some cells and not on others. Relative to the variant with the transport off and the amplification kept, BALLAST-C has higher worst-group accuracy in all ten splits on CelebA ViT-B/16 (median gain 0.038) and on CelebA RN101, and in eight of ten on Waterbirds ViT-L/14@336px (0.663 to 0.678, median gain 0.016) and on Waterbirds RN50. In the same comparison it has lower worst-group accuracy in all ten splits on CelebA ViT-L/14@336px (median loss 0.007), and the remaining cells change by at most 0.003 on average. On CelebA RN50 the calibration selected the identity on every split, so the three composite variants coincide. The amplification matters on one cell, Waterbirds RN50, and that cell's gain over Best SEM comes from the amplification. The no-edit variant exceeds Best SEM in three of the eight cells, and BALLAST-C does so in four. On the ViT aggregate the transport adds 0.011 with the amplification kept (0.687 to 0.698), and the whole gallery repair, transport and amplification together, adds 0.010 (0.688 to 0.698).

What shapes the transport. Table 22 lists the retrieval and classification sides of every ablation. On the retrieval side, the gallery repair on its own (Gallery repair only) has an aggregate KL of 0.135, below base CLIP (0.169) but above Best SEM (0.115), and it stays above Best SEM on nine of the twelve cells. The classification side shows what shapes the transport. Soft weights are the shared default of both tasks because they are the one choice that serves both: hard assignments instead of soft weights lower the aggregate retrieval KL to 0.032 (Table 22) but cost 8.4 worst-group points in the plain-BALLAST classification ablation. Removing the strength tier costs 1.6 worst-group points and a uniform instead of prior-weighted barycenter 1.5; these two costs are unrounded, and both variants display as 0.588. SAE-code instead of dense-CLIP pseudo-grouping raises aggregate worst-group accuracy from 0.604 to 0.623 but worsens both retrieval aggregates, with KL rising from 0.033 to 0.048, so dense-CLIP grouping is retained. On retrieval, the strength

Table 25: Per-split comparisons of BALLAST against the label-free baselines on the 24 retrieval cells: the number of the ten splits on which BALLAST's value is lower than the baseline's on the same split, for KL and MaxSkew (MS) at $k = 500$. Both methods are evaluated on the same ten splits; the counts are descriptive and the per-split ranges of both sides are in the accompanying artifact. PRISM and ZSDebias are omitted because their seeds are not paired with the splits.

	Base CLIP		SEM_i		SEM_b		SEM_{bi}		Orth-Proj	
Cell	KL	MS	KL	MS	KL	MS	KL	MS	KL	MS
ViT-L/14@336px										
FairFace Race	10	10	10	10	10	10	10	10	10	10
FairFace Gender	10	10	10	10	10	10	10	10	9	2
UTKFace Race	10	10	10	10	10	10	10	10	10	10
UTKFace Gender	10	10	10	10	10	10	10	10	10	10
CelebA Stereo.	10	10	10	10	10	10	10	10	10	10
CelebA Hair	10	10	10	10	10	10	10	10	10	10
ViT-B/16										
FairFace Race	10	10	10	10	10	10	10	10	10	10
FairFace Gender	10	10	10	10	10	10	10	10	10	10
UTKFace Race	10	10	10	10	10	10	10	10	10	10
UTKFace Gender	10	10	10	10	10	10	10	10	10	10
CelebA Stereo.	10	10	10	10	10	10	10	10	10	10
CelebA Hair	10	10	10	10	10	10	10	10	10	10
RN50										
FairFace Race	10	10	10	10	10	10	10	10	10	10
FairFace Gender	10	10	10	10	10	10	10	10	10	10
UTKFace Race	10	10	7	9	10	10	10	10	10	10
UTKFace Gender	10	10	10	10	10	10	10	10	10	10
CelebA Stereo.	10	10	0	0	8	0	8	0	10	10
CelebA Hair	10	10	10	10	10	10	10	10	10	10
RN101										
FairFace Race	10	10	10	10	10	10	10	10	10	10
FairFace Gender	10	10	10	10	10	10	10	10	10	10
UTKFace Race	10	10	10	10	10	10	10	10	10	10
UTKFace Gender	10	10	10	10	10	10	10	10	10	10
CelebA Stereo.	10	10	0	0	10	10	10	10	10	10
CelebA Hair	10	10	10	10	10	10	10	10	10	10

Table 26: Factorial retrieval ablation on the twelve ViT cells, KL (lower is better), with query equalization active in every column: BALLAST_{eq}, Eq. + Amp., Eq. + Tr. and BALLAST as in Table 4, each recalibrated without labels. Δ_{tr} is the per-split KL of Eq. + Tr. minus that of BALLAST_{eq}, as median [min, max] over the ten splits (negative favors the transport), and Splits counts the splits in which Eq. + Tr. is lower; descriptive ranges only. MaxSkew: Table 27.

	KL↓				Eq. + Tr. minus BALLAST_{eq}, per split	
Cell	BALLAST_{eq}	Eq. + Amp.	Eq. + Tr.	BALLAST	Median [min, max]	Splits
ViT-L/14@336px						
FairFace Race	0.036	0.036	0.034	0.034	-0.002 [-0.005, +0.004]	8/10
FairFace Gender	0.039	0.040	0.028	0.029	-0.010 [-0.016, -0.005]	10/10
UTKFace Race	0.017	0.017	0.014	0.014	-0.003 [-0.004, -0.001]	10/10
UTKFace Gender	0.007	0.007	0.008	0.008	+0.001 [-0.001, +0.004]	2/10
CelebA Stereo.	0.034	0.034	0.033	0.033	+0.000 [-0.003, +0.000]	2/10
CelebA Hair	0.097	0.097	0.070	0.070	-0.028 [-0.033, -0.018]	10/10
ViT-B/16						
FairFace Race	0.038	0.038	0.034	0.034	-0.004 [-0.009, +0.000]	9/10
FairFace Gender	0.048	0.049	0.037	0.038	-0.011 [-0.023, +0.000]	7/10
UTKFace Race	0.035	0.035	0.032	0.032	-0.003 [-0.005, +0.000]	8/10
UTKFace Gender	0.011	0.012	0.013	0.013	+0.002 [-0.001, +0.005]	1/10
CelebA Stereo.	0.050	0.049	0.050	0.049	+0.000 [+0.000, +0.000]	0/10
CelebA Hair	0.043	0.043	0.042	0.042	-0.001 [-0.003, +0.002]	7/10

tier and the prior-weighted barycenter are within 0.002 of the full method on KL and are kept for their classification effect, and the alignment switch is roughly neutral and is left to the calibration.

I PER-SPLIT RANGES AND THE FACTORIAL RETRIEVAL ABLATION

This section reports per-split results for the BALLAST family. The released benchmark records carry one value per configuration for the baselines, so per-split variability is available for the BALLAST family only. We first describe the factorial ablation that separates the two parts of the gallery edit and state its aggregate result. We then read the transport per split, give the per-split ranges for the two close comparisons against baselines, count the splits on which BALLAST is lower than each label-free baseline, and close with how these ranges should be read.

Table 27: Factorial retrieval ablation on the twelve ViT cells, MaxSkew (lower is better); columns, variants and conventions as in Table 26.

	MS↓				Eq. + Tr. minus BALLAST$_{eq}$, per split	
Cell	BALLAST$_{eq}$	Eq. + Amp.	Eq. + Tr.	BALLAST	Median [min, max]	Splits
ViT-L/14@336px						
FairFace Race	0.357	0.357	0.354	0.354	-0.006 [-0.019, +0.035]	9/10
FairFace Gender	0.212	0.215	0.182	0.182	-0.028 [-0.051, -0.014]	10/10
UTKFace Race	0.213	0.214	0.204	0.205	-0.008 [-0.024, +0.004]	9/10
UTKFace Gender	0.088	0.086	0.089	0.088	-0.001 [-0.009, +0.021]	6/10
CelebA Stereo.	0.229	0.229	0.227	0.228	+0.000 [-0.015, +0.000]	2/10
CelebA Hair	0.323	0.322	0.294	0.294	-0.036 [-0.041, -0.006]	10/10
ViT-B/16						
FairFace Race	0.352	0.354	0.350	0.350	-0.002 [-0.012, +0.012]	7/10
FairFace Gender	0.235	0.235	0.204	0.205	-0.030 [-0.067, +0.000]	7/10
UTKFace Race	0.235	0.236	0.226	0.225	-0.009 [-0.020, +0.000]	8/10
UTKFace Gender	0.109	0.109	0.117	0.115	+0.006 [-0.006, +0.029]	2/10
CelebA Stereo.	0.227	0.224	0.227	0.224	+0.000 [+0.000, +0.000]	0/10
CelebA Hair	0.231	0.232	0.208	0.208	-0.021 [-0.038, -0.012]	10/10

Factorial design. BALLAST$_{eq}$ disables both the transport and the protected-atom amplification, so the difference between the full method and BALLAST$_{eq}$ is the whole gallery edit. Within that edit, the calibration selected a nonzero amplification in 58 of the 120 ViT retrieval splits (released with the code). Table 26 separates the two parts by a factorial ablation on the twelve ViT retrieval cells, with the equalization active throughout. Its four variants are BALLAST$_{eq}$, Eq. + Amp., Eq. + Tr., and BALLAST, each recalibrated without labels.

Aggregate result. In aggregate, Eq. + Amp. is within 0.001 of BALLAST$_{eq}$ and Eq. + Tr. within 0.001 of BALLAST, on KL, on MaxSkew, and on hair precision, while individual cells differ slightly (Tables 26 and 27). Aggregated over the twelve cells, the per-split mean KL is lower with the transport than without it in all ten splits, and higher without the equalization than under the full method in all ten. On the four settings of Table 4, removing the equalization raises KL on all eight cells, by a factor above two on seven of them, and adding the amplification to BALLAST$_{eq}$ changes no KL of Table 4 by more than 0.002, with changes in both directions. On hair precision, BALLAST$_{eq}$ retains higher precision than BALLAST in all ten splits on both ViT hair cells, 0.608 and 0.592 against 0.572 and 0.585, while Eq. + Amp. moves it by at most 0.001. The transport is therefore the active part of the gallery edit in retrieval, and the amplification has almost no aggregate effect on retrieval.

Per-split effect of the transport. Per split, Eq. + Tr. lowers KL relative to BALLAST$_{eq}$ in at least seven of ten splits on eight of the twelve cells. It does so in all ten splits on FairFace-gender ViT-L/14@336px (0.039 to 0.028) and on hair ViT-L/14@336px (0.097 to 0.070); on the two UTKFace-gender cells it raises KL by at most 0.002. On the other two settings of Table 4 it lowers KL from 0.036 to 0.034 on FairFace-race ViT-L/14@336px (lower in eight of ten splits) and from 0.017 to 0.014 on UTKFace-race ViT-L/14@336px (lower in all ten); on ViT-B/16 it lowers FairFace race from 0.038 to 0.034 (lower in nine of ten) and FairFace gender from 0.048 to 0.037 (lower in seven of ten), and moves the other two settings of that table, UTKFace race and hair, by less than 0.005. On the two CelebA stereotype cells the differences are within 0.004 and most splits tie. On these cells the calibration switched the transport off in all ten ViT-B/16 splits and in eight of ten ViT-L/14@336px splits (Table 7). On the two UTKFace-gender cells Eq. + Tr. raises mean KL, from 0.007 to 0.008 on ViT-L/14@336px and from 0.011 to 0.013 on ViT-B/16; it lowers KL in only two of ten splits on ViT-L/14@336px and in one of ten on ViT-B/16. Table 26 gives KL and Table 27 MaxSkew, and Table 23 gives the paired comparisons of the classification composites.

Two close comparisons against baselines. For the two close cases named in the main text, we set BALLAST's per-split range next to the baseline's point estimate. BALLAST's per-split MaxSkew on FairFace-gender ViT-L/14@336px ranges from 0.164 to 0.224 around its mean of 0.182, and Orth-Proj's point estimate is 0.164. BALLAST's per-split KL on UTKFace-race RN50 ranges from 0.030 to 0.039 around 0.035, against 0.039 for SEM$_i$.

Per-split comparisons against the baselines. The regenerated baseline records carry the same ten splits, so the comparisons of Sec. 4.2 can be read split by split. Table 25 counts, for every retrieval cell and each label-free baseline, the splits on which BALLAST's KL or MaxSkew is lower. The table covers 240 comparisons, one per retrieval cell, baseline and metric. BALLAST is lower on all ten splits in 228 of them and on at least seven of ten in 233. It is lower on five or fewer splits in 7: one of these is on a ViT backbone, Orth-Proj on FairFace-gender ViT-L/14@336px MaxSkew,

Table 28: Transport strength on the two hair cells: α pinned to one value with the strength tier, the amplification and the alignment switch calibrated as usual; KL and MaxSkew (MS) at $k = 500$, lower is better, and precision@500 (Prec.), higher is better. The calibrated BALLAST and BALLAST$_{eq}$ are the values of Table 29.

Strength α	KL↓	MS↓	Prec.↑
ViT-L/14@336px			
0.25	0.088	0.300	0.600
0.50	0.081	0.277	0.592
0.75	0.075	0.279	0.581
1.00	0.070	0.294	0.572
BALLAST (calibrated)	0.070	0.294	0.572
BALLAST$_{eq}$	0.097	0.323	0.608
ViT-B/16			
0.25	0.043	0.212	0.591
0.50	0.042	0.209	0.587
0.75	0.042	0.203	0.582
1.00	0.041	0.191	0.576
BALLAST (calibrated)	0.042	0.208	0.585
BALLAST$_{eq}$	0.043	0.231	0.592

Table 29: Hair-color retrieval on CelebA with gender as the protected attribute: KL and MaxSkew (MS) at $k = 500$, lower is better, and precision@500 (Prec.), higher is better. Row groups distinguish access to reference images; gray rows use group-labeled reference images and are excluded from the ranking. Bold and underline mark the best and second-best values among the other rows, ties sharing the mark. PRISM and ZSDebias are means over three seeds (Table 16); conventions in App. D.

	ViT-L/14@336px			ViT-B/16		
Method	KL↓	MS↓	Prec.↑	KL↓	MS↓	Prec.↑
No task-specific reference images						
Base CLIP	0.148	0.359	0.622	0.179	0.409	0.630
RoboShot	0.275	0.429	0.674	0.144	0.244	0.632
Orth-Proj	0.144	0.373	0.692	0.189	0.380	0.660
Orth-Cali	0.105	0.308	0.687	0.150	0.376	0.685
PRISM	0.171	<u>0.299</u>	0.659	0.060	**0.186**	0.670
PRISM-mini	0.143	0.366	0.696	0.189	0.379	0.658
ZSDebias	0.116	0.353	0.509	0.190	0.375	0.519
SEM$_i$	0.236	0.458	0.698	0.191	0.344	0.678
SEM$_b$	0.135	0.365	<u>0.699</u>	0.172	0.396	<u>0.729</u>
SEM$_{bi}$	0.167	0.401	**0.732**	0.183	0.400	**0.734**
Unlabeled reference images						
BALLAST$_{eq}$	<u>0.097</u>	0.323	0.608	<u>0.043</u>	0.231	0.592
BALLAST	**0.070**	**0.294**	0.572	**0.042**	<u>0.208</u>	0.585
Group-labeled reference images						
BendVLM	0.028	0.163	0.680	0.028	0.174	0.656
BendSEM$_{bi}$	0.032	0.187	0.685	0.029	0.159	0.752

where BALLAST is lower in two of ten splits, and the other six are on the CelebA-stereotype cells of the two ResNets against the SEM variants. These counts describe the ten shared splits and make no significance statement.

Reading the ranges. The ten splits share overlapping evaluation halves, so these ranges describe the variability of the reported means and are not confidence intervals. We draw no significance conclusion from them.

J HAIR-COLOR RETRIEVAL AND THE STRENGTH SWEEPS

This section reports the hair-color retrieval results behind the trade-off between fairness and precision of Sec. 4.3. Table 29 reports every method on the labeled hair-color queries discussed there, Table 30 sweeps the query-equalization strength from 0.25 to 1.00, and Table 28 pins the transport strength. We read the two sweeps first, then compare them with BALLAST$_{eq}$, and close with where BALLAST and SEM sit on the trade-off.

Strength sweep. On ViT-B/16, fairness improves with strength while precision decreases from 0.602 to 0.585. On ViT-L/14@336px, precision decreases only from 0.573 to 0.572.

Transport strength. Table 28 pins the transport strength α and calibrates the other knobs as usual. Precision falls with α on both backbones, from 0.600 to 0.572 on ViT-L/14@336px and from 0.591 to 0.576 on ViT-B/16, while KL falls from 0.088 to 0.070 on ViT-L/14@336px and MaxSkew from

Table 30: Hair-color retrieval under a sweep of the query-equalization strength from 0.25 to 1.00 on the two ViT backbones: KL and MaxSkew (MS) at $k = 500$, lower is better, and precision@500 (Prec.), higher is better. Strength 1.00 is BALLAST as reported in Table 29.

Strength	KL↓	MS↓	Prec.↑
ViT-L/14@336px			
0.25	0.071	0.275	0.573
0.50	0.068	0.282	0.572
0.75	0.068	0.289	0.572
1.00	0.070	0.294	0.572
ViT-B/16			
0.25	0.095	0.299	0.602
0.50	0.071	0.265	0.596
0.75	0.054	0.228	0.591
1.00	0.042	0.208	0.585

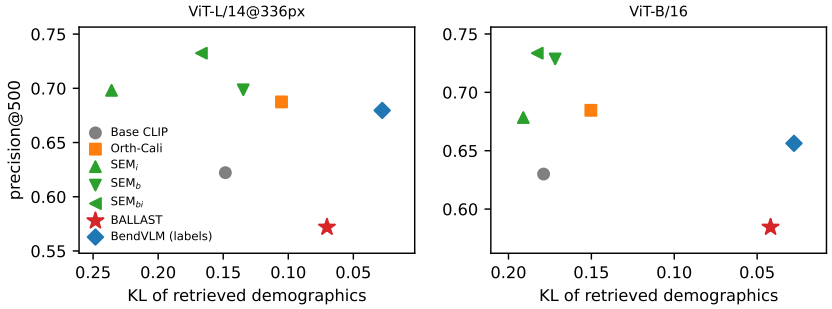


Figure 6: Fairness and precision on hair-color retrieval. BALLAST occupies a distinct operating point: markedly lower KL at a precision cost that persists across the tested equalization-strength range (Table 30).

0.212 to 0.191 on ViT-B/16. The transport therefore buys fairness with precision along the whole range of settings. The calibrated method sits at the strong end on ViT-L/14@336px, where the calibration selected $\alpha = 1$ in every split, and at $\alpha = 0.5$ in nine of ten splits on ViT-B/16.

BALLAST$_{eq}$ and precision. BALLAST$_{eq}$, which applies the query equalization without the gallery repair, retains higher precision than BALLAST, 0.608 against 0.572 on ViT-L/14@336px and 0.592 against 0.585 on ViT-B/16. This is consistent with the gallery repair contributing part of the cost, though the sweep does not isolate it.

Reading the trade-off. Balancing per se need not cost precision: with group labels, BendVLM and BendSEM$_{bi}$ balance without a precision cost relative to base CLIP (Table 29). Among the label-free methods, seed-mean PRISM keeps precision above base CLIP, at 0.659 on ViT-L/14@336px and 0.670 on ViT-B/16, with KL of 0.171 and 0.060 on the same two backbones, a different operating point again. No tested strength, however, combines SEM-level precision with BALLAST-level fairness. SEM resolves the same tension in the opposite direction, retaining precision at higher KL, so the two methods occupy different operating points of one trade-off (Fig. 6).

K Query-family transfer

This section reports the per-family split behind the transfer statement of the main text. The fit of each cell is calibrated on its first query family only, the seen family, and the unseen value averages the remaining families. Table 31 gives the seen and unseen KL of the twelve ViT cells next to Best SEM's average over all families, and Table 32 extends the same split to all twenty-four retrieval cells. We read the unseen-family KL first against the headline number of each cell and then against SEM, compare the seen family with the unseen ones, and close with the hair cells, which have fewer families.

Unseen families and the headline. Unseen-family KL tracks the headline number of each cell. On the twenty stereotype cells, where 24 of 25 families are unseen, it stays within 4.1% of the headline. This agreement is bookkeeping rather than evidence of transfer. The headline averages the seen family with the 24 unseen ones, so the deviation is bounded by $1/24$ by construction, and it is largest where the seed family's KL is near zero. The evidence of transfer is given per family in Table 33.

Table 31: Query-family transfer on the ViT backbones (KL@500; lower is better). BALLAST is fitted and calibrated on the first query family; unseen results average the remaining 24 stereotype or three hair families. SEM values average all families and are not a matched comparison on unseen queries. Full results appear in App. K.

		FairFace		UTKFace		CelebA	
Method	Query families	Race	Gender	Race	Gender	Stereo.	Hair
ViT-L/14@336px							
BALLAST	Seen	0.022	0.004	0.006	0.003	0.001	0.015
BALLAST	Unseen	0.034	0.030	0.015	0.008	0.034	0.089
Best SEM	All	0.147	0.097	0.058	0.033	0.153	0.135
ViT-B/16							
BALLAST	Seen	0.027	0.001	0.046	0.006	0.046	0.012
BALLAST	Unseen	0.034	0.039	0.032	0.013	0.050	0.052
Best SEM	All	0.170	0.087	0.096	0.064	0.173	0.172

Table 32: Query-family transfer per retrieval cell (KL@500, lower is better). Families: query families of the cell. Seen: the first family, on which the fit is calibrated; Unseen: mean over the remaining families; Headline: mean over all families of the same replication, within a relative 10^{-3} of the main-table record; Best SEM / Unseen: the best SEM variant's KL over all families divided by BALLAST's unseen KL. Cell names as in Tables 1 and 29; the ResNet cells are read against Tables 14 and 15.

		KL@500↓			Ratio
Cell	Families	Seen	Unseen	Headline	Best SEM / Unseen
ViT-L/14@336px					
FairFace Race	25	0.022	0.034	0.034	4.3×
FairFace Gender	25	0.004	0.030	0.029	3.3×
UTKFace Race	25	0.006	0.015	0.014	3.9×
UTKFace Gender	25	0.003	0.008	0.008	4.1×
CelebA Stereo.	25	0.001	0.034	0.033	4.4×
CelebA Hair	4	0.015	0.089	0.070	1.5×
ViT-B/16					
FairFace Race	25	0.027	0.034	0.034	4.9×
FairFace Gender	25	0.001	0.039	0.038	2.2×
UTKFace Race	25	0.046	0.032	0.032	3.0×
UTKFace Gender	25	0.006	0.013	0.013	4.8×
CelebA Stereo.	25	0.046	0.050	0.049	3.5×
CelebA Hair	4	0.012	0.052	0.042	3.3×
RN50					
FairFace Race	25	0.019	0.046	0.045	2.8×
FairFace Gender	25	0.003	0.012	0.012	2.6×
UTKFace Race	25	0.030	0.035	0.035	1.1×
UTKFace Gender	25	0.005	0.012	0.011	6.4×
CelebA Stereo.	25	0.021	0.108	0.105	0.5×
CelebA Hair	4	0.004	0.133	0.101	1.8×
RN101					
FairFace Race	25	0.014	0.063	0.061	1.8×
FairFace Gender	25	0.002	0.011	0.010	3.5×
UTKFace Race	25	0.027	0.041	0.041	2.0×
UTKFace Gender	25	0.009	0.013	0.012	3.3×
CelebA Stereo.	25	0.013	0.102	0.099	0.4×
CelebA Hair	4	0.005	0.081	0.062	2.9×

Comparison with SEM. On every ViT cell, unseen-family KL remains a factor 1.5 to 4.9 below the best SEM variant's average over all families. The released SEM records carry no per-family split, so this comparison is between BALLAST's unseen families and SEM's full family set. On the ResNets the ratio mirrors the per-cell outcomes of App. D, with seen and unseen moving together on the CelebA-stereotype cells where SEM leads.

Seen against unseen families. Normalizing each family's improvement by base CLIP shows no systematic advantage for the seen family: it improves more than the unseen average in fifteen of the twenty-four cells and less in the remaining nine. This ratio divides by base-CLIP values near zero on several cells, so the per-family counts of Table 33 are the better summary. Over all 24 cells, 437 of the 492 unseen family evaluations have BALLAST KL below base CLIP's. Across the ten ViT stereotype cells, BALLAST lowers KL below base CLIP on 214 of the 240 unseen family evaluations, with 18 to 24 of the 24 families improving per cell (median 22). On the ten ViT stereotype cells the unseen-family median KL is 0.004 to 0.031 under BALLAST against 0.016 to

Table 33: Per-family transfer on the 24 retrieval cells (KL@500, lower is better), grouped by backbone. Seen: the first query family, on which the fit is calibrated; the remaining families are unseen (24 of 25 stereotype families, 3 of 4 hair families). Unseen improved: number of unseen families on which BALLAST's KL is below Base CLIP's for the same family; median and range are over the unseen families. Descriptive per-family values from the records of Table 32.

	Seen-family KL		Unseen	Unseen median KL		Unseen range
Cell	Base CLIP	BALLAST	improved	Base CLIP	BALLAST	BALLAST
ViT-L/14@336px						
FairFace Race	0.351	0.022	23 of 24	0.234	0.027	0.008 to 0.173
FairFace Gender	0.054	0.004	20 of 24	0.077	0.015	0.001 to 0.129
UTKFace Race	0.158	0.006	23 of 24	0.122	0.013	0.002 to 0.036
UTKFace Gender	0.001	0.003	18 of 24	0.016	0.004	0.000 to 0.057
CelebA Stereo.	0.078	0.001	23 of 24	0.216	0.028	0.001 to 0.116
CelebA Hair	0.083	0.015	1 of 3	0.102	0.103	0.061 to 0.103
ViT-B/16						
FairFace Race	0.123	0.027	24 of 24	0.211	0.031	0.009 to 0.081
FairFace Gender	0.002	0.001	18 of 24	0.101	0.020	0.001 to 0.181
UTKFace Race	0.073	0.046	23 of 24	0.112	0.030	0.007 to 0.076
UTKFace Gender	0.006	0.006	21 of 24	0.089	0.006	0.001 to 0.056
CelebA Stereo.	0.038	0.046	21 of 24	0.268	0.023	0.001 to 0.329
CelebA Hair	0.056	0.012	3 of 3	0.234	0.045	0.002 to 0.110
RN50						
FairFace Race	0.111	0.019	22 of 24	0.189	0.041	0.006 to 0.119
FairFace Gender	0.004	0.003	18 of 24	0.156	0.003	0.000 to 0.047
UTKFace Race	0.046	0.030	22 of 24	0.089	0.035	0.005 to 0.073
UTKFace Gender	0.001	0.005	22 of 24	0.141	0.008	0.000 to 0.069
CelebA Stereo.	0.122	0.021	22 of 24	0.449	0.068	0.002 to 0.463
CelebA Hair	0.041	0.004	3 of 3	0.291	0.024	0.001 to 0.374
RN101						
FairFace Race	0.130	0.014	21 of 24	0.166	0.052	0.023 to 0.133
FairFace Gender	0.002	0.002	21 of 24	0.124	0.004	0.000 to 0.037
UTKFace Race	0.092	0.027	23 of 24	0.118	0.033	0.015 to 0.134
UTKFace Gender	0.005	0.009	22 of 24	0.101	0.007	0.001 to 0.057
CelebA Stereo.	0.048	0.013	20 of 24	0.283	0.062	0.002 to 0.364
CelebA Hair	0.020	0.005	3 of 3	0.305	0.071	0.037 to 0.135

0.268 under base CLIP, and the 26 unseen stereotype families on these ten cells that do not improve are mostly ones on which base CLIP is already near parity: 20 of them have a base KL below 0.02.

Hair cells. The hair cells have only four families, so the seen share is larger there. Hair ViT-L/14@336px is correspondingly the closest ViT cell in the comparison with SEM, at $1.5\times$.

L Evaluation depth

The main tables evaluate retrieval at $k = 500$. Table 34 repeats the evaluation of Base CLIP, the three SEM variants, BALLAST$_{eq}$ and BALLAST on the twelve ViT retrieval cells at $k = 100$ and $k = 1{,}000$, with every method fitted and calibrated exactly as in the main run, so that only the evaluation depth changes. The ranking is the same at every depth: BALLAST is below all three SEM variants on every cell for both KL and MaxSkew at $k = 100$, 500 and 1,000. The twelve-cell mean KL of BALLAST is 0.063 at $k = 100$, 0.033 at $k = 500$ and 0.024 at $k = 1{,}000$. At the same three depths Best SEM has 0.152, 0.115 and 0.099, and Base CLIP has 0.211, 0.169 and 0.149. The values fall with the depth for every method because a larger retrieved set averages toward the gallery composition.

M Provenance

This section records how the BALLAST-family results were produced and checked. We describe the runs and their manifests, the two checks that accompany the main run, and the generation of the table values. We close with the split protocol.

Runs and manifests. The BALLAST-family results were generated by one experiment code revision with two later changes. The first made the evaluator's output directory configurable. The second capped the retained rank of the equalization subspace at its exact value $G - 1$ (App. B): the uncapped SVD cutoff had kept one rounding-noise direction in 39 of the 240 per-split BALLAST fits, spread over 15 of the 24 retrieval cells, and in 26 of the 240 BALLAST$_{eq}$ fits. Every retrieval configuration reported in this version that uses the equalization was therefore rerun under the cap, 230 in all: 24 BALLAST, 24 BALLAST$_{eq}$, 24 anchor-difference controls, 24 factorial variants, 60 design ablations, 12 fixed configurations, 14 configurations of the two strength sweeps and 48 of the

Table 34: Evaluation depth. KL and MaxSkew (MS) of Base CLIP, the three SEM variants, BALLAST$_{eq}$ and BALLAST on the twelve ViT retrieval cells at $k = 100$, 500 (the main tables) and 1,000; lower is better. BALLAST is fitted and calibrated exactly as in the main run (the calibration depth is fixed), so only the evaluation changes. No ranking marks.

		Base CLIP		SEM$_i$		SEM$_b$		SEM$_{bi}$		BALLAST$_{eq}$		BALLAST	
Cell	k	KL	MS	KL	MS	KL	MS	KL	MS	KL	MS	KL	MS
ViT-L/14@336px													
FairFace Race	100	0.325	0.908	0.209	0.728	0.259	0.806	0.217	0.738	0.075	0.476	0.074	0.486
	500	0.244	0.798	0.147	0.625	0.194	0.706	0.155	0.623	0.036	0.357	0.034	0.354
	1,000	0.211	0.750	0.119	0.570	0.166	0.659	0.131	0.574	0.027	0.316	0.025	0.310
FairFace Gender	100	0.130	0.336	0.153	0.355	0.128	0.327	0.142	0.329	0.063	0.261	0.049	0.231
	500	0.114	0.326	0.123	0.328	0.097	0.298	0.110	0.299	0.039	0.212	0.029	0.182
	1,000	0.099	0.309	0.106	0.311	0.085	0.279	0.099	0.285	0.031	0.192	0.022	0.164
UTKFace Race	100	0.174	0.512	0.084	0.442	0.204	0.521	0.184	0.499	0.051	0.355	0.049	0.361
	500	0.124	0.461	0.058	0.451	0.137	0.446	0.118	0.420	0.017	0.213	0.014	0.205
	1,000	0.104	0.430	0.047	0.431	0.111	0.412	0.092	0.383	0.008	0.156	0.007	0.149
UTKFace Gender	100	0.072	0.254	0.061	0.248	0.077	0.258	0.086	0.271	0.022	0.157	0.022	0.153
	500	0.040	0.186	0.033	0.186	0.047	0.201	0.054	0.217	0.007	0.088	0.008	0.088
	1,000	0.028	0.159	0.022	0.155	0.032	0.171	0.039	0.189	0.004	0.064	0.004	0.063
CelebA Stereo.	100	0.222	0.511	0.176	0.440	0.204	0.475	0.216	0.496	0.038	0.227	0.038	0.226
	500	0.237	0.536	0.153	0.413	0.198	0.480	0.209	0.489	0.034	0.229	0.033	0.228
	1,000	0.231	0.535	0.136	0.392	0.193	0.480	0.203	0.487	0.031	0.223	0.030	0.221
CelebA Hair	100	0.176	0.382	0.310	0.501	0.140	0.343	0.163	0.377	0.132	0.349	0.096	0.341
	500	0.148	0.359	0.236	0.458	0.135	0.365	0.167	0.401	0.097	0.323	0.070	0.294
	1,000	0.136	0.340	0.218	0.442	0.139	0.363	0.164	0.391	0.084	0.309	0.055	0.244
ViT-B/16													
FairFace Race	100	0.310	0.884	0.247	0.839	0.310	0.871	0.296	0.870	0.085	0.490	0.079	0.485
	500	0.237	0.795	0.170	0.690	0.232	0.750	0.217	0.750	0.038	0.352	0.034	0.350
	1,000	0.199	0.725	0.144	0.630	0.199	0.698	0.181	0.686	0.027	0.301	0.024	0.297
FairFace Gender	100	0.165	0.368	0.107	0.289	0.131	0.318	0.124	0.307	0.079	0.270	0.064	0.239
	500	0.139	0.346	0.087	0.269	0.098	0.278	0.088	0.255	0.048	0.235	0.038	0.205
	1,000	0.126	0.329	0.080	0.262	0.086	0.264	0.078	0.243	0.038	0.218	0.029	0.188
UTKFace Race	100	0.199	0.556	0.175	0.505	0.240	0.595	0.238	0.597	0.095	0.362	0.090	0.350
	500	0.125	0.476	0.096	0.407	0.145	0.502	0.137	0.499	0.035	0.235	0.032	0.225
	1,000	0.092	0.425	0.064	0.342	0.104	0.445	0.098	0.443	0.020	0.190	0.018	0.182
UTKFace Gender	100	0.191	0.371	0.102	0.296	0.186	0.377	0.183	0.370	0.037	0.182	0.034	0.172
	500	0.134	0.321	0.064	0.241	0.124	0.320	0.119	0.317	0.011	0.109	0.013	0.115
	1,000	0.100	0.287	0.048	0.213	0.086	0.271	0.080	0.269	0.004	0.071	0.005	0.075
CelebA Stereo.	100	0.345	0.560	0.186	0.450	0.300	0.535	0.276	0.517	0.082	0.289	0.082	0.289
	500	0.314	0.555	0.173	0.443	0.240	0.495	0.224	0.489	0.050	0.227	0.049	0.224
	1,000	0.302	0.554	0.160	0.435	0.226	0.485	0.210	0.477	0.037	0.200	0.037	0.197
CelebA Hair	100	0.223	0.421	0.203	0.342	0.230	0.454	0.250	0.466	0.075	0.266	0.077	0.261
	500	0.179	0.409	0.191	0.344	0.172	0.396	0.183	0.400	0.043	0.231	0.042	0.208
	1,000	0.160	0.395	0.174	0.337	0.144	0.322	0.149	0.323	0.034	0.203	0.032	0.190

depth sweep. Every reported retrieval value of the family comes from these runs. The 182 configurations outside the depth sweep report 406 means. Relative to the uncapped records, 177 of these means changed, 86 for the better and 91 for the worse, by a median of 0.006 and at most 0.097. The changes fall on 87 configurations, exactly those in whose fits the cutoff had kept the extra direction; the other 95 configurations reproduced the uncapped values to the last digit. The calibration selected the same knobs in every split. The classification configurations and the gallery-repair-only variant do not use the equalization and were not rerun; the BendVLM-style composition, which this version does not report, was regenerated under the cap when the release was certified. The reruns were launched only after the prompt templates, split arrays, embeddings, and checkpoints had been verified byte-identical to the main run's manifest. Every run carries per-split records for every metric, and every input is SHA-256 manifested with the release.

Checks on the main run. Two checks accompany the main run. The first is an identity check. For every BALLAST$_{eq}$ configuration and split, the gallery returned by the evaluation path equals the L2-normalized base gallery exactly, with a maximum absolute difference of 0, so the variant is query equalization on the unrepaired gallery. The BALLAST-C variant with the gallery edit off pins the same two parameters to zero and therefore follows the same code path. The second is a determinism check. Rerunning the archived configurations of the main run reproduced every ablation, BALLAST-C, and sweep value to the last digit.

Table generation. Every benchmark-table value, including the new rows and tables of this revision, is produced programmatically from the records. No number was typed by hand.

Splits. Each dataset is split ten times by the released script into a reference half and an evaluation half. The script uses seed 42 and draws one random permutation per split, with no stratification. For odd N the evaluation half holds one more image. The ten evaluation halves therefore overlap, so the splits are repeated holdouts rather than disjoint folds and per-split ranges describe variability, not confidence intervals. We verified that the script reproduces the shipped index files exactly. The two halves play different roles. The reference half is the reference set, the unlabeled sample the method fits on; the evaluation half is the evaluation gallery, which every query is answered against; and the fairness prior of each split is that evaluation gallery's own group composition.

N Additional related work

Further post-hoc debiasing methods. Beyond the methods of Sec. 2, post-hoc corrections include prompt tuning, additive residuals, kernel-space corrections, feature pruning and subspace removal on both modalities, a rotation in place of the projection, and a concurrent closed-form both-modality correction from prompts (Berg et al., 2022; Seth et al., 2023; Dehdashtian et al., 2024; Jung et al., 2024; Zhao et al., 2026; Gerych et al., 2026; Lian et al., 2026). Among them, WRING (Gerych et al., 2026) argues that projecting out a subspace can amplify bias on concepts the subspace does not cover. This argument is one reading of the precision cost of Sec. 4.3.

Further gallery-aware precedents. Cached gallery-side score corrections exist without groups (Chowdhury et al., 2024). Proportional-representation constraints select the retrieved set directly (Geyik et al., 2019; Oesterling et al., 2024).

Precedents of difference-vector removal and soft groups. Removing the span of difference vectors is the classical neutralization device of word embeddings (Bolukbasi et al., 2016), whose neutralize step projects a vector off the bias subspace and renormalizes it, the form of Eq. 3; for vision-language models, Orth-Proj (Chuang et al., 2023) projects out the span of spurious prompt embeddings, whereas the image-search method of Wang et al. (2021) instead prunes the embedding coordinates most informative of gender. Soft group weights from CLIP zero-shot scores follow Ghosal & Li (2023).

O Computational and memory cost

This section measures the compute and memory cost of BALLAST against the SEM variants, the other method that operates in the same frozen MSAE space, with Base CLIP as the floor. The figures of Sec. 4.3 come from the instrumented run described here; that run is separate from the main run and changes no benchmark number.

What is measured. The harness evaluates each (cell, split) by calling every method through two entry points: a text correction for the query, or for the class prompts, and an image correction for the evaluation gallery. In retrieval the pair of calls is made once per query family ($Q = 25$ stereotype or $Q = 4$ hair families); in classification it is made once per split. CLIP image embeddings are precomputed for every method. We instrumented the harness's own call sequence: every call is timed between CUDA synchronizations, peak GPU memory is read from the PyTorch allocator (per process, excluding the CUDA context), and a warm-up pass precedes every configuration. All 32 cells of the main tables, 24 retrieval and 8 classification, were measured on their first split in three passes. The device was the NVIDIA RTX 5000 Ada Generation (32 GB) of the main run, with PyTorch 2.11 and CUDA 12.8. Table 35 reports the median over passes and, across cells, the median and range: Fit is the first text call of a split (which for BALLAST contains the fit), Gallery the first image call, Query the median over the remaining families of text plus image plus top-k, Split the total, and Peak the allocation during the method's calls with its excess over what was resident before. Every number is reproduced from the released records of the instrumented run.

Time. BALLAST does its work once per gallery. The fit takes 11.7 s in the median retrieval cell and 5.5 to 26.3 s over the 24 cells, with the maximum on FairFace race with $G = 7$; the uninstrumented worker logs of the main run give a median of 7 s. The gallery repair takes about one second, and the repaired gallery is cached. Each further query family then costs one text encode and a projection of rank at most $G - 1$: 14 ms in the median cell. This is the same order as Base CLIP (12 ms) and SEM$_i$ (12 ms). It is below SEM$_b$ (29 ms) and SEM$_{bi}$ (35 ms, up to 93 ms on the $G = 7$ cells), because those two variants encode and score their $20G$ bias prompts on every call.

Table 35: Measured cost per split of the SEM and BALLAST families on the evaluation harness, with Base CLIP as the floor: first split of every cell, median over three passes, entries median [min, max] over the 24 retrieval or 8 classification cells. Fit: first text call (contains the fit where there is one); Gallery: first image call; Query: per further query family (retrieval only; n/a in classification); Split: all method calls of the split plus the top-k or prediction call; Peak and Excess: PyTorch peak allocation during the method's calls and its excess over the allocation already resident, as defined in App. O. † fits TRIM on unlabeled evaluation-gallery statistics.

Method	Fit (s)	Gallery (s)	Query (ms)	Split (s)	Peak (GiB)	Excess (GiB)
Retrieval cells						
Base CLIP	0.0 [0.0, 0.0]	0.00 [0.00, 0.00]	12 [4.2, 16]	0.3 [0.1, 0.5]	0.8 [0.3, 1.7]	0.2 [0.0, 0.8]
SEM_i	0.0 [0.0, 0.0]	0.00 [0.00, 0.00]	12 [6.0, 19]	0.3 [0.1, 0.5]	0.9 [0.5, 1.9]	0.1 [0.0, 0.4]
SEM_b	0.0 [0.0, 0.1]	0.00 [0.00, 0.00]	29 [20, 77]	0.7 [0.1, 2.0]	0.9 [0.5, 1.9]	0.2 [0.1, 0.4]
SEM_{bi}	0.0 [0.0, 0.1]	0.00 [0.00, 0.00]	35 [22, 93]	0.9 [0.1, 2.4]	0.9 [0.5, 1.9]	0.2 [0.1, 0.4]
BALLAST_{eq}	4.0 [1.6, 11.7]	0.99 [0.14, 2.81]	14 [5.1, 23]	5.3 [2.1, 15.2]	13.3 [3.7, 19.1]	12.3 [3.0, 17.9]
BALLAST	11.7 [5.5, 26.3]	0.98 [0.20, 2.86]	14 [5.1, 19]	12.8 [6.1, 29.7]	12.9 [3.7, 19.1]	12.3 [3.0, 17.9]
Classification cells						
Base CLIP	0.0 [0.0, 0.0]	0.00 [0.00, 0.01]	n/a	0.1 [0.0, 0.1]	0.9 [0.3, 1.7]	0.2 [0.0, 0.8]
SEM_i	0.0 [0.0, 0.0]	0.00 [0.00, 0.01]	n/a	0.0 [0.0, 0.1]	1.1 [0.5, 2.2]	0.2 [0.1, 0.8]
SEM_b	0.0 [0.0, 0.1]	0.00 [0.00, 0.01]	n/a	0.0 [0.0, 0.1]	1.1 [0.5, 2.2]	0.2 [0.1, 0.8]
SEM_{bi}	0.0 [0.0, 0.0]	0.00 [0.00, 0.00]	n/a	0.0 [0.0, 0.0]	1.1 [0.5, 2.2]	0.2 [0.1, 0.8]
BALLAST	7.3 [2.9, 13.3]	1.04 [0.06, 2.18]	n/a	8.4 [3.0, 15.4]	10.5 [2.8, 19.1]	9.9 [2.0, 17.9]
BALLAST-C†	7.2 [2.8, 13.1]	1.06 [0.08, 2.22]	n/a	8.3 [2.8, 15.3]	10.6 [2.9, 19.2]	9.9 [2.0, 17.9]

Under the benchmark protocol, which fits every split afresh and asks 25 families, the order reverses: a split costs 12.8 s for BALLAST against 0.9 s for SEM_{bi} and 0.3 s for SEM_i, which do no image-side work. A repaired gallery pays for its fit when it serves many queries; at the benchmark's query counts SEM remains an order of magnitude cheaper. The classification cells show the same picture with one text call per split, 8.4 s for BALLAST against at most 0.1 s for SEM, and BALLAST-C adds nothing measurable (8.3 s). BALLAST_{eq} spends 4.0 s on a fit whose transport tables it discards; its necessary work is the membership weights and the centroids.

Memory. BALLAST's peak allocation during the fit is 12.9 GiB in the median retrieval cell and 19.1 GiB at most, on CelebA with $N \approx 101{,}000$, against 0.9 GiB for every SEM variant (1.9 GiB at most in retrieval, 2.2 GiB in classification). The excess over what was resident before the method ran is 12.3 GiB for BALLAST (3.0 GiB on UTKFace to 17.9 GiB on CelebA) and at most 0.8 GiB for SEM. The excess comes from the temporaries of the atom-chunked quantile pass, which grow with the gallery until a fixed chunk bound binds; on CelebA they sit on top of about 3 GiB of codes. The excess is transient: after the fit, the resident state is the quantile tables and the repaired gallery, and the per-query footprint is below SEM's. Fitting the largest galleries therefore needs a GPU with tens of gigabytes, or a smaller chunk constant.

Caveats. Timings are wall-clock on a shared node. The other process on the node had a recorded utilization with a median of 74% and a maximum of 89%, so individual timings are upper bounds and the medians over passes and cells are the robust quantities. Both families were measured in the same passes. The timings characterize the released implementations, not the best possible ones: SEM re-encodes its bias prompts per query, and BALLAST_{eq} runs the full fit but leaves its transport tables unused. Peak GPU memory is the allocator's high-water mark for our process and excludes the CUDA context. We report medians and ranges only, because the passes are repeated measurements of one computation on a shared device, not independent samples.

P THEORETICAL PROPERTIES OF THE REPAIR

This section collects properties of the two operators, gallery transport and query equalization, that follow from their definitions in Sec. 3. They are statements about the operators in population form, that is, with exact quantile functions and laws. Where the implementation departs from that form, through finite quantile grids, the out-of-support guard, or protected atoms, we say so explicitly. None of the statements below concerns the benchmark numbers of Sec. 4. They explain how the transport strength controls group discrepancy and what the equalization does to a query.

We first fix the setting and recall two facts about one-dimensional transport. The results are ordered as follows. The gallery repair comes first, with the contraction of group discrepancy (Theorem 1) and the bound on latent displacement in terms of the transport cost (Corollary 1). Query equalization follows (Theorem 2). The remarks after the results state where the implementation departs from the population form and what the statements do not cover.

Setting and two facts about one-dimensional transport. The main text calls the membership vector a soft weight and makes no calibration claim (Sec. 3.2). This appendix calls the same object a posterior because the population statements treat it as one; no calibration hypothesis is assumed unless a statement says so.

Fix a selected atom j and suppress the index. A reference sample consists of an activation $z \in [0, M]$, its dense pre-activation, and a posterior $w = (w_1, \dots, w_G)$ with $w_g \geq 0$ and $\sum_g w_g = 1$, drawn from a law P. Activations are bounded, being a ReLU of a bounded embedding. Write $\pi_g = \mathbb{E}[w_g]$ for the prior and assume $\pi_g > 0$ for every g. This holds for the softmax posteriors of Sec. 3.2, for which every $w_g > 0$. We also admit hard posteriors, w a vertex of the simplex. For a function f of the sample write $\mathbb{E}_g[f] = \mathbb{E}[w_g f]/\pi_g$ for the w_g-tilted expectation, and let ν_g be the law of z under $\mathbb{E}_g$, with quantile function Q_g. With hard posteriors ν_g is the law of z in group g, and the weighted empirical distributions of Sec. 3.3 estimate these laws. We assume each ν_g has no atom except possibly at 0, the zero-inflated form of Sec. 3.3. The barycenter quantile is $\bar{Q} = \sum_g \pi_g Q_g$ (Eq. 1) with law $\bar{\nu}$. Since $z \leq M$, all these laws have finite second moments and $\bar{Q} \leq M$.

Every sample carries a quantile position $U_g \in (0, 1)$ in each group. If $z > 0$ the position is $U_g = F_g(z)$. If $z = 0$ it is a position inside the zero mass of ν_g given by the rank of the sample's dense pre-activation among the zero samples of that group (Sec. 3.3). In population form the pre-activation is continuous, so U_g is uniform on $(0, 1)$ under $\mathbb{E}_g$ and $Q_g(U_g) = z$. The group map is $T_g(z) = \bar{Q}(U_g)$. It is the map $T_g = \bar{Q} \circ F_g$ of Sec. 3.3 understood through quantile positions, so that at $z = 0$ it depends on the sample through its rank. The partial map is $T_g^a = (1 - a)\,\mathrm{id} + a\,T_g$ for $a \in [0, 1]$. Terms with $w_h = 0$ are dropped from $\sum_h w_h T_h(z)$, so $T_h(z)$ is needed only where $w_h > 0$. When every $w_g > 0$, all the laws ν_h share the support of z under P, and U_h is defined for every sample and every h.

Every statement below concerns the edited coordinates z'. None concerns the code obtained by re-encoding a repaired embedding, which need not equal z', and none concerns the joint law of the decoded embeddings.

We use two classical facts about transport in one dimension (Santambrogio, 2015). First, $W_2^2(\omega, \omega') = \int_0^1 (Q_\omega(u) - Q_{\omega'}(u))^2\, du$. Second, the comonotone coupling, which pairs equal quantile positions, is optimal. Since $(z, T_g(z)) = (Q_g(U_g), \bar{Q}(U_g))$ with U_g uniform under $\mathbb{E}_g$, these facts give $\mathbb{E}_g[(T_g(z) - z)^2] = W_2^2(\nu_g, \bar{\nu})$.

Gallery repair.

Theorem 1 (Barycentric transport contracts group discrepancy)**.** *In the setting above:*

(i) *$\bar{\nu}$ minimizes $\sum_g \pi_g W_2^2(\nu_g, \omega)$ over laws ω on $\mathbb{R}$: it is the π-weighted Wasserstein barycenter (Agueh & Carlier, 2011).*

(ii) *For each g let $\nu_g^a = (T_g^a)_{\#}\nu_g$ be the law obtained when every sample of group g is moved by the partial map of its own group. Then $Q_{\nu_g^a} = (1 - a)Q_g + a\bar{Q}$, and for all g, h,*

$$W_2(\nu_g^a, \nu_h^a) = (1 - a)\, W_2(\nu_g, \nu_h), \qquad W_2(\nu_g^a, \bar{\nu}) = (1 - a)\, W_2(\nu_g, \bar{\nu}),$$

while $\sum_g \pi_g Q_{\nu_g^a} = \bar{Q}$. The barycenter is therefore invariant under the repair, the repaired laws ν_g^a again have no positive atom, and applying (ii) to them with strength b yields the laws $\nu_g^{1-(1-a)(1-b)}$: a second repair fitted on the repaired laws composes into a single one.

(iii) *(Soft posteriors.) Let $z' = (1 - a)z + a \sum_h w_h T_h(z)$ be the update of Eq. 2 and ν_g' the law of z' under $\mathbb{E}_g$. Then*

$$W_2(\nu_g', \nu_h') \;\leq\; (1 - a)\, W_2(\nu_g, \nu_h) + a\,(\varepsilon_g + \varepsilon_h),$$

where

$$\varepsilon_g^2 = \mathbb{E}_g\Big[\Big(\sum_h w_h\,(T_h(z) - T_g(z))\Big)^2\Big] \;\leq\; \mathbb{E}_g\big[(1 - w_g)^2 R^2\big],$$

and $R = \max_{g,h} |T_g(z) - T_h(z)| \leq M$. With hard posteriors $\varepsilon_g = 0$, so (ii) applies without slack to the update of Eq. 2 in population form (Remark 1 for the discretized tables).

Proof. (i) For any ω with quantile function Q_ω, $\sum_g \pi_g \int (Q_g - Q_\omega)^2 = \sum_g \pi_g \int (Q_g - \bar{Q})^2 + \int (\bar{Q} - Q_\omega)^2$, because the cross term vanishes pointwise in u, as it contains $\sum_g \pi_g (Q_g - \bar{Q}) = 0$, so the

identity holds in $[0,\infty]$. The right-hand side is minimized at $Q_\omega = \bar{Q}$, and $\bar{Q}$ is a quantile function since it is nondecreasing and left-continuous as a convex combination of quantile functions. (ii) ν_g is the image of the uniform law on $(0,1)$ under Q_g, so ν_g^a is the image of the uniform law under $u \mapsto (1-a)Q_g(u) + a\bar{Q}(u)$, which is nondecreasing and left-continuous and hence the quantile function of ν_g^a. Then $Q_{\nu_g^a} - Q_{\nu_h^a} = (1-a)(Q_g - Q_h)$ and $Q_{\nu_g^a} - \bar{Q} = (1-a)(Q_g - \bar{Q})$; take $L^2(0,1)$ norms. The barycenter identity is $\sum_g \pi_g[(1-a)Q_g + a\bar{Q}] = \bar{Q}$. For the composition, $(1-a)Q_g + a\bar{Q}$ is flat only where both Q_g and $\bar{Q}$ are flat, and a flat piece of $\bar{Q}$ forces every Q_h to be flat there, so ν_g^a has an atom only at 0 and the family $(\nu_g^a)_g$ satisfies the Setting with the same π and the same barycenter quantile $\bar{Q}$; applying (ii) to it with strength b gives the quantile functions $(1-b)[(1-a)Q_g + a\bar{Q}] + b\bar{Q} = (1-a)(1-b)Q_g + [1-(1-a)(1-b)]\bar{Q}$. (iii) By the triangle inequality, $W_2(\nu_g', \nu_h') \le W_2(\nu_g', \nu_g^a) + W_2(\nu_g^a, \nu_h^a) + W_2(\nu_h^a, \nu_h')$, and the middle term is $(1-a)W_2(\nu_g, \nu_h)$ by (ii). Under $\mathbb{E}_g$ the pair $(z', T_g^a(z))$ has marginals ν_g' and ν_g^a, so $W_2^2(\nu_g', \nu_g^a) \le \mathbb{E}_g[(z' - T_g^a(z))^2] = a^2\,\mathbb{E}_g[(\sum_h w_h T_h(z) - T_g(z))^2] = a^2\varepsilon_g^2$, using $\sum_h w_h = 1$. Finally $|\sum_h w_h(T_h(z) - T_g(z))| \le \sum_{h\neq g} w_h R = (1-w_g)R$, and $R \le M$ because $0 \le \bar{Q} \le M$. □

Remark 1 (Implementation). The implementation departs from the population form through its quantile tables, the out-of-support guard, and the protected atoms. It replaces Q_g and $\bar{Q}$ by 64-knot tables on $u \in [0.002, 0.998]$, with linear interpolation inside the knot range and clamping outside it. The identities of Theorem 1(ii) therefore hold up to an interpolation error that we do not analyze, and a sample beyond the outermost knots is moved toward $\bar{Q}$ at the edge knot. The zero-branch position uses a 16-knot table of the pre-activation among zero samples, scaled by the zero mass as represented by the Q_g table, so that zero samples never rank above barely positive ones. The out-of-support guard of Sec. 3.3 and the protected atoms set the coefficient of some samples or atoms to zero: the pointwise form $(1-a)\,\mathrm{id} + a\,T_g$ of the map is unchanged on every transported sample, the upper bound of Corollary 1 remains valid with a as an upper bound on the per-sample coefficient, and the W_2 equalities of Theorem 1(ii) hold only up to the guarded tail. Order preservation differs between the hard and the soft update. With hard posteriors each T_g^a is nondecreasing, so the order of activations within a group is preserved on the transported samples. The soft update mixes the group maps with sample-dependent weights and need not preserve order between samples with different posteriors. Finally, the cost-based bound of Corollary 1 uses the exact population maps, and the quantile-table approximation adds an error that is not bounded here.

Corollary 1 (Transport cost bounds latent displacement)**.** *Let $C = \sum_g \pi_g W_2^2(\nu_g, \bar{\nu})$ be the barycentric transport cost of the atom. In population form the update of Eq. 2 satisfies $\mathbb{E}[(z'-z)^2] \le a^2 C$, with equality for hard posteriors, where the expectation is over a sample from the reference law P that defines π_g, ν_g, and $\bar{Q}$. In particular an atom on which the posterior-tilted laws $\nu_1, \dots, \nu_G$ coincide ($C = 0$; with hard posteriors, an atom whose group-conditional laws coincide) is left unchanged even if it is selected: the shuffle screen of Sec. 3.3 decides which atoms the repair may move, not whether it can move an atom that is bias-free in the sense $C = 0$.*

Proof. $z' - z = a\sum_h w_h(T_h(z) - z)$, and by Jensen's inequality $(\sum_h w_h(T_h(z) - z))^2 \le \sum_h w_h(T_h(z) - z)^2$, with equality when w is a vertex of the simplex. Taking expectations under P, $\mathbb{E}[\sum_h w_h(T_h(z) - z)^2] = \sum_h \pi_h\,\mathbb{E}_h[(T_h(z) - z)^2] = \sum_h \pi_h W_2^2(\nu_h, \bar{\nu})$ by the comonotone-coupling fact of the Setting. If $C = 0$ then $\nu_g = \bar{\nu}$, hence $Q_g = \bar{Q}$, for every g; since $Q_g(U_g) = z$, we get $T_g(z) = \bar{Q}(U_g) = z$ for every g, so $z' = z$. □

Query equalization. Let $c_1, \dots, c_G$ be the repaired soft centroids, computed from the reference fold and held fixed, $\bar{c} = \sum_g \pi_g c_g$, $\mathcal{B} = \mathrm{span}\{c_g - \bar{c}\} = \mathrm{span}\{c_g - c_h\}$, and Π the orthogonal projector onto $\mathcal{B}$. For a query $t \neq 0$ write $\hat{t} = t/\|t\|$ and $\gamma = \|\Pi\hat{t}\| \in [0,1]$. For $\lambda \in [0,1]$ define the partial correction $t'_\lambda = \|t\|\,(\hat{t} - \lambda\Pi\hat{t})/\|\hat{t} - \lambda\Pi\hat{t}\|$, which is defined whenever $\gamma < 1$ or $\lambda < 1$; the implementation of Sec. 3.4 uses $\lambda = 1$, and $\lambda < 1$ is the equalization strength swept in App. J. Statements about t'_1 assume $\gamma < 1$, that is $t \notin \mathcal{B}$; in the degenerate case the implementation returns the vanishing direction (App. B) and nothing is claimed.

Theorem 2 (Query equalization)**.** *With the notation above:*

(i) For all g, h, $\langle t'_\lambda, c_g - c_h\rangle = \kappa_\lambda \langle t, c_g - c_h\rangle$ with $\kappa_\lambda = (1-\lambda)/\sqrt{1-\gamma^2+(1-\lambda)^2\gamma^2} \in [1-\lambda, 1]$. At $\lambda = 1$ the inner products of the corrected query with all repaired centroids coincide, $\langle t'_1, c_g\rangle = \langle t'_1, \bar{c}\rangle$ for every g.

(ii) $\hat{t} - \Pi\hat{t}$ *is the unique minimizer of* $\|v - \hat{t}\|$ *over* $\{v : \langle v, c_g - c_h\rangle = 0 \text{ for all } g, h\} = \mathcal{B}^{\perp}$*, a subspace of codimension at most* $G - 1$*.*

(iii) $\|t'_1 - t\| = \|t\|\sqrt{2 - 2\sqrt{1-\gamma^2}} \le \sqrt{2}\,\gamma\,\|t\|$*. Consequently, for gallery embeddings* x, y *of norm at most one, if* $\langle t, x\rangle - \langle t, y\rangle > 2\|t'_1 - t\|$ *then* $\langle t'_1, x\rangle > \langle t'_1, y\rangle$*: every pair of gallery items separated by more than* $2\sqrt{2}\,\gamma\|t\|$ *keeps its order.*

Proof. (i) Since $c_g - c_h \in \mathcal{B}$, $\langle \hat{t} - \lambda\Pi\hat{t}, c_g - c_h\rangle = \langle \hat{t}, c_g - c_h\rangle - \lambda\langle \Pi\hat{t}, c_g - c_h\rangle = (1-\lambda)\langle \hat{t}, c_g - c_h\rangle$, using $\langle \Pi\hat{t}, v\rangle = \langle \hat{t}, v\rangle$ for $v \in \mathcal{B}$. By orthogonality $\|\hat{t} - \lambda\Pi\hat{t}\|^2 = \|(I - \Pi)\hat{t}\|^2 + (1-\lambda)^2\|\Pi\hat{t}\|^2 = 1 - \gamma^2 + (1-\lambda)^2\gamma^2$, which lies in $[(1-\lambda)^2, 1]$; multiply by $\|t\|$. At $\lambda = 1$ the factor vanishes, and $c_g - \bar{c} \in \mathcal{B}$ gives the second statement. (ii) The constraint set is exactly $\mathcal{B}^{\perp}$, and the closest point of a closed subspace is the orthogonal projection; $\dim \mathcal{B} \le G - 1$ because the G vectors $c_g - \bar{c}$ satisfy one linear relation $\sum_g \pi_g(c_g - \bar{c}) = 0$. (iii) $t'_1 = \|t\|\,(I - \Pi)\hat{t}/\|(I - \Pi)\hat{t}\|$ with $\|(I - \Pi)\hat{t}\| = \sqrt{1-\gamma^2} > 0$, so $\|t'_1 - t\|^2 = \|t\|^2\,(2 - 2\langle \hat{t}, (I - \Pi)\hat{t}\rangle/\sqrt{1-\gamma^2}) = \|t\|^2(2 - 2\sqrt{1-\gamma^2})$, and $1 - \sqrt{1-\gamma^2} \le \gamma^2$. For the ranking statement, $|\langle t'_1 - t, v\rangle| \le \|t'_1 - t\|$ for $\|v\| \le 1$, hence $\langle t'_1, x\rangle - \langle t'_1, y\rangle \ge \langle t, x\rangle - \langle t, y\rangle - 2\|t'_1 - t\| > 0$. □

Remark 2. The transport's effect on the remaining score-distribution differences is measured, not derived (Sec. 4.3), and class prompts are not equalized (Sec. 3.5), so the theorem makes no statement about the classification path.