Last-Block Branch Reweighting: Training-Free Enhancement of CLIP-Based Cross-Modal Retrieval

Anonymous ACL submission

Abstract

Contrastively trained vision-language models such as CLIP encode images and text into compressed, anisotropic cones on the unit hypersphere, and this geometry limits cross-modal retrieval. We propose **Last-Block Branch Reweighting** (**LBR**), a training-free intervention that scales the residual and attention branches of the *final* encoder block with two scalar parameters, applied first to the vision encoder and then symmetrically to the text encoder. The mechanism behind LBR's gain is an implicit form of MLP attenuation mediated by the LayerNorm bottleneck: rebalancing attention rebalances how much the MLP contributes to the block output, without modifying any MLP weight. This account predicts—and we confirm—that the optimal direction of reweighting is determined by each encoder's branch balance (yielding opposite signs on CLIP and SigLIP), and that vision-side and text-side LBR contribute asymmetrically to the two retrieval directions (vision → T2I, text → I2T), combining near-additively into a joint variant that outperforms either alone. A label-free geometry score selects near-optimal hyperparameters without any retrieval labels, enabling plug-and-play deployment on frozen encoders.

1 Introduction

Contrastive vision-language models (Radford et al., 2021; Jia et al., 2021) produce aligned image and text embeddings comparable by cosine similarity, enabling zero-shot cross-modal retrieval without task-specific training. Despite their success, the geometry of these embedding spaces limits retrieval performance. Liang et al. (2022) showed that CLIP embeddings of both modalities cluster into narrow, anisotropic cones on the unit hypersphere, separated by a substantial *modality gap*; within each modality, a small number of principal directions dominate the variance, leaving many embedding dimensions nearly empty. Schrodi et al. (2025) connected this geometry to an information imbalance between vision and language: images carry richer low-level detail than typical captions, causing the two modalities to settle into separated regions during contrastive training.

Correcting this geometry without retraining is attractive because fine-tuning CLIP is expensive, many production systems treat the backbone as a frozen feature extractor, and post-hoc corrections preserve compatibility with cached embeddings. Existing training-free methods such as all-but-the-top normalization (Mu and Viswanath, 2018) and PCA-based whitening operate on the *output* embeddings, independently of how those embeddings were formed.

We take a different perspective. During contrastive training, gradient flow reaches the vision encoder's last block most directly, through the projection head, and shapes that block's attention-to-residual branch ratio in service of the training objective. This ratio is optimized for gradient-based learning dynamics, not for nearest-neighbor retrieval; by adjusting it at inference time—without any weight updates—we can reshape the image geometry toward one better suited to cosine-similarity ranking.

This motivates **Last-Block Branch Reweighting (LBR)**: for the final vision encoder transformer block, we replace the standard attention sub-layer update, $x \leftarrow x + \mathrm{Attn}(\mathrm{LN}_1(x))$, with

$$x \;\leftarrow\; \alpha \cdot x \;+\; \beta \cdot \mathrm{Attn}(\mathrm{LN}_1(x)), \qquad (1)$$

where $\alpha, \beta \in \mathbb{R}$ are scalar hyperparameters and the MLP branch is unchanged. The standard forward pass corresponds to $(\alpha, \beta) = (1, 1)$. The method adds two scalar multiplications per token per image and requires no labeled data.

Our contributions are threefold:

1. **A training-free, dual-encoder method.** We

introduce LBR and show that applying it symmetrically to both encoders yields complementary gains—vision-side primarily improves T2I, text-side primarily improves I2T—which combine near-additively into a joint variant that consistently outperforms vision-only LBR across CLIP-B/32, B/16, and L/14 on Flickr30k and MSCOCO val2017.

2. **A mechanism with cross-architecture predictive power.** We trace LBR's gain to a small cross-dimensional remix of the embedding space and identify its underlying cause as implicit MLP attenuation through the LayerNorm bottleneck. Direct γ-scaling of the MLP branch alone reproduces vision-LBR's effect when matched to the same MLP fraction, an observation that characterizes the mechanism. This account explains both SigLIP's opposite optimal direction and the otherwise puzzling rank of LBR's gain across model scales (smaller patches, tighter projection bottlenecks, larger gain).

3. **Label-free deployment.** We propose an unsupervised geometry score combining anisotropy, hubness, and spectral concentration, which selects near-optimal hyperparameters from an unlabeled set with Spearman correlation above 0.75 against the supervised optimum across all three datasets, enabling plug-and-play use on frozen encoders.

2 Related Work

Contrastive vision-language models and retrieval. CLIP (Radford et al., 2021) and ALIGN (Jia et al., 2021) established contrastive image-text pre-training; subsequent work expanded the recipe with generative objectives (Li et al., 2022; Yu et al., 2022), alternative losses such as SigLIP's sigmoid pairwise loss (Zhai et al., 2023), and large-scale curated datasets (Cherti et al., 2023; Schuhmann et al., 2022). Flickr30k (Young et al., 2014) and MSCOCO (Lin et al., 2014) are the standard intrinsic benchmarks; classic methods such as VSE++ (Faghri et al., 2018) and SCAN (Lee et al., 2018) train task-specific retrieval heads on top of these encoders, whereas LBR improves retrieval from a fully frozen encoder without any labels.

Modality gap and embedding geometry. Liang et al. (2022) formalized the modality gap as two narrow, anisotropic cones separated by a near-constant offset; Schrodi et al. (2025) attribute the gap and an associated object-orientation bias to information imbalance between images and captions. Huh et al. (2024) observe that diverse large-scale models converge toward a shared statistical structure, suggesting the anisotropy is systemic. Beyond vision-language, the broader literature on pre-trained embeddings finds that anisotropy correlates with degraded retrieval performance (Ethayarajh, 2019; Cai et al., 2021; Rajaee and Pilehvar, 2021), quantified e.g. by effective rank (Roy and Vetterli, 2007). Our analysis confirms LBR does *not* minimize the modality gap (it is nearly identical for B/32 and L/14; §4.2), arguing against gap closure as the mechanism.

Post-hoc correction and lightweight CLIP adaptation. All-but-the-top (Mu and Viswanath, 2018) and sentence whitening (Su et al., 2021) reduce anisotropy by operating on the final output embeddings of pre-trained encoders; SimCSE (Gao et al., 2021) achieves a similar effect via contrastive fine-tuning. Weight-space methods such as WiSE-FT (Wortsman et al., 2022) and patching (Ilharco et al., 2022) interpolate or merge full parameter vectors across all layers. Closest to our intervention, the Attention Ablation Technique (AAT) (Lin et al., 2025) refines CLIP's image representation by identifying and suppressing individual attention heads across the vision encoder—a training-free, head-level manipulation of attention weights. LBR sits in a different regime along several axes: its parameterization is continuous rather than discrete (two scalars (α, β) vs. a binary per-head decision), its intervention point is a single encoder block rather than multiple layers, it acts on the attention-to-residual *branch balance* rather than on individual heads, and it extends symmetrically to the text encoder. Like AAT, LBR acts *inside* the encoder rather than on its output, and requires no data at all.

3 Method

3.1 LBR Formulation

Let f_θ be a CLIP-style vision encoder comprising L transformer blocks. Block l has the standard

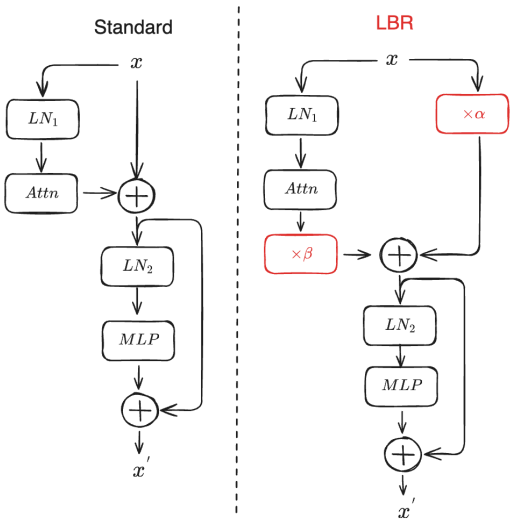

Figure 1: One transformer block under the standard forward pass (left) and under LBR (right). LBR inserts two scalars α (residual) and β (attention output) in the last block only; the MLP sub-layer and all other blocks are unchanged. At $(\alpha, \beta) = (1, 1)$ the two diagrams are identical.

residual form:

$$\begin{aligned} h &= \mathrm{LN}_1(x), \\ x &\leftarrow x + \mathrm{Attn}(h), \\ x &\leftarrow x + \mathrm{MLP}(\mathrm{LN}_2(x)). \end{aligned} \tag{2}$$

LBR replaces Eq. (2) in the *last* block only ($l = L$):

$$x \leftarrow \underbrace{\alpha \cdot x}_{\text{scaled residual}} + \underbrace{\beta \cdot \mathrm{Attn}(\mathrm{LN}_1(x))}_{\text{scaled attention}}. \tag{3}$$

The MLP sub-layer of block L and all preceding blocks are unchanged. At $(\alpha, \beta) = (1, 1)$, Eq. (3) reduces exactly to the original forward pass; all weights θ remain frozen. Figure 1 illustrates the modification.

We search over a discrete grid of 25 points: $\alpha, \beta \in \{0.8, 1.0, 1.2, 1.4, 1.6\}$.

3.2 Why the Last Block?

Contrastive training optimizes a temperature-scaled cross-entropy loss over batch cosine similarities of projected image and text features. Gradients flow through the projection head and directly into the last encoder block; layers closer to the loss receive stronger and more direct gradient signals than earlier layers. The last block's attention-to-residual ratio therefore converges toward a configuration optimized for training stability—balancing the scale of the gradient signal with representation quality—rather than for inference-time nearest-neighbor retrieval, which has different geometric requirements (uniform coverage of the embedding sphere vs. sharp positive/negative separation within a training batch).

This reasoning makes two testable predictions: (i) only layers close to the projection head should benefit from branch reweighting; (ii) early layers, whose representations serve intermediate processing rather than final output, should be harmed. Both are confirmed by the layer-position ablation (§4.6).

3.3 Symmetric Extension to the Text Encoder

The gradient-flow argument above is encoder-agnostic: a contrastive loss applied to the cosine similarity of projected image and text features backpropagates through both projection heads, locking branch ratios in the last block of *each* encoder. The standard CLIP text encoder has the same residual block structure as the vision encoder ($\mathrm{LN}_1 \rightarrow$ Attention $\rightarrow$ residual add $\rightarrow \mathrm{LN}_2 \rightarrow$ MLP $\rightarrow$ residual add), so LBR's formulation transfers without modification. We apply Eq. (3) to the text encoder's last block with its own pair of scalars (α_t, β_t):

$$u \leftarrow \alpha_t \cdot u + \beta_t \cdot \mathrm{Attn}_t(\mathrm{LN}_1(u)), \tag{4}$$

yielding a dual-encoder method governed by four scalars $(\alpha_v, \beta_v, \alpha_t, \beta_t)$ in total. The vision-only variant fixes $\alpha_t = \beta_t = 1$, the text-only variant fixes $\alpha_v = \beta_v = 1$, and the joint variant tunes all four. As we show in §4.5, the two sides contribute asymmetrically to retrieval directions (vision-side mostly improves T2I, text-side mostly improves I2T) and combine near-additively across CLIP variants.

3.4 Unsupervised Hyperparameter Selection

To avoid the need for a held-out retrieval set, we propose a label-free scoring function. For a candidate (α, β), we extract image embeddings $\{\boldsymbol{v}_i\}$ and text embeddings $\{\boldsymbol{u}_j\}$ from an unlabeled set and compute:

$$s(\alpha, \beta) = -\bar{c}_v - \bar{c}_u - G_{v \to u} - G_{u \to v} - r_v^{(1)} - r_u^{(1)}, \tag{5}$$

where $\bar{c}_v$ is the mean pairwise cosine similarity among image embeddings (anisotropy); $G_{v \to u}$ is the Gini coefficient of the image-to-text cosine similarity distribution (hubness (Radovanović

et al., 2010)); and $r_v^{(1)}$ is the fraction of variance captured by the leading principal component of image embeddings. Text-side terms $(\bar{c}_u, G_{u\to v}, r_u^{(1)})$ are defined symmetrically. All six terms measure geometric concentration; the score s is maximized when embeddings are spread uniformly. The optimal (α, β) is $\arg\max_{\alpha,\beta} s(\alpha, \beta)$.

4 Experiments

4.1 Setup

Datasets. We evaluate on Flickr30k (Young et al., 2014) (1,000 test images, 5 captions each) and MSCOCO val2017 (Lin et al., 2014) (5,000 images, ≈5 captions each).

Models. We evaluate CLIP-ViT-B/32 (Radford et al., 2021) and CLIP-ViT-L/14 trained on LAION-2B (Cherti et al., 2023) as primary models, and SigLIP-base-patch16-224 (Zhai et al., 2023) for cross-architecture generalization.

Evaluation protocol. We report R@k ($k \in \{1, 5, 10\}$) for image-to-text (I2T) and text-to-image (T2I) retrieval, and Mean Recall (MeanR = mean of all six values). We use the *multi-caption protocol*: for I2T, an image is deemed retrieved at rank k if it appears in the top-k for *any* of its five captions; for T2I, a caption is a hit if its ground-truth image is in the top-k. CLIP text features are extracted with `padding=max_length` (77 tokens) to preserve positional embedding alignment (Radford et al., 2021).

Statistical testing. We report 95% confidence intervals via 1,000-iteration paired bootstrap over image-level hit indicators (Koehn, 2004). A result is "significant" if its CI excludes zero.

Scope, baselines, and reproducibility. LBR is not designed to compete with parameter-efficient fine-tuning (LoRA (Hu et al., 2022), BitFit (Ben Zaken et al., 2022), VPT (Jia et al., 2022)), which require gradient access and labeled data. LBR targets the regime where the backbone is a black-box or frozen feature extractor, labels are unavailable, or no backward pass is affordable; its relevant baselines are therefore other zero-gradient corrections (all-but-the-top, PCA whitening; §4.4). LBR is in principle composable with fine-tuning methods; we leave a systematic study of LBR-on-top-of-LoRA to future work. All code is submitted as supplementary material and will be released publicly upon acceptance.

Table 1: Embedding geometry for CLIP-B/32 and CLIP-L/14. Both modalities exhibit anisotropy, but L/14 is markedly more isotropic. The modality gap is nearly identical (0.764 vs. 0.771), ruling it out as the mechanism for any retrieval difference.

	B/32		**L/14**	
Metric	Image	Text	Image	Text
Anisotropy	0.487	0.423	0.364	0.267
Top-1 SV share (%)	49.7	45.7	37.5	27.6
Eff. rank / D (%)	61.6	59.2	52.7	46.8
Modality gap	0.764		0.771	

4.2 Embedding Geometry

Table 1 characterizes the embedding geometry of CLIP-B/32 (evaluated on Flickr30k test) and CLIP-L/14 (MSCOCO val2017). We report: (1) anisotropy = mean pairwise cosine similarity (higher = more cone-like); (2) top-1 singular value share; (3) effective rank (Roy and Vetterli, 2007) normalized by embedding dimension; (4) modality gap = cosine distance between mean image and text embeddings.

B/32 shows high anisotropy on both modalities, with nearly half the variance concentrated in a single singular direction. L/14 is markedly flatter (image anisotropy 0.364 vs. 0.487). The modality gap is nearly identical for both models, indicating that gap size alone does not explain retrieval differences—a point we return to with the full anisotropy-gain trend in §4.7.

4.3 Retrieval Performance

We first compare four vision-only conditions that isolate different components of LBR's effect: **A** (baseline, $\alpha = \beta = 1$); **B** (diagonal-only: per-dimension scaling extracted from the best linear fit to LBR outputs, without cross-dimensional mixing); **C** (full LBR on the vision encoder: Eq. (3)); **D** (full linear map: best unconstrained $D \times D$ affine approximation to LBR's transformation of image features).

Results are in Table 2. The consistent pattern **B** ≈ **A** ≪ **C** ≈ **D** holds on both Flickr30k and MSCOCO, with C achieving +0.47 MeanR over A on Flickr30k and +0.67 on MSCOCO (95% CI: [+0.33, +0.84]). Condition D is an *oracle* that applies the optimal unconstrained linear map fitted on the evaluation data itself; the fact that C nearly matches D confirms that vision-only LBR captures the relevant transformation well with just two parameters. Adding text-side LBR on top of C (joint

variant; see §4.5) further improves Mean Recall by up to $+0.59$ on the same backbone.

4.4 Mechanism: Cross-Dimensional Remix

The A/B/C/D comparison already indicates that LBR's dominant diagonal component is irrelevant; here we quantify it precisely.

Diagonal is inert; off-diagonal carries the gain. We fit a $D \times D$ affine map W from LBR-transformed image features $\boldsymbol{v}_{\text{lbr}}$ to baseline features $\boldsymbol{v}_{\text{base}}$ (condition D); at $(\alpha, \beta) = (1.4, 1.4)$ on Flickr30k it achieves $R^2 = 0.998$, with $R^2_{\text{diag}} = 0.976$ from the diagonal of W alone. Condition B isolates this diagonal component, and the result B $\approx$ A on both datasets (MeanR gap $-0.05/+0.03\%$) shows the 97.6% diagonal piece carries no retrieval benefit; essentially all gain is in the 2.4% off-diagonal remix. The same pattern holds for SigLIP, whose stronger off-diagonal ($R^2_{\text{diag}} = 0.865$ vs. 0.976 for CLIP) tracks its larger relative C-over-B improvement (§4.7).

Anti-compression. To characterize what the cross-dimensional component does, we compute, for each singular direction of the baseline image embeddings, the Spearman correlation between (i) the degree to which the baseline suppresses that direction (ranked by ascending energy) and (ii) the energy gain after LBR. $\rho = 0.444$, $p < 10^{-25}$: LBR disproportionately amplifies previously-suppressed directions. This *anti-compression* partially counteracts the anisotropic clustering identified by Liang et al. (2022).

Branch ablation: MLP attenuation is the dominant path. We test four variants on Flickr30k (CLIP-B/32, $\alpha = 1.4$, $\beta = 1.6$): **full LBR** ($+0.85$ MeanR); **MLP-only attenuation** ($\gamma \cdot$MLP with $\alpha = \beta = 1$), which a fine sweep ($\gamma \in \{0.50, 0.55, \ldots, 1.00\}$, step 0.05) peaks at $\gamma^* = 0.70$ with $+0.82$ MeanR (**97% of full LBR**); **attention-only** (α, β as in LBR, MLP unchanged): $+0.17$; **residual-only** (α modified, $\beta = 1$): ≈ 0. Direct MLP attenuation alone nearly matches full LBR.

Mechanism equivalence: a key operative quantity is the MLP fraction. Instrumenting the last block, we measure the MLP's relative contribution to the block output, $f_{\text{MLP}} = \mathbb{E}_{\text{token}}\big[\|\text{MLP}(\text{LN}_2(x_1))\| / \|x_2\|\big]$: baseline 0.402, under LBR 0.286 (a 29% relative reduction) *without modifying any MLP weight*. Choosing γ so direct attenuation matches the same $f_{\text{MLP}} = 0.286$ (interpolated $\gamma \approx 0.66$) yields MeanR $= 83.68$, only $+0.07$ below LBR's 83.75. The two paths reach the same f_{MLP} by *opposite* means: direct γ shrinks the numerator ($\|\text{MLP}\|$: $6.67 \to 4.43$), while LBR inflates the denominator ($\|x_2\|$: $16.73 \to 22.92$). Under LBR, $\alpha, \beta > 1$ scale x_1 before LN_2 normalizes its variance, so the MLP input distribution and $\|\text{MLP}\|$ itself stay near baseline (6.50 vs. 6.67); only the MLP's relative share of the residual stream shrinks. The retrieval-relevant quantity is therefore the MLP's *relative* contribution. LBR can be read as an implicit MLP-attenuation operator mediated by the LayerNorm bottleneck, with the residual $+0.07$ beyond pure MLP attenuation reflecting a small complementary contribution from attention reweighting. Figure 2 visualizes this: MeanR is U-shaped in f_{MLP} with optimum near 0.30, and LBR's $(1.4, 1.6)$ falls on the curve at $f_{\text{MLP}} = 0.286$ with a $+0.07$ vertical offset over the γ-only path.

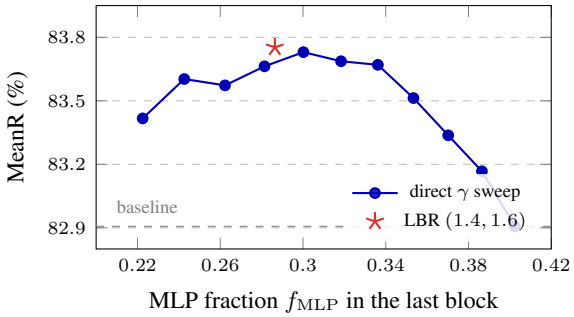

Figure 2: MeanR vs. MLP fraction f_{MLP} on Flickr30k (CLIP-B/32). Direct γ-only attenuation (blue, $\gamma \in [0.5, 1.0]$) traces a U-shaped curve with optimum near $f_{\text{MLP}} \approx 0.30$. LBR's $(\alpha, \beta) = (1.4, 1.6)$ (red $\star$) lands at $f_{\text{MLP}} = 0.286$ with MeanR 83.75, just $+0.07$ above the γ-only curve at the same fraction. The two parameterizations move along nearly the same response surface.

Comparison to post-hoc whitening. Since MLP attenuation reduces anisotropy, one might ask whether standard post-hoc techniques achieve similar results. On CLIP-B/32 Flickr30k, all-but-the-top (Mu and Viswanath, 2018) at its best $D = 2$ yields ΔMeanR $= -0.61$ and PCA whitening collapses recall entirely (MeanR ≈ 0.5); LBR achieves $+0.85$. CLIP's leading singular vectors encode cross-modal alignment rather than anisotropic frequency bias, so removing them discards the information retrieval depends on. LBR instead redistributes energy *inside* the encoder.

Table 2: Retrieval results for the four experimental conditions. **A** = baseline ($\alpha=\beta=1$); **B** = diagonal scaling only (per-dimension, no cross-dimensional remix); **C** = full LBR; **D** = oracle unconstrained linear map. Best non-oracle result per column in **bold**. †: bootstrap 95% CI [+0.33, +0.84] vs. A (MSCOCO, n=5K images). Flickr30k uses CLIP-B/32 ($\alpha=1.4$, $\beta=1.6$); MSCOCO val2017 uses CLIP-B/32 ($\alpha=1.5$, $\beta=1.75$).

	Flickr30k (1,000 test images)							*MSCOCO val2017* (5,000 images)			
	I2T R@k (%)			T2I R@k (%)			MeanR	I2T	T2I	MeanR	
Cond.	@1	@5	@10	@1	@5	@10	(%)	R@1	R@1	(%)	Δ
A	79.4	95.0	98.1	58.84	83.46	90.04	84.14	50.0	30.42	59.92	—
B	78.9	94.5	98.0	59.02	83.80	**90.30**	84.09	49.9	30.57	59.95	+0.03
C	**79.6**	**95.4**	**98.5**	**59.90**	**84.12**	90.16	**84.61**	**51.2**	**31.03**	**60.59**†	**+0.67**†
D	79.6	95.2	98.4	59.42	83.98	90.26	84.48	50.7	31.41	60.60	+0.68

4.5 Text-Side and Joint LBR

The mechanism is encoder-agnostic, so we apply LBR symmetrically to the text encoder's last block (Eq. (4)) and test combination with vision-side, running a 4×4 text grid $(\alpha_t, \beta_t) \in \{0.6, 0.8, 1.0, 1.2\} \times \{1.0, 1.2, 1.4, 1.6\}$ at the vision optimum on three CLIP variants and SigLIP on both datasets (Table 3).

Text-side direction mirrors vision-side; joint LBR is near-additive. Across CLIP variants and datasets, the text-side optimum sits in the $\alpha_t < 1$, $\beta_t > 1$ region: $(0.8, 1.2)$ for B/32/B/16 and SigLIP-Flickr30k, $(0.6, 1.0)$ for L/14-COCO—mirror of vision-side CLIP's $(\alpha_v, \beta_v) > 1$, again determined by each encoder's baseline branch balance. Joint LBR improves on each of the six CLIP cells over vision-only (+0.07 to +0.59 marginal MeanR), peaking at B/16-COCO (+0.97 → +1.45). The two sides decompose cleanly by direction (vision-side → T2I, text-side → I2T), targeting complementary failure modes (gallery hubness vs. query distortion). Notably, joint LBR raises L/14-COCO's gain from +0.17 to +0.57, *mostly via the text side*: even when the visual representation has little suppressed structure to redistribute, the text encoder still does. The exception is SigLIP (margin +0.08/0.00), discussed in §4.7.

4.6 Layer Position Ablation

Table 4 reports ΔMeanR for LBR applied to individual CLIP-B/32 layers on Flickr30k with $(\alpha, \beta) = (1.4, 1.6)$.

The pattern is monotone with layer depth: the last layer gives the largest gain (+0.85); third-to-last also helps (+0.54); early and middle layers hurt (−2.10 to −1.02). This matches the gradient-flow argument of §3.2: stronger training-time gradient lock leaves more room for beneficial adjustment.

Table 3: Vision-only and joint LBR gains on three CLIP variants and SigLIP. ΔMeanR is relative to the unmodified baseline. Joint LBR uses the model-specific vision optimum from Table 6 and the text optimum found by a 4×4 text-side grid; "Δ_t" is the text-side marginal contribution (joint − vision-only).

Model	Dataset	vision	joint	Δ_t
B/32	Flickr30k	+0.85	+1.12	+0.28
	COCO17	+0.59	+0.98	+0.40
B/16	Flickr30k	+1.00	+1.11	+0.24
	COCO17	+0.97	+1.45	+0.59
L/14	Flickr30k	+0.20	+0.27	+0.07
	COCO17	+0.17	+0.57	+0.40
SigLIP	Flickr30k	+0.54	+0.62	+0.08
	COCO17	+0.54	+0.54	+0.00

Table 4: Layer-position ablation on Flickr30k (CLIP-B/32, $\alpha = 1.4$, $\beta = 1.6$, 12 layers). MeanR and ΔMeanR relative to the ablation baseline (82.91).

Position	Layer	MeanR (%)	Δ MeanR
First	0	80.81	−2.10
Middle	6	81.89	−1.02
Third-to-last	9	83.45	+0.54
Second-to-last	10	83.03	+0.12
Last	**11**	**83.75**	**+0.85**
Baseline	—	82.91	—

4.7 Cross-Architecture Generalization

SigLIP and the architecture-dependent direction of LBR. Table 5 shows the A/B/C/D comparison for SigLIP-base-patch16-224 on Flickr30k with $(\alpha, \beta) = (1.4, 0.6)$. The B ≈ A < C ≈ D pattern holds, indicating the cross-dimensional remix generalizes across architectures. SigLIP's optimal $\beta = 0.6 < 1$ is *opposite* to CLIP's $\beta > 1$; the implicit-attenuation account of §4.4 explains why. Measuring the last block's MLP fraction f_{MLP} on identical inputs, SigLIP has $f_{\text{MLP}} = 0.72$ vs. CLIP's 0.40: its last

Table 5: A/B/C/D conditions for SigLIP-base-patch16-224 on Flickr30k ($\alpha = 1.4$, $\beta = 0.6$). Linear fit: $R^2 = 0.990$, $R^2_{\text{diag}} = 0.865$. Best non-oracle in **bold**.

Cond.	I2T@1	T2I@1	I2T@5	T2I@5	MeanR
A	89.1	74.66	98.0	92.30	91.50
B	89.3	73.98	98.0	92.20	91.37
C	**89.8**	**75.44**	**98.5**	**93.04**	**92.04**
D	90.1	75.04	98.4	93.00	91.99

Table 6: Anisotropy–gain trend across three CLIP-ViT variants on two datasets. "Aniso" is mean pairwise cosine similarity of image embeddings on Flickr30k (higher = more anisotropic). ΔMeanR is the best-grid gain over baseline, $\alpha, \beta \in \{0.8, 1.0, 1.2, 1.4, 1.6, 1.8\}$.

Model	Aniso	ΔMeanR Flickr30k	ΔMeanR COCO17
L/14	0.361	+0.20	+0.02
B/32	0.487	+0.85	+0.59
B/16	0.510	+1.00	+0.97

block is already MLP-dominated, with attention contributing roughly half as much relative to the residual stream ($\|\text{Attn}\|/\|\text{residual}\| = 0.42$ vs. CLIP's 0.69). Aggressively inflating x_1 via $\beta \gg 1$ would over-attenuate SigLIP's already-dominant MLP and discard useful signal. A milder inflation ($\alpha > 1$, $\beta < 1$) provides just enough f_{MLP} reduction without crossing into over-attenuation. The direction of LBR is thus predictable from a single pre-deployment measurement—the encoder's baseline MLP fraction—rather than being an architectural quirk: $\beta > 1$ when MLP is the minor branch (CLIP, 40%); $\beta < 1$ when MLP is dominant (SigLIP, 72%). Consistent with this milder, less-diagonal transformation, SigLIP has $R^2_{\text{diag}} = 0.865$ vs. CLIP's 0.976 and a larger relative improvement (C over B: +0.67 MeanR vs. +0.52 for CLIP). The $\beta < 1$ direction is robust across datasets: on MSCOCO val2017, the supervised optimum for SigLIP is $(\alpha, \beta) = (1.2, 0.8)$, again with $\beta < 1$, yielding ΔMeanR $= +0.54$. Eight of the top-10 grid points have $\beta \leq 1$, confirming that the architecture-determined direction predicted by SigLIP's MLP fraction holds independent of the evaluation set. Under *joint* LBR (Table 3), text-side adds only +0.08 (Flickr30k) and ~ 0 (MSCOCO) on SigLIP—an order of magnitude smaller than on CLIP. A plausible explanation: SigLIP's sigmoid pairwise loss removes the partition-function pressure CLIP's softmax exerts on the text branch, so text-side branch ratios may already be near retrieval-optimal at convergence; we leave a mechanism-level confirmation to future work.

Scale and anisotropy: a three-point trend on two datasets. We add CLIP-ViT-B/16 to the scale comparison and evaluate all three variants on both Flickr30k and MSCOCO val2017. Table 6 summarizes the supervised-best gain over the same 6×6 grid ($\alpha, \beta \in \{0.8, 1.0, 1.2, 1.4, 1.6, 1.8\}$). The LBR gain ranks *monotonically* with image-embedding anisotropy on both datasets: more isotropic models gain less, more anisotropic models gain more. This is consistent with the anti-compression view of §4.4: LBR redistributes energy toward suppressed singular directions, so a flatter spectrum (L/14) provides little to redistribute, while a more concentrated one (B/16) provides more. The trend is robust across datasets: B/16's gain is essentially identical on the two benchmarks (+1.00 vs. +0.97); B/32 retains a sizeable gain on both (+0.85 vs. +0.59); only L/14 collapses on MSCOCO (+0.02, bootstrap 95% CI: [−0.14, +0.18], n=5,000)—consistent with the picture that once anisotropy falls below a threshold, the signal-to-noise ratio of the redistribution becomes too small to manifest cleanly. The monotone trend supports using *baseline image anisotropy as a pre-deployment indicator* of expected LBR benefit; we recommend measuring it before deploying LBR on a new backbone.

Anisotropy is governed by the projection-head bottleneck. B/16 having *higher* anisotropy and *larger* LBR gain than B/32 is consistent with our mechanism: the two models share architecture and training data (12 layers, 768 hidden, 512-dim projection, OpenAI WIT-400M) and differ only in patch size, so B/16's 4× more patch tokens are aggregated into the same final 512-dim embedding, imposing a tighter projection-head bottleneck and raising anisotropy (0.51 vs. 0.49). L/14 is the opposite case: a 50% larger projection (768-dim) and ∼5× more training data (LAION-2B) relax the bottleneck and yield much lower anisotropy (0.36). LBR therefore benefits encoders under information-bottleneck pressure rather than weaker models in general.

4.8 Unsupervised Hyperparameter Selection

We evaluate the unsupervised score (Eq. (5)) on the 5×5 grid for CLIP-B/32 on Flickr30k (Figure 3).

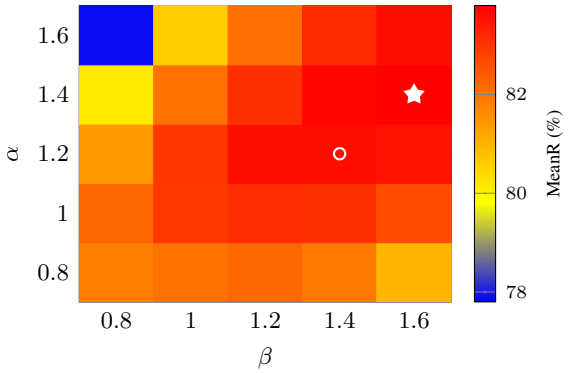

Figure 3: MeanR across the 5×5 (α, β) grid (CLIP-B/32, Flickr30k test). $\star$ = supervised optimum (1.4, 1.6), MeanR 83.75. $\circ$ = unsupervised best (1.2, 1.4), MeanR 83.58, gap 0.17. High MeanR concentrates in the $\beta \geq \alpha$ region.

Table 7: Unsupervised selection results across three datasets (CLIP-B/32). The supervised optimum $(\alpha = 1.4, \beta = 1.6)$ is identical on all three; Gap = MeanR below the supervised optimum.

Dataset	Spearman ρ	Gap	Unsup-best (α, β)
Flickr30k	0.753	0.17	(1.2, 1.4)
MSCOCO17	0.905	0.16	(1.2, 1.6)
MSCOCO14	0.898	0.06	(1.4, 1.8)

The Spearman correlation between unsupervised score and supervised MeanR across all 25 points is **0.753**. The best grid point selected by the unsupervised score ($\alpha = 1.2$, $\beta = 1.4$, MeanR $= 83.58$) lies only **0.17 MeanR points** below the supervised optimum ($\alpha = 1.4$, $\beta = 1.6$, MeanR $= 83.75$). The score correctly penalizes high-α/low-β configurations by detecting the resulting increase in anisotropy and hubness.

Table 7 extends this analysis to MSCOCO val2017 and MSCOCO val2014 (Karpathy test split). The supervised optimum is $(\alpha = 1.4, \beta = 1.6)$ on all three datasets, confirming it is determined by the model's representation geometry rather than the dataset. The unsupervised score achieves higher Spearman correlation on MSCOCO ($\rho \approx 0.90$) than on Flickr30k, and the unsup-best pick is within 0.06–0.17 MeanR of the supervised optimum across all settings. This confirms that our label-free score generalizes beyond the development dataset.

5 Discussion

Training-time LBR and the equivalent γ-only path. Treating (α, β) as learnable scalars and fine-tuning on MSCOCO train2017 (5,000 steps, batch 512, all CLIP weights frozen, lr$= 5 \times 10^{-2}$), (α, β) converges to $\approx (1.15, 1.48)$ with MeanR $+0.34$ (95% CI: [$+0.04$, $+0.63$], significant). Training only γ on the MLP branch under the same recipe converges to $\gamma \approx 0.70$—exactly the inference-time U-curve optimum (Figure 2)—with MeanR $+0.41$ (95% CI: [$+0.17$, $+0.63$]), statistically indistinguishable from the LBR-trained gain. Mechanism equivalence therefore extends to training time: SGD finds the same f_{MLP} operating point regardless of parameterization, and the (α, β) formulation is retained throughout this paper not for a training-time advantage but for its naturalness from the gradient-flow argument (§3.2).

6 Conclusion

We proposed LBR, a training-free reweighting of the residual and attention branches in the last block of CLIP's encoders. Its gain is carried by a small cross-dimensional remix of the embeddings, which we trace to implicit MLP attenuation through the LayerNorm bottleneck—direct γ-scaling of the MLP recovers 97% of vision-only LBR's effect at a matched MLP fraction. The mechanism is encoder-agnostic: applied symmetrically to the text encoder, LBR contributes asymmetrically by direction (vision $\rightarrow$ T2I, text $\rightarrow$ I2T) and combines near-additively, reaching $+1.45$ MeanR on CLIP-B/16 MSCOCO and raising CLIP-L/14's gain from $+0.17$ to $+0.57$. The same account predicts SigLIP's opposite vision-side direction from its MLP-dominated last block. A label-free geometry score selects near-optimal hyperparameters without retrieval labels, and the framework suggests two follow-ups: optimizing the last-block branch ratio during pretraining, and using the geometry score as a general diagnostic for compression-driven retrieval loss.

Limitations

LBR's absolute gains are modest ($\approx$0.2–1.5 Mean Recall depending on model anisotropy and on whether the joint variant applies). The optimal (α, β) is architecture-dependent (CLIP and SigLIP have opposite vision-side optimal directions), so hyperparameter re-search is required when switching backbones; our label-free geometry score mitigates this but does not eliminate it. We also find that text-side LBR does not generalize uniformly: it contributes an additive marginal gain

of $+0.07$ to $+0.59$ MeanR on CLIP but near-zero gain on SigLIP, and we have not pinned this down to a specific property of SigLIP's training objective or text-encoder geometry. All evaluations are on cross-modal retrieval; we have not measured whether the embedding redistribution helps or hurts downstream tasks such as zero-shot classification, compositional reasoning, or captioning, and the implicit-MLP-attenuation account suggests the trade-off may not be uniformly beneficial.

Ethics, Potential Risks, and License

Potential risks. LBR is a training-free, weight-preserving intervention; it introduces no new training data, no new model parameters, and no generation capability, so the most common risk vectors in modern NLP—toxic generation, misinformation, training-data leakage, and privacy violations from memorization—do not arise directly from our method. Two risks nonetheless deserve mention. First, because LBR redistributes embedding energy along previously suppressed singular directions of the underlying encoder, it may alter how known social biases in CLIP-family models (Agarwal et al., 2021; Birhane et al., 2021) surface in retrieval results. We have not measured whether LBR amplifies, attenuates, or merely shifts these biases, and we recommend a targeted bias audit before applying LBR in user-facing systems or in any setting where demographic-sensitive content is retrieved. Second, improvements in cross-modal retrieval quality are dual-use: the same performance gain that benefits search, accessibility, and assistive applications can also lower the cost of image-based content surveillance, person re-identification, and demographic profiling. Practitioners should weigh these trade-offs in the specific deployment context, particularly when LBR is composed with large-scale image indices crawled without consent.

License and use of existing artifacts. The source code accompanying this paper will be released under the MIT License upon acceptance, consistent with our training-free, plug-and-play positioning. All pretrained models and datasets used in our experiments are publicly available and are used in compliance with their respective licenses: CLIP (Radford et al., 2021) under the MIT License; SigLIP (Zhai et al., 2023) under Apache 2.0; LAION-2B (Schuhmann et al., 2022) under CC-BY 4.0; Flickr30k (Young et al., 2014) under its research-use terms; MSCOCO (Lin et al., 2014) under CC-BY 4.0. Our use—inference-time hyperparameter search and feature extraction on standard cross-modal retrieval benchmarks—falls within the stated intended use of these artifacts.

Use of AI assistants. We used AI assistants (large language model chat tools) for code scaffolding on selected experiment scripts and for proofreading and tightening the manuscript. All scientific content—the LBR method, mechanism analysis, experimental design, baselines, interpretation, and numerical results—is the authors' own work, and all released code and reported numbers were audited and verified by the authors.

References

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

Elad Ben Zaken, Yoav Goldberg, and Shauli Ravfogel. 2022. BitFit: Simple parameter-efficient fine-tuning for transformer-based masked language-models. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 1–9.

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

Xingyu Cai, Jiaji Huang, Yuchen Bian, and Kenneth Church. 2021. Isotropy in the contextual embedding space: Clusters and manifolds. In *International Conference on Learning Representations (ICLR)*.

Mehdi Cherti, Romain Beaumont, Ross Wightman, Mitchell Wortsman, Gabriel Ilharco, Cade Gordon, Christoph Schuhmann, Ludwig Schmidt, and Jenia Jitsev. 2023. Reproducible scaling laws for contrastive language-image learning. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2818–2829.

Kawin Ethayarajh. 2019. How contextual are contextualized word representations? Comparing the geometry of BERT, ELMo, and GPT-2 embeddings. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 55–65.

Fartash Faghri, David J Fleet, Jamie Ryan Kiros, and Sanja Fidler. 2018. VSE++: Improving visual-semantic embeddings with hard negatives. In *Proceedings of the British Machine Vision Conference (BMVC)*.

Tianyu Gao, Xingcheng Yao, and Danqi Chen. 2021. SimCSE: Simple contrastive learning of sentence embeddings. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6894–6910.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations (ICLR)*.

Minyoung Huh, Brian Cheung, Tongzhou Wang, and Phillip Isola. 2024. The platonic representation hypothesis. In *Proceedings of the 41st International Conference on Machine Learning (ICML)*. PMLR.

Gabriel Ilharco, Mitchell Wortsman, Samir Yitzhak Gadre, Shuran Song, Hannaneh Hajishirzi, Simon Kornblith, Ali Farhadi, and Ludwig Schmidt. 2022. Patching open-vocabulary models by interpolating weights. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 35.

Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. 2021. Scaling up visual and vision-language representation learning with noisy text supervision. In *Proceedings of the 38th International Conference on Machine Learning (ICML)*, pages 4904–4916. PMLR.

Menglin Jia, Luming Tang, Bor-Chun Chen, Claire Cardie, Serge Belongie, Bharath Hariharan, and Ser-Nam Lim. 2022. Visual prompt tuning. In *European Conference on Computer Vision (ECCV)*, pages 709–727.

Philipp Koehn. 2004. Statistical significance tests for machine translation evaluation. In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 388–395.

Kuang-Huei Lee, Xi Chen, Gang Hua, Houdong Hu, and Xiaodong He. 2018. Stacked cross attention for image-text matching. In *European Conference on Computer Vision (ECCV)*, pages 212–228. Springer.

Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. 2022. BLIP: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *Proceedings of the 39th International Conference on Machine Learning (ICML)*, pages 12888–12900.

Victor Weixin Liang, Yuhui Zhang, Yongchan Kwon, Serena Yeung, and James Y Zou. 2022. Mind the gap: Understanding the modality gap in multi-modal contrastive representation learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 35, pages 17612–17625.

Feng Lin, Marco Chen, Haokui Zhang, Xiaotian Yu, Guangming Lu, and Rong Xiao. 2025. Not all attention heads are what you need: Refining CLIP's image representation with attention ablation. *arXiv preprint arXiv:2507.00537*.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. 2014. Microsoft COCO: Common objects in context. In *European Conference on Computer Vision (ECCV)*, pages 740–755. Springer.

Jiaqi Mu and Pramod Viswanath. 2018. All-but-the-top: Simple and effective postprocessing for word representations. In *International Conference on Learning Representations (ICLR)*.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, and 1 others. 2021. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning (ICML)*, pages 8748–8763. PMLR.

Miloš Radovanović, Alexandros Nanopoulos, and Mirjana Ivanović. 2010. Hubs in space: Popular nearest neighbors in high-dimensional data. *Journal of Machine Learning Research*, 11:2487–2531.

Sara Rajaee and Mohammad Taher Pilehvar. 2021. How does fine-tuning affect the geometry of embedding space: A case study on isotropy. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP*, pages 2329–2338.

Olivier Roy and Martin Vetterli. 2007. The effective rank: A measure of effective dimensionality. In *15th European Signal Processing Conference (EUSIPCO)*, pages 606–610.

Simon Schrodi, Bernt Schiele, Thomas Brox, and 1 others. 2025. Two effects, one trigger: On the modality gap, object bias, and information imbalance in contrastive vision-language representation learning. In *International Conference on Learning Representations (ICLR)*.

Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, and 1 others. 2022. LAION-5B: An open large-scale dataset for training next generation image-text models. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 35, pages 25278–25294.

Jianlin Su, Jiarun Cao, Weijie Liu, and Yangyiwen Ou. 2021. Whitening sentence representations for better semantics and faster retrieval. *arXiv preprint arXiv:2103.15316*.

Mitchell Wortsman, Gabriel Ilharco, Jong Wook Kim, Mike Li, Simon Kornblith, Rebecca Roelofs, Raphael Gontijo Lopes, Hannaneh Hajishirzi, Ali Farhadi, and Ludwig Schmidt. 2022. Robust fine-tuning of zero-shot models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7959–7971.

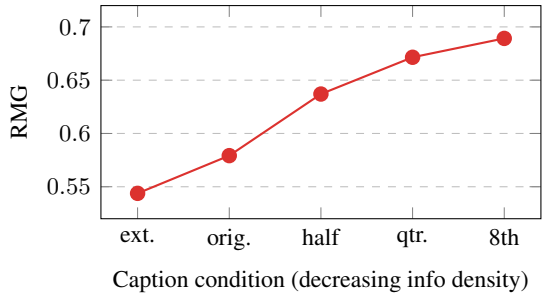

Figure 4: Relative modality gap (RMG) vs. caption information density (CLIP-B/32, MSCOCO val2017). RMG is monotonically increasing as captions lose information, confirming Schrodi et al. (2025)'s prediction.

Peter Young, Alice Lai, Micah Hodosh, and Julia Hockenmaier. 2014. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *Transactions of the Association for Computational Linguistics (TACL)*, 2:67–78.

Jiahui Yu, Zirui Wang, Vijay Vasudevan, Leah Yeung, Mojtaba Seyedhosseini, and Yonghui Wu. 2022. CoCa: Contrastive captioners are image-text foundation models. *Transactions on Machine Learning Research*.

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

A Connection to Information Asymmetry

To connect LBR to the information-asymmetry account of the modality gap proposed by Schrodi et al. (2025), we constructed five caption sets for MSCOCO val2017 by LLM-based expansion/truncation: *extended*, *original*, *half*, *quarter*, *eighth* (progressively less informative). For each set we measured the *relative modality gap* (RMG), defined as the modality gap normalized by the average intra-modality spread of image and text embeddings.

Figure 4 shows RMG monotonically increases as captions become less informative ($0.544 \rightarrow 0.579 \rightarrow 0.637 \rightarrow 0.672 \rightarrow 0.689$), confirming that information asymmetry drives geometric separation as TEOT predicts. This suggests LBR operates in the regime where the asymmetry is greatest: by reweighting the vision branch, it partially compensates for the imbalance between the rich image cone and the narrow text cone.

B Training-Time LBR: Learning-Rate Sensitivity

The training-time LBR experiment of §5 reports a significant $+0.34$ MeanR gain at lr$= 5 \times 10^{-2}$. At the more typical lr$= 10^{-2}$, the gain drops to $+0.14$ (95% CI: $[-0.17, +0.44]$) and is not significant. This LR sensitivity is informative: the (α, β) loss landscape is flat near the training basin, requiring a large step to escape the region where gradient-based pretraining has placed the branch ratios. This is consistent with the picture that contrastive training drives (α, β) toward a near-optimal but not retrieval-optimal configuration, and that LBR's gain comes from the small residual misalignment between training and retrieval objectives.