NC-CLIP: GEOMETRIC FEATURE DECOUPLING FOR ROBUST VISION-LANGUAGE ALIGNMENT

Anonymous authors
Paper under double-blind review

ABSTRACT

Contrastive vision–language learning effectively matches paired images and text, yet pairwise similarity objectives leave the geometry of the aligned representation under-specified and may preserve distribution-specific correlations. We introduce NC−CLIP, a SmartCLIP-initialized dual-encoder framework that combines symmetric shared routing with fixed geometric anchoring. Image and text representations are mapped into a learned routing basis, where modality-conditioned observers predict coordinate-importance logits; averaging the two logits yields a shared fine/coarse routing mask for the matched pair. This construction enforces shared support in the learned routing space without assuming coordinate-wise semantic equivalence. The routed representation is organized by Sinkhorn-balanced fixed anchors, while shared-only cross-modal reconstruction preserves predictive common information without reconstructing the entire opposite-modality embedding. On Waterbirds and CelebA, NC−CLIP improves worst-group accuracy over SmartCLIP by 6.19 and 10.56 percentage points, respectively. Three-seed COCO-5K evaluation further yields consistent gains across all six retrieval metrics. A controlled comparison with random fixed anchors shows that the exact simplex ETF primarily improves routing balance rather than retrieval itself. Together, these results support shared routing and fixed geometric anchoring as effective inductive biases for robust vision–language alignment.

1 INTRODUCTION

Vision–language models are commonly trained with contrastive objectives that pull matched image–text pairs together while separating mismatched pairs. CLIP (Radford et al., 2021) established the effectiveness of this paradigm, and subsequent work has extended it through larger-scale data, longer text, and selective alignment (Alayrac et al., 2022; Zhang et al., 2024; Xie et al., 2025). However, global alignment treats all representation components as equally eligible for cross-modal matching, even though some may encode visual context, lexical regularities, or other distribution-dependent cues. Moreover, because InfoNCE depends on pairwise inner products, applying a shared orthogonal transformation to the embedding space leaves the objective unchanged. Pairwise matching therefore does not uniquely determine the internal geometry of the learned representation.

Selective alignment addresses the first issue by restricting which information participates in cross-modal matching. SmartCLIP (Xie et al., 2025), for example, uses asymmetric text-conditioned visual feature selection, with text determining which visual dimensions are emphasized. Building on its pretrained representation, NC−CLIP replaces asymmetric selection with symmetric routing. Each modality is first mapped into a learned routing basis and processed by a modality-conditioned observer. For each matched pair, the two observer logits are averaged to form a single mask that is applied to both modalities. This guarantees shared support in the *learned routing basis*, while making no assumption that individual coordinates have fixed or human-interpretable semantics.

We further organize the routed representation with a fixed geometric reference. Sinkhorn balancing assigns matched groups to frozen anchors, and our default anchor set is a simplex Equiangular Tight Frame (ETF), which provides a symmetric and maximally separated geometry inspired by Neural Collapse. The role of the ETF is deliberately separated from that of anchoring itself: fixed anchors improve retrieval and routing balance, whereas replacing random fixed anchors with the exact simplex leaves retrieval essentially unchanged but reduces Route KL. We therefore interpret the simplex

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.

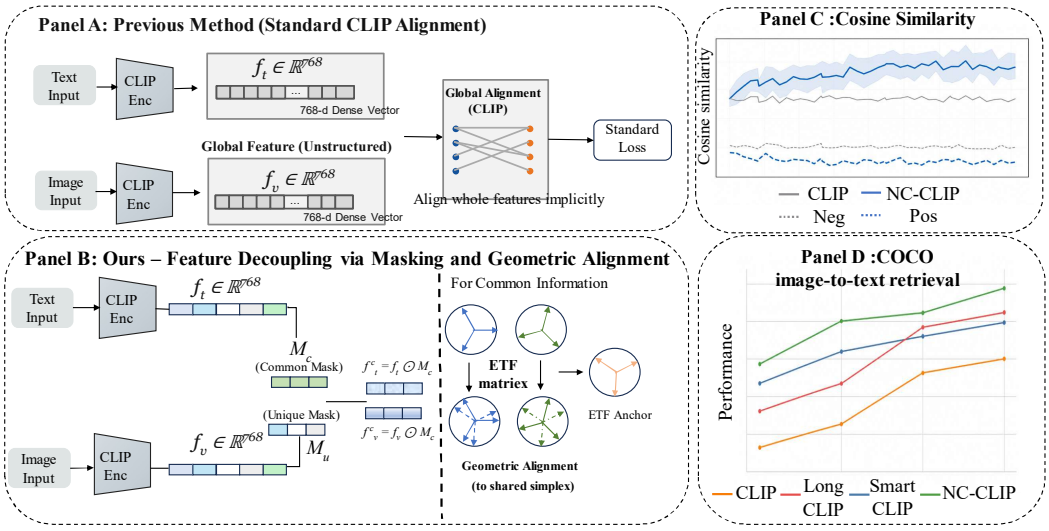


Figure 1: **Geometric comparison between standard CLIP-style alignment and `NC-CLIP`. Panel A:** global contrastive alignment acts on the full representation. **Panel B:** `NC-CLIP` performs shared routing in a learned common basis and applies the same routed support to both modalities before geometric regularization. **Panel C:** training diagnostics track pair separation and alignment to the fixed anchor system. **Panel D:** retrieval improves along the training trajectory as the routed representation becomes more structured. This trajectory is correlational; our controlled exact-ETF-versus-random-anchor ablation later shows that the exact simplex primarily improves route balance rather than retrieval by itself.

primarily as a regularizer of route utilization, rather than as a semantic class system or an independent source of retrieval gains.

Our evaluation separates domain-shift retrieval from explicit spurious-correlation testing. ShareGPT4V and Urban1k measure retrieval under longer captions and compositional or domain variation, while Waterbirds and CelebA provide group-labeled tests of worst-group robustness. We additionally report three-seed SmartCLIP comparisons, joint-versus-independent routing, anchor-geometry ablations, reconstruction robustness, and tuning sensitivity analyses.

Our contributions are:

1. **Symmetric shared routing in a learned basis.** Modality-conditioned observers produce routing logits whose average defines one fine/coarse support for each matched pair. The construction enforces shared support without assuming semantic equivalence between raw image and text coordinates. Under a matched budget, joint routing preserves retrieval performance relative to independent masks while improving cross-modal geometry.

2. **Fixed geometric anchoring with controlled ETF analysis.** Sinkhorn-balanced frozen anchors organize the routed common space. Fixed anchoring improves retrieval and route balance, while the exact simplex further reduces Route KL relative to random fixed anchors without materially changing retrieval.

3. **Direct robustness and statistical evaluation.** On group-labeled benchmarks, `NC-CLIP` improves worst-group accuracy over SmartCLIP by 6.19 points on Waterbirds and 10.56 points on CelebA. Across three matched COCO-5K seeds, all six retrieval metrics improve consistently, with paired gains of 2.57–5.75 percentage points; we report sample standard deviations and paired-difference intervals rather than relying on a single run.

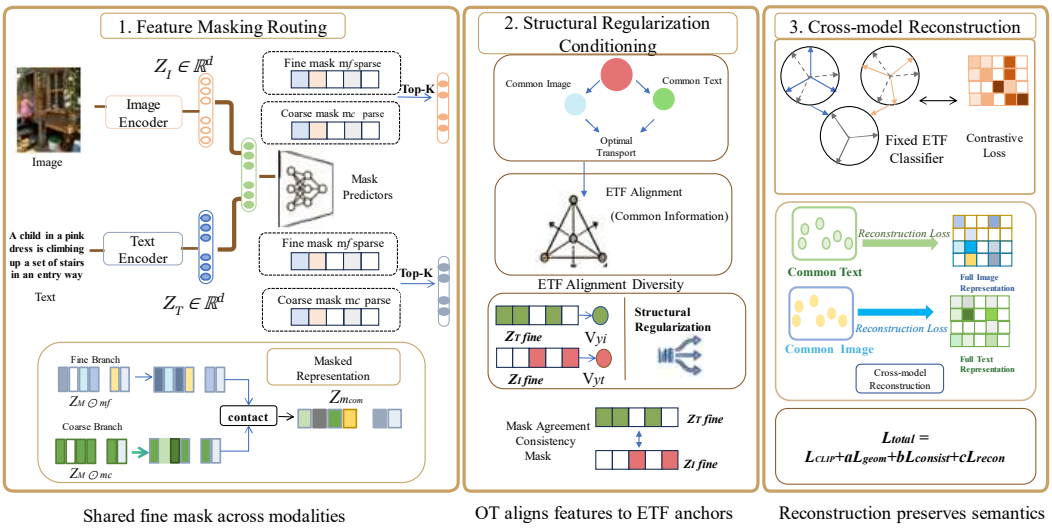


Figure 2: Overview of NC−CLIP. (1) **Shared routing:** image and text representations enter a learned routing basis; modality-conditioned observers predict logits that are averaged into one fine/coarse mask for the matched pair. (2) **Geometric regularization:** the routed common representation is assigned to frozen anchors through Sinkhorn balancing. (3) **Shared-only reconstruction:** the routed representation predicts the opposite modality's shared component rather than its full embedding. The shared mask guarantees common support in the learned routing basis, but does not by itself imply coordinate-wise semantic equivalence.

2 Related Work

2.1 Selective Alignment and Representation Decomposition

CLIP (Radford et al., 2021) aligns global image and text representations, motivating methods that select which information participates in cross-modal matching. SmartCLIP (Xie et al., 2025) performs asymmetric text-conditioned visual feature selection, whereas other work considers text-conditioned patch selection (Zohra et al., 2025; Wang et al., 2025c), spatial masking, or optimization-level modulation (Wu et al., 2024). In contrast, NC−CLIP introduces symmetric shared routing in a learned common basis and directly compares it with an independent-mask alternative.

Information-theoretic approaches provide a complementary perspective. Partial Information Decomposition (Dufumier et al., 2025) and multiview bottleneck formulations (Poklukar et al., 2022; Wang et al., 2025b) characterize shared and unique information at the distribution level. Our focus is operational: we learn a capacity-controlled routed support and then impose a fixed reference geometry on that routed space.

2.2 Neural Collapse–Inspired Representation Alignment

Neural Collapse (NC) describes the late-training geometry of supervised classifiers, where class means and classifier weights approach a simplex ETF (Papyan et al., 2020). Fixed ETF classifiers and related geometric priors have been used in incremental learning (Yang et al., 2023), federated learning (Huang et al., 2023), out-of-distribution generalization (Chen et al., 2024), and vision–language prompt tuning (Zhu et al., 2024). NC−CLIP does not claim that supervised Neural Collapse naturally emerges in vision–language pretraining, and its anchors are not semantic class prototypes. We use the simplex as a fixed geometric reference for balanced routing of unlabeled image–text groups. The exact-simplex ablation in Sec. 5 further separates the benefit of fixed anchoring from the additional effect of the ETF arrangement itself.

3 PRELIMINARIES

3.1 CONTRASTIVE VISION–LANGUAGE ALIGNMENT

We consider a standard vision–language pre-training framework utilizing N paired samples $\mathcal{D} = \{(x_i, t_i)\}_{i=1}^{N}$. A dual-encoder architecture employs an image encoder $f_I : \mathcal{X} \to \mathbb{R}^d$ and a text encoder $f_T : \mathcal{T} \to \mathbb{R}^d$ to project modalities into a shared ℓ_2-normalized embedding space:

$$z_I = \frac{f_I(x)}{\|f_I(x)\|}, \quad z_T = \frac{f_T(t)}{\|f_T(t)\|}, \quad \text{s.t.} \quad z_I, z_T \in \mathbb{S}^{d-1}, \tag{1}$$

where $\mathbb{S}^{d-1}$ denotes the unit hypersphere. Given a mini-batch of size B, alignment is typically learned via the symmetric InfoNCE loss (van den Oord et al., 2018), which maximizes the mutual information between paired samples while minimizing that of negative pairs:

$$\mathcal{L}_{\text{ctr}} = \frac{1}{2B} \sum_{i=1}^{B} \left(\ell_i^{I \to T} + \ell_i^{T \to I} \right), \quad \text{with} \quad \ell_i^{I \to T} = -\log \frac{\exp(\langle z_I^i, z_T^i \rangle / \tau)}{\sum_{j=1}^{B} \exp(\langle z_I^i, z_T^j \rangle / \tau)}, \tag{2}$$

where τ is a learnable temperature scaling parameter.

3.2 NEURAL COLLAPSE AND THE FIXED ETF GEOMETRY

Neural Collapse (NC) describes a striking geometric phenomenon observed in the terminal phase of training deep classifiers (Papyan et al., 2020). As training progresses, the within-class feature variability vanishes, and class prototypes converge to a *Simplex Equiangular Tight Frame* (ETF). This structure represents the most efficient way to pack C vectors in a d-dimensional space such that they are maximally separated.

A set of vectors $\{v_c \in \mathbb{R}^d\}_{c=1}^{C}$ constitutes a simplex ETF if its Gram matrix $G = V^\top V \in \mathbb{R}^{C \times C}$ satisfies: $G = \frac{C}{C-1} I_C - \frac{1}{C-1} \mathbf{1}_C \mathbf{1}_C^\top$, where $V = [v_1, \ldots, v_C]$ and $\mathbf{1}_C$ is an all-ones vector. In this configuration, all vectors have equal norm and share a constant pairwise inner product of $-1/(C-1)$.

While NC is typically an emergent property, recent work has proposed leveraging it as an *explicit inductive bias* by using a *fixed ETF classifier* (Yang et al., 2023). By pre-computing and freezing the classifier weights as an ETF, one can force the backbone to organize representation spaces around a rigid, maximally separated geometric scaffold, bypassing the instability of co-evolving learnable heads.

3.3 GEOMETRIC UNDER-SPECIFICATION IN ALIGNMENT

A limitation of the InfoNCE objective in Eq. (2) is geometric under-specification. Because the loss depends on pairwise inner products, a shared orthogonal transformation $Q \in \mathcal{O}(d)$ leaves the pairwise objective unchanged. This invariance does not by itself prove that a model will exploit spurious features; rather, it shows that pairwise matching alone does not choose a unique internal coordinate system. We therefore treat routing and anchoring as additional inductive biases, and evaluate their consequences empirically rather than presenting geometric identifiability as a guarantee of semantic invariance.

The method targets four concrete desiderata: shared support through modality-conditioned observers averaged into a single route (§4.1); bounded routing capacity through fine/coarse top-k selection (§4.1); stable geometry through frozen anchors with Sinkhorn-balanced assignment (§4.2); and preservation of predictive common information through shared-only cross-modal reconstruction (§4.2).

4 METHODOLOGY

Initialization and inherited components. `NC-CLIP` is initialized from the official SmartCLIP checkpoint (Xie et al., 2025). It retains the dual-encoder architecture, the 248-token text encoder, SIDM/DISM components, and SmartCLIP mask information for bootstrap supervision. On top

of this initialization, we introduce modality-conditioned image/text observers, a shared fine/coarse route for each matched pair, frozen simplex anchors with a shared Sinkhorn target, geometric route regularization, and cross-modal reconstruction.

4.1 Shared Routing and Dual-Granularity Selection

Given a matched image–text pair, the encoders produce normalized representations z_I^i, z_T^i. Rather than assuming coordinate-wise correspondence between the raw image and text embeddings, we first map them through modality-specific learnable projections into a common routing basis,

$$u_I^i = P_I(z_I^i), \qquad u_T^i = P_T(z_T^i). \tag{3}$$

For each routing granularity $r \in \{f, c\}$, modality-conditioned observers predict coordinate-importance logits

$$h_I^{i,r} = g_I^r(u_I^i), \qquad h_T^{i,r} = g_T^r(u_T^i), \tag{4}$$

and the matched pair produces one shared routing vector by averaging:

$$s^{i,r} = \frac{h_I^{i,r} + h_T^{i,r}}{2}. \tag{5}$$

Before averaging, a lightweight observer-logit agreement regularizer (weight 0.05) encourages compatible routing scores across modalities. The resulting construction enforces *shared support in a learned coordinate system*; it does not require individual coordinates to share a fixed semantic interpretation, nor does it assume that the routing space contains only modality-invariant information.

During training, shared logits are converted to differentiable soft gates with a temperature-scaled sigmoid,

$$m_{\text{soft}}^{i,r} = \sigma(s^{i,r}/\tau_m), \tag{6}$$

with τ_m annealed from 2.0 to 0.1. At evaluation, top-k routing yields a binary mask

$$m_{\text{hard},j}^{i,r} = \mathbf{1}\big[j \in \text{TopK}_{k_r}(s^{i,r})\big]. \tag{7}$$

The same $m^{i,r}$ is applied to both modalities in the learned routing basis. Shared support therefore follows by construction, whereas semantic faithfulness remains an empirical property rather than an architectural guarantee.

Dual-granularity selection. The fine route retains $\rho_f = 0.30$ of the dimensions and the coarse route retains $\rho_c = 0.50$. For $M \in \{I, T\}$,

$$\tilde{u}_M^{i,r} = u_M^i \odot m^{i,r}, \qquad r \in \{f, c\}. \tag{8}$$

The two routed branches are normalized independently and concatenated:

$$z_{\text{com},M}^i = \left[\frac{\tilde{u}_M^{i,f}}{\|\tilde{u}_M^{i,f}\|_2} \,\middle\|\, \frac{\tilde{u}_M^{i,c}}{\|\tilde{u}_M^{i,c}\|_2} \right]. \tag{9}$$

The fine route provides a tighter capacity budget, whereas the coarse route preserves broader context. The final retrieval representation is normalized before the contrastive objective is computed.

4.2 Geometric Anchoring, Reconstruction, and Training

Shared routing determines which coordinates participate in alignment, but does not constrain how routed samples are organized geometrically. We therefore introduce C frozen anchors $V = [v_1, \ldots, v_C]$, using a simplex ETF by default,

$$V^\top V = \frac{C}{C-1} I_C - \frac{1}{C-1} \mathbf{1}\mathbf{1}^\top. \tag{10}$$

For each matched group, we form a joint routed representation $\bar{z}_i$ from the fine image/text features and score it against the anchors. Entropy-regularized Sinkhorn normalization (Cuturi, 2013; Caron et al., 2020) yields a balanced assignment target shared by the image and its captions. ETF barycentric route-NCE and a sample-level conditional-collapse constraint then encourage matched groups to share a route while maintaining balanced anchor utilization.

Anchor indices have no predefined class semantics. Moreover, the controlled exact-ETF-versus-random-fixed-anchor comparison isolates the effect of the simplex geometry: Rsum is essentially unchanged (448.92 vs. 448.94), whereas Route KL decreases from 0.0488 to 0.0351. These results indicate that *fixed anchoring* supports retrieval and balanced routing, while the *simplex arrangement* primarily regularizes route utilization.

Shared-only reconstruction. Cross-modal reconstruction prevents routing from discarding predictive information shared across modalities. We use *shared-only* reconstruction: each routed fine representation predicts the shared component of the opposite modality rather than its full embedding. Denoting this target by $z_{\mathrm{shared},M}$,

$$\mathcal{L}_{\mathrm{recon}} = \frac{1}{B}\sum_i \left[\|\mathcal{R}_{I\to T}(\tilde{u}_I^{i,f}) - \mathrm{sg}(z^i_{\mathrm{shared},T})\|_2^2 + \|\mathcal{R}_{T\to I}(\tilde{u}_T^{i,f}) - \mathrm{sg}(z^i_{\mathrm{shared},I})\|_2^2 \right]. \quad (11)$$

This choice is supported by the group-robustness ablation: shared-only reconstruction gives higher worst-group accuracy than full-embedding reconstruction on both Waterbirds and CelebA.

Training objective. Training retains the SmartCLIP objectives and adds the routing and geometric terms described above. Model-selection coefficients are chosen once on a held-out tuning subset and then fixed for all full-scale evaluations. Appendix E reports the selected coefficients and tuning-only sensitivity analyses; no coefficient is retuned for a specific benchmark or backbone. Training uses 500 warm-up steps followed by 3 epochs, with optimizer and batch details given in Sec. 5 and Appendix F.

5 Experiments

We evaluate `NC-CLIP` on standard retrieval, long-caption and domain-shift retrieval, group robustness, zero-shot transfer across eight classification benchmarks, statistical reliability, and targeted design ablations. All results in the main paper use the corresponding full evaluation sets; tuning-only subset sweeps are reported separately in Appendix E.

5.1 Setup and Main Results

Experimental setup. We train on CC3M and initialize `NC-CLIP` from the official SmartCLIP checkpoint, retaining its dual-encoder backbone and 248-token text encoder. Unless otherwise stated, we use $C = 20$ anchors, fine/coarse sparsities of $(0.30, 0.50)$, AdamW, batch size 1024, 500 warm-up steps, and 3 training epochs on 4 NVIDIA A100 40GB GPUs. For the three-seed SmartCLIP comparison, both methods use the same COCO-5K evaluation protocol, sampler schedule, training length, batch size, and random seeds.

Retrieval across standard and shifted captions. Table 1 summarizes retrieval on the full COCO, Flickr30k, ShareGPT4V, and Urban1k evaluation sets. COCO and Flickr30k measure conventional short-caption retrieval, whereas ShareGPT4V and Urban1k stress longer descriptions and compositional or domain variation. Because Urban1k does not provide group annotations, we use it as a domain-shift retrieval benchmark rather than as direct evidence of spurious-correlation mitigation.

Statistical design. The seed-matched comparison starts both methods from the same public SmartCLIP ViT-B/16 checkpoint and holds the sampler schedule, training length, batch size, and three seed identities fixed. The pairing is meaningful for measuring run-to-run differences, although it is not described as a strict common-random-number experiment. With only three seeds, the raw direction and magnitude of the paired effects are emphasized over distributional assumptions.

Claim scope. The evaluation suites serve different roles: COCO and Flickr30k test standard retrieval; ShareGPT4V and Urban1k test longer or shifted captions; Waterbirds and CelebA provide the group labels needed for worst-group analysis. All coefficient and sparsity choices are selected on the fixed tuning protocol in Appendix E and then frozen for the full-set evaluations.

Table 1: Retrieval R@1 (%) on full evaluation sets. I2T/T2I denote image-to-text/text-to-image; SG4V denotes ShareGPT4V.

Backbone	Method	COCO I2T	COCO T2I	Flickr I2T	Flickr T2I	SG4V	Urban1k
B/16	CLIP	51.8	32.7	44.1	24.7	78.9	60.9
	Direct FT	37.4	21.8	25.7	17.9	93.9	–
	LongCLIP	57.6	40.4	46.8	34.1	94.0	79.2
	SmartCLIP	61.9	42.4	55.6	36.3	98.4	88.7
	NC-CLIP	**63.7**	**44.1**	**56.7**	**38.4**	**98.8**	**90.1**
L/14	CLIP	56.1	35.4	48.5	28.0	82.9	60.8
	Direct FT	37.9	23.1	26.0	17.9	95.4	–
	LongCLIP	62.8	46.3	53.4	41.2	95.7	84.4
	SmartCLIP	66.0	48.5	63.9	43.8	98.2	91.6
	NC-CLIP	**68.4**	**49.9**	**65.2**	**45.7**	**98.3**	**93.2**

Matched-seed results. Table 2 reports sample mean $\pm$ standard deviation together with paired gains and intervals. Every metric improves for all three seeds; the raw I2T R@1 values are 60.48/60.44/60.62 for SmartCLIP and 65.38/65.54/65.36 for NC-CLIP.

Table 2: Matched-seed COCO-5K retrieval (%, mean $\pm$ std over three seeds).

Direction	Method / statistic	R@1	R@5	R@10
I2T	SmartCLIP	60.51±0.09	82.96±0.09	89.86±0.09
	NC-CLIP	**65.43±0.10**	**86.79±0.04**	**92.43±0.08**
	Paired gain	4.91	3.83	2.57
	95% CI	[4.47,5.36]	[3.51,4.16]	[2.21,2.94]
	p	0.00045	0.00038	0.00109
T2I	SmartCLIP	42.43±0.06	68.05±0.05	77.78±0.06
	NC-CLIP	**47.89±0.02**	**73.80±0.07**	**82.52±0.02**
	Paired gain	5.47	5.75	4.75
	95% CI	[5.31,5.63]	[5.46,6.04]	[4.55,4.95]
	p	0.00005	0.00014	0.00010

5.2 Analysis and Ablations

Targeted mechanisms. Table 3 separates three questions. Shared routing leaves COCO-5K I2T R@1 effectively unchanged relative to independent masks but improves matched-group cosine and modality gap. Fixed anchoring contributes to retrieval, whereas replacing random fixed anchors with the exact ETF leaves Rsum unchanged and lowers Route KL. Shared-only reconstruction improves worst-group accuracy relative to full-embedding reconstruction on both Waterbirds and CelebA. The routing comparison matches initialization, sparsity, data, losses, training length, and batch size. The reconstruction values are reported descriptively because they use three seeds and a non-smooth worst-group statistic, with the smallest CelebA group containing 180 examples.

Table 3: Group robustness and targeted design ablations on full evaluation sets.

(a) Group robustness

Dataset	Model	Overall	Bal.	Worst
Waterb.	SmartCLIP	60.96	61.97	28.29
Waterb.	NC-CLIP	**61.41**	**62.86**	**34.48**
CelebA	SmartCLIP	92.16	90.23	52.22
CelebA	NC-CLIP	**93.41**	**90.35**	**62.78**

(b) Shared vs. independent routing

Variant	I2T R@1	Matched cosine	Modality gap
Independent	65.43	0.366	0.779
Shared	65.37	**0.379**	**0.754**

(c) Exact ETF vs. random fixed anchors

Variant	Rsum	Route KL
Random fixed	**448.94**	0.0488
Exact ETF	448.92	**0.0351**

(d) Reconstruction robustness

Variant	Waterbirds	CelebA
Full embedding	30.48	58.49
Shared-only	**33.81**	**62.78**

Transfer and component attribution. Table 4 reports all eight zero-shot datasets together with the full-set component ablation. The zero-shot results are mixed rather than uniformly dominated by NC-CLIP; the component ablation answers a different question by testing whether the proposed routing and geometric terms contribute under the retrieval objective for which they were introduced.

Table 4: Zero-shot transfer and full-set component ablation.

(a) Zero-shot classification I

Dataset	Long CLIP	Smart CLIP	NC-CLIP
ImageNet	**75.3**	72.9	72.5	73.3
ImageNet-V2	69.7	66.9	66.6	**70.2**
GTSRB	50.2	48.9	52.4	**53.8**
VOC2007-M	79.0	82.1	83.7	**84.1**

(b) Zero-shot classification II

Dataset	Long CLIP	Smart CLIP	NC-CLIP
Country211	–	–	26.9	**27.2**
FER2013	–	–	**58.6**	57.6
FGVC-Air.	–	–	30.4	**31.0**
SUN397	–	–	**72.1**	72.0

(c) Component ablation, ViT-B/16

Configuration	R@1	Δ
NC-CLIP	63.7	–
w/o anchoring	60.8	−2.9
w/o shared mask	61.9	−1.8
w/o capacity control	63.0	−0.7
Only contrastive	60.1	−3.6

(d) Component ablation, ViT-L/14

Configuration	R@1	Δ
NC-CLIP	68.4	–
w/o anchoring	65.5	−2.9
w/o shared mask	66.1	−2.3
w/o capacity control	65.5	−2.9
Only contrastive	64.3	−4.1

Interpretation. Removing anchoring, shared routing, or capacity control consistently lowers COCO I2T R@1, but the random-anchor control shows that the retrieval contribution of fixed anchoring should not be conflated with the route-balancing effect of the exact simplex. Appendix I reports complementary optimization diagnostics; these characterize training behavior rather than serving as standalone causal evidence.

6 Discussion

The evidence assigns distinct roles to the three added mechanisms. Shared routing is supported primarily by tighter cross-modal geometry: relative to independent masks, it improves matched-group cosine and modality gap while leaving retrieval within run-to-run variation. Fixed anchoring is supported by the component ablation, where removing the anchor mechanism lowers COCO retrieval on both backbones. The exact simplex arrangement is a narrower geometric choice: it reduces Route KL relative to random fixed anchors but does not produce an additional retrieval gain in that

controlled comparison. Shared-only reconstruction is selected for a third reason, namely its stronger worst-group performance on Waterbirds and CelebA.

The transfer results also delimit the scope of the method. NC−CLIP improves several shifted-retrieval and zero-shot benchmarks, but it does not dominate every classification dataset: CLIP remains strongest on ImageNet, while SmartCLIP is slightly better on FER2013 and SUN397. We therefore interpret the method as a structured alignment bias rather than a universally superior representation. The strongest evidence concerns full-set retrieval, seed-matched COCO-5K gains, and group-labeled robustness; Urban1k is retained as a domain-shift retrieval benchmark and is not used as a substitute for group-robustness evaluation.

Finally, the architecture guarantees shared *support* only after the modality-specific projections. It does not guarantee coordinate-wise semantic identity, human interpretability, or causal removal of every spurious feature. These distinctions motivate the conservative terminology used throughout the paper and the additional diagnostic analyses in the appendix.

Several limitations remain operational rather than merely terminological. The anchor count and routing budgets are fixed globally, and the reported sensitivity range does not establish that the same configuration is optimal at substantially larger data or model scales. Training is conducted on one primary corpus, and the statistical comparison uses three seeds. The group-robustness reconstruction results are therefore presented descriptively, and the optimization figures are treated as diagnostics rather than causal explanations of the retrieval gains.

The current framework is also restricted to image–text alignment. A natural extension is to replace the flat anchor set with hierarchical or data-adaptive fixed geometries while preserving a held-out model-selection protocol. Evaluating the same routing principle on larger pretraining corpora, additional backbones, and video–language or audio–visual settings would test whether the observed separation between support selection and geometric organization persists beyond the present setting.

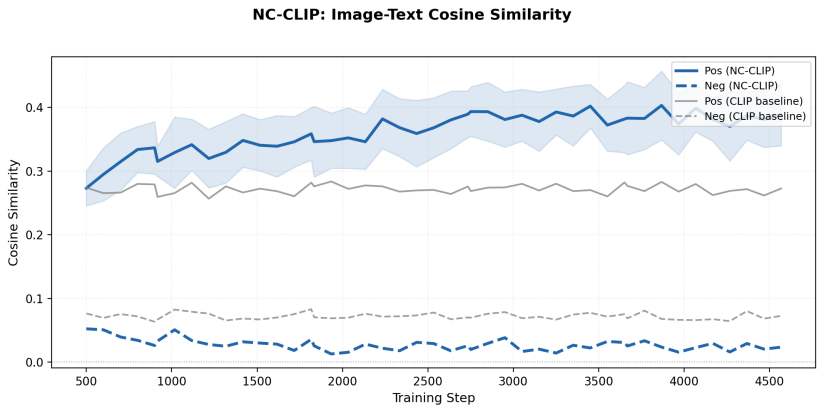


Figure 3: Positive and negative image–text cosine similarity during training. NC−CLIP increases positive-pair similarity while keeping negative-pair similarity low relative to the CLIP baseline.

7 Conclusion

We introduced NC−CLIP, which combines shared routing in a learned basis with fixed geometric anchoring. Shared routing enforces common support without assuming coordinate-level semantic equivalence; fixed anchoring improves retrieval and route balance, while the exact simplex mainly regularizes route utilization. Across full-set retrieval, three-seed COCO-5K evaluation, and group-labeled robustness benchmarks, NC−CLIP consistently improves the targeted metrics. Several limitations remain: the current formulation assumes that modality-specific projections expose a useful coordinate-wise structure for routing, and it relies on fixed anchor counts and sparsity budgets. Moreover, the empirical study is limited to one primary training corpus, three seeds for repeated analyses, and image–text alignment; validation across larger corpora, additional backbones, and other modality pairs remains future work. Overall, these results support separating feature selection from geometric organization in vision–language alignment.

AI USE STATEMENT

Generative AI tools were used to assist with language editing, manuscript organization, and consistency checking, including improving readability and helping reconcile terminology and experimental descriptions across sections. They were also used to suggest wording for presenting and discussing experimental results that had already been computed by the authors. Generative AI was not used to originate the research idea, formulate the method, design the experiments, implement the method, process or generate data, derive mathematical claims or proofs, perform translation, or conduct qualitative data analysis. All AI-assisted text and result descriptions were manually reviewed against the underlying experiments by the authors. The authors take full responsibility for the final content of the paper, including all claims, analyses, and artifacts produced with the aid of generative AI.

REFERENCES

Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, Roman Ring, Eliza Rutherford, Serkan Cabi, Tengda Han, Zhitao Gong, Sina Samangooei, Marianne Monteiro, Jacob L. Menick, Sebastian Borgeaud, Andy Brock, Aida Nematzadeh, Sahand Sharifzadeh, Mikolaj Binkowski, Ricardo Barreira, Oriol Vinyals, Andrew Zisserman, and Karén Simonyan. Flamingo: a visual language model for few-shot learning. In Sanmi Koyejo, S. Mohamed, A. Agarwal, Danielle Belgrave, K. Cho, and A. Oh (eds.), *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*, 2022. URL http://papers.nips.cc/paper_files/paper/2022/hash/960a172bc7fbf0177ccccbb411a7d800-Abstract-Conference.html.

Mathilde Caron, Ishan Misra, Julien Mairal, Priya Goyal, Piotr Bojanowski, and Armand Joulin. Unsupervised learning of visual features by contrasting cluster assignments. In Hugo Larochelle, Marc'Aurelio Ranzato, Raia Hadsell, Maria-Florina Balcan, and Hsuan-Tien Lin (eds.), *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020. URL https://proceedings.neurips.cc/paper/2020/hash/70feb62b69f16e0238f741fab228fec2-Abstract.html.

Zhikang Chen, Min Zhang, Sen Cui, Haoxuan Li, Gang Niu, Mingming Gong, Changshui Zhang, and Kun Zhang. Neural collapse inspired feature alignment for out-of-distribution generalization. In Amir Globersons, Lester Mackey, Danielle Belgrave, Angela Fan, Ulrich Paquet, Jakub M. Tomczak, and Cheng Zhang (eds.), *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024*, 2024. URL http://papers.nips.cc/paper_files/paper/2024/hash/aa56c74513a5e35768a11f4e82dd7ffb-Abstract-Conference.html.

Zhikang Chen, Sen Cui, Deheng Ye, Min Zhang, Gang Niu, Yu Zhang, Masashi Sugiyama, and Tingting Zhu. From coefficients to directions: Rethinking model merging with directional alignment. *CoRR*, abs/2512.00391, 2025. doi: 10.48550/ARXIV.2512.00391. URL https://doi.org/10.48550/arXiv.2512.00391.

Marco Cuturi. Sinkhorn distances: Lightspeed computation of optimal transport. In Christopher J. C. Burges, Léon Bottou, Zoubin Ghahramani, and Kilian Q. Weinberger (eds.), *Advances in Neural Information Processing Systems 26: 27th Annual Conference on Neural Information Processing Systems 2013. Proceedings of a meeting held December 5-8, 2013, Lake Tahoe, Nevada, United States*, pp. 2292–2300, 2013. URL https://proceedings.neurips.cc/paper/2013/hash/af21d0c97db2e27e13572cbf59eb343d-Abstract.html.

Benoit Dufumier, Javiera Castillo Navarro, Devis Tuia, and Jean-Philippe Thiran. What to align in multimodal contrastive learning? In *The Thirteenth International Conference on Learning Representations, ICLR 2025, Singapore, April 24-28, 2025*. OpenReview.net, 2025. URL https://openreview.net/forum?id=Pe3AxLq6Wf.

Sara Ghazanfari, Alexandre Araujo, Prashanth Krishnamurthy, Siddharth Garg, and Farshad Khorrami. EMMA: efficient visual alignment in multi-modal llms. *Trans. Mach. Learn. Res.*, 2025, 2025. URL https://openreview.net/forum?id=lbrO3bGpeO.

Chenxi Huang, Liang Xie, Yibo Yang, Wenxiao Wang, Binbin Lin, and Deng Cai. Neural collapse inspired federated learning with non-iid data. *CoRR*, abs/2303.16066, 2023. doi: 10.48550/ARXIV.2303.16066. URL https://doi.org/10.48550/arXiv.2303.16066.

Songtao Jiang, Yan Zhang, Ruizhe Chen, Tianxiang Hu, Yeying Jin, Qinglin He, Yang Feng, Jian Wu, and Zuozhu Liu. Modality-fair preference optimization for trustworthy MLLM alignment. In *Proceedings of the Thirty-Fourth International Joint Conference on Artificial Intelligence, IJCAI 2025, Montreal, Canada, August 16-22, 2025*, pp. 403–411. ijcai.org, 2025. doi: 10.24963/IJCAI.2025/46. URL https://doi.org/10.24963/ijcai.2025/46.

Zexi Li, Xinyi Shang, Rui He, Tao Lin, and Chao Wu. No fear of classifier biases: Neural collapse inspired federated learning with synthetic and fixed classifier. In *IEEE/CVF International Conference on Computer Vision, ICCV 2023, Paris, France, October 1-6, 2023*, pp. 5296–5306. IEEE, 2023. doi: 10.1109/ICCV51070.2023.00490. URL https://doi.org/10.1109/ICCV51070.2023.00490.

Ji Lin, Hongxu Yin, Wei Ping, Pavlo Molchanov, Mohammad Shoeybi, and Song Han. VILA: on pre-training for visual language models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2024, Seattle, WA, USA, June 16-22, 2024*, pp. 26679–26689. IEEE, 2024. doi: 10.1109/CVPR52733.2024.02520. URL https://doi.org/10.1109/CVPR52733.2024.02520.

Xinlei Liu, Tao Hu, Peng Yi, Baolin Li, Jichao Xie, and Hailong Ma. Gradient semi-masking for improving adversarial robustness. *Pattern Recognit.*, 172:112433, 2026. doi: 10.1016/J.PATCOG.2025.112433. URL https://doi.org/10.1016/j.patcog.2025.112433.

Ming Y. Lu, Bowen Chen, Drew F. K. Williamson, Richard J. Chen, Ivy Liang, Tong Ding, Guillaume Jaume, Igor Odintsov, Andrew Zhang, Long Phi Le, Georg Gerber, Anil V. Parwani, and Faisal Mahmood. Towards a visual-language foundation model for computational pathology. *CoRR*, abs/2307.12914, 2023. doi: 10.48550/ARXIV.2307.12914. URL https://doi.org/10.48550/arXiv.2307.12914.

Shiyin Lu, Yang Li, Qing-Guo Chen, Zhao Xu, Weihua Luo, Kaifu Zhang, and Han-Jia Ye. Ovis: Structural embedding alignment for multimodal large language model. *CoRR*, abs/2405.20797, 2024. doi: 10.48550/ARXIV.2405.20797. URL https://doi.org/10.48550/arXiv.2405.20797.

Vardan Papyan, X. Y. Han, and David L. Donoho. Prevalence of neural collapse during the terminal phase of deep learning training. *CoRR*, abs/2008.08186, 2020. URL https://arxiv.org/abs/2008.08186.

Can Peng, Yuyuan Liu, Yingyu Yang, Pramit Saha, Qianye Yang, and J. Alison Noble. Neural collapse-inspired multi-label federated learning under label-distribution skew. *CoRR*, abs/2509.12544, 2025. doi: 10.48550/ARXIV.2509.12544. URL https://doi.org/10.48550/arXiv.2509.12544.

Petra Poklukar, Miguel Vasco, Hang Yin, Francisco S. Melo, Ana Paiva, and Danica Kragic. Geometric multimodal contrastive representation learning. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvári, Gang Niu, and Sivan Sabato (eds.), *International Conference on Machine Learning, ICML 2022, 17-23 July 2022, Baltimore, Maryland, USA*, Proceedings of Machine Learning Research, pp. 17782–17800. PMLR, 2022. URL https://proceedings.mlr.press/v162/poklukar22a.html.

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, ICML 2021, 18-24 July 2021, Virtual Event*, Proceedings of Machine Learning Research, pp. 8748–8763. PMLR, 2021. URL http://proceedings.mlr.press/v139/radford21a.html.

Aäron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *CoRR*, abs/1807.03748, 2018. URL http://arxiv.org/abs/1807.03748.

Jingchao Wang, Zhijian Wu, Dingjiang Huang, Yefeng Zheng, and Hong Wang. Unlocking the potential of mllms in referring expression segmentation via a light-weight mask decoder. *CoRR*, abs/2508.04107, 2025a. doi: 10.48550/ARXIV.2508.04107. URL https://doi.org/10.48550/arXiv.2508.04107.

Jingyao Wang, Siyu Zhao, Wenwen Qiang, Jiangmeng Li, Changwen Zheng, Fuchun Sun, and Hui Xiong. Towards the causal complete cause of multi-modal representation learning. In Aarti Singh, Maryam Fazel, Daniel Hsu, Simon Lacoste-Julien, Felix Berkenkamp, Tegan Maharaj, Kiri Wagstaff, and Jerry Zhu (eds.), *Forty-second International Conference on Machine Learning, ICML 2025, Vancouver, BC, Canada, July 13-19, 2025*, Proceedings of Machine Learning Research. PMLR / OpenReview.net, 2025b. URL https://proceedings.mlr.press/v267/wang25ey.html.

Zining Wang, Tongkun Guan, Pei Fu, Chen Duan, Qianyi Jiang, Zhentao Guo, Shan Guo, Junfeng Luo, Wei Shen, and Xiaokang Yang. Marten: Visual question answering with mask generation for multi-modal document understanding. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2025, Nashville, TN, USA, June 11-15, 2025*, pp. 14460–14471. Computer Vision Foundation / IEEE, 2025c. doi: 10.1109/CVPR52734.2025.01348. URL https://openaccess.thecvf.com/content/CVPR2025/html/Wang_Marten_Visual_Question_Answering_with_Mask_Generation_for_Multi-modal_Document_CVPR_2025_paper.html.

Tao Wu, Mengze Li, Jingyuan Chen, Wei Ji, Wang Lin, Jinyang Gao, Kun Kuang, Zhou Zhao, and Fei Wu. Semantic alignment for multimodal large language models. In Jianfei Cai, Mohan S. Kankanhalli, Balakrishnan Prabhakaran, Susanne Boll, Ramanathan Subramanian, Liang Zheng, Vivek K. Singh, Pablo César, Lexing Xie, and Dong Xu (eds.), *Proceedings of the 32nd ACM International Conference on Multimedia, MM 2024, Melbourne, VIC, Australia, 28 October 2024 - 1 November 2024*, pp. 3489–3498. ACM, 2024. doi: 10.1145/3664647.3681014. URL https://doi.org/10.1145/3664647.3681014.

Shaoan Xie, Lingjing, Yujia Zheng, Yu Yao, Zeyu Tang, Eric P. Xing, Guangyi Chen, and Kun Zhang. Smartclip: Modular vision-language alignment with identification guarantees. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2025, Nashville, TN, USA, June 11-15, 2025*, pp. 29780–29790. Computer Vision Foundation / IEEE, 2025. doi: 10.1109/CVPR52734.2025.02772. URL https://openaccess.thecvf.com/content/CVPR2025/html/Xie_SmartCLIP_Modular_Vision-language_Alignment_with_Identification_Guarantees_CVPR_2025_paper.html.

Yibo Yang, Haobo Yuan, Xiangtai Li, Zhouchen Lin, Philip H. S. Torr, and Dacheng Tao. Neural collapse inspired feature-classifier alignment for few-shot class-incremental learning. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net, 2023. URL https://openreview.net/forum?id=y5W8tpojhtJ.

Beichen Zhang, Pan Zhang, Xiaoyi Dong, Yuhang Zang, and Jiaqi Wang. Long-clip: Unlocking the long-text capability of CLIP. In Ales Leonardis, Elisa Ricci, Stefan Roth, Olga Russakovsky, Torsten Sattler, and Gül Varol (eds.), *Computer Vision - ECCV 2024 - 18th European Conference, Milan, Italy, September 29-October 4, 2024, Proceedings, Part LI*, Lecture Notes in Computer Science, pp. 310–325. Springer, 2024. doi: 10.1007/978-3-031-72983-6_18. URL https://doi.org/10.1007/978-3-031-72983-6_18.

Pengfei Zhao, Rongbo Luan, Wei Zhang, Peng Wu, and Sifeng He. Guiding cross-modal representations with MLLM priors via preference alignment. *CoRR*, abs/2506.06970, 2025. doi: 10.48550/ARXIV.2506.06970. URL https://doi.org/10.48550/arXiv.2506.06970.

Didi Zhu, Zexi Li, Min Zhang, Junkun Yuan, Jiashuo Liu, Kun Kuang, and Chao Wu. Neural collapse anchored prompt tuning for generalizable vision-language models. In Ricardo Baeza-Yates

and Francesco Bonchi (eds.), *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, KDD 2024, Barcelona, Spain, August 25-29, 2024*, pp. 4631–4640. ACM, 2024. doi: 10.1145/3637528.3671690. URL https://doi.org/10.1145/3637528.3671690.

Fatimah Zohra, Chen Zhao, Hani Itani, and Bernard Ghanem. β-clip: Text-conditioned contrastive learning for multi-granular vision-language alignment. *CoRR*, abs/2512.12678, 2025. doi: 10.48550/ARXIV.2512.12678. URL https://doi.org/10.48550/arXiv.2512.12678.

A Related Work

A.1 Vision–Language Contrastive Learning

Vision–language contrastive learning is a standard paradigm for multimodal representation learning. CLIP (Radford et al., 2021) aligns images and text through large-scale contrastive pretraining and enables strong zero-shot transfer. Subsequent work improves data scaling, architecture, and training objectives (Alayrac et al., 2022; Lu et al., 2023; Lin et al., 2024), while a complementary line studies finer-grained or selective alignment. Region- and patch-level objectives associate local visual features with textual phrases (Zohra et al., 2025), and token- or dimension-level mechanisms restrict which representation components participate in matching (Xie et al., 2025). Other approaches use preference signals, optimization-aware objectives, structural alignment, or lightweight adaptation modules to improve cross-modal correspondence (Zhao et al., 2025; Jiang et al., 2025; Lu et al., 2024; Ghazanfari et al., 2025). Collectively, these methods motivate controlling *where* alignment occurs rather than treating the full representation uniformly. `NC-CLIP` follows this direction but additionally imposes a fixed geometric organization on the routed common space.

A.2 Selective Alignment and Representation Decomposition

Recent work increasingly uses *selective mechanisms* to control which features participate in multimodal alignment. One line of work applies *feature-level masking* to isolate task-relevant or semantically consistent information. For instance, SmartCLIP formulates vision–language alignment as a latent-variable identification problem and introduces learnable masks over representation dimensions to selectively align only the visual features corresponding to textual concepts, thereby mitigating information misalignment and representation entanglement (Xie et al., 2025). Similarly, text-conditioned contrastive frameworks such as β-CLIP dynamically select image patches based on textual queries, enabling fine-grained alignment while suppressing irrelevant visual context (Zohra et al., 2025). Beyond feature dimensions, masking has also been explored at the *spatial level* to enforce localized and interpretable vision–language correspondence. Methods such as Marten introduce auxiliary mask generation objectives to explicitly align visual regions with textual elements, improving spatially-aware representation learning and reducing hallucinations in document understanding scenarios (Wang et al., 2025c). Related efforts in referring expression segmentation further demonstrate that selectively leveraging visual details through lightweight mask decoders is critical for effective multimodal grounding (Wang et al., 2025a). These approaches highlight the importance of region-level selection in bridging visual perception and language semantics. More recently, selective mechanisms have been extended to the *optimization level*. Adaptive Gradient Masking dynamically modulates gradient updates across different representation branches, preventing dominant modalities from overwhelming weaker ones and promoting balanced multimodal learning (Liu et al., 2026). From a complementary perspective, semantic alignment frameworks for multimodal large language models emphasize selectively preserving *linking information* across inputs, as indiscriminate fusion of all visual tokens may obscure meaningful cross-instance or cross-image relations (Wu et al., 2024). Together, these methods show that masking can control information flow and reduce indiscriminate cross-modal alignment. `NC-CLIP` complements this perspective by combining selective routing with an explicit geometric prior on the shared representation.

A.3 Neural Collapse–Inspired Representation Alignment

Neural Collapse (NC) characterizes a striking geometric phenomenon in the terminal phase of training deep classifiers, where features of the same class collapse to their class means, and the class means together with classifier weights form a simplex Equiangular Tight Frame (ETF), yielding maximally separated and optimally discriminative representations (Papyan et al., 2020). This geometric structure has been shown to correspond to an optimal solution under both cross-entropy and mean-squared-error losses, motivating a growing body of work that leverages NC as an explicit inductive bias for representation learning. Early studies exploit neural collapse to address feature–classifier misalignment in incremental and continual learning. By pre-assigning fixed classifier prototypes as a simplex ETF and driving features toward these targets, NC-inspired methods effectively mitigate catastrophic forgetting and preserve feature–classifier alignment in few-shot class-incremental learning settings (Yang et al., 2023). Similar principles have been extended to federated learning, where data heterogeneity across clients leads to biased classifiers and misaligned features. By introducing fixed or synthetic ETF

classifiers during training, several works demonstrate that enforcing NC-consistent geometry can substantially improve both global generalization and client-level personalization under non-IID data distributions (Li et al., 2023; Peng et al., 2025). More recent efforts generalize neural collapse beyond standard multi-class classification. In out-of-distribution (OOD) generalization, NC-inspired methods align semantic features across environments to a shared simplex ETF while separating spurious components via learned masks, thereby improving robustness to distribution shifts and spurious correlations (Chen et al., 2024). From a geometric perspective, neural collapse has also been employed to guide directional alignment in model merging, where deviations from the ideal ETF structure reflect inconsistencies across independently trained models, and explicitly restoring ETF-aligned directions leads to improved merged model performance (Chen et al., 2025). Neural collapse has further been explored in the context of vision–language models. Recent work shows that text representations learned via prompt tuning in CLIP exhibit neural-collapse-like behavior, and that the degree of alignment between text and image representations to a shared simplex ETF is strongly correlated with downstream generalization performance, especially under class imbalance (Zhu et al., 2024). By introducing neural-collapse-anchored regularization terms that explicitly encourage both language-modality collapse and multi-modality isomorphism, these methods improve the generalizability of vision–language models without modifying the backbone architecture. Collectively, these studies highlight neural collapse as a powerful geometric principle for structuring representation spaces across diverse learning paradigms. However, existing NC-inspired approaches predominantly focus on class-level, task-level, or modality-level alignment, and typically assume a single semantic space. In contrast, our work extends neural collapse to the multimodal setting by routing matched image–text representations through a shared support before applying a fixed ETF reference, while leaving unselected coordinates outside that geometric regularizer.

B URBAN1K

Urban1k contains urban scenes with substantial variation in background composition, lighting, viewpoint, and object co-occurrence, together with long and compositionally rich captions. We use it as a *domain-shift retrieval* benchmark. Because it lacks group annotations, direct worst-group evaluation of spurious correlations is instead performed on Waterbirds and CelebA.

C DETAILED ANALYSIS, DERIVATIONS, AND ADDITIONAL RESULTS

This appendix provides additional derivations, implementation details, and analyses for `NC-CLIP`. A key distinction is between *shared support* and *semantic equivalence*: the method enforces shared routed support after learnable modality-specific projections, but does not assume that raw image/text coordinates or individual routed coordinates have identical semantics. Likewise, the geometric anchor system is used to regularize route organization rather than to define supervised semantic classes.

Section D restates the routing and anchoring formulation consistently with the main text. Sections E–F document model selection and implementation. Sections G–H specify the training/evaluation procedure and reporting conventions. Section I analyzes optimization, and Sections J–N study routing behavior and fine/coarse complementarity. Sections O–Q discuss robustness, overhead, and limitations.

D MATHEMATICAL FORMULATION AND DERIVATIONS

D.1 CONTRASTIVE OBJECTIVE

Given normalized image and text representations z_I^i, z_T^i, the symmetric contrastive objective uses pairwise logits

$$\ell_{ij} = \frac{\langle z_I^i, z_T^j \rangle}{\tau}. \tag{12}$$

The corresponding image-to-text and text-to-image cross-entropies are averaged, following CLIP-style training. This objective constrains pairwise similarities but is invariant to a shared orthogonal change of basis, motivating additional routing and geometric priors.

D.2 Learned Routing Basis and Shared Support

The method does not place a common mask directly on raw dual-encoder coordinates. Instead,

$$u_I^i = P_I(z_I^i), \qquad u_T^i = P_T(z_T^i), \tag{13}$$

where P_I and P_T are learnable modality-specific projections. For route $r \in \{f, c\}$, observers produce

$$h_I^{i,r} = g_I^r(u_I^i), \qquad h_T^{i,r} = g_T^r(u_T^i), \tag{14}$$

and the shared logits are

$$s^{i,r} = \frac{h_I^{i,r} + h_T^{i,r}}{2}. \tag{15}$$

The same mask derived from $s^{i,r}$ is applied to both modalities. This guarantees shared support in the *learned routing basis*; it does not imply that $u_{I,j}$ and $u_{T,j}$ have a fixed interpretable semantic identity.

The independent-mask ablation instead uses

$$m_I^r = \mathrm{Top}_{\rho_r}(h_I^r), \qquad m_T^r = \mathrm{Top}_{\rho_r}(h_T^r), \tag{16}$$

with all other training conditions held fixed. The joint and independent variants have essentially identical COCO-5K I2T R@1 (65.37 versus 65.43), while the joint variant yields better matched-group cosine and modality gap. This is the empirical basis for using shared routing.

D.3 Fine and Coarse Routes

The default budgets are $\rho_f = 0.30$ and $\rho_c = 0.50$. Soft masks are used during training and top-k masks at evaluation:

$$m_{\mathrm{soft}}^{i,r} = \sigma(s^{i,r}/\tau_m), \qquad m_{\mathrm{hard},j}^{i,r} = \mathbf{1}\left[j \in \mathrm{TopK}_{k_r}(s^{i,r})\right]. \tag{17}$$

For $M \in \{I, T\}$,

$$\tilde{u}_M^{i,r} = u_M^i \odot m^{i,r}. \tag{18}$$

The normalized fine and coarse routes are concatenated for retrieval. The sensitivity sweep shows only 0.34 Rsum total variation over $(0.20, 0.40)$, $(0.30, 0.50)$, and $(0.40, 0.60)$.

D.4 Fixed Anchors and Balanced Assignment

The default frozen anchor matrix forms a simplex ETF:

$$V^\top V = \frac{C}{C-1} I_C - \frac{1}{C-1} \mathbf{1}\mathbf{1}^\top. \tag{19}$$

A joint fine routed embedding $\bar{z}_i$ is compared with the anchors,

$$a_{ic} = \langle \bar{z}_i, v_c \rangle, \tag{20}$$

and Sinkhorn normalization is applied to $\exp(a_{ic}/\epsilon)$ to obtain a balanced assignment target. The image and its captions share this target. The implementation combines this assignment with ETF barycentric route-NCE and a sample-level conditional-collapse constraint.

The anchors should be interpreted as latent route directions, not predefined semantic categories. In a separate full-COCO-5K ablation, exact ETF and random fixed anchors produce nearly identical Rsum (448.92 and 448.94), whereas exact ETF improves Route KL (0.0351 versus 0.0488). Thus, the exact simplex chiefly regularizes route utilization.

D.5 Shared-Only Reconstruction

The final model reconstructs the opposite modality's shared component rather than its full embedding:

$$\mathcal{L}_{\mathrm{recon}} = \frac{1}{B} \sum_i \left[\|\mathcal{R}_{I\to T}(\tilde{u}_I^{i,f}) - \mathrm{sg}(z_{\mathrm{shared},T}^i)\|_2^2 + \|\mathcal{R}_{T\to I}(\tilde{u}_T^{i,f}) - \mathrm{sg}(z_{\mathrm{shared},I}^i)\|_2^2 \right]. \tag{21}$$

This choice is motivated by the full group-robustness evaluation: shared-only reconstruction outperforms full reconstruction on Waterbirds and CelebA worst-group accuracy. The earlier in-distribution tuning comparison is reported separately in Appendix E.

Table 5: Coefficient settings selected on the tuning subset.

Term	Weight / scale
Mask ranking	0.50
Observer agreement	0.05
SmartCLIP bootstrap	0.25
Coarse objective	0.25
Route-NCE	0.50
ETF-to-backbone gradient scale	0.25
Shared-only reconstruction	0.25

Table 6: Sensitivity analysis on the fixed COCO-1K tuning subset.

C	Rsum	(ρ_f, ρ_c)	Rsum
10	526.70	(0.20,0.40)	527.34
20	527.20	(0.30,0.50)	527.20
50	527.66	(0.40,0.60)	527.00

D.6 Optimization Terms

The implementation retains the SmartCLIP objectives and adds the weighted terms summarized in Table 5. Each coefficient is selected on the tuning-only protocol in Appendix E and then frozen for all full-scale evaluations. We report only objective terms whose implementation is fully specified by the available training configuration, avoiding additional unverified formulations.

E Tuning-Only Model Selection Experiments

The experiments in this section are used only for model selection and sensitivity analysis. They are intentionally separated from the full-scale evaluations in the main paper. All final coefficients and structural choices selected here are frozen before evaluation on the full benchmark sets.

E.1 Coefficient Selection

Each major coefficient was screened independently on a fixed COCO-1K tuning subset while the remaining coefficients were held fixed. The selected values were then frozen across datasets and backbones.

E.2 Prototype Count and Sparsity Sensitivity

Prototype count and fine/coarse sparsity are screened on the same fixed COCO-1K tuning subset: The total spread is 0.96 Rsum across anchor counts and 0.34 Rsum across sparsity settings. We retain $C = 20$ and $(0.30, 0.50)$ as the default used in all full-scale evaluations.

E.3 In-Distribution Reconstruction Tuning

On the fixed COCO-1K tuning subset, removing reconstruction gives an Rsum of 527.20, while full-embedding reconstruction gives 528.60, a gain of 1.40 Rsum. This tuning result is not used as the robustness claim. The final shared-only reconstruction choice is instead supported by the full Waterbirds and CelebA worst-group evaluation reported in the main paper.

F Implementation Details and Hyperparameters

Table 7 summarizes the training configuration used throughout the experiments. We use a single batch-size convention of 1024 in all reported experiments. The embedding dimension depends on the backbone: the ViT-B/16 routing space uses 512 dimensions (154 and 256 retained by the 30%

Table 7: Corrected implementation details and hyperparameters.

Hyperparameter	Value
Embedding dimension d	512 (B/16), 768 (L/14)
ETF anchors C	20
Fine sparsity ρ_f	0.30
Coarse sparsity ρ_c	0.50
Mask temperature τ_m	$2.0 \rightarrow 0.1$
Sinkhorn ϵ	0.05
Images / GPU	32
GPUs	$4 \times$ A100 40GB
Captions / image	2
Gradient accumulation	4
Batch size	1024
Seeds for repeated analyses	3
Optimizer	AdamW
Warmup steps	500
Training epochs	3

and 50% routes), while the ViT-L/14 configuration uses the 768-dimensional setting reported in the original implementation table.

All repeated statistical analyses use three independent random seeds. Matched-seed SmartCLIP and `NC-CLIP` runs share the data split, sampler schedule, training length, batch size, and Python/PyTorch seeding protocol. They start from the same public SmartCLIP checkpoint. This matching does not constitute a strict common-random-number design, which we state explicitly when reporting paired statistics.

G Training and Evaluation Procedure

The implementation follows the same sequence for every full-scale run. First, the image and text encoders produce normalized embeddings. Second, modality-specific projections map the embeddings into the learned routing basis, and the image/text observers produce fine- and coarse-route logits. Third, the two modality logits are averaged to define one support for the matched pair. During training, temperature-scaled soft gates are used; evaluation uses the corresponding hard top-k supports with budgets $(0.30, 0.50)$. Fourth, the routed fine and coarse features are normalized and concatenated for contrastive retrieval. Fifth, the routed representation is compared with the frozen anchor set and Sinkhorn normalization supplies the balanced assignment target. Finally, the contrastive, routing, geometric, bootstrap, and shared-only reconstruction terms are combined and optimized with AdamW.

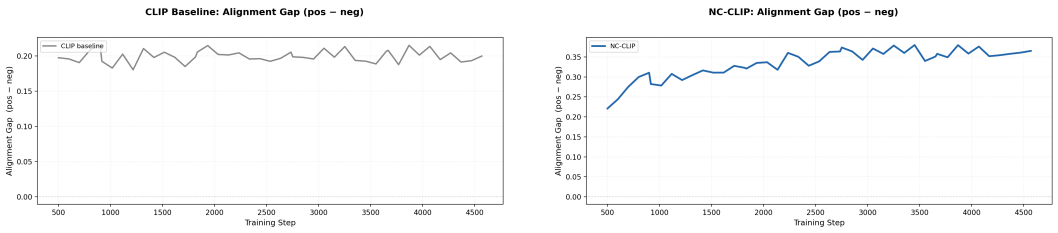
At evaluation time, all selected coefficients, the anchor count, routing budgets, and preprocessing choices remain fixed. Retrieval and zero-shot evaluations do not reuse the COCO-1K tuning subset. The matched-seed comparison additionally holds the public initialization, sampler schedule, training length, batch size, and seed identities fixed across SmartCLIP and `NC-CLIP`.

H Evaluation Metrics and Statistical Reporting

For retrieval, Recall@K is the fraction of queries for which the paired target appears among the top-K ranked candidates. The reported retrieval Rsum is the sum of the six I2T/T2I recalls at $K \in \{1, 5, 10\}$. For group-labeled datasets, overall accuracy aggregates all examples, balanced accuracy averages the relevant label-conditioned accuracies, and worst-group accuracy is the minimum accuracy over the predefined groups. Route KL is used only as a route-balance diagnostic; lower values indicate route usage closer to the balanced target, and it is not treated as a task metric.

Table 8: Roles and primary metrics of the evaluation suites.

Suite	Evaluation role	Primary reporting
COCO, Flickr30k	Standard bidirectional retrieval	I2T/T2I Recall@K
ShareGPT4V, Urban1k	Long-caption or domain-shift retrieval	Benchmark-average R@1
Waterbirds, CelebA	Group-labeled robustness	Overall, balanced, worst-group accuracy
Eight classification datasets	Zero-shot transfer	Top-1 accuracy (VOC2007-Multi follows the reported multi-label protocol)
Routing diagnostics	Geometry and route utilization	Matched cosine, modality gap, Route KL



(a) CLIP alignment gap.

(b) NC−CLIP alignment gap.

Figure 4: Positive–negative alignment gap during training. NC−CLIP develops a larger and more persistent separation than the CLIP baseline.

For a metric measured under three paired seeds, let d_s denote the NC−CLIP–SmartCLIP difference for seed s. We report

$$\bar{d} = \frac{1}{3}\sum_{s=1}^{3} d_s, \qquad s_d = \sqrt{\frac{1}{2}\sum_{s=1}^{3}(d_s - \bar{d})^2}. \tag{22}$$

The two-sided 95% interval is $\bar{d} \pm t_{0.975,2}\, s_d/\sqrt{3}$, and the paired p-value uses the corresponding two-sided paired t statistic. Because $n = 3$ does not permit a meaningful normality assessment, the raw direction of all seed-wise differences and the magnitude of $\bar{d}$ are emphasized over the p-values.

I Training Dynamics and Optimization Behavior

This section analyzes the optimization dynamics of NC−CLIP and the interaction among its regularization terms.

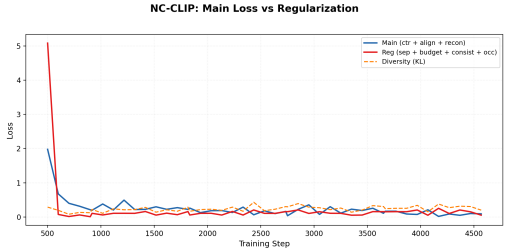
I.1 Alignment Diagnostics

Figure 4 reports the positive–negative alignment gap for CLIP and NC−CLIP. The corresponding cosine-similarity trajectory is shown in Fig. 3 in the main paper. NC−CLIP develops a larger positive–negative margin during training, while positive-pair similarity rises without a comparable increase in negative-pair similarity. These diagnostics characterize optimization behavior rather than serving as standalone evidence for the retrieval gains.

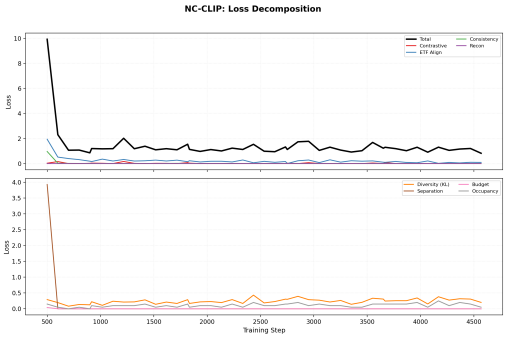
I.2 Loss Dynamics and Objective Interplay

Figure 5 summarizes the evolution of the main objective and the associated regularization terms. The dominant losses decrease rapidly in the early phase of optimization and then stabilize, indicating that the model quickly identifies a viable shared structure before refining it further. The regularization terms remain well-behaved throughout training and do not exhibit oscillatory or divergent dynamics.

These observations suggest that the different components of NC−CLIP operate cooperatively rather than competitively. In particular, diversity-related terms remain active even after the core contrastive

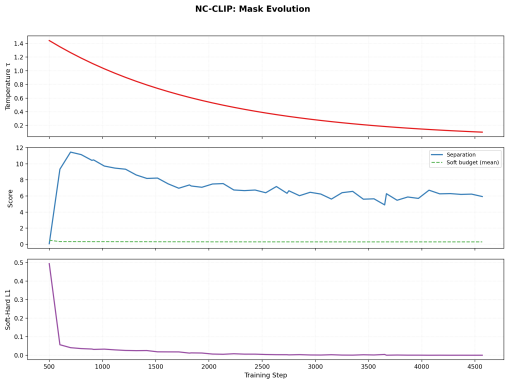


(a) Main loss versus regularization terms.

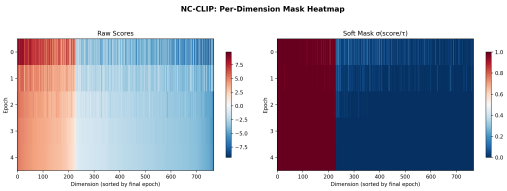


(b) Detailed loss decomposition.

Figure 5: Optimization dynamics of NC−CLIP. Left: the main objective and aggregated regularization terms quickly stabilize after the initial training phase. Right: a detailed decomposition of the individual loss components shows that the different objectives remain well-behaved and operate in a compatible regime throughout optimization.



(a) Mask evolution during training.



(b) Per-dimension mask structure across epochs.

Figure 6: Evolution of the learned routed representation in NC−CLIP. The annealed mask temperature gradually sharpens feature selection, while the heatmap reveals a stable and structured coordinate pattern emerging over the course of training.

objective has largely converged, indicating that balanced prototype usage continues to refine the learned geometry at later stages of training.

I.3 Training Dynamics of NC−CLIP

Figure 6 presents additional training dynamics of NC−CLIP, shedding light on how the proposed objectives jointly shape representation geometry over the course of optimization.

The mask evolution curve shows that the temperature schedule gradually sharpens feature selection while keeping the separation and budget-related quantities stable. The heatmap reveals that a coherent coordinate-selection pattern emerges early and remains largely preserved across epochs. This suggests that the routed shared representation is not formed abruptly, but learned progressively through a smooth optimization trajectory.

J Mask Distribution Analysis

Figure 7 shows the evolution of observer scores and the resulting shared soft-mask values. As temperature decreases, the shared routing distribution becomes more polarized, indicating progressively

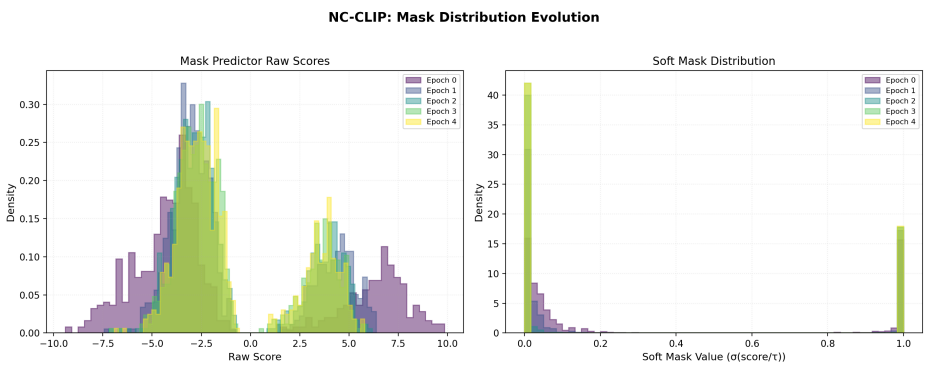


Figure 7: Distribution of observer scores and corresponding shared soft-mask values across training epochs. The distribution becomes increasingly polarized as routing sharpens.

sharper selection. Because the final mask is produced by averaging the image/text observer logits, there are not two independently applied masks whose agreement must be inferred after the fact.

The observer outputs remain a useful diagnostic, but similar observer distributions should not be interpreted as proof that individual coordinates carry the same semantic concept in both modalities. Our interpretation is narrower: the learned projections and observers produce a routing basis in which a shared support can be used without measurable retrieval loss, and the joint-mask ablation shows improved cross-modal geometry relative to independent masks.

K Geometric and Structural Analysis of NC-CLIP

Beyond downstream retrieval, we analyze how the objectives shape anchor utilization, assignment confidence, and the geometry of the routed common space.

K.1 Emergence of ETF Structure During Optimization

Figure 8 tracks several statistics associated with the ETF anchor structure during training. The number of occupied vertices remains consistently high, indicating that the assignment process makes use of a broad portion of the anchor set rather than collapsing onto a few dominant prototypes. At the same time, the alignment angle decreases steadily and the top-2 assignment margin increases, showing that the learned common representations become progressively better aligned with anchor directions and more confidently assigned.

Together, these trends indicate that the ETF prior acts as a stable geometric regularizer: the routed representation becomes more confidently aligned to the anchor system while maintaining broad anchor utilization.

L Mask Stability Analysis

Figure 9 characterizes the stability of the *shared routing support*. The left panel reports inter-sample Jaccard overlap; the right panel reports overlap within groups assigned to the same anchor. High overlap indicates that the routing mechanism contains a stable global component, while non-identical masks leave room for conditional routing.

These statistics do not establish a fixed semantic label for each selected coordinate. They instead show that routing is neither arbitrary nor purely instance-specific, which is consistent with the interpretation of the mask as conditional coordinate routing.

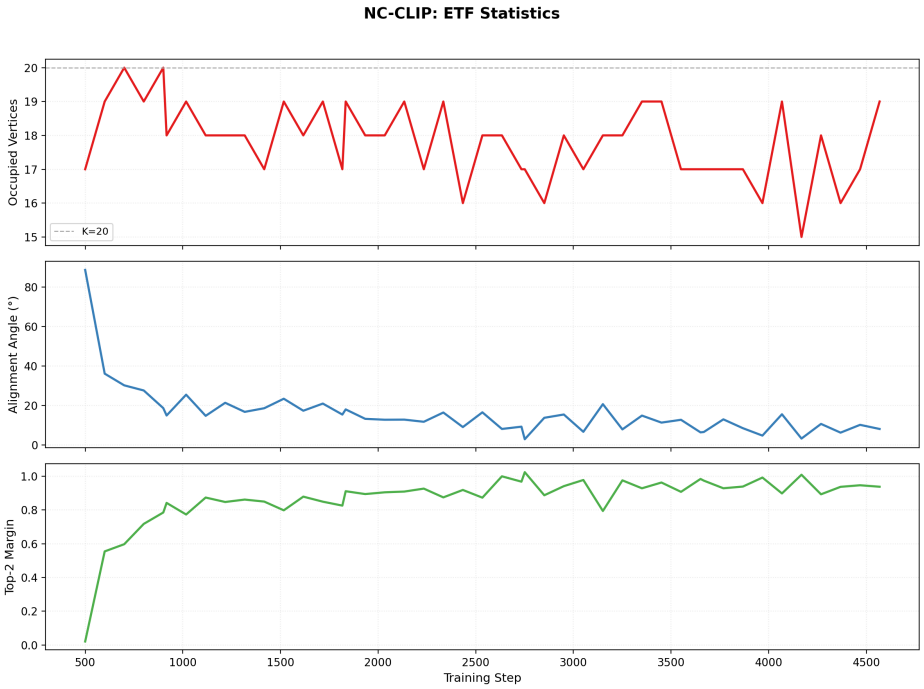


Figure 8: ETF-related statistics during training. NC−CLIP maintains broad prototype utilization, progressively decreases the alignment angle to ETF anchors, and increases top-2 assignment margin over time.

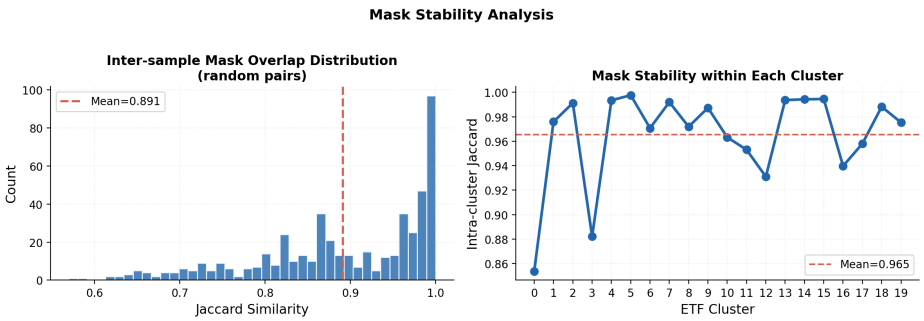


Figure 9: **Shared-route stability.** *Left:* inter-sample Jaccard overlap. *Right:* within-anchor-cluster overlap. High overlap indicates a stable shared component without implying coordinate-level semantic identity.

M Case Study: Routing by Anchor Cluster

Figure 10 visualizes mean routing scores by ETF anchor cluster. A block of coordinates receives high scores across many clusters, indicating a common routing core, while other coordinates vary with anchor assignment. We describe this as *cluster-conditional routing specialization*, not as evidence that an individual coordinate has a stable human-readable semantic meaning.

The pairwise Jaccard matrix in the lower panel shows substantial off-diagonal overlap together with cluster-dependent differences. This supports a hierarchical routing interpretation: a broadly reused core plus conditional refinements.

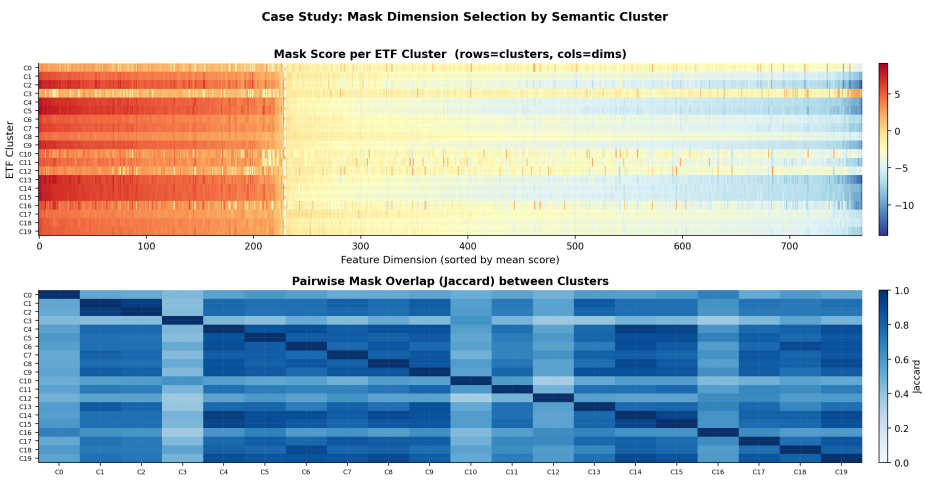


Figure 10: **Routing by anchor cluster.** *Top:* mean routing scores per ETF cluster. *Bottom:* pairwise Jaccard similarity of binary routes. The visualization shows a shared routing core together with cluster-dependent refinements.

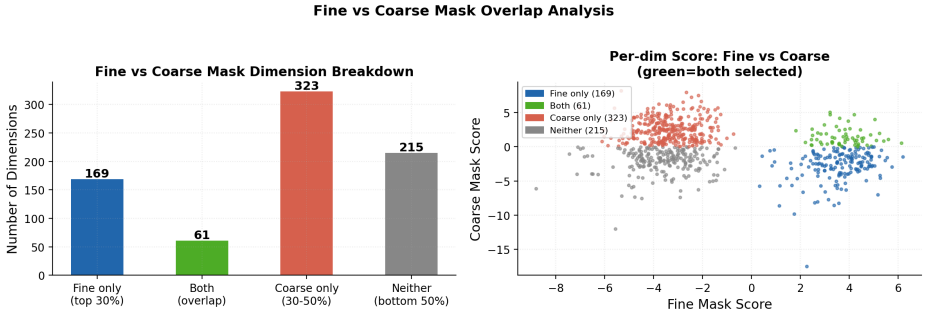


Figure 11: **Fine vs. Coarse Mask Overlap Analysis.** *Left:* Breakdown of feature dimensions by mask membership. *Right:* Per-dimension scatter of fine vs. coarse mask scores. The limited overlap confirms that the two predictors capture complementary portions of the routed representation space.

N Fine vs. Coarse Mask Complementarity

Figure 11 provides a quantitative breakdown of how the feature dimensions are partitioned between the fine mask m^f and the coarse mask m^c.

Partition statistics. A substantial number of dimensions are selected exclusively by the fine predictor, while an even larger group is selected exclusively by the coarse predictor. Only a relatively small subset is selected by both. This confirms that the two predictors are not redundant: the fine mask does not simply select a strict subset of the coarse mask, but instead captures a qualitatively different subset of the routed representation space.

Score geometry. The scatter plot in the right panel reveals the per-dimension relationship between fine and coarse mask scores. Fine-only and coarse-only dimensions occupy distinct score regions, confirming that the two predictors learn qualitatively different routing criteria although both are derived from the same matched pair in the learned routing basis.

Implication for retrieval. This complementarity provides a direct mechanistic explanation for the ablation trend: removing the two-level design discards either the tighter routed core or the broader routed context, both of which are necessary for strong retrieval and robust generalization.

O WHY NC−CLIP IMPROVES ROBUSTNESS

The robustness evidence operates at two complementary levels. ShareGPT4V and Urban1k show that the learned representation remains effective under longer captions and compositional/domain changes, but these datasets do not provide group-labeled evidence about spurious correlations. Waterbirds and CelebA supply that direct test: relative to SmartCLIP, NC−CLIP improves worst-group accuracy by 6.19 and 10.56 percentage points, respectively.

Mechanistically, shared routing limits the coordinates that participate in matching, while the fixed anchor system regularizes how routed groups occupy the common space. The independent-mask ablation is important here: joint routing does not improve COCO retrieval over independent routing within run-to-run variation, but it improves matched-group cosine and reduces modality gap. We therefore interpret shared routing as a geometry-tightening inductive bias rather than as proof of semantic disentanglement.

The reconstruction ablation provides a second mechanistic constraint. Full reconstruction can preserve correlations that are useful for in-distribution retrieval but undesirable for group robustness. Shared-only reconstruction increases worst-group accuracy from 30.48 to 33.81 on Waterbirds and from 58.49 to 62.78 on CelebA in the reported three-seed comparison. These values are descriptive; with three seeds and a non-smooth worst-group statistic, we do not claim significance for this ablation.

Finally, the exact ETF arrangement should not be credited with all robustness or retrieval gains. Random fixed anchors match its retrieval almost exactly, whereas the ETF lowers Route KL. The evidence therefore supports *fixed geometric anchoring* plus *balanced route organization*; the exact simplex contributes primarily to the latter.

P COMPLEXITY AND COMPUTATIONAL OVERHEAD

NC−CLIP adds lightweight routing heads, reconstructors, and batch-level anchor operations on top of the image and text encoders.

The first source of overhead is the pair of mask predictors. These are lightweight MLPs applied to encoder outputs and produce feature-level scores that are transformed into masks. Because they operate on already computed embeddings and involve only shallow feed-forward layers, their structural cost is small relative to the backbone encoders, although exact wall-clock overhead depends on the implementation and hardware. The same is true for the cross-modal reconstructors, which are likewise lightweight MLPs used only to regularize the routed shared representation.

The second source of overhead arises from the ETF-based assignment and regularization. For each sample in a mini-batch, NC−CLIP computes similarities between the joint embedding and a fixed set of C anchors. Since C is small relative to the embedding dimension and batch size, this step adds only a modest number of dot products. The Sinkhorn normalization further introduces a small iterative procedure on the batch-by-anchor similarity matrix. However, this matrix is low-dimensional compared with the full cost of the encoders, and the number of Sinkhorn iterations is fixed and small in practice.

The geometry-related losses are also inexpensive. The anchor-alignment loss requires only inner products between masked embeddings and assigned anchors, while the diversity term depends only on the empirical assignment histogram. Both are lightweight relative to the contrastive objective itself.

Thus, the additional computation is concentrated in shallow heads and low-dimensional batch-level geometric operations rather than in the backbone encoders.

Q LIMITATIONS AND OPEN QUESTIONS

Several limitations remain. First, shared routing assumes that learnable modality-specific projections can expose a useful diagonal structure in a common routing basis. The joint-mask ablation supports this operationally, but does not establish coordinate-level semantic equivalence or human interpretability.

Second, the anchor count and fine/coarse routing budgets are fixed. The reported sweeps show modest sensitivity over $C \in \{10, 20, 50\}$ and sparsity pairs $(0.20, 0.40)$, $(0.30, 0.50)$, and $(0.40, 0.60)$, but they do not cover larger data scales or substantially different backbones.

Third, the exact simplex ETF is a specific geometric bias. Our random-anchor ablation shows that it mainly improves route balance rather than retrieval, so other fixed or hierarchical anchor geometries may be equally or more appropriate for other tasks.

Fourth, matched-seed statistical analysis uses only three seeds. The per-seed effects are consistent and the paired-difference intervals exclude zero, but normality cannot be assessed with $n = 3$. Group-robustness reconstruction results are therefore presented descriptively rather than as significance claims.

Fifth, training is evaluated from one primary training dataset (CC3M) and the method is currently restricted to image–text alignment. Broader evidence across training corpora, video–language, and audio–visual settings remains future work. Attribution visualizations may further clarify routing behavior and remain a useful direction for future analysis.

R Evaluation Protocol and Reporting Conventions

The experiments are organized around a strict separation between model selection and final evaluation. Hyperparameter and sparsity sweeps use the fixed COCO-1K tuning subset in Appendix E. Once the default anchor count, routing budgets, and loss coefficients are selected, they are held fixed for the full benchmark evaluations in the main paper. This prevents benchmark-specific retuning from being mixed with the final comparisons.

For retrieval, the main tables use the corresponding full evaluation sets. COCO and Flickr30k represent the standard short-caption setting, while ShareGPT4V and Urban1k probe longer or more compositionally shifted text. Urban1k is interpreted only as a retrieval/domain-shift benchmark because it does not contain the group annotations needed for direct worst-group evaluation.

For robustness, Waterbirds and CelebA are treated separately because they provide explicit group labels. We report overall, balanced, and worst-group accuracy in the main paper. The reconstruction comparison is also evaluated on these group-labeled benchmarks, which is why shared-only reconstruction is selected as the final configuration even though the earlier in-distribution tuning comparison favors full reconstruction on COCO-1K.

For the matched-seed COCO-5K study, SmartCLIP and `NC-CLIP` use the same public initialization, data split, sampler schedule, training length, batch size, and three random seeds. The pairing is therefore meaningful for comparing run-to-run effects, but it is not described as a strict common-random-number experiment. With only three seeds, normality cannot be established; the raw per-seed direction and paired effect sizes are consequently the primary statistical evidence.

S Ablation Interpretation Guide

The ablations answer different questions and should not be collapsed into a single attribution claim. The joint-versus-independent routing comparison isolates the effect of enforcing one shared support after the learned modality-specific projections. Retrieval is effectively unchanged under the matched budget, while matched-group cosine and modality gap improve, so this result is interpreted as evidence for tighter cross-modal geometry rather than a direct retrieval advantage.

The no-anchoring component ablation evaluates the contribution of fixed geometric anchoring as a whole. By contrast, the random-fixed-anchor control keeps anchoring but changes the anchor geometry. Its nearly unchanged Rsum together with the lower Route KL of the exact ETF shows that the retrieval contribution of anchoring should be separated from the additional route-balancing role of the simplex arrangement.

The reconstruction ablation addresses a third mechanism. Full-embedding reconstruction can preserve information that is useful for in-distribution retrieval but undesirable for worst-group robustness. Shared-only reconstruction is therefore justified by the Waterbirds and CelebA worst-group comparison rather than by the earlier COCO-1K tuning result. The three mechanisms—routing, anchoring,

and reconstruction—are thus supported by different measurements and are not treated as interchangeable sources of the same gain.

T Reproducibility Notes

The default training configuration is intentionally fixed across the reported full-scale experiments: $C = 20$ anchors, fine/coarse routing budgets of $(0.30, 0.50)$, batch size 1024, 500 warm-up steps, and 3 epochs using AdamW. The main implementation uses four NVIDIA A100 40GB GPUs. Repeated statistical analyses use three independent seeds. The detailed batch decomposition and backbone-dependent embedding dimensions are listed in Table 7.

The paper also distinguishes quantities that are architectural guarantees from those that are empirical observations. Applying the same mask to both modalities guarantees shared support in the learned routing basis, but it does not guarantee that individual coordinates carry identical semantics. Similarly, the ETF anchors are fixed geometric references rather than semantic class prototypes. These distinctions are maintained throughout the main text and appendix so that the geometric interpretation does not exceed what is established by the implementation and ablations.

U Summary of Findings

Taken together, the evidence supports a set of well-separated conclusions. `NC-CLIP` improves standard retrieval under the reported benchmark protocol and shows consistent three-seed gains over SmartCLIP on COCO-5K. It also improves worst-group accuracy on Waterbirds and CelebA, providing direct group-labeled evidence for improved robustness to spurious correlations.

Shared routing should be interpreted as an inductive bias for common support in a learned routing basis. Compared with independent masks, it leaves retrieval essentially unchanged while improving matched-group cosine and modality gap. Fixed geometric anchoring contributes to retrieval and route balance; the exact ETF arrangement primarily reduces Route KL relative to random fixed anchors. Shared-only reconstruction is preferred in the final configuration because it gives stronger worst-group accuracy than full reconstruction.

Accordingly, shared support should not be conflated with semantic identity, Urban1k should be interpreted as a domain-shift rather than group-robustness benchmark, and the exact simplex should be credited primarily for route balancing rather than for the retrieval gains of fixed anchoring as a whole.

V Devices

All experiments are implemented in PyTorch on a Linux server with an AMD EPYC 7742 64-Core Processor (128 logical threads). The main training configuration uses four NVIDIA A100 40GB GPUs with batch size 1024. All repeated statistical analyses use three independent random seeds.