ACA: COUPLING RELATIONAL ALIGNMENT AND COMPLEMENTARY REGULARIZATION FOR VISION ENCODER FUSION

Anonymous authors
Paper under double-blind review

ABSTRACT

Combining multiple vision foundation models has become an important strategy for improving visual representations. A common approach concatenates their [CLS] tokens and trains a task-specific predictor, but does not explicitly constrain patch-level relations. Inspired by DINOv3's Gram anchoring, which preserves patch relations during self-supervised training, we study relational alignment across encoders. Uniform Gram matching penalizes every pairwise discrepancy, whereas fusion calls for aligning selected relations while retaining differences between representations. We introduce Alignment-guided Complementary Aggregation (ACA), which couples selective relational alignment with regularization of the complementary residual. Global features provide importance weights, while projected patch features determine a shared gate for weighted alignment, a mean-gate penalty, and a variance reward on the weighted complementary residual. Both encoders are jointly trained and retained for prediction. Across four classification settings on COCO and ImageNet-1K, ACA improves over matched-pair concatenation by 0.25–0.94 percentage points and outperforms uniform Gram alignment in three. On ImageNet-Hard, the full objective reaches 69.98% and exceeds all six one- and two-term variants. ACA also exceeds direct summation of the importance, residual, and gate signals by 1.01 points on ImageNet-Hard and 0.67 points on COCO. These results support coupling relational alignment with complementary residual regularization for encoder fusion.

1 INTRODUCTION

Vision Transformers (ViTs) underpin many leading vision foundation models, including CLIP and DINOv3 (Dosovitskiy et al., 2021; Radford et al., 2021; Siméoni et al., 2025). A ViT divides an image into patches and represents them as spatially indexed tokens, while its [CLS] token provides an image-level summary. Combining encoders trained with different objectives gives a downstream model access to distinct feature spaces. A common feature-fusion approach concatenates their [CLS] representations and trains a prediction head. This supervises the joint prediction, but leaves the relations among the encoders' patch tokens implicit.

Global and local feature quality need not improve together. CLIPSelf identifies a gap between representations of image crops and regions pooled from dense CLIP features (Wu et al., 2024). In large ViTs, register and [CLS] tokens can decouple global summaries from local features (Lappe & Giese, 2025). DINOv3 further shows that global classification can improve while dense prediction deteriorates during training (Siméoni et al., 2025). These findings motivate a relation-level question for fusion: which patch relations should agree across encoders, and which differences should remain available during adaptation? A classifier over concatenated summaries, even when augmented with patch means, does not explicitly specify this allocation.

We examine this problem by comparing the similarities that two encoders assign to the same patch pairs. DINOv3's Gram anchoring expresses these relations independently of feature dimension (Siméoni et al., 2025). For CLIP–DINOv3, 88.63% of patch pairs on ImageNet-Hard and 91.73% on COCO have squared Gram discrepancy below 0.1. Most relations thus show small discrepancies, while a smaller fraction differ more; checkpoint matrices locate larger differences along particular

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.

rows and columns (Appendix A). Uniform Gram matching penalizes every squared residual with the same coefficient. We therefore study a learned allocation for alignment and complementary regularization, conditioned on global context and local features.

We introduce Alignment-guided Complementary Aggregation (ACA), a loss that treats relational agreement and difference as coupled quantities in encoder fusion. Global features provide pairwise importance, while projected patch features determine an alignment gate. Their product weights the squared Gram residual; the same importance and complementary gate define a residual whose variance we reward. A separate penalty controls mean gate activation. This formulation extends fusion beyond combining encoder outputs to shaping their relations during joint adaptation. Both encoders remain in the predictor, while the relational modules operate only during training.

ACA reaches 87.99% accuracy on ImageNet-1K at 224-pixel resolution and improves all four matched-scale concatenation references by 0.25–0.94 pp. It exceeds uniform Gram alignment in three settings. With the architecture fixed, the full objective reaches 69.98% on ImageNet-Hard, outperforming all six one- and two-term subsets; it also exceeds direct summation on ImageNet-Hard and COCO. Gram extensions improve C-RADIOv4–DINOv3 and SAM3–DINOv3 fusion. C-RADIOv4 already includes a DINOv3 teacher, so the former tests whether fusion remains useful despite this shared training source. Together, these experiments evaluate both classification performance and the relational construction underlying ACA.

2 Related Work

Fusion of retained encoders. One approach to combining pretrained models keeps multiple encoders available to the predictor. Cambrian-1 and Eagle combine visual experts for multimodal understanding (Tong et al., 2024; Shi et al., 2025). In few-shot classification, CaFo combines CLIP and DINO caches (Zhang et al., 2023); AMU-Tuning and VRGAdapter adapt auxiliary global predictions while keeping the visual backbones frozen (Tang et al., 2024; Jiang et al., 2026a). Their central question is how to combine the available outputs. We study how the retained encoders should be adapted together, using their patch relations to define the training objective.

Consolidation into one encoder. Multi-teacher distillation instead trains a single deployable student. AM-RADIO matches global and dense teacher features, and RADIOv2.5 improves their transfer across resolutions (Ranzinger et al., 2024b; Heinrich et al., 2024). Balancing teacher distributions is important because their feature scales differ (Ranzinger et al., 2024a). C-RADIOv4 updates the teacher set to SigLIP2, DINOv3, and SAM3 (Ranzinger et al., 2026); SigLino also preserves geometry among image summaries through relational distillation (Chaybouti et al., 2026). These methods consolidate teacher capabilities before downstream use. ACA jointly adapts two encoders that remain in the predictor. The distinction changes supervision: distillation asks a student to reproduce teacher-derived targets, while ACA uses downstream labels to adapt both retained encoders. Its relational objective operates between two changing representations, without assigning either a fixed teacher role. Our C-RADIOv4 experiment connects the two approaches by testing relational fusion after multi-teacher consolidation.

Local structure and relational objectives. Global recognition need not preserve local feature quality. CLIPSelf and FineCLIP use region-level supervision (Wu et al., 2024; Jing et al., 2024), while DINOv3 preserves patch relations through Gram anchoring (Siméoni et al., 2025). Relational distillation and second-order matching compare representation structure (Park et al., 2019; Gatys et al., 2015; Sun & Saenko, 2016). Spatial Gram Alignment transfers spatial relations in image synthesis (Zhang et al., 2026); Omnivorous aligns features across visual input modalities (Kabra et al., 2026). DinoRankCLIP couples DINO feature injection with vision–language ranking supervision (Jiang et al., 2026b). Statistical regularizers control representation variation (Bardes et al., 2022; Zbontar et al., 2021). For joint fusion, the design question is how agreement and difference should enter the same objective. ACA uses shared importance and a gate to weight alignment and define the complementary residual. The gate mean controls allocation, whereas variance acts on that residual. Our loss-subset and direct-sum comparisons test this shared construction.

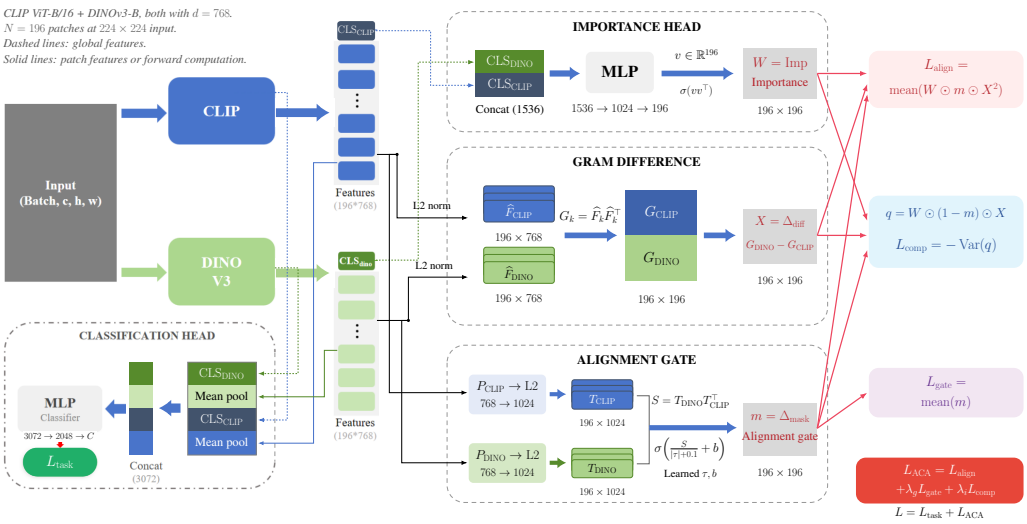

Figure 1: ACA with DINOv3-B (encoder A) and CLIP ViT-B/16 (encoder B). Both have width 768 and provide 196 spatial tokens. Class features concatenate to dimension 1536; adding both patch means gives a 3072-dimensional classifier input. Global features produce importance; patch features produce the Gram residual and gate, with separate projections and L2 normalization in the latter. Eqs. (6)–(8) define the scalar reductions; Eq. (12) adds task supervision. Both encoders and the classifier are retained for prediction.

3 Alignment-Guided Complementary Aggregation

ACA is a loss for joint encoder adaptation. We first construct a relational residual, a global importance head, and a local alignment gate, then combine their outputs into an auxiliary objective for the downstream predictor. For one image, encoder $k \in \{A, B\}$ provides a class feature $c_k \in \mathbb{R}^{d_k}$ and spatial features $F_k \in \mathbb{R}^{N \times d_k}$ on corresponding patch grids. We exclude class and register tokens from F_k and suppress batch indices until introducing the loss reductions.

3.1 Gram Difference: A Common Relational Quantity

The channels of independently pretrained encoders need not correspond. Relational distillation compares structural information across representations (Park et al., 2019); DINOv3's Gram anchoring preserves a student's patch relations relative to a teacher (Siméoni et al., 2025). We compare the similarities assigned to the same image locations by two encoders being adapted together. This gives a common quantity even when their feature dimensions differ.

After normalizing each patch vector, we form

$$\widehat{F}_{k,i} = \frac{F_{k,i}}{\|F_{k,i}\|_2}, \qquad G_k = \widehat{F}_k \widehat{F}_k^\top, \qquad X = \Delta_{\text{diff}} = G_A - G_B, \quad r_{ij} = X_{ij}^2. \tag{1}$$

Here $G_{k,ij}$ is the within-encoder cosine similarity between patches i and j. Small r_{ij} indicates agreement on this relation, rather than proximity of token vectors across encoders. The sign of X_{ij} identifies which encoder assigns the larger similarity. Both Gram matrices are $N \times N$ even when $d_A \neq d_B$, and are invariant to independent orthogonal coordinate changes: $(\widehat{F}_k Q_k)(\widehat{F}_k Q_k)^\top = G_k$ for orthogonal Q_k. For nonzero patch vectors, $X_{ij} \in [-2, 2]$ and $r_{ij} \in [0, 4]$; X is symmetric with a zero diagonal.

Uniform Gram matching minimizes $N^{-2}\|X\|_F^2$. We retain X so that subsequent modules can determine how each relation contributes to alignment and complementary regularization. The residual itself has no trainable parameters and passes gradients to both encoders. For the C-RADIOv4 and SAM3 component extensions, learned token-axis maps first bring heterogeneous spatial grids to a common length (Appendix E).

3.2 Importance Head: Weighting Pairs from Global Features

The residual measures geometric disagreement without using the image summaries to distinguish relations. We introduce global context through an MLP conditioned on both class features:

$$v = \mathrm{MLP}_{\phi_W}([c_A; c_B]) \in \mathbb{R}^N, \qquad W = \mathrm{Imp} = \sigma(vv^\top). \tag{2}$$

Each coordinate of v corresponds to a grid position. Its outer product produces N^2 pair weights from N image-conditioned values; sigmoid bounds them in $(0, 1)$ and preserves symmetry. The logits have rank at most one, although W generally has higher rank. Since v is unrestricted, its coordinates are not patch probabilities.

We share W between alignment and complementary regularization. It determines how strongly a pair enters the residual objectives; the gate determines how that weight is divided between alignment and complementary regularization.

3.3 Alignment Gate: Allocating Relations from Patch Features

We learn the allocation from patch features instead of fixing it by a residual threshold. To compare the two encoders directly, separate affine projections P_A, P_B map their features into a common channel dimension d_p:

$$T_k = \mathrm{normalize}(P_k(F_k)) \in \mathbb{R}^{N \times d_p}, \qquad S = T_A T_B^\top. \tag{3}$$

The projected similarity S_{ij} compares patch i from encoder A with patch j from encoder B. We convert it into a gate using

$$m = \Delta_{\mathrm{mask}} = \sigma\left(\frac{S}{|\tau| + \epsilon} + b\right), \qquad \epsilon = 0.1, \tag{4}$$

where temperature τ and bias b are learned scalars initialized to 1 and 0. Row normalization makes S a cosine-similarity matrix; $|\tau| + \epsilon$ keeps the denominator positive.

The residual and gate measure different relationships. X_{ij} compares two within-encoder similarities in the original feature spaces, whereas S_{ij} compares projected features across encoders. Thus m depends on the features, without being constrained to vary monotonically with residual magnitude. It can also be asymmetric. Larger m_{ij} assigns more weight to alignment, and $1 - m_{ij}$ supplies the complementary allocation. These are soft relation weights; they neither remove tokens nor impose discrete correspondences.

3.4 Combining the Modules into a Relational Objective

We now combine the discrepancy X, importance W, and allocation m on their common pair grid:

$$W \odot X = W \odot m \odot X + W \odot (1 - m) \odot X. \tag{5}$$

Changing W reweights both parts; changing m reallocates the residual between them. We minimize weighted squared discrepancy in the alignment part, reward variation in the complementary part, and penalize the mean allocation separately. For a batch of B images, we flatten the full $B \times N \times N$ grid into $M = BN^2$ entries indexed by a. The reductions include both pair orders and the diagonal.

Weighted alignment energy. We weight each squared discrepancy by the product of global importance and local allocation:

$$L_{\mathrm{align}} = \frac{1}{M} \sum_{a=1}^{M} W_a m_a X_a^2. \tag{6}$$

For fixed W, m, this term contracts the residual toward zero. Both weights enter linearly: $L_{\mathrm{align}} = M^{-1} \| \sqrt{W \odot m} \odot X \|_F^2$. The denominator M lets the learned weights change both total alignment strength and its distribution over pairs.

Mean penalty on the gate. To regularize the amount of alignment, we add

$$L_{\text{gate}} = \frac{1}{M}\sum_{a=1}^{M} m_a. \tag{7}$$

Since $m_a \geq 0$, this term equals $\|m\|_1/M$: it regularizes the overall alignment allocation. The same mean gate value can produce different alignment energies depending on which residuals and importance weights it selects. Gate mean therefore controls allocation separately from the weighted error that allocation incurs.

Variance of the complementary residual. Variance regularization controls feature spread in self-supervised learning (Bardes et al., 2022). Here we apply it to the complementary relational residual $q_a = W_a(1 - m_a)X_a$ and reward its sample variance:

$$L_{\text{comp}} = -\frac{1}{M-1}\sum_{a=1}^{M}(q_a - \bar{q})^2, \qquad \bar{q} = \frac{1}{M}\sum_{a=1}^{M} q_a. \tag{8}$$

We retain the residual's sign so that positive and negative differences contribute relative to a common mean. Since $\text{Var}(q) = \frac{M}{M-1}(\mathbb{E}[q^2] - \mathbb{E}[q]^2)$, this rewards dispersion rather than uncentered residual energy. We compute one variance over all M entries, including the batch axis. Complementary here refers to the residual weighted by $1 - m$; its usefulness is evaluated through the downstream task.

Coupling through shared variables. The relational auxiliary is

$$L_{\text{ACA}} = L_{\text{align}} + \lambda_g L_{\text{gate}} + \lambda_i L_{\text{comp}}. \tag{9}$$

The two residual objectives share the importance and allocation functions. A gate change simultaneously changes a pair's alignment coefficient and its contribution to q, so both residual paths are trained through the same allocation. Centering q over the batch also links its entries through $\bar{q}$. The mean-gate penalty regulates allocation directly, separately from both residual statistics. Removing a term changes the objective applied to these shared variables. Section 4.4.1 tests these interactions with the architecture fixed; the Direct-sum control tests whether the intermediate signals alone reproduce the coupled construction.

The variance reward is bounded: $|X_a| \leq 2$ and $W_a, m_a \in [0, 1]$ imply $|q_a| \leq 2$ and $\text{Var}(q) \leq 4M/(M-1)$. Hence $L_{\text{ACA}} \geq -4\lambda_i M/(M-1)$ for nonnegative coefficients. We give the derivation and the residual and gate derivatives in Appendix D.

3.5 Downstream Fusion and Joint Optimization

The downstream predictor receives both global and pooled spatial features:

$$\overline{F}_k = \frac{1}{N}\sum_{i=1}^{N} F_{k,i}, \qquad z = [c_A; \overline{F}_A; c_B; \overline{F}_B], \qquad \hat{y} = g_\psi(z). \tag{10}$$

Here $z \in \mathbb{R}^{2(d_A+d_B)}$, and g_ψ is a two-layer MLP producing class logits. The importance and gate shape encoder features through training; the classifier reads those features directly (Figure 1).

For labels $y_1, \ldots, y_B$, we use mean cross-entropy,

$$L_{\text{task}} = -\frac{1}{B}\sum_{b=1}^{B} \log\big[\text{softmax}(g_\psi(z_b))\big]_{y_b}. \tag{11}$$

We train with

$$L = L_{\text{task}} + L_{\text{align}} + \lambda_g L_{\text{gate}} + \lambda_i L_{\text{comp}}, \qquad \lambda_g = 0.01, \quad \lambda_i = 0.1. \tag{12}$$

We jointly optimize both encoders and all heads. Task gradients reach the encoders through the classifier; relational gradients reach them through X, W, m. The auxiliary acts on encoder features, with task-specific supervision supplied by L_{task}.

At inference, we retain both encoders and the classifier and omit the auxiliary modules. Algorithm 1 gives the training procedure, including microbatch reductions and gradient accumulation. Appendix E details the implementation and computational cost.

4 EXPERIMENTS

We evaluate the resulting classifiers against fixed-pair and published references, then isolate the ACA terms within a fixed architecture. C-RADIOv4 and SAM3 test Gram alignment with additional encoder families; native-task experiments combine feature fusion with Gram supervision.

4.1 EXPERIMENTAL SETUP

Datasets and tasks. We train separate classifiers on COCO, ImageNet-1K, and ImageNet-Hard (Lin et al., 2014; Deng et al., 2009; Taesiri et al., 2023). COCO uses 80 categories and one label per image, selected from the first instance annotation encountered by the loader. COCO and ImageNet-1K use their training and validation partitions; ImageNet-Hard uses a within-dataset, class-stratified split. The CLIP native-task experiments train separately on COCO and Flickr30k (Plummer et al., 2015); the DINOv3 self-supervised experiments use COCO.

Encoders and optimization. Our classification recipe uses AdamW (Loshchilov & Hutter, 2019), batch size 16, two gradient-accumulation steps, and 20 epochs with cosine decay. Both encoders and all task and auxiliary modules are jointly optimized. Learning rates are 10^{-6} for encoders and 5×10^{-4} for task and auxiliary modules. Full ACA uses $\lambda_g = 0.01$ and $\lambda_i = 0.1$ throughout. We evaluate it on CLIP–DINOv3 B/B and L/L pairs, and use DINOv3-S+ with CLIP ViT-B/16 for the ImageNet-Hard loss study. C-RADIOv4 and SAM3 use the Gram extension with learned token-grid projections. Their original input resolutions are retained within each comparison. Appendix E specifies features, heads, and reductions.

Comparisons and metrics. *Concatenation* combines the encoders' features with task supervision. *Gram* adds MSE between normalized patch Gram matrices. *ACA* uses Eq. (12); its loss subsets retain the same predictor and auxiliary heads. *Direct-sum* directly adds the importance, residual, and gate signals. The plain Gram and ACA implementations have different prediction heads, so their comparison concerns the complete training configurations. The subset study isolates term inclusion within ACA. We report top-1 accuracy and changes in percentage points (pp).

4.2 CLASSIFICATION RESULTS AND ANALYSIS

Table 1 organizes results by encoder pair and training objective. Each row compares a fusion configuration with its stronger constituent and with concatenation. The two panels distinguish the complete ACA loss from heterogeneous-grid experiments that use only Gram alignment.

Table 1: Classification accuracy (%). **(a)** Full-ACA comparisons at 224-pixel resolution; Best single denotes DINOv3. Δ is ACA minus concatenation. **(b)** COCO Gram extensions, with input sizes shown in encoder order; Δ is Gram minus concatenation. Bold marks the best fusion result in each row. Additional baselines appear in Appendix C.

(a) Complete objective: fixed CLIP–DINOv3 pairs					
Encoder pair	Best single	Concat.	Gram	ACA	Δ
COCO					
CLIP-B + DINOv3-B	73.64	76.04	76.33	**76.98**	+0.94
CLIP-L + DINOv3-L	75.71	77.60	77.47	**77.85**	+0.25
ImageNet-1K					
CLIP-B + DINOv3-B	84.40	84.64	85.04	**85.37**	+0.73
CLIP-L + DINOv3-L	86.92	87.36	**88.08**	87.99	+0.63
(b) Gram component: heterogeneous encoders on COCO					
Encoder pair (input sizes)	Best single	Concat.	Gram		Δ
C-RADIOv4 + DINOv3-L (512/224)	79.68	79.95	**80.04**		+0.09
SAM3 + DINOv3-L (252/224)	77.75	78.76	**79.13**		+0.37

Constituent and fusion performance. Concatenation improves on the stronger constituent in all four matched-scale settings. On COCO, these gains are 2.40 pp for the B/B pair and 1.89 pp for L/L; ACA adds another 0.94 and 0.25 pp. On ImageNet-1K, concatenation adds 0.24 and 0.44 pp, whereas the ACA configuration adds 0.73 and 0.63 pp. Most of the COCO gain accompanies feature combination; on ImageNet, the larger increment occurs between concatenation and the ACA configuration. This comparison addresses the gap identified in Section 1: giving the classifier both representations leaves room to improve how those representations are adapted together.

Uniform matching versus selective alignment. For the B/B pair, Gram improves over concatenation by 0.29/0.40 pp on COCO/ImageNet-1K; ACA improves over Gram by a further 0.65/0.33 pp. For L/L on COCO, uniform matching lowers accuracy by 0.13 pp, while ACA raises it by 0.25 pp. The ordering reverses on ImageNet-1K L/L, where Gram reaches 88.08% and ACA 87.99%. We therefore find that ACA improves all four concatenation references, but exceeds uniform alignment in three. The COCO reversal supports learning how relations enter the objective; the ImageNet result shows that this choice does not improve every encoder–dataset combination. Section 4.4.1 examines the loss construction with head capacity held fixed.

Fusion with a distilled encoder. C-RADIOv4 already distills DINOv3-7B alongside SigLIP2 and SAM3 (Ranzinger et al., 2026). We fuse it with DINOv3-L to test whether a member of that teacher family still contributes during downstream adaptation. Concatenation raises COCO accuracy from 79.68% to 79.95%; Gram alignment then reaches 80.04%, the highest among our COCO configurations at their stated resolutions. The two increments, 0.27 and 0.09 pp, distinguish adding the encoder from introducing a relational training objective. Thus, shared teacher lineage does not remove the measured benefit of fusion in this setting. SAM3 + DINOv3-L follows the same ordering, gaining 1.01 pp from concatenation and another 0.37 pp from Gram alignment. These extensions support relational training across encoder families alongside the full-ACA evaluations in Table 1(a).

Scale and pairing. Changing from the B/B to the L/L pair raises ACA accuracy by 0.87/2.62 pp on COCO/ImageNet-1K, while its gain over concatenation decreases from 0.94/0.73 to 0.25/0.63 pp. A stronger final classifier therefore need not show a larger gain from the loss. The cross-scale concatenation results in Appendix C.1 reinforce this distinction: adding CLIP-B to DINOv3-L changes ImageNet accuracy by -0.10 pp. We use within-pair comparisons to assess the training objective and report final accuracy alongside them.

4.3 Comparison with Published ImageNet-1K Classifiers

Table 2(a) compares full-data supervised adaptation on ImageNet-1K at 224×224. FT-CLIP refines the adaptation recipe for CLIP; BEiT-3 uses masked multimodal pretraining and a vision-only classification path (Dong et al., 2022; Wang et al., 2023).

ACA attains competitive ImageNet-1K accuracy: the L/L configuration reaches 87.99%, compared with 88.0% for FT-CLIP-L and 87.6% for BEiT-3-L. At B/B scale, its 85.37% is close to BEiT-3-B's 85.4% and below FT-CLIP-B's 85.7%. These are comparisons of complete training configurations. FT-CLIP uses EMA, layer-wise learning-rate decay, and 50/30 epochs for B/L; we use a common 20-epoch recipe. Our single CLIP scores of 81.35/85.19% further illustrate why the adaptation recipe matters when interpreting these differences. Table 2 establishes the performance reference, while the fixed-pair comparisons and fixed-architecture subsets evaluate the ACA construction.

Recent agglomerative models require a separate comparison. At 512 pixels, SigLino reports 86.10% kNN accuracy for its dense model and 85.89% for its MoE model (Chaybouti et al., 2026). Our C-RADIOv4 result at that resolution is 87.95% after supervised adaptation. Table 2(b) retains both evaluation modes in separate columns: equal resolution does not make frozen-feature kNN and backbone fine-tuning equivalent. C-RADIOv4's own report evaluates zero-shot and kNN transfer (Ranzinger et al., 2026); Appendix F records those results and the recent few-shot fusion comparisons. We assess ACA's gain within our training protocol.

FT-CLIP and BEiT-3 retain one visual encoder, and multi-teacher distillation retains one student. ACA retains both encoders and the classifier, with inference cost $C_A + C_B + C_g$. Its importance and gate heads are used only during training. Appendix E.6 gives the operation and parameter counts for these designs.

Table 2: ImageNet-1K top-1 accuracy under specified protocols. **(a)** Full-data supervised adaptation at 224 pixels: published recipes and ours. **(b)** Recent agglomerative encoders at 512 pixels: frozen-feature kNN and supervised fine-tuning occupy separate columns. SigLino results are from its 2026 report; C-RADIOv4 fine-tuning is our experiment.

(a) Full-data supervised adaptation, 224 pixels			
Visual encoder(s)	Adaptation recipe	Epochs	Accuracy (%)
CLIP ViT-B/16	FT-CLIP	50	85.7
CLIP ViT-L/14	FT-CLIP	30	88.0
BEiT-3 B/16	Released fine-tuning	50	85.4
BEiT-3 L/16	Released fine-tuning	50	87.6
CLIP-B + DINOv3-B	Ours, task loss + ACA	20	85.37
CLIP-L + DINOv3-L	Ours, task loss + ACA	20	87.99
(b) Recent agglomerative encoders, 512 pixels			
Encoder	Output	kNN	Fine-tuned
SigLino-0.6B	Head ensemble	86.10	–
SigLino-MoE-0.3-0.6B	Head ensemble	85.89	–
C-RADIOv4	Our task head	–	87.95

4.4 Ablation Studies

We test the shared-objective construction by activating every nonempty subset of its three terms and by replacing their construction with direct summation of the intermediate signals.

4.4.1 Loss Components and Their Interactions

The ImageNet-Hard study in Table 3 holds the encoder pair, prediction and auxiliary heads, and active-term coefficients fixed across seven subsets. Plain Gram MSE is a separate baseline without learned importance or allocation.

Table 3: ImageNet-Hard accuracy with DINOv3-S+ and CLIP-B. The seven ACA subsets share the same architecture. The separate Gram baseline uses unweighted MSE; L_{align} is ACA's learned weighted alignment term.

Configuration	L_{align}	L_{comp}	L_{gate}	Accuracy (%)
Gram (separate baseline)	n/a	n/a	n/a	69.39
Alignment	✓	–	–	69.44
Complementary variance	–	✓	–	68.95
Gate mean	–	–	✓	69.01
Alignment + variance	✓	✓	–	69.32
Alignment + gate	✓	–	✓	68.83
Variance + gate	–	✓	✓	68.95
Full ACA	✓	✓	✓	**69.98**

Removing terms from the full objective. Full ACA reaches 69.98%, exceeding the best single term by 0.54 pp and the best two-term subset by 0.66 pp. Removing gate mean, complementary variance, or alignment lowers accuracy by 0.66, 1.15, or 1.03 pp. Each term thus improves the objective when the other two are present. Weighted alignment alone reaches 69.44%, close to plain Gram MSE at 69.39%; the larger improvement requires the combined statistics.

Conditional contributions. Adding variance to alignment changes accuracy by -0.12 pp, but adding it to alignment plus gate mean gives $+1.15$ pp. Alignment likewise adds 0.37 pp to variance alone and 1.03 pp when gate mean is present. These conditional comparisons concern the shared variables in Eqs. (6)–(8): gate mean penalizes m, while variance rewards dispersion in

$q = W(1 - m)X$. Removing either term changes the objective that trains the same allocation. All two-term subsets fall below the best single term, yet the full objective ranks first. Selecting terms only by their gains over a single-term reference would therefore miss the best measured configuration. Appendix B reports the conditional contrasts.

4.4.2 Direct Summation of the Intermediate Signals

The `mp + diff + mask` control includes the importance, residual, and gate signals but adds them directly. Table 5 (Appendix B) compares it with ACA and the reference objectives.

ACA exceeds Direct-sum by 1.01 pp on ImageNet-Hard and 0.67 pp on COCO. Direct-sum falls 0.42 pp below Gram on ImageNet-Hard and differs from it by -0.02 pp on COCO. Its 0.27 pp gain over COCO concatenation shows that including the signals can improve accuracy without reproducing the complete result.

ACA uses importance and allocation to weight squared disagreement and to define the complementary residual whose centered variation is rewarded. Both contributions depend on the same learned weights, as their derivatives make explicit (Appendix D). The direct-sum comparison evaluates this objective construction; the subset experiment identifies the contribution of each term within it.

Learned allocation. The four ImageNet-Hard checkpoint panels in Appendix A.2 show nonzero gates that vary over patch pairs. At epoch 3, several high-residual columns receive lower gate values. Each column groups relations involving one patch, making the allocation spatially indexed even though the importance head is conditioned on global summaries. The panels depict different images and describe m, whereas the variance term acts on $q = W(1 - m)X$. They provide examples of the learned allocation; the fixed-head ablations measure how its training objectives affect accuracy.

4.5 Extensions to Native Training Objectives

We extend CLIP's image–text and DINOv3's self-supervised training with feature fusion and Gram supervision (Appendix C). These comparisons evaluate the combined extension. CLIP evaluation measures matching within each batch. In the primary experiments, the extension changes COCO I2T/T2I accuracy by $-0.02/+0.06$ pp and lowers Flickr30k accuracy by 0.66/0.23 pp. The additional COCO and Flickr30k experiments each gain 0.49/0.37 pp over their own Native reference, with opposite changes in total loss: $+0.0335$ and -0.0206, respectively. DINOv3's total loss decreases by 0.02 at B scale and 0.03 at L scale, with mixed component changes. The CLIP results therefore show setting-dependent retrieval effects; the DINOv3 records measure objective values. Because fusion and Gram supervision both change the training configuration, these experiments complement the fixed-architecture classification ablations rather than isolate the Gram term.

5 Limitations

ACA retains both encoders at inference, so deployment cost depends on the chosen pair. Our evaluation of the complete objective focuses on classification with corresponding patch grids. Extending it to dense prediction, more than two encoders, or tasks without spatial correspondence requires adapting the relation construction and task supervision.

6 Conclusion

We introduced ACA, which couples relational alignment and complementary residual regularization through shared importance and allocation. It reaches 87.99% on ImageNet-1K and improves all four matched-scale concatenation references by 0.25–0.94 pp. On ImageNet-Hard, the complete objective exceeds the best one- and two-term subsets by 0.54 and 0.66 pp. The C-RADIOv4 extension shows a benefit from relational training even when the fused encoders share a teacher family. The formulation separates the fusion objective from the choice of backbones: Gram residuals compare relations across unequal channel dimensions, while the task loss supplies prediction supervision. This separation provides a basis for extending the coupled objective to other cross-model fusion tasks with suitable feature correspondences.

AI USE STATEMENT

Generative AI assisted with language polishing and editorial revision of the manuscript written by the authors.

REPRODUCIBILITY STATEMENT

We specify the auxiliary and its reductions in Section 3. Section 4.1 defines the tasks and comparisons, and Appendix E describes encoder features, optimization, evaluation, and computational cost. The source package provides the tabulated observations, derived comparisons, original checkpoint images, figure-generation scripts, and source-record audit. These materials reproduce the descriptive analysis from the recorded experiments.

REFERENCES

Adrien Bardes, Jean Ponce, and Yann LeCun. VICReg: Variance-invariance-covariance regularization for self-supervised learning. In *International Conference on Learning Representations*, 2022. URL https://arxiv.org/abs/2105.04906.

Nicolas Carion, Laura Gustafson, Yuan-Ting Hu, Shoubhik Debnath, Ronghang Hu, Didac Suris, et al. SAM 3: Segment anything with concepts. *arXiv preprint arXiv:2511.16719*, 2025. URL https://arxiv.org/abs/2511.16719.

Sofian Chaybouti, Sanath Narayan, Yasser Dahou, Phúc H. Lê Khac, Ankit Singh, Ngoc Dung Huynh, Wamiq Reyaz Para, Hilde Kuehne, and Hakim Hacid. SigLino: Efficient multi-teacher distillation for agglomerative vision foundation models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2026. URL https://arxiv.org/abs/2512.20157.

Timothée Darcet, Maxime Oquab, Julien Mairal, and Piotr Bojanowski. Vision transformers need registers. In *International Conference on Learning Representations*, 2024. URL https://arxiv.org/abs/2309.16588.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. ImageNet: A large-scale hierarchical image database. In *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 248–255, 2009. URL https://www.image-net.org/static_files/papers/imagenet_cvpr09.pdf.

Xiaoyi Dong, Jianmin Bao, Ting Zhang, Dongdong Chen, Shuyang Gu, Weiming Zhang, Lu Yuan, Dong Chen, Fang Wen, and Nenghai Yu. CLIP itself is a strong fine-tuner: Achieving 85.7% and 88.0% top-1 accuracy with ViT-B and ViT-L on ImageNet. *arXiv preprint arXiv:2212.06138*, 2022. URL https://arxiv.org/abs/2212.06138.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, et al. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2021. URL https://arxiv.org/abs/2010.11929.

Leon A. Gatys, Alexander S. Ecker, and Matthias Bethge. A neural algorithm of artistic style. *arXiv preprint arXiv:1508.06576*, 2015. URL https://arxiv.org/abs/1508.06576.

Greg Heinrich, Mike Ranzinger, Hongxu Yin, Yao Lu, Jan Kautz, Andrew Tao, Bryan Catanzaro, and Pavlo Molchanov. RADIOv2.5: Improved baselines for agglomerative vision foundation models. *arXiv preprint arXiv:2412.07679*, 2024. URL https://arxiv.org/abs/2412.07679.

Minyoung Huh, Brian Cheung, Tongzhou Wang, and Phillip Isola. Position: The platonic representation hypothesis. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235, pp. 20617–20642. PMLR, 2024. URL https://proceedings.mlr.press/v235/huh24a.html.

Bo Jiang, Xueyang Ze, Beibei Wang, Xixi Wang, Xixi Wan, and Bin Luo. Beyond graph model: Reliable VLM fine-tuning via random graph adapter. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11664–11673, 2026a. URL https://arxiv.org/abs/2507.10355.

Shuyang Jiang, Nan Yu, Yiming Zhang, Zenghui Ding, and Zhenyu Wu. DINORANKCLIP: DINOv3 distillation and injection for vision-language pretraining with high-order ranking consistency. *arXiv preprint arXiv:2605.06592*, 2026b. URL https://arxiv.org/abs/2605.06592.

Dong Jing, Xiaolong He, Yutian Luo, Nanyi Fei, Guoxing Yang, Wei Wei, Huiwen Zhao, and Zhiwu Lu. FineCLIP: Self-distilled region-based CLIP for better fine-grained understanding. In *Advances in Neural Information Processing Systems*, volume 37, 2024. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/3122aaa22b2fe83f9cead1a696f65ceb-Abstract-Conference.html.

Rishabh Kabra, Maks Ovsjanikov, Drew A. Hudson, Ye Xia, Skanda Koppula, Andre Araujo, Joao Carreira, and Niloy J. Mitra. A mixed diet makes DINO an omnivorous vision encoder. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2026. URL https://arxiv.org/abs/2602.24181.

Simon Kornblith, Mohammad Norouzi, Honglak Lee, and Geoffrey Hinton. Similarity of neural network representations revisited. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97, pp. 3519–3529. PMLR, 2019. URL https://proceedings.mlr.press/v97/kornblith19a.html.

Alexander Lappe and Martin A. Giese. Register and [CLS] tokens induce a decoupling of local and global features in large ViTs. In *Advances in Neural Information Processing Systems*, volume 38, 2025. URL https://proceedings.neurips.cc/paper_files/paper/2025/hash/01bcbd34b02f3da9700a3ddd0480c156-Abstract-Conference.html.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft COCO: Common objects in context. In *European Conference on Computer Vision*, pp. 740–755, 2014. URL https://doi.org/10.1007/978-3-319-10602-1_48.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2019. URL https://arxiv.org/abs/1711.05101.

Wonpyo Park, Dongju Kim, Yan Lu, and Minsu Cho. Relational knowledge distillation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3967–3976, 2019. URL https://arxiv.org/abs/1904.05068.

Bryan A. Plummer, Liwei Wang, Chris M. Cervantes, Juan C. Caicedo, Julia Hockenmaier, and Svetlana Lazebnik. Flickr30k Entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 2641–2649, 2015. URL https://arxiv.org/abs/1505.04870.

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 *Proceedings of the 38th International Conference on Machine Learning*, volume 139, pp. 8748–8763. PMLR, 2021. URL https://proceedings.mlr.press/v139/radford21a.html.

Mike Ranzinger, Jon Barker, Greg Heinrich, Pavlo Molchanov, Bryan Catanzaro, and Andrew Tao. PHI-S: Distribution balancing for label-free multi-teacher distillation. *arXiv preprint arXiv:2410.01680*, 2024a. URL https://arxiv.org/abs/2410.01680.

Mike Ranzinger, Greg Heinrich, Jan Kautz, and Pavlo Molchanov. AM-RADIO: Agglomerative vision foundation model—reduce all domains into one. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 12490–12500, 2024b. URL https://arxiv.org/abs/2312.06709.

Mike Ranzinger, Greg Heinrich, Collin McCarthy, Jan Kautz, Andrew Tao, Bryan Catanzaro, and Pavlo Molchanov. C-RADIOv4 (tech report). *arXiv preprint arXiv:2601.17237*, 2026. URL https://arxiv.org/abs/2601.17237.

Mert Bulent Sariyildiz, Philippe Weinzaepfel, Thomas Lucas, Diane Larlus, and Yannis Kalantidis. UNIC: Universal classification models via multi-teacher distillation. In *European Conference on Computer Vision*, 2024. URL https://arxiv.org/abs/2408.05088.

Mert Bulent Sariyildiz, Philippe Weinzaepfel, Thomas Lucas, Pau de Jorge, Diane Larlus, and Yannis Kalantidis. DUNE: Distilling a universal encoder from heterogeneous 2D and 3D teachers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2025. URL https://arxiv.org/abs/2503.14405.

Jinghuan Shang, Karl Schmeckpeper, Brandon B. May, Maria Vittoria Minniti, Tarik Kelestemur, David Watkins, and Laura Herlant. Theia: Distilling diverse vision foundation models for robot learning. In *Proceedings of The 8th Conference on Robot Learning*, volume 270, pp. 724–748. PMLR, 2025. URL https://proceedings.mlr.press/v270/shang25a.html.

Min Shi, Fuxiao Liu, Shihao Wang, Shijia Liao, Subhashree Radhakrishnan, Yilin Zhao, et al. Eagle: Exploring the design space for multimodal LLMs with mixture of encoders. In *International Conference on Learning Representations*, 2025. URL https://arxiv.org/abs/2408.15998.

Oriane Siméoni, Huy V. Vo, Maximilian Seitzer, Federico Baldassarre, Maxime Oquab, Cijo Jose, et al. DINOv3. *arXiv preprint arXiv:2508.10104*, 2025. URL https://arxiv.org/abs/2508.10104.

Baochen Sun and Kate Saenko. Deep CORAL: Correlation alignment for deep domain adaptation. *arXiv preprint arXiv:1607.01719*, 2016. URL https://arxiv.org/abs/1607.01719.

Mohammad Reza Taesiri, Giang Nguyen, Sarra Habchi, Cor-Paul Bezemer, and Anh Nguyen. ImageNet-Hard: The hardest images remaining from a study of the power of zoom and spatial biases in image classification. In *Advances in Neural Information Processing Systems*, 2023. URL https://arxiv.org/abs/2304.05538.

Yuwei Tang, Zhenyi Lin, Qilong Wang, Pengfei Zhu, and Qinghua Hu. AMU-Tuning: Effective logit bias for CLIP-based few-shot learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 23323–23333, 2024. URL https://arxiv.org/abs/2404.08958.

Shengbang Tong, Ellis Brown, Penghao Wu, Sanghyun Woo, Manoj Middepogu, Sai Charitha Akula, et al. Cambrian-1: A fully open, vision-centric exploration of multimodal LLMs. In *Advances in Neural Information Processing Systems*, 2024. URL https://arxiv.org/abs/2406.16860.

Michael Tschannen, Alexey Gritsenko, Xiao Wang, Muhammad Ferjad Naeem, Ibrahim Alabdulmohsin, Nikhil Parthasarathy, et al. SigLIP 2: Multilingual vision-language encoders with improved semantic understanding, localization, and dense features. *arXiv preprint arXiv:2502.14786*, 2025. URL https://arxiv.org/abs/2502.14786.

Wenhui Wang, Hangbo Bao, Li Dong, Johan Bjorck, Zhiliang Peng, Qiang Liu, Kriti Aggarwal, Owais Khan Mohammed, Saksham Singhal, Subhojit Som, and Furu Wei. Image as a foreign language: BEiT pretraining for vision and vision-language tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 19175–19186, 2023. URL https://openaccess.thecvf.com/content/CVPR2023/html/Wang_Image_as_a_Foreign_Language_BEiT_Pretraining_for_Vision_and_CVPR_2023_paper.html.

Size Wu, Wenwei Zhang, Lumin Xu, Sheng Jin, Xiangtai Li, Wentao Liu, and Chen Change Loy. CLIPSelf: Vision transformer distills itself for open-vocabulary dense prediction. In *International Conference on Learning Representations*, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/hash/e7947b5e1d30864ebbe8714dbdd611d9-Abstract-Conference.html.

Feyza Yavuz, Mert Bülent Sarıyıldız, and Diane Larlus. IDeaL: Data-free multi-teacher distillation via improved dead leaves. In *European Conference on Computer Vision*, 2026. URL https://arxiv.org/abs/2608.24759.

Jure Zbontar, Li Jing, Ishan Misra, Yann LeCun, and Stephane Deny. Barlow twins: Self-supervised learning via redundancy reduction. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139, pp. 12310–12320. PMLR, 2021. URL https://proceedings.mlr.press/v139/zbontar21a.html.

Jinjin Zhang, Xiefan Guo, and Di Huang. Spatial gram alignment for ultra-high-resolution image synthesis. *arXiv preprint arXiv:2605.20808*, 2026. URL https://arxiv.org/abs/2605.20808.

Renrui Zhang, Xiangfei Hu, Bohao Li, Siyuan Huang, Hanqiu Deng, Yu Qiao, Peng Gao, and Hongsheng Li. Prompt, generate, then cache: Cascade of foundation models makes strong few-shot learners. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 15211–15222, 2023. URL https://arxiv.org/abs/2303.02151.

A PATCH-RELATION DIAGNOSTICS

We compare cross-encoder patch relations to examine the geometry left implicit by global feature fusion. Similarity-based analyses compare learned representation structure across examples (Kornblith et al., 2019; Huh et al., 2024); here we examine patch pairs within each image. The histograms summarize discrepancy magnitudes, while the checkpoint matrices retain pair indices and show how learned gates vary over them.

Measuring relational inconsistency. We measure the cosine similarities assigned to corresponding patch pairs by CLIP ViT-B/16 and DINOv3-B. A pair's squared discrepancy is $r_{ij} = [\cos(F_{A,i}, F_{A,j}) - \cos(F_{B,i}, F_{B,j})]^2$, written in Gram form in Eq. (1). Smaller values indicate closer agreement between the encoders on the relation between patches i and j. At 224×224 resolution, each spatial grid contains 196 tokens after excluding class and register tokens, giving 19,110 distinct off-diagonal pairs per image. Table 4 summarizes the ten-bin distributions; Figure 2 gives the histograms.

Table 4: Counts and relative frequencies of squared Gram discrepancies r_{ij}. The first bin is $0 \leq r_{ij} < 0.1$. Percentages count unordered, off-diagonal patch pairs. Figure 2 shows all ten bins.

Statistic	ImageNet-Hard	COCO val2017
Counted patch pairs	209,827,800	94,174,080
Lowest bin	88.63%	91.73%
Remaining nine bins	11.37%	8.27%
Last seven bins	6.60%	4.16%
Last three bins	0.90%	0.69%
First-to-second bin ratio	≈ 36	≈ 45

On ImageNet-Hard, 88.63% of patch pairs have $r_{ij} < 0.1$; on COCO, the fraction is 91.73%. These counts exceed the next bin by approximately 36 and 45 times. Most pairwise relations therefore have small discrepancies across encoders, while the other bins contain 11.37% and 8.27% of pairs. The checkpoint matrices in Figure 3 add spatial information: larger residuals occur along particular rows and columns, each grouping relations involving one patch. The distributions and matrices thus describe the magnitude and spatial organization of cross-encoder relational disagreement.

From discrepancy to a fusion objective. These observations place relational agreement and residual differences within the same fusion problem. Global concatenation leaves their treatment implicit, while uniform Gram matching assigns every squared residual the same coefficient. We learn how to allocate alignment across these relations and regularize the complementary residual through the same weights. Global summaries condition importance, patch features determine allocation, and task supervision adapts both encoders. Section 3 develops this objective; the fixed-architecture ablations in Section 4.4.1 test the joint use of its terms.

A.1 PATCH-PAIR COUNTS AND HISTOGRAM SCALE

The two distributions contain 209,827,800 counted pairs on ImageNet-Hard and 94,174,080 on COCO val2017. With 196 spatial tokens, one complete upper triangle excluding the diagonal contains 19,110 pairs. The totals correspond to 10,980 and 4,928 such image grids, respectively. These totals count patch pairs, with each token participating in many pairs.

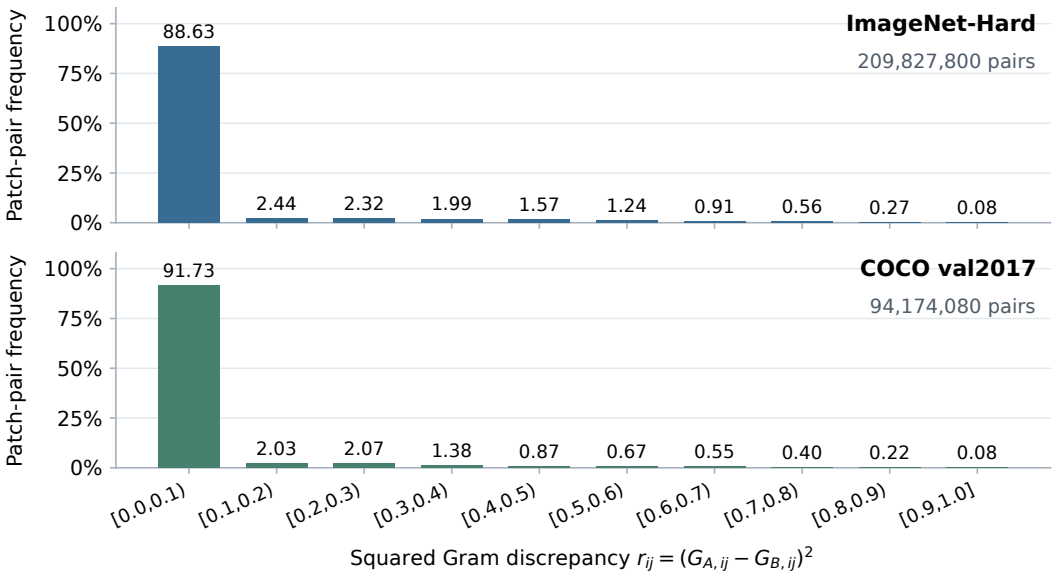

Figure 2: Squared Gram discrepancies between CLIP ViT-B/16 and DINOv3-B, reconstructed from the original integer counts. Each patch vector is L2-normalized before Gram construction; the horizontal axis plots $r_{ij} = (G_{A,ij} - G_{B,ij})^2$. Bar labels give percentages of unordered, off-diagonal patch pairs. The first bin, $0 \leq r_{ij} < 0.1$, contains 88.63% of ImageNet-Hard pairs and 91.73% of COCO pairs.

Normalization acts on the patch vectors, $\widehat{F}_{k,i} = F_{k,i}/\|F_{k,i}\|_2$, before computing the Gram matrices. The histograms plot the squared residual r_{ij} from Eq. (1) in the original ten bins spanning $[0, 1]$. The theoretical bound $r_{ij} \leq 4$ follows from the cosine range $[-1, 1]$ and does not prescribe the displayed range. In the first bin, the two cosine similarities differ by less than $\sqrt{0.1} \approx 0.316$. These frequencies describe agreement in relational geometry. We sum integer counts before rounding percentages.

ImageNet-Hard places 11.37% of counted pairs outside the first bin, compared with 8.27% on COCO. Its last seven and last three bins also carry more mass. A patch participates in many relations, so these fractions cannot be converted into a number of foreground or divergent tokens without identifying their shared indices. The checkpoint matrices below display those indices for individual samples.

The Gram comparison measures within-image relation geometry. CLIP and DINOv3 use different pretraining objectives, but this does not establish disjoint training data: CLIP uses web image–text pairs, and DINOv3's pretraining collection includes ImageNet (Radford et al., 2021; Siméoni et al., 2025). The histogram is a descriptive distribution, not an estimate of the effect of pretraining-data overlap. The classification comparisons in Section 4.2 evaluate the downstream effect of relational training with fixed encoder pairs.

A.2 CHECKPOINT RESIDUAL AND GATE PANELS

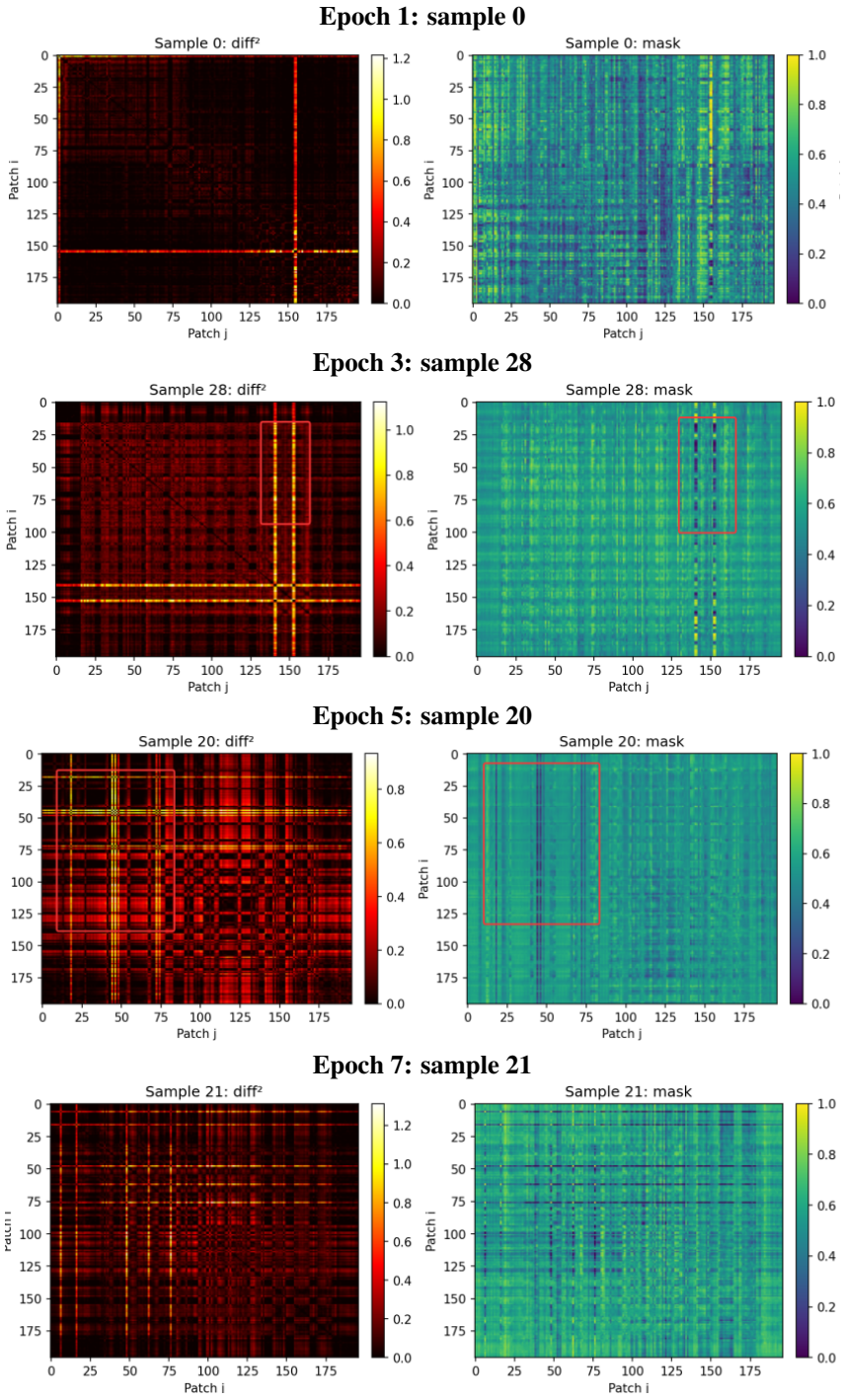

Figure 3: Squared residuals (left) and alignment gates (right) for DINOv3-S+ + CLIP ViT-B/16 on ImageNet-Hard. Sample and epoch identifiers and original color scales are retained.

The gates are nonzero and vary across relations. At epoch 3 (sample 28), several columns with larger residuals receive lower gate values. The panels preserve the pair indices lost in the histograms and show the learned allocation at those indices.

B Loss Controls and Conditional Contributions

B.1 Direct Summation Control

Table 5 reports the direct-sum comparison analyzed in Section 4.4.2. The three intermediate signals are included, while ACA constructs alignment, gate-mean, and complementary-variance statistics from their shared variables.

Table 5: Direct-sum control and reference accuracies (%). ImageNet-Hard uses DINOv3-S+ + CLIP-B; COCO uses DINOv3-B + CLIP-B.

Configuration	ImageNet-Hard	COCO
Concatenation	–	76.04
Gram	69.39	76.33
Direct-sum	68.97	76.31
Full ACA	**69.98**	**76.98**
ACA − Direct-sum	+1.01	+0.67

B.2 Conditional Contributions

Figure 4 displays the six conditional additions in Section 4.4.1. All configurations retain the encoder pair, prediction and auxiliary heads, and the coefficients of active terms. The horizontal axis gives the accuracy change relative to the indicated starting subset.

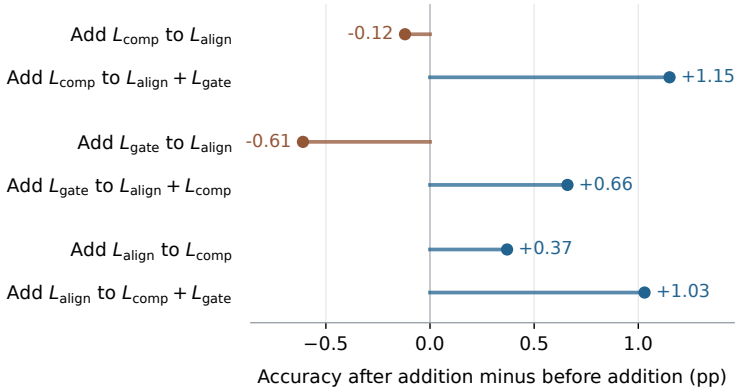

Figure 4: Conditional contributions from Table 3. Each row gives the accuracy change after adding one objective to the indicated starting subset. The same added objective can raise or lower accuracy depending on the other active terms.

For terms u, v and a fixed third term w, we calculate the conditional interaction as

$$J_{u,v|w} = f(\{u, v, w\}) - f(\{u, w\}) - f(\{v, w\}) + f(\{w\}), \tag{13}$$

where $f(S)$ is accuracy with active terms S. Table 6 expands each contrast using measured nonempty subsets. For example, $J_{I,g|G} = 1.27$ pp is the change in the variance contribution from -0.12 to $+1.15$ pp when gate regularization is present.

Table 6: Conditional interactions. G, I, and g denote alignment, variance, and gate mean. Scores are percentages; contrasts are pp.

Interaction	Calculation	Value (pp)
$J_{I,g\|G}$	$69.98 - 69.32 - 68.83 + 69.44$	$+1.27$
$J_{G,g\|I}$	$69.98 - 69.32 - 68.95 + 68.95$	$+0.66$
$J_{G,I\|g}$	$69.98 - 68.83 - 68.95 + 69.01$	$+1.21$

C Additional Baselines and Native-Objective Results

Table 7: Single-encoder accuracies (%) under our classification recipe. A dash denotes an unreported result.

Configuration	COCO	ImageNet-1K	ImageNet-Hard
DINOv3-S+	72.35	81.65	67.55
DINOv3-B	73.64	84.40	–
DINOv3-L	75.71	86.92	–
CLIP-L	75.47	85.19	–
CLIP-B	73.58	81.35	68.14
ResNet-50	–	–	64.27
C-RADIOv4	79.68	87.95	–
SAM3	77.75	–	–

C.1 Encoder Pairing and Heterogeneous Grids

Table 8: Additional encoder-pair accuracies (%). Cross-scale CLIP–DINOv3 rows use concatenation. C-RADIOv4 and SAM3 extensions add the Gram component. Best single is the stronger constituent on the named dataset.

Pair	Dataset	Best single	Concat.	Gram
DINOv3-B + CLIP-L	COCO	75.47	76.68	–
	ImageNet-1K	85.19	85.54	–
DINOv3-L + CLIP-B	COCO	75.71	76.08	–
	ImageNet-1K	86.92	86.82	–
C-RADIOv4 + DINOv3-L	COCO	79.68	79.95	80.04
SAM3 + DINOv3-L	COCO	77.75	78.76	79.13

The cross-scale rows compare the benefit of adding encoders of different strengths. DINOv3-B + CLIP-L improves over its stronger constituent by 1.21 pp on COCO and 0.35 pp on ImageNet-1K. For DINOv3-L + CLIP-B, the changes are $+0.37/-0.10$ pp. Thus the accuracy of a constituent and the benefit of pairing it with another encoder are different quantities: the stronger DINOv3-L does not ensure a larger concatenation gain. This makes matched-pair comparisons necessary when assessing ACA.

C-RADIOv4 and SAM3 extend the relational comparison to heterogeneous token grids, mapped to the DINOv3 grid through learned projections. C-RADIOv4 + DINOv3-L gains 0.27 pp over C-RADIOv4 through concatenation and another 0.09 pp through Gram alignment; SAM3 + DINOv3-L gains 1.01 pp over SAM3 and another 0.37 pp from alignment. These are the heterogeneous-grid Gram extensions discussed alongside the full-ACA results in Section 4.2. They test whether relational training adds to feature fusion with stronger or differently trained encoders.

C.2 Native Image–Text and Self-Supervised Objectives

The native-objective experiments train CLIP separately on COCO and Flickr30k (Plummer et al., 2015), and DINOv3 on COCO. Each extended configuration is compared with its Native reference from the same dataset and training experiment.

Image–text evaluation samples one caption per image and reports top-1 image-to-text (I2T) and text-to-image (T2I) matching within each batch, averaged over batches. Candidate sets follow each experiment's batch configuration. Self-supervised experiments report training losses and their components.

Both native-task extensions combine cross-encoder features and add Gram alignment. CLIP combines projected DINO and CLIP image features before caption scoring; the DINOv3 extension mixes encoder features within its DINO/iBOT training stack. The comparisons with Native therefore evaluate feature fusion together with relational supervision.

Table 9: CLIP native-task experiments, grouped by training dataset and experiment. I2T/T2I are batch retrieval top-1 percentages. Fusion + Gram combines image features and Gram supervision. Each experiment has its own Native reference. Loss is the total training objective.

Training dataset	Experiment	Configuration	Loss	I2T	T2I
COCO	Primary	Native	0.9446	72.61	72.77
		Fusion + Gram	0.9342	72.59	72.83
		Direct-sum	0.9533	72.43	72.80
Flickr30k	Primary	Native	0.5043	88.95	88.66
		Fusion + Gram	0.5386	88.29	88.43
		Direct-sum	0.5271	88.30	87.98
COCO	Additional training	Native	0.2377	93.89	93.26
		Fusion + Gram	0.2712	94.38	93.63
Flickr30k	Additional training	Native	0.2919	93.89	93.26
		Fusion + Gram	0.2713	94.38	93.63

Retrieval depends on the training setting. In the primary COCO experiment, Fusion + Gram changes Native I2T/T2I by $-0.02/+0.06$ pp. On Flickr30k, both directions decrease, by 0.66 and 0.23 pp. Direct-sum gives $-0.18/+0.03$ pp on COCO and $-0.65/-0.68$ pp on Flickr30k. Fusion + Gram exceeds Direct-sum in Flickr30k T2I by 0.45 pp, but both remain below Native. The comparison with Native is therefore less favorable than the comparison between extended configurations.

The additional COCO and Flickr30k experiments each obtain gains of 0.49 pp in I2T and 0.37 pp in T2I over their own Native reference. The primary and additional Flickr30k experiments differ in the direction of the extension's effect. We compare configurations within each experiment because its batch configuration defines the retrieval candidates.

Patch geometry and prediction quality require separate measurements. Gram alignment constrains visual patch relations; retrieval evaluates image–text rankings. The additional COCO and Flickr30k experiments improve those rankings while their total losses move in opposite directions. Lower total loss is therefore insufficient to establish a retrieval benefit (Appendix C.3). For DINOv3, total training loss decreases by 0.02 at B scale and 0.03 at L scale, while DINO and iBOT components move in opposite directions (Appendix C.4). Retrieval accuracy evaluates the CLIP extension; the DINOv3 records describe objective values only.

C.3 Image–Text Loss and Retrieval Accuracy

The retrieval records also let us compare objective values with prediction accuracy. In the additional Flickr30k experiment (Table 9), loss decreases by 0.0206 as retrieval improves. In the additional COCO experiment, loss increases by 0.0335 despite the same directional gains. Primary COCO loss decreases by 0.0104 while the two retrieval directions move oppositely. The extended objective

includes the auxiliary penalty, whereas top-1 matching depends on candidate rankings. Lower total loss consequently need not accompany better retrieval.

C.4 DINOv3 Self-Supervised Objectives

Table 10: DINOv3 SSL on COCO: native training and its feature-fusion plus Gram extension. The total includes DINO, iBOT, KoLeo, and Gram terms where active.

Configuration	Total loss	DINO loss	iBOT loss
DINOv3-B native	19.47	8.50	10.97
DINOv3-B Fusion + Gram	19.45	8.52	10.93
DINOv3-L native	20.45	9.38	10.98
DINOv3-L Fusion + Gram	20.42	9.36	10.99

At B scale, the total/DINO/iBOT changes are $-0.02/+0.02/-0.04$; at L scale, they are $-0.03/-0.02/+0.01$. The DINO and iBOT components move in different directions at each scale; their changes should be read alongside the total objective.

D Analytical Properties of the Auxiliary

Residual dependence. Holding W and m fixed makes the two residual contributions explicit:

$$\frac{\partial L_{\mathrm{ACA}}}{\partial X_a} = \frac{2W_a m_a X_a}{M} - \frac{2\lambda_i W_a (1-m_a)(q_a - \bar{q})}{M-1}. \tag{14}$$

The first contribution contracts the residual toward zero; the second depends on the centered complementary value. This partial derivative holds importance and allocation fixed. The complete training gradient also includes their dependence on the encoder features.

Bounded variance reward. The normalized Gram residual lies in $[-2, 2]$. With $W, m \in [0, 1]$, the same bound holds for $q = W \odot (1-m) \odot X$. Its empirical variance with denominator M is at most the squared half-range, 4. For the sample denominator $M-1$, and nonnegative classification, alignment, and gate penalties,

$$0 \le \mathrm{Var}(q) \le \frac{4M}{M-1}, \qquad L \ge -\frac{4\lambda_i M}{M-1}. \tag{15}$$

The bound follows from the residual range, irrespective of the relative magnitudes of the loss terms.

Gate dependence. Holding W and X fixed gives

$$\frac{\partial L_{\mathrm{ACA}}}{\partial m_a} = \frac{W_a X_a^2 + \lambda_g}{M} + \frac{2\lambda_i W_a X_a (q_a - \bar{q})}{M-1}. \tag{16}$$

The mean penalty acts directly on m, while the variance reward reaches it through q. For a constant gate $m = t\mathbf{1}$, $0 \le t \le 1$, the fixed-W, X auxiliary becomes

$$A(t) = t\mathbb{E}[W \odot X^2] + \lambda_g t - \lambda_i (1-t)^2 \,\mathrm{Var}(W \odot X), \tag{17}$$

$$A'(t) = \mathbb{E}[W \odot X^2] + \lambda_g + 2\lambda_i (1-t)\,\mathrm{Var}(W \odot X) > 0. \tag{18}$$

Here $\mathbb{E}$ denotes the full-grid mean and Var the sample variance. Smaller constant gates lower the auxiliary, approaching $t = 0$ at the sigmoid limit; no positive occupancy is imposed. This fixes W, X, whereas joint training updates them alongside a nonconstant gate (Figure 3).

Symmetry and reduction. The logits $vv^\top$ have rank at most one. Sigmoid preserves symmetry but generally increases rank, with $W_{ii} = \sigma(v_i^2) \ge 0.5$. The gate can be asymmetric. For one image, the residual's full-grid mean is

$$\mathbb{E}[\Delta_{\mathrm{diff}}] = \left\| \frac{1}{N} \sum_i \widehat{F}_{A,i} \right\|_2^2 - \left\| \frac{1}{N} \sum_i \widehat{F}_{B,i} \right\|_2^2. \tag{19}$$

The signed mean permits cancellation; alignment squares the residual, while complementary variance centers its weighted values.

E Training Algorithm and Implementation

E.1 Classification Training Algorithm

Algorithm 1 implements the full classification objective. It computes the complementary variance within each microbatch, before accumulating gradients. All encoder features remain attached to the computation graph.

Algorithm 1 ACA classification training

Inputs: labeled data; encoders E_A, E_B; importance MLP ϕ_W; gate projections P_A, P_B, temperature τ, bias b; classifier g_ψ.
Settings: $\lambda_g = 0.01$, $\lambda_i = 0.1$, $\epsilon = 0.1$, accumulation factor $K = 2$; optimizer and schedule from Table 11.

Initialize AdamW parameter groups and clear gradients.
for each training epoch **do**
  for each microbatch (x, y) of size B **do**
    Extract $(c_A, F_A) = E_A(x)$ and $(c_B, F_B) = E_B(x)$, excluding class and register tokens from the spatial grids F_A, F_B.
    Row-normalize F_k; compute $G_k = \widehat{F}_k \widehat{F}_k^\top$ and $X = G_A - G_B$.
    Set $v = \mathrm{MLP}_{\phi_W}([c_A; c_B])$ and $W = \sigma(vv^\top)$.
    Set $T_k = \mathrm{normalize}(P_k(F_k))$ and $m = \sigma(T_A T_B^\top / (|\tau| + \epsilon) + b)$.
    Flatten the full batch and pair axes into $M = BN^2$ entries, including the diagonal and both pair orders.
    Set $L_{\mathrm{align}} = M^{-1} \sum_a W_a m_a X_a^2$ and $L_{\mathrm{gate}} = M^{-1} \sum_a m_a$.
    Set $q_a = W_a(1 - m_a)X_a$, $\bar{q} = M^{-1} \sum_a q_a$, and $L_{\mathrm{comp}} = -(M-1)^{-1} \sum_a (q_a - \bar{q})^2$.
    Form $z = [c_A; \mathrm{mean}(F_A); c_B; \mathrm{mean}(F_B)]$ and compute mean cross-entropy L_{task} from $g_\psi(z)$ and y.
    Set $L = L_{\mathrm{task}} + L_{\mathrm{align}} + \lambda_g L_{\mathrm{gate}} + \lambda_i L_{\mathrm{comp}}$; backpropagate L/K through the complete computation.
    if at an accumulation boundary or the last microbatch of the epoch **then**
      Update all parameter groups with AdamW and clear gradients.
    end if
  end for
  Advance the cosine learning-rate schedule.
end for
return E_A, E_B, g_ψ for prediction using Eq. (10).

E.2 Encoder Features and Prediction

The B/B implementation uses CLIP ViT-B/16 and DINOv3-B, both with width 768 and 16-pixel patches. At an input resolution of 224×224, each encoder provides $N = 196$ spatial patch tokens. The importance MLP maps $1536 \to 1024 \to 196$, the two independent gate projections each map $768 \to 1024$, and the classifier maps $3072 \to 2048 \to C$ for C classes. We exclude CLIP's class token and DINOv3's class and four register tokens from the spatial grids; registers provide separate storage for global information (Darcet et al., 2024). The resulting features are $F_k \in \mathbb{R}^{B \times N \times d_k}$, accompanied by the class-token features c_k.

The classifier receives $[c_A; \overline{F}_A; c_B; \overline{F}_B]$, where the bar denotes mean pooling over spatial tokens. Both encoders are optimized during ACA classification. The Importance Head uses the class-token features, while the Gram residual and Alignment Gate use patch features. Channel normalization is applied separately before Gram construction and after the learned gate projections. These operations retain gradients to the encoders.

E.3 Optimization and Loss Reduction

Table 11 lists the classification hyperparameters. Each configuration uses one training seed, as specified in the table. We fix $\lambda_g = 0.01$ and $\lambda_i = 0.1$ across the reported full-ACA classification

experiments. The ImageNet-Hard subsets retain these coefficients for active terms and use weight decay 0.05 and a minimum learning rate of 10^{-8}. The image-conditioned importance and gate are learned, as are the gate temperature and bias; $\epsilon = 0.1$ is fixed.

Table 11: Classification hyperparameters. The full-ACA auxiliary coefficients are shared across the reported classification settings. Loss subsets change term inclusion while retaining the coefficients of active terms.

Setting	Main classification	Hard loss subsets
Optimizer	AdamW	AdamW
Batch size	16	16
Gradient accumulation	2 steps	2 steps
Training duration	20 epochs	20 epochs
Learning-rate schedule	Cosine	Cosine
Encoder learning rate	10^{-6}	10^{-6}
Task and auxiliary learning rate	5×10^{-4}	5×10^{-4}
λ_g, λ_i	0.01, 0.1	0.01, 0.1
Seed	42	0
Full-ACA encoder pairs	DINOv3-B + CLIP-B; DINOv3-L + CLIP-L	DINOv3-S+ + CLIP-B

In each forward pass, we compute the classifier logits and auxiliary statistics from the same encoder features. Classification cross-entropy is added to the weighted alignment, gate-mean, and complementary-variance terms in Eq. (12). Encoder parameters use the encoder learning rate; the classifier, importance MLP, gate projections, temperature, and bias use the task and auxiliary learning rate.

The auxiliary reductions operate on the full $B \times N \times N$ grids. Alignment energy and gate mass use arithmetic means. For $q = W \odot (1 - m) \odot X$, the implementation computes one sample variance over $M = BN^2$ entries, with denominator $M - 1$. The gate-mean penalty acts directly on m, while the variance reward acts on q; the allocation variable therefore enters both residual objectives.

We divide the loss by the accumulation factor before backpropagation and update parameters with AdamW at each accumulation boundary or the last microbatch of the epoch. Each microbatch computes its own variance. With batch size 16 and two accumulation steps, a complete update combines two microbatches; each variance is computed over $16 \times 196^2 = 614{,}656$ entries.

The classification loader drops incomplete batches in both training and evaluation. The trainer evaluates after each epoch and saves a checkpoint whenever validation accuracy improves.

E.4 Component Extensions and Evaluation

Plain Gram and heterogeneous encoders. The plain Gram baseline applies MSE to the two normalized patch Gram matrices and uses its own classifier. The ACA loss-subset study retains the architecture and active-term coefficients while varying which auxiliary terms are included. For heterogeneous Gram extensions, learned token-axis projections map C-RADIOv4's 1,024 tokens or SAM3's 324 tokens to the 196-token DINOv3 grid. Channel normalization and Gram construction follow this mapping.

Prediction heads in the supplied implementation. For CLIP-B/16 + DINOv3-B, plain Gram uses a $3072 \to 3072 \to C$ classifier with dropout 0.1; ACA uses $3072 \to 2048 \to C$ with dropout 0.2. Both include LayerNorm and GELU. Consequently, the Gram-to-ACA accuracy difference includes changes to the classifier as well as the auxiliary. Table 3 holds the ACA architecture fixed to compare its objective terms. The C-RADIOv4 extension uses a linear classifier over concatenated summary and pooled patch features; its token-axis projection is used for Gram alignment.

CLIP image–text training. The CLIP extension combines CLIP class-token features with projected DINO features, normalizes the combined representation, and optimizes symmetric image–

text cross-entropy with Gram MSE. Its optimizer updates CLIP and the DINO projection. Each image is paired with one sampled caption. For the batch similarity matrix S, the retrieval scores are

$$a_{\mathrm{I2T}} = \frac{100}{B}\sum_i \mathbf{1}[\arg\max_j S_{ij} = i], \qquad a_{\mathrm{T2I}} = \frac{100}{B}\sum_j \mathbf{1}[\arg\max_i S_{ij} = j]. \tag{20}$$

Incomplete batches are dropped, and scores are averaged over evaluated batches. COCO and Flickr30k results in Table 9 are grouped by training dataset and experiment; each extended configuration is compared with its corresponding Native reference.

DINOv3 self-supervised training. The self-supervised branch trains the student and its DINO/iBOT heads using an EMA teacher and a Gram teacher. Its cross-encoder extension includes Gram alignment and feature combination. Table 10 reports the total objective and the DINO and iBOT components; the total also includes the remaining terms in the training objective.

E.5 Cost of the Relational Auxiliary

Let d_A, d_B be encoder channel dimensions, d_p the gate projection dimension, and h the hidden width of the importance MLP. Given the encoder features, the two channel projections cost $O(BN(d_A + d_B)d_p)$. Forming the Gram matrices costs $O(BN^2(d_A + d_B))$, and the projected cross-encoder similarity costs $O(BN^2 d_p)$. The importance MLP costs $O(B(d_A+d_B)h+BhN)$; its outer product and the loss reductions cost $O(BN^2)$. Combining these terms gives

$$\begin{aligned} T_{\mathrm{aux}} = O\big(&BN(d_A + d_B)d_p + BN^2(d_A + d_B + d_p) \\ &+ B(d_A + d_B)h + BhN\big). \end{aligned} \tag{21}$$

The pairwise relation tensors require $O(BN^2)$ storage, and the projected patch features require $O(BNd_p)$ storage. Encoder and classifier computations contribute separately to the complete model cost. Inference uses both encoders and the classification head; the Gram, importance, and gate computations belong to the training auxiliary.

E.6 Comparison of Training and Inference Cost

Table 12 contrasts single-encoder adaptation, multi-teacher consolidation, and two-encoder training with ACA. Consolidation also supports 2D/3D representations (Sariyildiz et al., 2025), robot learning (Shang et al., 2025), and training with synthetic images (Yavuz et al., 2026). These uses vary the data or deployment task while retaining a student for inference. Parameter counts refer to the named vision models, excluding discarded pretraining modules and teachers. Section 4.3 separates full-data supervised adaptation at 224 pixels from recent 512-pixel frozen-feature results.

The released BEiT-3 classifiers contain 87M/305M vision parameters at B/L scale (Wang et al., 2023). UNIC's reported ViT-B/16 student has 85.8M parameters after removing its expendable projectors (Sariyildiz et al., 2024); C-RADIOv4-H has 631M (Ranzinger et al., 2026). SigLino's dense 0.6B model retains teacher-specific output heads (Chaybouti et al., 2026). These architectural examples distinguish retained inference computation from training supervision. Our C-RADIOv4 classification scores come from supervised adaptation; the published report evaluates frozen representations. Parameter counts above refer to the explicitly named model variants.

Where the training cost occurs. Let c_k^{f} and c_k^{b} denote the forward and backward costs of network k, and let c_{heads} collect projection-head and loss computations. The core online-distillation path with J frozen teachers has cost

$$C_{\mathrm{distill}} = c_S^{\mathrm{f}} + c_S^{\mathrm{b}} + \sum_{j=1}^{J} c_{T_j}^{\mathrm{f}} + c_{\mathrm{heads}}. \tag{22}$$

An EMA model adds a forward pass when it supplies self-distillation targets. Only the student and its retained output heads are needed for prediction. Caching teacher targets changes when the teacher computation is paid. ACA instead backpropagates through both constituent encoders during task adaptation:

$$\begin{aligned} C_{\mathrm{ACA,train}} &= c_A^{\mathrm{f}} + c_A^{\mathrm{b}} + c_B^{\mathrm{f}} + c_B^{\mathrm{b}} + c_{g,\mathrm{train}} + c_{\mathrm{aux,train}}, \\ C_{\mathrm{ACA,test}} &= c_A^{\mathrm{f}} + c_B^{\mathrm{f}} + c_g^{\mathrm{f}}. \end{aligned} \tag{23}$$

Table 12: Design and deployment of the comparison methods. Distillation targets are computed during training; teacher networks are not part of student inference.

Method	Training design	Vision computation at inference
FT-CLIP	Supervised adaptation of CLIP with EMA and layer-wise learning-rate decay.	One CLIP encoder.
BEiT-3 B/L	Masked prediction over visual and textual inputs with a Multiway Transformer; vision-only classification fine-tuning.	One vision path: 87M / 305M parameters.
UNIC	Intermediate-layer projectors and teacher dropping; the selected model uses four teachers.	One ViT-B/16 (85.8M); projectors removed.
AM-RADIO / RADIOv2.5	Teacher-specific mappings supervise summary and dense features. RADIOv2.5 adds resolution handling and teacher-loss balancing.	One student; adapters retained for teacher-compatible outputs.
C-RADIOv4-H	SigLIP2, DINOv3 and SAM3 teachers; spatial and summary-feature supervision.	One 631M student with output adapters.
SigLino 0.6B	DINOv3/SigLIP2 feature targets and asymmetric cross-example relations.	One dense 0.6B student; ensemble of its teacher-specific heads.
Two encoders + ACA	Joint task adaptation; importance and allocation couple alignment and complementary residual statistics.	Two encoders and classifier; auxiliary heads omitted.

The auxiliary term includes differentiation through the operations counted in Eq. (21). These expressions account for one update or prediction; comparing complete training budgets additionally requires the number of updates and the pretrained checkpoints being reused.

Resolution and token count. For a standard dense-attention ViT with ℓ blocks, width d, and sequence length n, the leading per-image forward terms scale as $O(\ell(nd^2+n^2d))$: channel projections and feed-forward layers contribute the first term, and pairwise attention contributes the second. In the C-RADIOv4 extension, we process 512×512 images with 1,024 spatial tokens and map their token axis to the 196-token DINOv3 grid. For N_R C-RADIOv4 tokens and channel width d_R, this linear map costs $O(Bd_RN_RN)$ before Gram construction. It changes the grid used by the auxiliary after encoding; it does not reduce the C-RADIOv4 backbone's input sequence. At 224 pixels, a patch-14 encoder has 256 spatial tokens and a patch-16 encoder has 196. Equal image resolution therefore does not imply equal sequence length or attention cost. Width, depth, and retained heads also enter the comparison.

Parameter accounting for the B/B heads. For $d_A = d_B = 768$, gate projection width $d_p = 1024$, importance hidden width 1,024, and classifier hidden width 2,048, Table 13 gives exact counts from the linear and LayerNorm layers. Dropout and GELU introduce no parameters. The two gate scalars are included. This distinguishes the training-only auxiliary from the task head, whose size depends on the number of classes C.

Table 13: Parameter counts for the B/B head dimensions above. These are architectural counts, separate from the encoder checkpoints.

Module	Parameters	Retained at inference
Importance MLP, including LayerNorm	1,776,836	No
Two gate projections, temperature and bias	1,574,914	No
Auxiliary subtotal	3,351,750	No
Classification MLP, $C = 80$	6,461,520	Yes
Classification MLP, $C = 1000$	8,346,600	Yes

F Correspondence of the Comparison Protocols

F.1 ImageNet Adaptation and Frozen Representations

The 224-pixel block of Table 2 uses ImageNet-1K validation top-1 after full-data supervised adaptation. FT-CLIP and BEiT-3 retain their published pretraining and fine-tuning recipes. Our rows use the classification recipe in Table 11. Table 14 makes the optimization differences explicit. The scores compare the resulting classifiers; the paired and subset experiments test our relational objectives.

Table 14: ImageNet-1K adaptation recipes for the 224-pixel comparison. FT-CLIP settings follow its Table 10; BEiT-3 settings follow the released classification instructions. Published batch sizes are global; ours is the microbatch size, with two accumulation steps. Paired values correspond to B/L.

Setting	FT-CLIP	BEiT-3	Ours
Epochs	50 / 30	50 / 50	20
Batch size	2048	1024	16
Backbone learning rate	6×10^{-4} / 4×10^{-4}	7×10^{-4} / 2×10^{-4}	10^{-6}
Layer-wise decay	0.60 / 0.65	0.65 / 0.80	None
Warmup epochs	10 / 5	5 / 5	None

FT-CLIP additionally evaluates EMA weights with decay 0.9998. Its recipe is designed for CLIP adaptation; our shared recipe keeps the encoder learning rate fixed across model families. A shorter schedule does not by itself establish lower training cost, since batch size, encoder count, and resolution also determine computation. Nor does the shared recipe establish each model's best achievable accuracy. Our conclusions about ACA use the measured changes for a fixed encoder pair and the fixed-head loss subsets.

F.2 Recent Agglomerative Encoders

C-RADIOv4 and SigLino both consolidate multiple teachers into one student, but differ in their teachers and relational targets (Ranzinger et al., 2026; Chaybouti et al., 2026). C-RADIOv4 uses SigLIP2 (Tschannen et al., 2025), DINOv3-7B, and SAM3 (Carion et al., 2025); SigLino uses SigLIP2 and DINOv3 with teacher-specific heads and asymmetric relations among image summaries. Our fusion extension adds DINOv3-L, a member of the teacher family, to C-RADIOv4. ACA optimizes relations among patches within an image and retains both adapted encoders.

The C-RADIOv4 report lists ImageNet-1K zero-shot/kNN accuracies of 82.01/85.76% for SO400M and 83.09/86.59% for H. Its Figures 3–4 vary input resolution; these summary values are not labeled as 224-pixel supervised fine-tuning results. SigLino's Table 1 explicitly evaluates at 512×512: the dense 0.6B ensemble obtains 86.10% kNN accuracy, and the MoE variant with 0.3B active parameters obtains 85.89%. These evaluations assess transfer through pretrained features.

We obtain 87.95% with C-RADIOv4 after supervised ImageNet adaptation, using a classifier on its summary and mean patch features. This differs from the report's text-aligned zero-shot classifier and frozen-feature kNN. Likewise, its COCO result uses our one-label-per-image task. Our experiment measures task adaptation of the pretrained encoder.

F.3 Global Prediction Fusion and Other Tasks

CaFo, AMU-Tuning, and VRGAdapter combine global predictions from frozen encoders under few-shot adaptation (Zhang et al., 2023; Tang et al., 2024; Jiang et al., 2026a). CaFo uses CLIP and DINO caches; AMU-Tuning learns an auxiliary predictor with uncertainty-dependent fusion; VRGAdapter adds a graph-based text adapter and multiple auxiliary predictions. Their ImageNet-1K results illustrate improvements within that setting: with a CLIP ViT-B/16 branch and 16 labeled examples per class, the reported accuracies are 74.48%, 74.98%, and 76.78%, respectively. The last method is a CVPR 2026 result. Full-data backbone fine-tuning uses different supervision and trainable parameters, so these values are not baselines for a claimed ACA accuracy margin.