ALIGNMENT STRENGTH IS NOT ENOUGH: BACKWARD INFLUENCE ALLOCATION IN HETEROGENEOUS MULTIMODAL LEARNING

Anonymous authors
Paper under double-blind review

ABSTRACT

Multimodal alignment is commonly controlled through a scalar strength, yet paired modalities need not contribute equally useful information. This paper studies a complementary control dimension: *backward influence allocation*. We introduce encoder-specific admission variables that determine how strongly a shared alignment gradient enters each encoder while leaving the forward alignment objective unchanged. We show that isotropic admission is equivalent to scalar tuning, whereas asymmetric admission accesses encoder updates that a single scalar weight cannot in general reproduce. Dense multi-seed comparisons reveal residual representation trade-offs under heterogeneous sensing, while matched sensing can be nearly scalar-sufficient, indicating that the value of allocation is condition dependent. Motivated by this observation, we develop a lightweight training-time policy that maps task–alignment gradient interaction to encoder-specific admission levels and learns from training-only progress, without preassigning a teacher modality or adding inference-time components. On ViT-B/16 across RGB–event and RGB–sonar retrieval, the learned controller attains the highest mean validation and ID R@5 among five specialized multimodal optimization baselines on all three datasets. Against validation-tuned scalar alignment, it achieves higher mean ID retrieval on all three datasets and higher mean validation retrieval on two. These results establish backward influence allocation as a distinct and learnable control mechanism for heterogeneous multimodal alignment.

1 INTRODUCTION

Multimodal representation learning seeks to encode shared and complementary structure across heterogeneous sources, spanning vision, language, audio, depth, and other sensor streams; surveys commonly organize this problem around representation, alignment, fusion, translation, and co-learning (Baltrušaitis et al., 2019; Ramachandram & Taylor, 2017). Early deep approaches learned correlated or transferable latent spaces across image–text, speech, audio, and video, while self-supervised systems later exploited natural cross-modal synchronization to transfer semantic structure without manual labels (Andrew et al., 2013; Aytar et al., 2016; Owens & Efros, 2018). Modern multimodal pretraining extends shared-space learning across increasingly heterogeneous inputs, but recent work also studies the geometry and scalability of alignment itself, asking how more than two modalities can be coupled, when modality gaps persist inside nominally shared spaces, and how hierarchical structure changes alignment geometry (Cicchetti et al., 2025; Eslami & de Melo, 2025; Wu et al., 2026). Recent pretraining work also broadens the objective toward context-aware adaptation and universal omni-modal representations across modality sets (Roth et al., 2025; Zhang et al., 2025).

A common premise behind alignment-based learning is that paired observations share underlying content, so their representations should be coupled in a common semantic space; task-agnostic vision–language pretraining operationalized this through cross-modal attention and explicit image–text alignment (Lu et al., 2019; Chen et al., 2020). Recent analyses show that a nominally shared space can remain only partially aligned even when instance-level matching is strong, while unified multimodal training can exhibit modality-specific gradient conflict and convergence imbalance across network depth and optimization stages (Grassucci et al., 2026; Hao et al., 2026; Chen et al., 2026).

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.

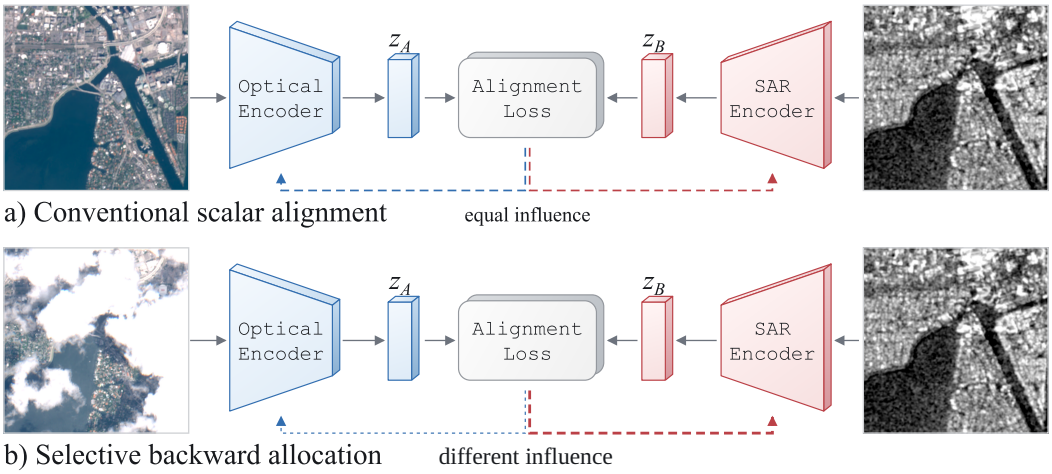


Figure 1: **Paired does not imply equally informative.** A SEN12 SAR–optical pair remains correctly matched when cloud obscures the optical view, while the information available to the two branches becomes asymmetric. Conventional scalar alignment couples the two encoder-side backward influences (top); backward allocation allows them to differ while keeping the forward alignment objective unchanged (bottom). The illustrated asymmetry is schematic rather than a prescribed teacher direction (Schmitt et al., 2019; Ebel et al., 2021).

Yet valid correspondence does not imply equal informativeness: multimodal networks can rely disproportionately on one modality or learn paired streams at different rates, indicating that paired signals can contribute unequally during optimization and leaving a sharper question for heterogeneous sensing—when correspondence remains valid but information quality differs, should one shared alignment signal reshape both encoders to the same extent (Wu et al., 2022; Yang et al., 2024)?

Figure 1 makes this distinction concrete. The clear and cloud-obscured optical views remain correctly paired with SAR, but they no longer provide equally informative evidence. Discarding the cloudy pair would remove still-valid cross-modal supervision; treating it exactly like the clear pair, however, leaves a different assumption untouched—that the shared alignment signal should reshape both encoders in the same way. Under conventional scalar alignment (Figure 1a), making alignment stronger or weaker changes the two encoder-side effects together. Figure 1b illustrates the missing degree of freedom: the forward alignment objective can remain the same while its backward influence is distributed differently across the encoders. This leads to the broader question we study: *when paired modalities contribute unequally, should alignment exert equal backward influence on both encoders?*

We study this missing degree of freedom as *backward influence allocation*. Instead of modifying the forward alignment objective, we allow its backward influence to differ across the two encoders. This separates two questions that are coupled by a conventional scalar weight: how much alignment to apply, and where that influence should be absorbed. We show that symmetric admission is equivalent to scalar tuning, whereas asymmetric admission opens off-diagonal update control. More importantly, this additional freedom is not uniformly useful. Under dense scalar tuning, heterogeneous sensing still exposes representation trade-offs outside the scalar family, while matched sensing can be nearly scalar-sufficient. Stronger alignment pressure makes allocation more consequential, and reversing the allocation redistributes encoder-side representation retention even at similar cross-modal utility.

These observations motivate making allocation conditional rather than fixing an asymmetric direction in advance. We therefore develop a lightweight training-time controller that infers two admission levels from the current task–alignment gradient interaction and learns from training-only progress, without assigning a permanent teacher modality or adding an inference-time module. In full-scale ViT-B/16 experiments on RGB–event and RGB–sonar retrieval, the learned controller achieves the highest mean validation and ID retrieval among the five specialized multimodal optimization baselines we test on all three datasets. Compared with validation-tuned scalar alignment, it is higher on mean ID retrieval on all three datasets, while the broader results show that useful allocation is condition dependent rather than universally asymmetric.

The contributions of this paper are summarized as follows:

- We identify encoder-side backward influence as a control dimension of multimodal alignment that is distinct from conventional scalar alignment strength, and introduce a forward-identical mechanism for controlling how the shared alignment gradient enters each encoder.
- We characterize the relation between alignment strength and backward allocation: isotropic admission is scalar-equivalent, whereas asymmetric admission accesses encoder updates that a single alignment weight cannot in general reproduce.
- We show empirically that the value of this additional degree of freedom depends on the learning condition. Dense scalar tuning leaves residual representation trade-offs under heterogeneous sensing, while matched sensing can be nearly scalar-sufficient; gradient-pressure and retention analyses further reveal when and how allocation becomes consequential.
- We develop a lightweight training-time controller that selects encoder-specific admissions from task–alignment gradient interaction and train-only progress, and validate it on heterogeneous RGB–event and RGB–sonar retrieval against validation-tuned scalar alignment and specialized multimodal optimization baselines.

2 Related Work

Multimodal alignment. Multimodal representation learning uses correspondence to learn shared structure across deep multimodal, audio–visual, and contrastive multiview settings (Ngiam et al., 2011; Arandjelović & Zisserman, 2017; Tian et al., 2020). Vision–language models scale this principle with paired image–text supervision (Radford et al., 2021; Jia et al., 2021; Singh et al., 2022). Any-to-any models broaden it across sensor and task spaces (Girdhar et al., 2023; Bachmann et al., 2024). Recent analyses distinguish agreement from modality-specific, redundant, unique, and synergistic information (Dufumier et al., 2025; Tjandrasuwita et al., 2025; Qian et al., 2026). We instead fix the forward alignment objective and study how its backward influence enters each encoder.

Multimodal optimization. Unequal modality learning has motivated gradient blending, on-the-fly modulation, and prototype rebalancing (Wang et al., 2020; Peng et al., 2022; Fan et al., 2023). Later work addresses imbalance through alternating unimodal adaptation, reconcilement, and data remixing (Zhang et al., 2024; Hua et al., 2024; Ma et al., 2025). Recent work diverges: ARL rejects universal balance, while Data Remixing targets balanced, sufficient learning (Wei et al., 2025a; Ma et al., 2025). These approaches modify modality-level learning more broadly; our branch objectives remain intact, and only the derivative of the shared alignment term is independently admitted.

Gradient optimization and routing. General multi-objective methods balance gradient magnitudes or alter conflicting update directions (Chen et al., 2018; Yu et al., 2020; Liu et al., 2021). Multimodal methods likewise use Pareto integration, classifier-guided modulation, or gradient truncation to reshape joint optimization (Wei & Hu, 2024; Guo et al., 2024; Wei et al., 2025b). We isolate a narrower object—the two encoder-side derivatives of one unchanged alignment loss—and characterize scalar tuning as the diagonal case of this allocation space.

3 Backward Influence Allocation

3.1 From scalar alignment to encoder-specific influence

Cross-modal representation learning commonly couples paired representations through a shared alignment objective (Radford et al., 2021), while recent multimodal methods also preserve modality-specific structure alongside cross-modal coupling (Qian et al., 2026). We study this broad setting through the generic two-branch objective

$$\mathcal{L}(\theta_A, \theta_B) = \mathcal{L}_A + \mathcal{L}_B + \lambda \mathcal{L}_{\text{align}}, \tag{1}$$

where $\mathcal{L}_A$ and $\mathcal{L}_B$ act on their respective branches and $\mathcal{L}_{\text{align}}$ couples the paired representations. The scalar $\lambda \geq 0$ is the usual alignment strength. Equation 1 is not a new loss; it is the abstraction whose backward structure we analyze. Our experiments instantiate all three terms with contrastive objectives, but the argument below does not depend on that particular choice (Appendix B).

For paired observations (x_A, x_B), let $z_m = h_m(f_m(x_m))$ denote the projected representation of branch $m \in \{A, B\}$. Writing $g_m^{\text{task}} = \nabla_{\theta_m} \mathcal{L}_m$ and $g_m^{\text{align}} = \nabla_{\theta_m} \mathcal{L}_{\text{align}}$, differentiating Eq. 1 gives

$$g_A = g_A^{\text{task}} + \lambda g_A^{\text{align}}, \qquad g_B = g_B^{\text{task}} + \lambda g_B^{\text{align}}. \tag{2}$$

Thus λ changes the two encoder-side alignment contributions together. To separate their backward influence without changing the forward objective, we introduce one admission variable per branch,

$$\tilde{g}_A = g_A^{\text{task}} + \lambda \alpha_A g_A^{\text{align}}, \qquad \tilde{g}_B = g_B^{\text{task}} + \lambda \alpha_B g_B^{\text{align}}, \tag{3}$$

with $\alpha_A, \alpha_B \in [0, 1]$. The alignment objective is unchanged; only the derivative admitted into each encoder is altered.

We realize Eq. 3 with a stop-gradient gate. Stop-gradient has long been used to alter optimization paths without changing the numerical activations in the forward pass (Chen & He, 2021), and recent multimodal work likewise exploits explicit gradient routing between model components (Wei et al., 2025b). Immediately before $\mathcal{L}_{\text{align}}$, we replace each z_m by

$$\tilde{z}_m = \text{sg}(z_m) + \alpha_m \big(z_m - \text{sg}(z_m)\big), \qquad m \in \{A, B\}, \tag{4}$$

where sg denotes stop-gradient. At a fixed parameter state, $\tilde{z}_m = z_m$ numerically, whereas $\partial \tilde{z}_m / \partial z_m = \alpha_m I$. The forward alignment value is therefore identical for all admissions at that state, while its derivative into branch m is multiplied by α_m. The branch-specific losses bypass this gate.

3.2 Scalar alignment as a special case

Multi-objective optimization is commonly analyzed through the gradient contributions supplied by different objectives (Chen et al., 2018; Yu et al., 2020); more recent multimodal methods additionally manipulate gradient paths to control cross-component influence (Wei et al., 2025b). In our two-encoder setting, these observations lead to a simple characterization of what scalar alignment can and cannot express.

Define the effective alignment coefficients $\beta_A = \lambda \alpha_A$ and $\beta_B = \lambda \alpha_B$. Figure 2 summarizes the coefficient geometry in (β_A, β_B): scalar tuning stays on the diagonal, whereas independent admission reaches the surrounding square. The first result identifies the entire isotropic admission family with ordinary scalar tuning.

Proposition 1 (isotropic equivalence). At a common parameter state, admission $(\alpha_A, \alpha_B) = (c, c)$ with weight λ supplies the same model gradients as scalar alignment with $\lambda' = c\lambda$. With matched optimizer state, data, and update operations, the resulting parameter update is therefore identical in exact arithmetic.

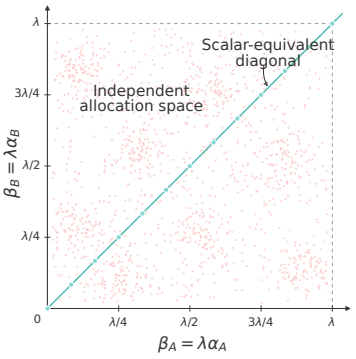


Figure 2: **Strength and allocation.** Scalar alignment occupies the diagonal $\beta_A = \beta_B$. Independent admission accesses off-diagonal coefficients within $[0, \lambda]^2$ at fixed λ.

Proposition 2 (non-reducibility of asymmetric gradients). Suppose $\lambda > 0$ and both g_A^{align} and g_B^{align} are nonzero. If $\alpha_A \neq \alpha_B$, no single scalar λ' reproduces both pre-optimizer branch gradients. Equality on branch A requires $\lambda' = \lambda \alpha_A$, whereas equality on branch B requires $\lambda' = \lambda \alpha_B$.

Together, the two propositions give the geometry in Figure 2: scalar tuning explores a one-dimensional diagonal in (β_A, β_B), whereas independent admission opens the off-diagonal plane. Hard one-way routing, such as $(\alpha_A, \alpha_B) = (0, 1)$, is only a boundary case; intermediate asymmetric points represent influence patterns that no single scalar weight can match at the gradient level.

The distinction is specific to the shared alignment contribution. A branch-specific learning-rate multiplier rescales both g_m^{task} and g_m^{align}, while reweighting directional retrieval losses generally still differentiates through both encoders. Proofs, degenerate cases, and numerical equivalence checks are given in Appendix A. Whether the additional off-diagonal degree of freedom changes useful representation trade-offs is an empirical question, which we address next.

4 Empirical Characterization of Backward Allocation

Section 3 establishes a distinction at the gradient level: independent admission can express encoder-side alignment coefficients that no single scalar weight can reproduce. The remaining question is empirical—whether this extra degree of freedom leaves an observable consequence after training. We study three questions in turn: whether allocation reaches representation trade-offs beyond a densely tuned scalar family, when those additional trade-offs become pronounced, and what changes on the two encoder sides when the allocation direction is reversed. We first introduce the structural view used to make these effects visible.

4.1 Structural evaluation: cross-modal utility and encoder-side retention

Cross-modal agreement is only one part of a multimodal representation. Prior work shows that multimodal representations can retain modality-specific information and that alignment need not track downstream utility (Liang et al., 2023; Tjandrasuwita et al., 2025; Qian et al., 2026). This matters for backward allocation: two models can reach similar cross-modal retrieval while leaving different amounts of structure in the individual encoders.

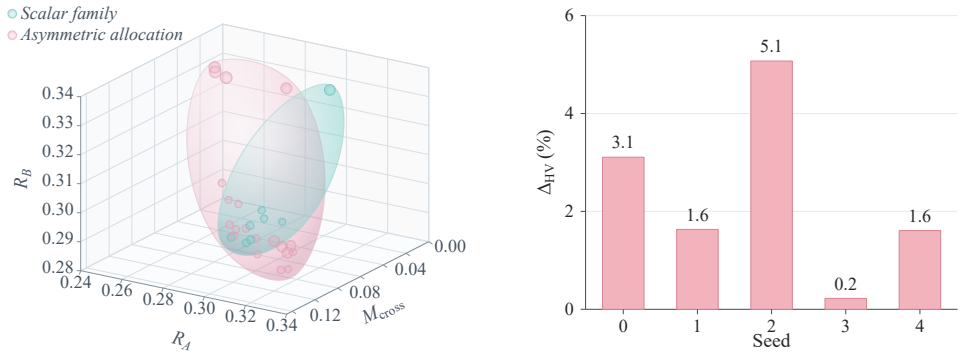
We therefore describe each trained model by $p = (M_{\text{cross}}, R_A, R_B)$, where M_{cross} is bidirectional cross-modal R@5 and R_A, R_B are within-modality retention scores for the two branches. The retention diagnostic asks whether a clean representation can still retrieve its degraded counterpart under fixed corruptions; it is an observable robustness measure rather than an information-decomposition estimate. Similar cross-modal utility therefore need not imply the same encoder-side outcome.

To compare families of solutions rather than a single operating point, we summarize their coverage in this three-dimensional space by relative hypervolume expansion,

$$\Delta_{\text{HV}} = 100\,\frac{\text{HV}(S \cup A) - \text{HV}(S)}{\text{HV}(S)}, \tag{5}$$

where S denotes the tested scalar family and A the asymmetric family. A positive value means that allocation reaches trade-offs not already covered by the scalar set. The exact retention corruptions, grids, reference point, and training protocol are deferred to Appendix D.

4.2 Beyond dense scalar tuning



(a) Outcome space.

(b) Seed-wise hypervolume expansion.

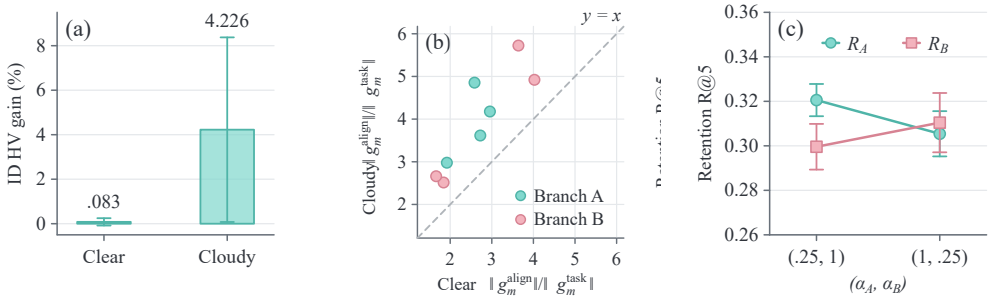
Figure 3: **Beyond dense scalar tuning.** (a) Seed 2 outcome space. (b) Seed-wise ID hypervolume expansion. Asymmetric allocation reaches a larger set of outcome trade-offs than the scalar family.

Does the off-diagonal degree of freedom survive once scalar strength is tuned aggressively? We test this on cloudy SEN12 (Schmitt et al., 2019), comparing a 44-weight scalar sweep with matched off-diagonal allocations over five seeds. Figure 3 gives two views of the same comparison. In panel (a), the seed 2 outcomes overlap over much of the space, but asymmetric allocations extend beyond the scalar family along encoder-side retention trade-offs. Panel (b) shows that this is not specific to seed 2: ID hypervolume expansion is positive in all five matched seeds. The full search grid is given in Appendix D.

Across the five seeds, ID hypervolume increases by 0.223–5.072%, with a mean of $2.329\% \pm 1.842\%$. Mean expansion is also positive on validation (3.498%) and OOD (2.897%). A tolerance-based coverage view gives the same qualitative result: even when scalar points are allowed to be slightly worse on every coordinate, part of the asymmetric family remains outside scalar coverage.

Dense scalar tuning substantially strengthens the frontier, but it does not absorb every observed off-diagonal trade-off. Isotropic controls recover their scalar counterparts (Appendix A), tying the residual to independent allocation rather than to another symmetric scale.

4.3 When does allocation matter more?



(a) Clear–Cloudy HV. (b) Matched pressure shift. (c) Reversed-direction retention.

Figure 4: **Condition dependence and direction.** Cloudy sensing amplifies both the residual beyond scalar tuning and the relative alignment pressure, while reversing allocation redistributes encoder-side retention at similar cross-modal utility.

The residual beyond scalar tuning is not equally pronounced across sensing conditions. We compare matched Clear and Cloudy SEN12 runs using the same scalar and allocation families. As shown in Figure 4a, the scalar family already covers nearly all of the observed outcome space under Clear sensing, where the mean ID hypervolume expansion is only 0.083%. Under Cloudy sensing, the expansion rises to 4.226%, leaving substantially more room for off-diagonal allocation.

Figure 4b provides an optimization-side view of this change. For each seed and encoder branch, we compare the unweighted alignment-to-task gradient norm ratio under the matched Clear and Cloudy runs. Every point lies above the $y = x$ line: the ratio increases from Clear to Cloudy for both branches in all four seeds. Averaged across seeds, it rises from 2.55 to 3.91 for branch A, and from 2.79 to 3.96 for branch B.

These observations suggest a simple interpretation. When the shared alignment signal becomes large relative to the branch-specific learning signal, forcing both encoders to absorb it with the same coefficient becomes more restrictive. Independent admission is therefore most useful when heterogeneous sensing produces stronger relative alignment pressure on the encoder updates. The matched comparison does not establish gradient pressure as the sole cause of the larger frontier, but it identifies an optimization regime in which the additional allocation degree of freedom becomes consequential.

4.4 What does allocation direction change?

A larger allocation space is useful only if different directions lead to meaningfully different encoder outcomes. We therefore reverse the two admission coefficients while keeping the overall alignment setting fixed. At $\lambda = 10$, $(\alpha_A, \alpha_B) = (.25, 1)$ and $(1, .25)$ reach almost identical mean cross-modal R@5, 0.0983 and 0.0985, respectively. Their encoder-side behavior, however, is different.

Figure 4c shows that reversing the admission direction shifts retention in opposite directions. Moving from $(.25, 1)$ to $(1, .25)$ changes (R_A, R_B) from $(.3206, .2996)$ to $(.3054, .3104)$: the branch receiving less alignment influence retains more of its own modality structure, while the other branch moves in the opposite direction. Thus, two models can reach essentially the same cross-modal utility while distributing the representational cost of alignment differently across the encoders.

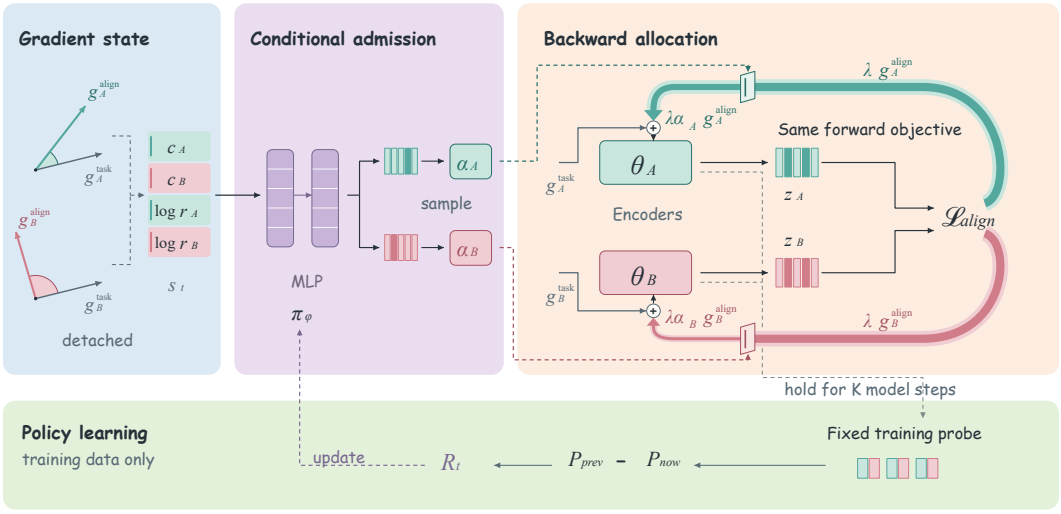


Figure 5: **Conditional backward allocation.** Gradient interaction forms a detached state, a policy samples encoder-specific admissions, and the same forward alignment objective receives separately gated backward paths. The bottom loop uses training data only.

5 Learning Conditional Backward Allocation

Section 4 shows that useful allocation is condition dependent: stronger alignment pressure makes off-diagonal control more consequential, while reversing the direction can redistribute encoder-side retention at nearly unchanged cross-modal utility. Rather than replace scalar coupling with a fixed asymmetric rule, we therefore condition the allocation on the current task–alignment interaction. Figure 5 gives the full loop: detached gradient statistics form the state, a policy predicts two admissions, the admissions gate only the two backward paths of the unchanged alignment objective, and training-only progress updates the policy.

5.1 Conditional admission from gradient state

Gradient magnitude and direction are standard signals for adaptive optimization. GradNorm uses gradient magnitudes to diagnose objective imbalance, while PCGrad and CAGrad reason about directional conflict (Chen et al., 2018; Yu et al., 2020; Liu et al., 2021); multimodal methods such as OGM-GE similarly adapt optimization to unequal modality learning (Peng et al., 2022). We use these signals only as *state*: branch gradients are neither projected nor directly rescaled by the controller.

For branch $m \in \{A, B\}$, let g_m^{task} and g_m^{align} denote the branch and unweighted alignment gradients. The left panel of Figure 5 summarizes their interaction by

$$c_m = \frac{\langle g_m^{\text{task}}, g_m^{\text{align}} \rangle}{\|g_m^{\text{task}}\| \|g_m^{\text{align}}\| + \epsilon}, \qquad r_m = \frac{\|g_m^{\text{align}}\|}{\|g_m^{\text{task}}\| + \epsilon}. \tag{6}$$

Here c_m measures local directional compatibility and r_m measures alignment pressure relative to branch-specific learning. Because either statistic alone is ambiguous, we combine both branches in the detached state

$$\hat{s}_t = [c_A, c_B, \log \max(r_A, 10^{-8}), \log \max(r_B, 10^{-8})], \qquad s_t = 0.9 s_{t-1} + 0.1 \hat{s}_t, \tag{7}$$

updated every $K = 10$ model steps. State construction is detached from the model update.

The center panel maps s_t through a small multilayer perceptron (Rumelhart et al., 1986). Our main policy uses two $\tanh$ hidden layers of widths 64 and 32 and two categorical heads over $\mathcal{A}_{10} = \{0, 1/9, \ldots, 1\}$. We sample α_A and α_B independently and hold them for the next decision interval, so neither modality is assigned a permanent teacher role.

The right panel shows the intervention itself. The representations are evaluated by the same forward $\mathcal{L}_{\text{align}}$; only its encoder-side derivatives are gated:

$$\tilde{g}_A = g_A^{\text{task}} + \lambda \alpha_A g_A^{\text{align}}, \qquad \tilde{g}_B = g_B^{\text{task}} + \lambda \alpha_B g_B^{\text{align}}. \tag{8}$$

The task gradients bypass the gate. Thus $\alpha_A = \alpha_B$ remains on the scalar-equivalent diagonal of Section 3, whereas $\alpha_A \neq \alpha_B$ uses the additional allocation degree of freedom. Training begins with 20 symmetric steps, $(\alpha_A, \alpha_B) = (1, 1)$, before policy sampling.

5.2 Learning admission from training progress

There is no target label for the desired (α_A, α_B). The lower panel of Figure 5 therefore learns from what happens after an allocation has acted for one interval. To make successive evaluations comparable, we use a fixed probe drawn entirely from the training split: 96 training pairs are selected once and six views per pair are cached. Probe evaluation reuses these tensors without gradients and computes

$$P_t = \mathcal{L}_A^{\text{probe}}(\theta_t) + \mathcal{L}_B^{\text{probe}}(\theta_t) + \lambda \mathcal{L}_{\text{align}}^{\text{probe}}(\theta_t), \qquad R_t = P_{t-1} - P_t. \tag{9}$$

Positive R_t means that the probe objective decreased while the previous admissions were active. We use it as a credit signal, not as a causal estimate of the action effect, because the model state and stochastic optimization also evolve across the interval.

We update the policy with a score-function estimator in the spirit of REINFORCE (Williams, 1992). With $b_t = 0.9 b_{t-1} + 0.1 R_t$ and $A_t = \text{clip}(10(R_t - b_t), -1, 1)$,

$$\mathcal{L}_\pi = -A_t \sum_{m \in \{A,B\}} \log \pi_{\phi,m}(\alpha_{m,t-1} \mid s_{t-1}) - 0.01 \sum_{m \in \{A,B\}} H(\pi_{\phi,m}(\cdot \mid s_{t-1})). \tag{10}$$

The entropy term discourages premature collapse. Model and policy use separate optimizers, the policy does not backpropagate through the trajectory that produced R_t, and validation or ID data never enter the controller update. Exact update order, optimizer settings, and simpler linear and deterministic controllers are given in Appendices E.2 and E.

5.3 Full-scale evaluation of conditional allocation

Table 1: **ViT-B/16 retrieval.** Bidirectional R@5 (%, mean ± 95% CI over five seeds).

Method	DSEC		Eilat		Sunboat	
	Val	ID	Val	ID	Val	ID
No alignment	0.63 ± 0.37	0.65 ± 0.30	0.18 ± 0.09	0.13 ± 0.09	1.70 ± 0.55	1.31 ± 0.79
Scalar	7.88 ± 1.58	3.91 ± 0.70	**2.01** ± 0.46	0.97 ± 0.28	3.30 ± 1.39	1.38 ± 0.61
Gradient Blending	0.97 ± 0.53	0.69 ± 0.10	0.18 ± 0.09	0.11 ± 0.06	1.34 ± 0.94	1.57 ± 0.76
OGM-GE	7.82 ± 1.80	3.49 ± 0.88	1.87 ± 0.27	0.90 ± 0.15	2.30 ± 0.99	1.64 ± 0.75
MMPareto	3.15 ± 1.18	1.86 ± 0.75	0.29 ± 0.09	0.18 ± 0.13	1.34 ± 0.77	1.33 ± 0.90
ReconBoost	6.57 ± 1.07	2.88 ± 0.57	1.43 ± 0.23	0.75 ± 0.21	3.25 ± 0.57	1.64 ± 1.23
DGL	5.69 ± 2.37	2.03 ± 0.25	0.51 ± 0.13	0.34 ± 0.10	1.65 ± 0.82	1.15 ± 0.42
Ours	**7.92** ± 1.56	**4.60** ± 1.19	1.88 ± 0.30	**0.99** ± 0.31	**3.40** ± 0.94	**1.67** ± 0.74

Section 4 establishes that off-diagonal allocation can matter; here we ask whether the conditional controller can turn that degree of freedom into a useful training rule. We evaluate ViT-B/16 (Dosovitskiy et al., 2021) on DSEC RGB–event (Gehrig et al., 2021), Eilat RGB–sonar (Treibitz et al., 2024a), and Sunboat RGB–sonar (Treibitz et al., 2024b). Scalar and Ours each use five seeds at $\lambda \in \{2.5, 5, 10\}$, selecting one λ^* per dataset by mean validation R@5 before ID evaluation; No alignment is the $\lambda = 0$ reference. We additionally compare with Gradient Blending (Wang et al., 2020), OGM-GE (Peng et al., 2022), MMPareto (Wei & Hu, 2024), ReconBoost (Hua et al., 2024), and DGL (Wei et al., 2025b). These provide two complementary tests: specialized multimodal optimizers ask whether conditional allocation is competitive with methods designed for imbalance or gradient interaction, while validation-tuned Scalar tests whether allocation adds useful control beyond choosing alignment

strength. All methods share the backbone, data, training budget, and evaluator; exact adaptations are in Appendices B and C.

Against the five specialized baselines, Ours has the highest mean validation and ID R@5 on all three datasets. Against validation-tuned Scalar, the pattern is narrower: validation favors Ours on DSEC and Sunboat but Scalar on Eilat, whereas held-out ID changes from $3.91 \rightarrow 4.60$, $0.97 \rightarrow 0.99$, and $1.38 \rightarrow 1.67$ on DSEC, Eilat, and Sunboat, respectively, or +0.69, +0.02, and +0.29 percentage points. Paired 95% intervals in Appendix F all include zero, so these results support a consistent mean-level pattern rather than a claim of statistically significant superiority.

The policy also uses the control dimension introduced in Section 3 rather than collapsing to an implicit scalar rule. Mean logged $|\alpha_A - \alpha_B|$ is 0.333, 0.346, and 0.332 on the three datasets, confirming sustained off-diagonal actions; this gap establishes use of the action space, not that asymmetry alone causes the retrieval gains. At a common $\lambda = 5$, linear and deterministic rule-based controllers remain competitive and can outperform the MLP in individual settings (Appendix E). Together with Section 4, this supports the interpretation that conditional backward allocation is the useful mechanism, while the MLP is one practical realization.

6 Discussion

The main implication of our results is not that scalar alignment should be replaced by asymmetric control. Rather, alignment strength and backward allocation describe two different aspects of the same optimization process. Scalar weighting determines how much shared alignment enters training, whereas allocation determines how that influence is distributed across encoders. When the two admissions are equal, the resulting update remains on the scalar-equivalent diagonal; the additional degree of freedom appears only when the encoder-side influences differ. This distinction explains why stronger scalar tuning can recover part, but not necessarily all, of the behavior exposed by allocation.

Our empirical results further suggest when this additional control becomes useful. Under matched sensing, dense scalar tuning already covers nearly all of the observed outcome space. Under heterogeneous sensing, the residual outside the scalar family becomes substantially larger, together with stronger alignment-to-task gradient pressure. Reversing the allocation direction can also redistribute encoder-side retention while leaving cross-modal utility nearly unchanged. These observations support a view in which allocation matters when the shared alignment signal interacts differently with the two branch-specific learning processes. They also explain why a fixed asymmetric rule is insufficient: the useful direction depends on the current optimization state rather than on a permanently designated teacher modality.

This conditionality is itself part of the result. Backward allocation is not universally preferable, nor is asymmetry an objective in its own right. Instead, it provides an additional control axis for regimes in which isotropic coupling becomes restrictive. The learned controller operationalizes this view by selecting admissions from current gradient interaction, while the competitive linear and rule-based alternatives indicate that the underlying benefit is not tied to a particular policy architecture. A natural next step is therefore to understand more precisely which optimization states predict not only whether off-diagonal control is useful, but also which allocation should be chosen.

7 Conclusion

Multimodal alignment is usually controlled by the strength of a shared objective. This paper shows that a second question remains: where should its backward influence be absorbed? Backward influence allocation exposes this encoder-specific degree of freedom while leaving the forward alignment objective unchanged. We characterize scalar alignment as its isotropic special case and show that the value of off-diagonal control depends on the learning condition. A lightweight training-time controller makes this additional degree of freedom usable online. Together, these results establish backward allocation as a distinct, condition-dependent optimization variable for heterogeneous multimodal learning.

REFERENCES

Galen Andrew, Raman Arora, Jeff Bilmes, and Karen Livescu. Deep canonical correlation analysis. In *Proceedings of the 30th International Conference on Machine Learning*, volume 28 of *Proceedings of Machine Learning Research*, pp. 1247–1255. PMLR, 2013.

Relja Arandjelović and Andrew Zisserman. Look, listen and learn. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 609–617, 2017. doi: 10.1109/ICCV.2017.73.

Yusuf Aytar, Carl Vondrick, and Antonio Torralba. SoundNet: Learning sound representations from unlabeled video. In *Advances in Neural Information Processing Systems*, volume 29, 2016.

Roman Bachmann, Oğuzhan Fatih Kar, David Mizrahi, Ali Garjani, Mingfei Gao, David Griffiths, Jiaming Hu, Afshin Dehghan, and Amir Zamir. 4M-21: An any-to-any vision model for tens of tasks and modalities. In *Advances in Neural Information Processing Systems*, volume 37, 2024.

Tadas Baltrušaitis, Chaitanya Ahuja, and Louis-Philippe Morency. Multimodal machine learning: A survey and taxonomy. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41(2): 423–443, 2019. doi: 10.1109/TPAMI.2018.2798607.

Xinlei Chen and Kaiming He. Exploring simple siamese representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 15750–15758, 2021. doi: 10.1109/CVPR46437.2021.01549.

Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. UNITER: Universal image-text representation learning. In *European Conference on Computer Vision*, pp. 104–120. Springer, 2020.

Zhao Chen, Vijay Badrinarayanan, Chen-Yu Lee, and Andrew Rabinovich. GradNorm: Gradient normalization for adaptive loss balancing in deep multitask networks. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 794–803. PMLR, 2018.

Zijie Chen, Kejing Yin, Wenfang Yao, William Kwok-wai Cheung, and Jing Qin. MASAM: Multimodal adaptive sharpness-aware minimization for heterogeneous data fusion. In *International Conference on Learning Representations*, 2026.

Giordano Cicchetti, Eleonora Grassucci, Luigi Sigillo, and Danilo Comminiello. Gramian multimodal representation learning and alignment. In *International Conference on Learning Representations*, 2025.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2021.

Benoit Dufumier, Javiera Castillo Navarro, Devis Tuia, and Jean-Philippe Thiran. What to align in multimodal contrastive learning? In *International Conference on Learning Representations*, 2025.

Patrick Ebel, Andrea Meraner, Michael Schmitt, and Xiao Xiang Zhu. Multi-sensor data fusion for cloud removal in global and all-season sentinel-2 imagery. *IEEE Transactions on Geoscience and Remote Sensing*, 59(7):5866–5878, 2021. doi: 10.1109/TGRS.2020.3024744.

Sedigheh Eslami and Gerard de Melo. Mitigate the gap: Improving cross-modal alignment in CLIP. In *International Conference on Learning Representations*, 2025.

Yunfeng Fan, Wenchao Xu, Haozhao Wang, Junxiao Wang, and Song Guo. PMR: Prototypical modal rebalance for multimodal learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 20029–20038, 2023.

Mathias Gehrig, Willem Aarents, Daniel Gehrig, and Davide Scaramuzza. DSEC: A stereo event camera dataset for driving scenarios. *IEEE Robotics and Automation Letters*, 6(3):4947–4954, 2021. doi: 10.1109/LRA.2021.3068942.

Rohit Girdhar, Alaaeldin El-Nouby, Zhuang Liu, Mannat Singh, Kalyan Vasudev Alwala, Armand Joulin, and Ishan Misra. ImageBind: One embedding space to bind them all. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 15180–15190, 2023. doi: 10.1109/CVPR52729.2023.01457.

Eleonora Grassucci, Giordano Cicchetti, Emanuele Frasca, Aurelio Uncini, and Danilo Comminiello. Closing the modality gap aligns group-wise semantics. In *International Conference on Learning Representations*, 2026.

Zirun Guo, Tao Jin, Jingyuan Chen, and Zhou Zhao. Classifier-guided gradient modulation for enhanced multimodal learning. In *Advances in Neural Information Processing Systems*, volume 37, 2024.

Jitai Hao, Hao Liu, Xinyan Xiao, Qiang Huang, and Jun Yu. Uni-X: Mitigating modality conflict with a two-end-separated architecture for unified multimodal models. In *International Conference on Learning Representations*, 2026.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 770–778, 2016.

Cong Hua, Qianqian Xu, Shilong Bao, Zhiyong Yang, and Qingming Huang. ReconBoost: Boosting can achieve modality reconcilement. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 19573–19597. PMLR, 2024.

Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc V. Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 4904–4916. PMLR, 2021.

Paul Pu Liang, Zihao Deng, Martin Q. Ma, James Zou, Louis-Philippe Morency, and Ruslan Salakhutdinov. Factorized contrastive learning: Going beyond multi-view redundancy. In *Advances in Neural Information Processing Systems*, volume 36, 2023. doi: 10.52202/075280-1430.

Bo Liu, Xingchao Liu, Xiaojie Jin, Peter Stone, and Qiang Liu. Conflict-averse gradient descent for multi-task learning. In *Advances in Neural Information Processing Systems*, volume 34, 2021.

Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. ViLBERT: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

Xiaoyu Ma, Hao Chen, and Yongjian Deng. Improving multimodal learning balance and sufficiency through data remixing. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 41827–41836. PMLR, 2025.

Jiquan Ngiam, Aditya Khosla, Mingyu Kim, Juhan Nam, Honglak Lee, and Andrew Y. Ng. Multimodal deep learning. In *Proceedings of the 28th International Conference on Machine Learning*, pp. 689–696, 2011.

Andrew Owens and Alexei A. Efros. Audio-visual scene analysis with self-supervised multisensory features. In *Proceedings of the European Conference on Computer Vision*, pp. 631–648, 2018.

Xiaokang Peng, Yake Wei, Andong Deng, Dong Wang, and Di Hu. Balanced multimodal learning via on-the-fly gradient modulation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8238–8247, 2022.

Chengxuan Qian, Shuo Xing, Shawn Li, Yue Zhao, and Zhengzhong Tu. DecAlign: Hierarchical cross-modal alignment for decoupled multimodal representation learning. In *International Conference on Learning Representations*, 2026.

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 of *Proceedings of Machine Learning Research*, pp. 8748–8763. PMLR, 2021.

Dhanesh Ramachandram and Graham W. Taylor. Deep multimodal learning: A survey on recent advances and trends. *IEEE Signal Processing Magazine*, 34(6):96–108, 2017. doi: 10.1109/MSP.2017.2738401.

Karsten Roth, Zeynep Akata, Dima Damen, Ivana Balazevic, and Olivier J. Henaff. Context-aware multimodal pretraining. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4267–4279, 2025.

David E. Rumelhart, Geoffrey E. Hinton, and Ronald J. Williams. Learning representations by back-propagating errors. *Nature*, 323:533–536, 1986. doi: 10.1038/323533a0.

Michael Schmitt, Lloyd Haydn Hughes, Chunping Qiu, and Xiao Xiang Zhu. SEN12MS: A curated dataset of georeferenced multi-spectral sentinel-1/2 imagery for deep learning and data fusion. *ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, IV-2/W7: 153–160, 2019. doi: 10.5194/isprs-annals-IV-2-W7-153-2019.

Amanpreet Singh, Ronghang Hu, Vedanuj Goswami, Guillaume Couairon, Wojciech Galuba, Marcus Rohrbach, and Douwe Kiela. FLAVA: A foundational language and vision alignment model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 15638–15650, 2022.

Yonglong Tian, Dilip Krishnan, and Phillip Isola. Contrastive multiview coding. In *Computer Vision – ECCV 2020*, pp. 776–794. Springer, 2020. doi: 10.1007/978-3-030-58621-8_45.

Megan Tjandrasuwita, Chanakya Ekbote, Liu Ziyin, and Paul Pu Liang. Understanding the emergence of multimodal representation alignment. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 59723–59760. PMLR, 2025.

Tali Treibitz, Izhak Fabian, Amir Dayan, Nir Zagdanski, Deborah Levy, Peal Naama, Amit Peleg, Opher Bat Nathan, Ohad Inbar, and Yevgeni Gutnik. ROV-based multi-sensor dataset: Synchronized camera and sonar images taken in the tropical waters of the red sea, eilat. Zenodo, 2024a.

Tali Treibitz, Yevgeni Gutnik, Izhak Fabian, Amir Dayan, Nir Zagdanski, Opher Bar Nathan, Assaf Zvuloni, and Merav Keidar. AUV-based multi-sensor dataset: Forward-looking camera (FLC) and forward-looking sonar (FLS) observations in the red sea. Zenodo, 2024b.

Weiyao Wang, Du Tran, and Matt Feiszli. What makes training multi-modal classification networks hard? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 12695–12705, 2020.

Shicai Wei, Chunbo Luo, and Yang Luo. Improving multimodal learning via imbalanced learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 2250–2259, 2025a.

Shicai Wei, Chunbo Luo, and Yang Luo. Boosting multimodal learning via disentangled gradient learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 22879–22888, 2025b.

Yake Wei and Di Hu. MMPareto: Boosting multimodal learning with innocent unimodal assistance. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 52559–52572. PMLR, 2024.

Ronald J. Williams. Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine Learning*, 8:229–256, 1992. doi: 10.1007/BF00992696.

Nan Wu, Stanislaw Jastrzebski, Kyunghyun Cho, and Krzysztof J. Geras. Characterizing and overcoming the greedy nature of learning in multi-modal deep neural networks. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 24043–24055. PMLR, 2022.

Wei Wu, Xiaomeng Fan, Yuwei Wu, Zhi Gao, Pengxiang Li, Yunde Jia, and Mehrtash Harandi. Modality alignment across trees on heterogeneous hyperbolic manifolds. In *International Conference on Learning Representations*, 2026.

Yang Yang, Fengqiang Wan, Qing-Yuan Jiang, and Yi Xu. Facilitating multimodal classification via dynamically learning modality gap. In *Advances in Neural Information Processing Systems*, volume 37, pp. 62108–62122, 2024. doi: 10.52202/079017-1984.

Tianhe Yu, Saurabh Kumar, Abhishek Gupta, Sergey Levine, Karol Hausman, and Chelsea Finn. Gradient surgery for multi-task learning. In *Advances in Neural Information Processing Systems*, volume 33, 2020.

Xiaohui Zhang, Jaehong Yoon, Mohit Bansal, and Huaxiu Yao. Multimodal representation learning by alternating unimodal adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 27456–27466, 2024.

Yiyuan Zhang, Handong Li, Jing Liu, and Xiangyu Yue. Scaling omni-modal pretraining with multimodal context: Advancing universal representation learning across modalities. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 1336–1348, 2025.

AI USE STATEMENT

Generative AI tools were used in a limited capacity to aid or polish writing and to support literature retrieval and discovery, including locating and reading relevant related work.

A GRADIENT IDENTITIES AND NUMERICAL CONTROLS

Scope of the parameter blocks. We assume disjoint branch parameters θ_A, θ_B, including their projection heads. The branch loss $\mathcal{L}_m$ depends only on θ_m, whereas the alignment term depends on both. The admission coefficient is an externally supplied scalar during each model backward pass. In the learned version, state features are detached and categorical actions are sampled; the policy is trained separately.

Proof of Proposition 1. The chain rule applied to Eq. 4 gives

$$\nabla_{\theta_m} \mathcal{L}_{\text{align}}(\tilde{z}_A, \tilde{z}_B) = \alpha_m \nabla_{\theta_m} \mathcal{L}_{\text{align}}(z_A, z_B). \tag{11}$$

When $\alpha_A = \alpha_B = c$, both branch gradients are $g_m^{\text{task}} + c\lambda g_m^{\text{align}}$. These equal the gradients supplied by ordinary scalar alignment at weight $c\lambda$. Equal gradients and equal optimizer states give equal updates for the same update rule, including clipping, momentum, and weight decay. With matched data and stochastic operations, the argument applies inductively to the full trajectory in exact arithmetic. Finite precision can introduce small discrepancies because the two computational graphs perform multiplication in different orders.

Proof of Proposition 2. Equality on branch A requires $(\lambda\alpha_A - \lambda')g_A^{\text{align}} = 0$. A nonzero alignment gradient implies $\lambda' = \lambda\alpha_A$. The same argument on B requires $\lambda' = \lambda\alpha_B$. For $\lambda > 0$ and unequal admissions, these conditions cannot both hold. If $\lambda = 0$ or an alignment gradient vanishes, the corresponding constraint is absent. The result is about pre-optimizer gradients: clipping or adaptive normalization can map some different gradients to the same parameter update.

Relation to learning rates and directional losses. A branch-specific learning-rate multiplier q_m gives $q_m g_m^{\text{task}} + q_m \lambda g_m^{\text{align}}$ before any adaptive optimizer transformation. Admission leaves the coefficient of g_m^{task} at one. Likewise, the two directional cross-entropy terms in symmetric InfoNCE each depend on both encoders, so weighting retrieval directions does not generally reproduce independent branch derivatives. These comparisons establish what the intervention controls; they do not imply that alternative optimizers cannot reach similar final representations.

Table 2: **Recorded isotropic-equivalence controls.** Cloudy SEN12, four matched seeds. Relative distance is $\|\theta_{\text{admission}} - \theta_{\text{scalar}}\|_2 / \|\theta_{\text{scalar}}\|_2$ over the stored model-state tensors. The final row reports the mean and range across seeds.

Admission at $\lambda = 5$	Scalar weight	Relative distance
$(0, 0)$	0	0 in every seed
$(.25, .25)$	1.25	0 in every seed
$(.5, .5)$	2.5	0 in every seed
$(.75, .75)$	3.75	3.343×10^{-4} $[3.203, 3.620] \times 10^{-4}$

The control records are consistent with the exact identity, with a small relative discrepancy in the $c = .75$ BF16 computation. During structural training, the implementation additionally checks the routed and unrouted forward losses at the same parameter state, with absolute tolerance 10^{-7}.

B DATA, OBJECTIVES, AND IMPLEMENTATION

B.1 FINAL VIT RETRIEVAL SPLITS

DSEC. RGB images are paired with the prepared event-frame images by instance identifier. Training uses sequences `zurich_city_00_a` and `zurich_city_03_a`; validation uses `zurich_city_02_a` and `zurich_city_04_b`; ID uses `zurich_city_01_a`. The OOD sequence is `interlaken_00_c`. These sequence sets are disjoint. The same 3,544-pair manifest is used in the controller search and final DSEC evaluation.

Eilat. Training, validation, and ID partition frames within three acquisition sequences: `20221211_092506`, `20221211_133252`, and `20221212_095821`. OOD comprises

Table 3: **Pair counts and full retrieval galleries.** Each evaluation row is one paired instance; all pairs in that split form the gallery in either retrieval direction. SEN12 counts apply separately to Clear and Cloudy in the structural study.

Dataset	Train	Validation	ID	OOD
DSEC RGB–event	1,822	504	681	537
Eilat RGB–sonar	22,091	3,156	3,158	19,746
Sunboat RGB–sonar	2,679	382	383	5,010
SEN12 SAR–optical	4,200	600	600	600

`20221212_141308` and `20221213_102542`. The first three sequences contribute 8,170/1,167/1,168, 6,261/895/895, and 7,660/1,094/1,095 train/validation/ID pairs, respectively. The OOD sequences contain 9,458 and 10,288 pairs.

Sunboat. Training, validation, and ID partition the acquisition at `2023-09-03-07-58-37`; OOD contains the 5,010 pairs from `2023-09-03-11-37-57`. Thus the sonar ID evaluations measure within-sequence generalization and OOD measures transfer to held-out sequences. Across all three final manifests, pair identifiers and exact image-path pairs are unique across splits. The sonar within-sequence partitions do not imply independence of neighboring frames.

B.2 Contrastive objectives and retrieval

Let $U, V \in \mathbb{R}^{N\times d}$ contain normalized paired embeddings. With temperature $\tau = .07$, define

$$\ell(U,V) = -\frac{1}{2N}\sum_{i=1}^{N}\left[\log\frac{\exp(u_i^\top v_i/\tau)}{\sum_j \exp(u_i^\top v_j/\tau)} + \log\frac{\exp(v_i^\top u_i/\tau)}{\sum_j \exp(v_i^\top u_j/\tau)}\right]. \tag{12}$$

For the final ViT study, $\mathcal{L}_A = \ell(Z_A^{(1)}, Z_A^{(2)})$ and $\mathcal{L}_B = \ell(Z_B^{(1)}, Z_B^{(2)})$ use independent within-modality augmentations, while $\mathcal{L}_{\text{align}} = \ell(Z_A^c, Z_B^c)$ uses deterministic paired views. The structural trainer divides each contrastive loss by $\max(\log N, 1)$; its batch size is fixed across scalar and admission configurations. The final ViT trainer uses the unscaled definition above. Both protocols use in-batch negatives.

For evaluation, cosine similarities rank the paired index among all N candidates. The reported bidirectional recall is

$$\text{R@}\,k = \frac{1}{2N}\sum_{i=1}^{N}\left[\mathbb{1}\{i \in \text{Top}_k(u_i V^\top)\} + \mathbb{1}\{i \in \text{Top}_k(v_i U^\top)\}\right]. \tag{13}$$

The exact top-k calculation is chunked over queries (chunk size 1,024); every query still compares with the entire opposite-modality gallery. There is no approximate nearest-neighbor search or negative subsampling. Gallery size therefore affects recall: an uninformative ranking has expected R@5 of $5/N$ for $N \geq 5$. For final ID galleries, these references are .00734 (DSEC), .00158 (Eilat), and .01305 (Sunboat). Method comparisons use the same gallery within a dataset and split.

B.3 Final ViT training details

We use two independent torchvision ViT-B/16 branches initialized with ImageNet-1K V1 weights, replace the classification heads by identity mappings, and attach a linear $768 \to 256$ projection followed by LayerNorm and normalization. All branch and projection parameters are trainable. Each prepared sensor image is loaded as a three-channel image. A training view uses RandomResizedCrop to 224×224, scale $[.75, 1]$, aspect ratio $[.9, 1.1]$, horizontal flip, and brightness/contrast jitter .15. Deterministic views resize the shorter side to 256 and center-crop to 224. Both use ImageNet mean and standard deviation. Augmentations are sampled independently within and across modalities; the deterministic cross-modal views preserve pairing.

AdamW uses learning rate and weight decay 10^{-4}, default momentum coefficients, and global gradient-norm clipping at 5. The final ViT runs use a constant learning rate, BF16 autocasting, and

batch size 32 with a potentially smaller last batch. Data are shuffled with the run seed. DSEC and Sunboat use 600 updates; Eilat uses 700. This corresponds to about 10.5, 7.1, and 1.0 training-set passes, respectively, so the evaluation is explicitly a fixed-budget comparison. The final model checkpoint supplies the reported retrieval values.

Selection and run counts. For each dataset, the final matrix contains 15 MLP runs, 15 scalar runs, 25 runs for five optimization adaptations, five alignment-off runs, and six Linear/Rule runs, totaling 66. Across three datasets this gives 198 ViT runs. The five-seed families use seeds 1–5 and the simpler policies use seeds 1–3. One λ is selected per dataset and per MLP/scalar family using the mean validation R@5 of its five seeds. Test results do not select the weight. Architecture selection preceded this matrix and is described in Appendix E.

Compute. The MLP adds 3,060 trainable parameters during training and none during inference. Its extra work consists mainly of task/alignment gradient-statistic calculations every ten steps and a cached-probe evaluation at policy decisions. Table 10 reports recorded elapsed durations, including final evaluation; they are not isolated measurements of controller overhead. In particular, gallery size and shared-machine scheduling affect runtime, and Gradient Blending's offline estimation occurs before its formal-training timer.

C Retrieval adaptations of optimization baselines

All five adaptations use the same two-branch backbone, projection space, data, training budget, and retrieval evaluator as the final MLP/scalar study. Their alignment weight is fixed at 5. The following specifies the implemented retrieval rules; the results compare these rules under a common task rather than reproducing each method's original supervised benchmark.

Gradient Blending. An offline procedure estimates nonnegative weights for three signals: A-self, B-self, and cross-modal contrastive learning. It uses deterministic subsets of up to 1,024 training and validation pairs and five estimation epochs per signal. For signal s, define the overfitting gap $O = \ell_{\text{val}} - \ell_{\text{train}}$ and validation improvement $G = \ell_{\text{val},0} - \ell_{\text{val},1}$. The raw weight is $\max(0, G)/((O_1 - O_0)^2 + 10^{-8})$, normalized across signals, with a uniform fallback if their sum is zero. The estimation self-signals use $\ell(Z_A, Z_A)$ and $\ell(Z_B, Z_B)$ on deterministic views; the formal training loss is $w_A\mathcal{L}_A + w_B\mathcal{L}_B + w_F\lambda\mathcal{L}_{\text{align}}$. The initial model and optimizer state are restored before formal training. Validation participates only in this baseline's offline weight estimation and in the stated selection procedures; ID/OOD do not.

OGM-GE. The unmodified total loss is differentiated first. Branch strengths are $s_A = \exp(-\mathcal{L}_A)$ and $s_B = \exp(-\mathcal{L}_B)$. The stronger branch receives coefficient $1 - \tanh(.3\, s_{\text{strong}}/s_{\text{weak}})$; the other receives one. For ViT parameter tensors of dimension at least two, the implementation scales the gradient and adds Gaussian noise whose standard deviation is the original gradient's standard deviation plus 10^{-8}. Lower-dimensional tensors are unchanged by this modulation; the common global clipping operation follows.

MMPareto. For each branch, let g_1 be its task gradient and g_2 the λ-weighted alignment gradient. The minimum-norm convex coefficient is

$$w = \text{clip}\left(\frac{\|g_2\|^2 - \langle g_1, g_2\rangle}{\|g_1\|^2 + \|g_2\|^2 - 2\langle g_1, g_2\rangle}, 0, 1\right), \tag{14}$$

with $w = .5$ for a degenerate denominator. The assigned gradient is $3[wg_1 + (1-w)g_2]$, followed by the common clipping and optimizer update.

ReconBoost. A modality-focused stage lasts four data-loader epochs, followed by an equally long joint stage. The selected modality alternates between cycles. A focused A stage optimizes $5\mathcal{L}_A + \lambda\ell(Z_A^c, \text{sg}(Z_B^c))$ and sets all B gradients to absent; the focused B stage is symmetric. A joint stage uses $\mathcal{L}_A + \mathcal{L}_B + \lambda\mathcal{L}_{\text{align}}$. The fixed step budgets can end within a stage, particularly on the larger Eilat training set; this is the schedule used by the reported adaptation.

DGL. The two full branches receive $4(\mathcal{L}_A + \mathcal{L}_B)$. For the alignment term, backbone features are detached before the projections, so $\lambda\mathcal{L}_{\text{align}}$ updates only the projection heads. This separates the backbone and cross-modal projection updates, whereas our admission scales the alignment derivative through the entire branch.

D STRUCTURAL METRICS AND ADDITIONAL DIAGNOSTICS

D.1 REPRESENTATION AND CORRUPTION PROTOCOL

The main structural study uses separate ImageNet-initialized ResNet-18 encoders (He et al., 2016) with $512 \to 256 \to 128$ projection MLPs and a GELU activation. Inputs are resized to 224×224 while preserving the full prepared field of view. Optical arrays use channels corresponding to B4/B3/B2; two-channel SAR arrays are represented by the two channels and their mean. Arrays are mapped to image intensity using the 1st and 99th percentiles. Optical images use ImageNet normalization and SAR uses mean/std .5. The study uses deterministic seed-indexed batches and augmentations, AdamW at 10^{-4} with weight decay 10^{-4}, cosine learning-rate decay, BF16, and gradient clipping at 5.

For each branch, evaluation computes R@5 from a fixed clean/mildly augmented reference view to three degraded views of the same instances. The retention score is

$$R_m = \frac{1}{3}\sum_{\ell=1}^{3} \mathrm{R@5}(Z_m^c \to Z_m^{(\ell)}). \tag{15}$$

The within-modality direction is reference-to-degraded; cross-modal recall averages both directions. Corruption random seeds are tied to instance identifiers, fixing the views across configurations.

Table 4: **Structural evaluation corruptions.** Triples list levels 1/2/3. Brightness and contrast are multiplicative factors; Gaussian noise is added in $[0, 1]$ intensity space and clipped. The final occlusion is a square with side length equal to the listed fraction of the shorter image side.

Operation	RGB/optical	Other sensor images
Brightness	.72 / .48 / .28	.90 / .76 / .62
Contrast	.82 / .58 / .38	.75 / .52 / .34
Blur radius	.60 / 1.40 / 2.70	.70 / 1.55 / 2.80
Noise standard deviation	.035 / .075 / .130	.055 / .105 / .170
Occlusion side fraction	.10 / .20 / .30	.10 / .20 / .30

These scores make encoder-side robustness observable under a controlled perturbation family. They do not estimate information decomposition terms or substitute for downstream supervised performance. In SEN12 records, branch A is SAR and branch B is optical; Figure 1 uses optical/SAR order only as an introductory illustration.

D.2 HYPERVOLUME, GRID COVERAGE, AND SEED-LEVEL RESULTS

All three coordinates of $p = (M_{\text{cross}}, R_A, R_B)$ are recalls in $[0, 1]$ and are maximized. The hypervolume is the volume of the union of boxes $[0, p_1] \times [0, p_2] \times [0, p_3]$ from the origin to each point, with no additional coordinate normalization. Exact axis slicing computes this union. Comparisons are performed separately within each seed and split, then aggregated across seeds. No single test-selected configuration is exported from this family-level diagnostic.

The dense comparison uses scalar weights $\{0, .25, .5, \ldots, 10, 12.5, 15, 20\}$ and asymmetric admissions from $\{0, .25, .5, .75, 1\}^2$ excluding the diagonal, at fixed $\lambda = 5$. Seeds 0–3 combine the original runs with additional dense scalar configurations; seed 4 provides a complete independently seeded grid. Each seed contains 44 scalar and 20 asymmetric configurations, for 320 comparisons in total. The mean ID expansion remains 2.329% after this strengthening. With coordinate tolerances .0025 and .005, Table 6 tests whether small recall differences alone account for frontier escape.

The matched Clear/Cloudy analysis in Figure 4 uses the identical scalar set $\{0, .5, 1.25, 2.5, 3.75, 5, 7.5, 10\}$ and the six non-diagonal points of $\{0, .5, 1\}^2$ at $\lambda = 5$, restricted to

Table 5: **Dense-scalar structural results by seed.** Relative hypervolume expansion (%) for 44 scalar weights plus 20 asymmetric configurations.

Seed	Validation	ID	OOD
0	2.352	3.107	4.464
1	7.089	1.632	2.024
2	3.664	5.072	2.578
3	0.791	0.223	0.765
4	3.594	1.608	4.656
Mean ± s.d.	3.498±2.323	2.329±1.842	2.897±1.655

Table 6: **Frontier coverage with coordinate tolerance.** Mean fraction (%) of asymmetric configurations not covered by a scalar configuration, across five seeds. A scalar point covers an asymmetric point when it is at most ϵ worse on every recall coordinate.

ϵ	Validation	ID	OOD
0	57.0	40.0	51.0
0.0025	45.0	35.0	39.0
0.005	32.0	24.0	25.0

Table 7: **Scalar-grid strengthening.** Relative increase in scalar-only hypervolume (%), averaged over five seeds. Grid comparisons use $\lambda \leq 10$; the tail adds $12.5, 15, 20$. The sparse grid is $\{0, .5, 1.25, 2.5, 3.75, 7.5, 10\}$.

Comparison	Validation	ID	OOD
Sparse to full dense	6.81	9.54	10.84
Spacing .5 to .25	3.08	5.34	4.30
Add high-weight tail	0.92	1.58	0.88

seeds 0–3. It is a reanalysis of existing runs, with no added training. Clear yields mean ID expansion $0.083\% \pm 0.165\%$ and Cloudy $4.226\% \pm 4.149\%$. This comparison isolates the sensing condition more directly than comparing Clear with a differently sized dense search.

Supporting sensor families. Three-seed sweeps use scalar weights $\{0, .5, 1.25, 2.5, 5, 10\}$ and the same six asymmetric settings at $\lambda = 5$. Mean ID expansions (sample s.d.) are $21.14\% \pm 7.64\%$ for Eilat, $24.17\% \pm 16.34\%$ for Sunboat, $64.32\% \pm 35.62\%$ for DSEC, and $.80\% \pm .97\%$ for NYU RGB–depth. These earlier structural galleries differ from the final ViT galleries: ID sizes are 1,000, 383, 24, and 144, respectively. The corresponding validation/OOD sizes are 1,000/1,000, 382/1,000, 24/47, and 144/147. These results describe within-protocol diversity; their percentages are not comparable effect sizes across gallery sizes or scalar grids.

A ViT structural boundary. An existing SEN12 ViT study uses 1,200 steps, batch size 32, seeds 0–2, 44 scalar weights, and four asymmetric settings. Its mean hypervolume expansion is .022% on validation, .182% on ID, and .092% on OOD. This provides an additional case in which dense scalar tuning captures most of the tested fixed-allocation trade-off. It concerns fixed admission on SEN12, distinct from the learned-policy retrieval study on DSEC and sonar data.

Pressure and direction. Figure 4 uses the step-600 probes from four symmetric-$\lambda = 5$ runs per Clear/Cloudy condition. The ratios are computed before multiplication by λ. For the direction comparison, four seeds at $\lambda = 10$ compare $(.25, 1)$ with $(1, .25)$; ID cross-modal recalls average .0983/.0985, and retention averages $(.3206, .2996)$/$(.3054, .3104)$. This is a matched-utility descriptive slice, not an optimization over teacher identities.

E CONTROLLER SELECTION AND REALIZATIONS

E.1 ARCHITECTURE SELECTION

The prior DSEC search uses seed 0 and the fixed 3,544-pair manifest. Its selection score averages validation R@5 over auxiliary ResNet-18 and ViT-B/16 search runs. MLP-H2-K10 ranks first under that score and also has the highest ViT validation score. The final three-dataset performance study reports ViT only and uses fresh seeds 1–5; the search scores are not additional test replicates.

Table 8: **DSEC controller search.** Seed-0 ViT validation R@5 on the frozen manifest. Controller candidates use $\lambda = 5$; the scalar reference is the best searched scalar ($\lambda = 10$). Hidden widths specify $\tanh$ MLP layers.

Family	Candidate	Validation R@5
Reference	Scalar ($\lambda = 10$)	.0764
Reference	OGM-GE	.0774
Linear	$K = 5$	.0724
Linear	$K = 10$	.0942
Linear	$K = 20$	.0784
Rule	$\tau = 2$	.0893
Rule	$\tau = 5$	.1032
Rule	$\tau = 10$	.0546
MLP	32, $K = 10$	.0645
MLP	$64, 32$, $K = 10$	**.1081**
MLP	$64, 64, 32$, $K = 10$	.0744

Rule policy. Using the smoothed state components from Eq. 7, define

$$u_A = \bar{c}_A - \tfrac{1}{2}\overline{\log r_A}, \qquad u_B = \bar{c}_B - \tfrac{1}{2}\overline{\log r_B}, \tag{16}$$

$$b = u_A + u_B, \qquad d = \tfrac{1}{2}(u_A - u_B), \tag{17}$$

$$\alpha_A = \mathrm{clip}(\sigma(b + \tau d), .02, .98), \qquad \alpha_B = \mathrm{clip}(\sigma(b - \tau d), .02, .98). \tag{18}$$

The shared term controls overall admission and the difference term controls asymmetry. Rule-$\tau 5$ fixes $\tau = 5$. It uses the same 20-step symmetric warm-up and ten-step decision interval as the learned policies, but no reward optimizer.

Table 9: **Alternative policies at fixed $\lambda = 5$.** R@5 fractions on the same three seeds (1–3) for every policy. The common weight and seed set separate parameterization comparisons from the validation-selected results in Table 1.

Dataset	Policy	Seeds	Validation	ID	OOD
DSEC	MLP-H2-K10	3	.0552±.0124	.0377±.0069	.0326±.0065
DSEC	Linear-K10	3	.0777±.0068	.0340±.0089	.0295±.0030
DSEC	Rule-$\tau 5$	3	.0866±.0115	.0426±.0060	.0255±.0028
Eilat	MLP-H2-K10	3	.0192±.0017	.0102±.0012	.0038±.0002
Eilat	Linear-K10	3	.0212±.0043	.0092±.0025	.0040±.0010
Eilat	Rule-$\tau 5$	3	.0217±.0022	.0093±.0026	.0038±.0012
Sunboat	MLP-H2-K10	3	.0318±.0225	.0135±.0020	.0027±.0006
Sunboat	Linear-K10	3	.0240±.0046	.0139±.0027	.0029±.0008
Sunboat	Rule-$\tau 5$	3	.0253±.0131	.0170±.0069	.0038±.0020

At a common alignment weight, the simple alternatives retain competitive retrieval. This shows that the allocation framework supports both deterministic and learned implementations. We retain the MLP chosen before the full-scale study rather than reselecting a controller from the final test results.

Reading the admission traces. The reported gap is first averaged over the saved training history of each run, then averaged across seeds. Histories contain every ten-step probe plus the last step and include symmetric warm-up. A nonzero gap establishes asymmetric actions; it is not a separate measure of how much useful state information the policy learned. For calibration, independent

Table 10: **Controller behavior and recorded runtime.** Admissions summarize the logged training steps; gap includes warm-up. Runtime is the median recorded duration in seconds over five selected runs, including final evaluation. These shared-machine timings are descriptive, not a controlled overhead benchmark.

Dataset	Mean α_A	Mean α_B	$\lvert\alpha_A - \alpha_B\rvert$	MLP (s)	Scalar (s)
DSEC	0.537	0.499	0.333±0.049	539.5	517.3
Eilat	0.506	0.463	0.346±0.013	674.9	676.1
Sunboat	0.561	0.490	0.332±0.041	825.7	816.6

uniform choices from $\mathcal{A}_{10}$ have $\mathbb{E}|\alpha_A - \alpha_B| = 11/30$. Performance and action-space usage are therefore reported as distinct observations. The available histories contain sampled admissions and gradient statistics, rather than full categorical distributions.

E.2 Exact update order

The probe consists of 96 evenly spaced indices in training-manifest order. Its augmented and deterministic tensors are generated once per run and then cached. Probe evaluation uses model evaluation mode and no gradients. At zero-based step 20, the first policy action is sampled; the first policy update occurs at step 30, using the preceding interval's progress. The first 20 model updates are symmetric. State statistics are updated at steps 0, 10, 20, and every ten steps thereafter.

Algorithm 1 Conditional backward allocation: model and policy updates

Initialize model θ, policy ϕ, separate optimizers, $s = 0$, $b = 0$, and cached training probe
Set $(\alpha_A, \alpha_B) = (1, 1)$ and previous-action record to empty
for $i = 0, \ldots, T - 1$ **do**
  Compute branch losses $\mathcal{L}_A, \mathcal{L}_B$ and unrouted alignment loss
  if $i \bmod 10 = 0$ **then**
    Compute detached gradient statistics and update s by Eq. 7
    if $i \geq 20$ **then**
      Evaluate probe loss P using Eq. 9
      if a previous action record exists **then**
        Set $R = P_{\text{prev}} - P$; update $b \leftarrow .9b + .1R$ and $A \leftarrow \text{clip}(10(R - b), -1, 1)$
        Update ϕ with the saved log probabilities and entropies using Eq. 10
      end if
      Sample new admissions from $\pi_\phi(s)$; save their log probabilities, entropies, and $P_{\text{prev}} \leftarrow P$
    end if
  end if
  Recompute alignment through Eq. 4; backpropagate the total loss
  Clip assembled model gradients and update θ with AdamW
end for

F FULL RETRIEVAL RESULTS AND SEED VARIABILITY

F.1 ALIGNMENT-WEIGHT SENSITIVITY

Table 11: **Weight sensitivity.** All three tested weights for the scalar and MLP families, using five seeds per row. Entries are R@5 fractions. Selection uses validation only; this table includes unselected settings.

Dataset	Method	λ	Validation	ID
DSEC	Scalar	2.5	**.0788±.0127**	.0391±.0056
DSEC	Scalar	5	.0667±.0150	.0396±.0083
DSEC	Scalar	10	.0685±.0275	.0407±.0201
DSEC	MLP-H2-K10	2.5	**.0792±.0125**	.0460±.0096
DSEC	MLP-H2-K10	5	.0635±.0161	.0349±.0062
DSEC	MLP-H2-K10	10	.0687±.0160	.0430±.0034
Eilat	Scalar	2.5	.0178±.0018	.0095±.0025
Eilat	Scalar	5	**.0201±.0037**	.0097±.0023
Eilat	Scalar	10	.0174±.0031	.0091±.0017
Eilat	MLP-H2-K10	2.5	.0181±.0018	.0080±.0019
Eilat	MLP-H2-K10	5	.0184±.0016	.0086±.0023
Eilat	MLP-H2-K10	10	**.0188±.0024**	.0099±.0025
Sunboat	Scalar	2.5	**.0330±.0112**	.0138±.0049
Sunboat	Scalar	5	.0296±.0106	.0146±.0064
Sunboat	Scalar	10	.0309±.0057	.0125±.0027
Sunboat	MLP-H2-K10	2.5	.0322±.0080	.0136±.0035
Sunboat	MLP-H2-K10	5	.0285±.0181	.0123±.0029
Sunboat	MLP-H2-K10	10	**.0340±.0076**	.0167±.0059

Table 11 reports every weight in the final search. The selected setting is a dataset-level choice based on mean validation R@5, not a different choice for each seed or for ID evaluation. Keeping this selection fixed preserves the distinction between validation tuning and held-out evaluation.

F.2 PAIRED DIFFERENCES

Table 12: **Paired MLP–scalar R@5 differences.** Values are percentage points; positive favors the MLP. Intervals are two-sided 95% paired t intervals across five seed pairs. Wins count positive seed differences. Weights are selected as in Table 1.

Dataset	Split	Mean difference	95% interval	Wins
DSEC	Validation	+0.040	[-2.496, +2.576]	2/5
DSEC	ID	+0.690	[-0.493, +1.873]	4/5
Eilat	Validation	-0.127	[-0.698, +0.445]	3/5
Eilat	ID	+0.025	[-0.525, +0.575]	3/5
Sunboat	Validation	+0.105	[-2.071, +2.280]	2/5
Sunboat	ID	+0.287	[-0.797, +1.371]	3/5

For each selected MLP/scalar pair, define d_i as the R@5 difference at seed i. We report $\bar{d}$ and the interval

$$\bar{d} \pm t_{.975,4}\, s_d/\sqrt{5}, \qquad t_{.975,4} = 2.77645. \tag{19}$$

All intervals in Table 12 include zero. Accordingly, the main text reports mean rankings and observed effect sizes without claiming significant superiority. These intervals describe variation over training seeds for the selected settings; they do not incorporate uncertainty from architecture or weight selection, or resampling of acquisition sequences. Validation intervals are descriptive because validation also selects the weights.

F.3 RETRIEVAL AT RANK ONE

Table 13: **R@1 for validation-selected configurations.** Mean ± sample s.d. over five seeds, in raw fractions. We retain the weights selected by R@5 and do not retune for R@1.

Method	λ	Validation	ID
DSEC			
Alignment off	0	.0014±.0009	.0019±.0010
Scalar	2.5	.0183±.0052	.0090±.0025
Gradient Blending	5	.0012±.0016	.0015±.0007
OGM-GE	5	.0196±.0042	.0084±.0030
MMPareto	5	.0083±.0049	.0051±.0023
ReconBoost	5	.0155±.0015	.0066±.0014
DGL	5	.0121±.0043	.0044±.0021
MLP-H2-K10	2.5	.0161±.0044	.0094±.0013
Eilat			
Alignment off	0	.0003±.0002	.0002±.0002
Scalar	5	.0038±.0009	.0022±.0005
Gradient Blending	5	.0003±.0002	.0002±.0001
OGM-GE	5	.0036±.0005	.0018±.0003
MMPareto	5	.0005±.0002	.0005±.0002
ReconBoost	5	.0026±.0007	.0014±.0003
DGL	5	.0009±.0002	.0008±.0006
MLP-H2-K10	10	.0038±.0004	.0017±.0006
Sunboat			
Alignment off	0	.0034±.0035	.0029±.0017
Scalar	2.5	.0073±.0035	.0031±.0012
Gradient Blending	5	.0026±.0016	.0042±.0045
OGM-GE	5	.0050±.0022	.0023±.0011
MMPareto	5	.0037±.0030	.0031±.0024
ReconBoost	5	.0047±.0020	.0023±.0027
DGL	5	.0031±.0018	.0023±.0017
MLP-H2-K10	10	.0060±.0022	.0037±.0011

R@1 provides a stricter retrieval endpoint using the same final embeddings, galleries, and R@5-selected weights. It is included for completeness rather than used to change the primary metric or select configurations.