LEARNING HIERARCHICAL FUSION OF INTERMEDIATE REPRESENTATIONS ACROSS VISION FOUNDATION MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Modern vision foundation models expose rich hierarchies of intermediate representations, yet downstream transfer commonly reduces each encoder to its final feature; this discards potentially useful information across depth within an encoder and across independently trained encoders. We introduce Hierarchical Fusion Attention (HIFA), a hierarchical attention mechanism that first aggregates intermediate representations within each frozen vision backbone and then fuses the resulting summaries across backbones, operating on compact global layer descriptors rather than dense spatial maps so as to separate the selection of useful representation levels from the combination of heterogeneous encoders. On VTAB-1k, HIFA achieves substantially higher mean accuracy than concatenation of final-layer representations, with particularly large gains on Structured tasks: conventional concat reaches 0.660, a packed-final concat control 0.671, final-only HIFA 0.702, an architecture-matched final-token-only HIFA 0.704, and full HIFA 0.719. The near-equality of final-only and matched final-token indicates that the remaining numerical lift to full HIFA is associated with access to intermediate depth under the same hierarchical architecture; this depth-related difference is numerical and task-dependent rather than a statistically established uniform cross-task improvement. Beyond predictive performance, intra-backbone attention reveals allocation across network depth and inter-backbone attention reveals encoder contributions; combined with test-time masking interventions, these analyses show that the trained fusion mechanism depends strongly on specific depths and encoder summaries.

1 INTRODUCTION

Modern vision foundation models expose rich hierarchies of representations, yet downstream transfer commonly reduces each encoder to its final representation. This discards two sources of potentially useful information: representations across depth within an encoder, and representations across independently pretrained encoders.

Self-supervised Vision Transformers trained with different objectives induce distinct inductive biases. Discriminative methods such as DINOv3 (Siméoni et al., 2025) emphasize transferable global features; masked reconstruction (He et al., 2022) preserves fine spatial detail; latent predictive architectures such as I-JEPA (Assran et al., 2023) operate in embedding space. When only the last layer of each model is retained, multi-model fusion collapses to concatenating or voting over a few vectors. Intermediate layers can carry task-relevant structure for transfer (Evci et al., 2022; Tu et al., 2023; Dukler et al., 2023; Ciernik et al., 2026), and independently trained encoders can provide heterogeneous representation sources (Ramtoula et al., 2025).

Prior work typically addresses one of these axes in isolation. Methods such as Head2Toe, Visual Query Tuning, and attentive multi-layer fusion show that intermediate features are useful, but usually within a single backbone. Multi-model probing studies combine several foundation models, yet often at fixed final-layer embeddings. Few approaches jointly learn which internal levels to retain and how to combine heterogeneous encoders under one hierarchical mechanism. Retaining dense

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.

spatial maps from every layer of several encoders is informative, but it inflates the representation volume and the cost of fusion.

The scientific question is not only whether intermediate representations are useful, but *where* the gains of multi-backbone intermediate-representation fusion come from. In particular: how much comes from a learned fusion interface over last-layer tokens; how much additional gain comes from opening intermediate depth; whether sample-dependent intra attention is needed or a static layer combination suffices; and whether preserving a per-backbone intra stage before cross-backbone inter fusion matters relative to a collapsed flat attention.

We propose HIFA, a shared hierarchical attention head over frozen vision encoders. For each backbone b, an intra-backbone aggregator maps the layer stack to a compact summary z_b; an inter-backbone aggregator then fuses $\{z_b\}$ into a single representation for classification. Architecture hyperparameters are shared across backbones (projection dimension, query count, depth/heads, vector policy), so adding an encoder does not introduce new search axes; weights need not be tied across backbones. Empirically, last-layer fusion already helps: confidence-based voting of isolated linear probes reaches 0.636, and concatenating final features reaches 0.660, versus 0.616 for the best single encoder. A packed-final concat control that feeds the same $[\mathrm{CLS} \,\|\, \mathrm{mean}]$ / mean descriptors used by final-only HIFA reaches only 0.671, so richer final content alone explains little of the concat gap. A hierarchical final-only control reaches 0.702; an architecture-matched final-token-only HIFA (full θ^* with memory length one) reaches 0.704; full HIFA reaches 0.719. An optional query-key normalized scoring rule (HIFA-QK) is competitive numerically but not significantly better than HIFA on the full protocol.

Contributions.

1. A hierarchical factorization of aggregation: layer representations $\rightarrow$ per-backbone summary $\rightarrow$ cross-backbone fused representation, with a compact shared hyperparameter structure that does not grow with the number of backbones.
2. A structured empirical decomposition of VTAB-1k gains from final representation content, learned hierarchical fusion, intermediate depth, and encoder composition, with packed-final concat, final-only / matched final-token, weighted-sum, flat, linear-inter, and single-backbone intra-only controls.
3. Evidence that encoder composition matters: single-backbone intra-only probes, retrained backbone-pair HIFA, and the three-encoder system show non-uniform contributions and that adding encoders need not improve accuracy monotonically.
4. A representation-utilization analysis combining attention allocation with test-time masking of highly attended depth regions and encoder summaries.
5. An exploratory query-key normalized overlay with input-conditioned head-wise temperature; it raises mean development accuracy, but the paired gain over HIFA on the 19-task held-out test protocol is not statistically significant.

2 Related Work

2.1 Intermediate Representations

Head2Toe (Evci et al., 2022) demonstrates that probing intermediate layers improves transfer relative to final-layer features alone. Visual Query Tuning (Tu et al., 2023) introduces learnable queries over intermediate representations. InCA (Dukler et al., 2023) uses cross-attention over intermediate activations to access representations distributed across network depth. Attentive Multi-Layer Fusion (Ciernik et al., 2026) learns to aggregate representations across transformer depth, providing a closely related depth-aware fusion mechanism. Structured Model Probing (Wu et al., 2024) studies efficient transfer by structured regularization over intermediate features under VTAB-style evaluation. SMP addresses structured probing and feature selection within a pretrained model, whereas HIFA's contribution is an explicit two-level factorization (intermediate layers $\rightarrow$ per-backbone summary $\rightarrow$ cross-backbone fusion) across heterogeneous frozen encoders. Related analyses of depth-wise representation structure include the Tunnel Effect (Ostaszewski et al., 2023). These works

motivate depth-aware aggregation within a single encoder; HIFA additionally introduces an explicit second stage for fusion across independently pretrained encoders.

2.2 Multi-Model Feature Fusion

Combining independently trained models can improve coverage of visual concepts. In few-shot classification, Chowdhury et al. (2021) combine a library of pretrained feature extractors with a simple classifier, and the Universal Representation Transformer (Liu et al., 2021) learns to blend multiple pretrained representations; both operate on final-layer embeddings rather than intermediate hierarchies. ComBo (Ramtoula et al., 2025) is the closest methodological multi-backbone intermediate-representation neighbor: a probing adapter over layers and spatial locations from multiple foundation models. ComBo retains dense spatial layer representations, whereas HIFA deliberately compresses each layer to a compact global descriptor before fusion. The methods also use different encoder sets and probing protocols, so published VTAB-1k scores are reported only as unmatched context (Appendix O). Section 5.5 compares representation footprint and analytically estimated fusion compute; we do not claim a matched accuracy comparison.

2.3 Attention and Normalization

Scaled dot-product attention (Vaswani et al., 2017) is sensitive to the scale of queries and keys. When keys come from heterogeneous projected layer stacks, magnitude differences can be conflated with relevance. Query-Key Normalization (Henry et al., 2020) normalizes queries and keys before attention scoring and replaces the fixed dot-product scaling with a learned temperature. We study an input-conditioned, head-wise temperature as an optional overlay within HIFA. Vision Transformers (Dosovitskiy et al., 2021) provide the frozen feature extractors we fuse.

3 Hierarchical Representation Fusion

3.1 Problem Formulation

Let $\mathcal{B} = \{1, \dots, B\}$ index frozen vision encoders. For an image x and backbone b, let $\{h_b^1, \dots, h_b^{L_b}\}$ denote one packed global token per native layer. For DINOv3 and MAE, the CLS token and mean-pooled patch vector from each layer are concatenated along the feature dimension, yielding a single memory token per native layer. The I-JEPA checkpoint used does not provide a dedicated CLS stream, so we use the mean-pooled patch representation as its global layer descriptor. Thus, under the selected all-available policy, the pre-projection memory width is $2D_b$ for DINOv3/MAE and D_b for I-JEPA, while the number of memory tokens remains one per native layer. Intra-backbone aggregation answers: *which representation levels matter for this task?*

$$\{h_b^1, \dots, h_b^{L_b}\} \xrightarrow{\text{intra}} z_b. \tag{1}$$

Inter-backbone fusion then answers: *how should heterogeneous encoder summaries be combined for this task?*

$$\{z_1, \dots, z_B\} \xrightarrow{\text{inter}} z. \tag{2}$$

A linear classifier maps z to class logits. Only projections, attention parameters, and the classifier are trained; encoder weights remain frozen. Encoder activations are cached once per image.

3.2 Intra-Backbone Aggregation

For backbone b, write $D_b^{\text{mem}} = 2D_b$ for DINOv3 and MAE and $D_b^{\text{mem}} = D_b$ for I-JEPA. Let the selected memory be $M_b(x) \in \mathbb{R}^{N_b \times D_b^{\text{mem}}}$. A backbone-specific projection $P_b : \mathbb{R}^{D_b^{\text{mem}}} \to \mathbb{R}^d$ maps tokens to a shared fusion dimension d:

$$\tilde{M}_b(x) = P_b\big(M_b(x)\big) \in \mathbb{R}^{N_b \times d}. \tag{3}$$

With independent queries per backbone, $U_b \in \mathbb{R}^{Q \times d}$, cross-attention produces

$$Z_b(x) = A_{\text{intra},b}\big(U_b, \tilde{M}_b, \tilde{M}_b\big) \in \mathbb{R}^{Q \times d}. \tag{4}$$

We enforce $Q_D = Q_M = Q_J = Q$ and $d_D = d_M = d_J = d$. Architecture hyperparameters are shared; weights may remain backbone-specific. In the selected configuration, $Q{=}1$, so Z_b reduces to a vector $z_b \in \mathbb{R}^d$.

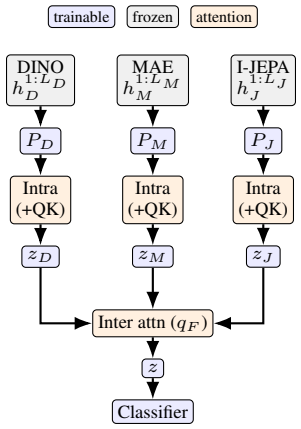


Figure 1: Overview of HIFA. Frozen encoders (gray) yield intermediate representations that are projected and summarized by intra-backbone attention (orange), optionally with query-key normalization. Inter-backbone attention fuses $\{z_D, z_M, z_J\}$ into z for classification. Architecture hyperparameters are shared across backbones; weights need not be tied.

3.3 Inter-Backbone Fusion

Concatenate backbone summaries

$$Z(x) = \big[Z_D(x); Z_M(x); Z_J(x)\big] \in \mathbb{R}^{3Q\times d}. \tag{5}$$

A single learned query q_F reads Z as key/value memory:

$$z(x) = A_{\text{inter}}\big(q_F, Z(x), Z(x)\big) \in \mathbb{R}^d, \qquad \hat{y} = Wz + b. \tag{6}$$

The search space is independent of the number of backbones. Intra hyperparameters (depth, heads, fusion dimension d, vector policy) and inter hyperparameters (depth, heads) are global. Adding an encoder increases fusion-memory length, $Z(x) \in \mathbb{R}^{BQ\times d}$, and compute, but does not introduce new axes such as a per-backbone query count, pooling policy, head count, or projection dimension.

3.4 Query-Key Normalized Attention

When intermediate representations from different depths and encoders are projected into a shared space, feature magnitude can dominate scaled dot-product logits. HIFA-QK optionally replaces scoring with query-key normalization (Henry et al., 2020): L2-normalized queries and keys with a learned temperature. For head h and key index j,

$$a_{ij} = \text{softmax}_j\left(\tau_h(x)\frac{q_{ih}^\top k_{jh}}{\|q_{ih}\|_2\|k_{jh}\|_2}\right). \tag{7}$$

The selected overlay uses a dynamic head-wise temperature on intra-backbone attention, $\tau_h(x) = \text{Linear}(\bar{m}_b)$, where $\bar{m}_b$ is the mean intra memory token for backbone b, and leaves inter-backbone attention as standard scaled dot-product. Optional magnitude (radial) terms and static temperatures are ablated in Appendix D.

4 Experimental Setup

4.1 VTAB-1k

We evaluate on VTAB-1k (Zhai et al., 2019): 19 classification tasks with 1,000 labeled examples each, grouped into Natural (7), Specialized (4), and Structured (8). Architecture selection uses validation accuracy only on a balanced six-task development subset $\mathcal{D}_{\text{dev}} = \{\text{SUN397}, \text{DTD}, \text{Patch Camelyon}, \text{RESISC45}, \text{CLEVR-count}, \text{SmallNORB-azi}\}$ (two

Table 1: Frozen backbones.

Backbone	Pretraining objective	Dim D_b	Blocks	Streams
DINOv3	discriminative SSL	1280	32	CLS, mean
MAE	masked reconstruction	1024	24	CLS, mean
I-JEPA	latent prediction	1408	40	mean

tasks per group). The remaining 13 task identities are not used during architecture selection; held-out test labels are never used for selection. Final evaluation reports means over all 19 tasks; Appendix I additionally separates the six development identities from the thirteen non-development identities as a post-hoc check. The full task list is given in Appendix L.

4.2 Frozen Vision Encoders

We use public DINOv3, MAE, and I-JEPA checkpoints as frozen feature extractors (Table 1), representing the three pretraining paradigms discussed in Section 1. Encoder weights are never updated; only the projection, hierarchical attention, and classification layers are trained. Intermediate representations are extracted from all native transformer blocks according to the policy described in Section 3.1.

4.3 Training and Model Selection

We train the fusion head with AdamW (weight decay 10^{-4}, cosine schedule, 10-epoch warmup, 100 epochs, LR 10^{-3}, batch size 64, grad clip 5.0; Appendix A, L). Architecture selection is performed on $\mathcal{D}_{\text{dev}}$ with validation only; final claims use all 19 VTAB-1k tasks and seeds $\{41, 42, 43\}$. Sequential validation on $\mathcal{D}_{\text{dev}}$ selects a compact shared configuration: one intra cross-attention layer (4 heads), d=512, a single independent query per backbone, all available intermediate streams, and a one-head inter-backbone query (Appendix B). "Shared" refers to a common hyperparameter structure across backbones, not necessarily weight tying.

4.4 Baselines

We compare against: (i) linear probes on each backbone's final representation; (ii) confidence-based voting of those isolated probes; (iii) concatenation of final representations followed by a linear classifier; (iv) a packed-final concat control that feeds the same $[\text{CLS} \,\|\, \text{mean}]$ / mean final descriptors used by final-only HIFA into a linear classifier (width 6016 vs. 3712 for conventional concat); (v) structural ablations of HIFA (final-layer tokens with the selected inter head; an architecture-matched final-token-only variant that keeps the full θ^* attention stack with memory length one; weighted intra aggregation; a flat single-head over all projected layer tokens; a linear inter readout; and θ^* intra alone on each encoder); (vi) backbone-pair HIFA retrained under the same θ^* with two encoders; and (vii) HIFA-QK. For conventional final-representation baselines, DINOv3 and MAE contribute CLS tokens and I-JEPA contributes patch-mean; the concatenation baseline therefore operates on a 3712-dimensional vector. These last-layer vectors are distinct from the `all_available` layer memory of HIFA (CLS concatenated with mean per native layer for DINOv3/MAE; mean per layer for I-JEPA). Final-only HIFA retains exactly the final packed HIFA memory token for each encoder ($[\text{CLS} \,\|\, \text{mean}]$ for DINOv3 and MAE, and patch mean for I-JEPA), but replaces the selected intra attention stack with a lighter final-token path. The architecture-matched final-token-only control instead keeps the complete θ^* intra$\rightarrow$inter path while restricting each backbone's memory to that same final packed token. Each isolated probe stores only its predicted class and associated confidence. For decision-level fusion, we reconstruct a class distribution by assigning the stored confidence to the predicted class and distributing the remaining mass uniformly across the other classes, then average these reconstructed distributions across backbones. Section 5.5 compares HIFA with ComBo (Ramtoula et al., 2025) on fusion representation and compute; accuracy is not matched across encoder sets.

Table 2: VTAB-1k test accuracy (mean over three seeds). Group means equally weight tasks within Natural / Specialized / Structured / Overall. Packed-final concat uses the same $[\mathrm{CLS} \,\|\, \mathrm{mean}]$ / mean finals as final-only HIFA with a linear classifier. Final-only uses the last packed HIFA memory token with the selected inter head; matched final-token keeps full θ^* attention with memory length one.

Method	Natural	Spec.	Struct.	Overall
DINOv3 linear	0.770	0.850	0.365	0.616
I-JEPA linear	0.651	0.797	0.368	0.563
MAE linear	0.591	0.794	0.371	0.541
Confidence-based voting	0.769	0.851	0.411	0.636
Concat + linear	0.793	0.863	0.443	0.660
Packed-final concat + linear	0.800	**0.868**	0.459	0.671
Final-only HIFA	0.788	0.857	0.548	0.702
Matched final-token HIFA	0.781	0.857	0.561	0.704
HIFA	**0.818**	0.856	0.563	0.719
HIFA-QK	0.815	0.862	**0.575**	**0.724**

5 Results

5.1 Main VTAB-1k Protocol

Table 2 reports mean test accuracy over seeds $\{41, 42, 43\}$ on all 19 VTAB-1k tasks. Group means average task accuracies within Natural (7), Specialized (4), Structured (8), and Overall (19).

Last-layer transfer follows a clear progression: the best isolated linear probe (DINOv3, 0.616), confidence-based voting (0.636), and concatenated final features (0.660). Packed-final concat reaches 0.671 overall (+0.011 vs. conventional concat; Structured 0.459), so supplying the richer HIFA final descriptors alone accounts for only a small fraction of the concat→final-only gap. Final-only HIFA reaches 0.702 overall and 0.548 Structured; an architecture-matched final-token-only control that keeps the full θ^* attention stack with memory length one reaches 0.704 / 0.561, nearly identical to final-only. Full HIFA reaches 0.719 overall (+0.015 vs. matched final-token) and 0.563 Structured; under this matched comparison the additional numerical difference is associated with opening intermediate depth while holding the hierarchical processing architecture fixed. The task-level pattern of this Δ is uneven and is not a uniform Structured lift (Appendix G). Natural and Specialized scores remain close to the concat baseline throughout. HIFA-QK is numerically highest overall (0.724) but is not significantly better than HIFA (Section 5.4). On the 13 task identities not used during architecture selection, HIFA remains numerically ahead of concat (0.727 vs. 0.668), packed-final concat (0.679), final-only (0.704), and flat attention (0.707); Appendix I reports the complete development/non-development split.

5.2 Architecture Ablations

Table 3 decomposes the hierarchical design with controls that are structurally matched with respect to the complementary stage from θ^* (not equalized for parameter count). These controls share the global AdamW protocol and inherit θ^* except for the intentional mode change; none receives an independent architecture search (Appendix H). Packed-final concat changes only final descriptor content under a linear classifier. Final-only removes intermediate-depth memory; matched final-token keeps full θ^* attention with memory length one. Weighted-sum, flat, inter-linear, and intra-only are one-factor controls that keep complementary θ^* stages (Table 3).

Relative to concat (0.660 / Struct. 0.443), packed-final reaches 0.671 / 0.459; final-only 0.702 / 0.548; matched final-token 0.704 / 0.561; full HIFA 0.719 / 0.563 (+0.015 vs. matched). Weighted-sum / flat / inter-linear stay at 0.704 / 0.705 / 0.714; DINOv3 intra-only is strong (0.709 / 0.539 vs. linear 0.616 / 0.365). Paired 19-task tests (Appendix F) show numerical advantages over these controls, none with Wilcoxon $p<0.05$ on all 19 tasks. Design choices: Appendix B (Table 10).

Table 3: Structural ablations (mean test accuracy, 19 tasks). Packed-final concat uses HIFA packed finals with a linear classifier. Final-only uses the last packed HIFA token with the selected inter head (no intermediate-depth intra aggregation); matched final-token keeps full θ^* attention with memory length one. One-factor intra/inter variants keep the complementary stage from θ^*; flat collapses both stages. Intra-only applies θ^* intra to a single encoder with a linear classifier.

Variant	Natural	Spec.	Struct.	Overall
Concat (final-layer fusion)	0.793	0.863	0.443	0.660
Packed-final concat + linear	0.800	**0.868**	0.459	0.671
Final-only, selected inter	0.788	0.857	0.548	0.702
Matched final-token (full θ^*)	0.781	0.857	0.561	0.704
Intra weighted-sum, selected inter	0.804	0.863	0.537	0.704
Flat single-head (no inter)	0.819	0.846	0.535	0.705
Intra-only (DINOv3)	**0.820**	0.852	0.539	0.709
Intra-only (I-JEPA)	0.694	0.806	0.499	0.635
Intra-only (MAE)	0.650	0.807	0.534	0.634
Selected intra, linear inter	0.814	0.850	0.559	0.714
HIFA (attention intra+inter)	0.818	0.856	0.563	0.719
HIFA-QK	0.815	0.862	**0.575**	**0.724**

Table 4: Backbone-pair HIFA under frozen θ^* (mean test accuracy, 19 tasks). Pairs are retrained from scratch with two encoders; the trio is the standard three-encoder HIFA.

Encoders	Natural	Spec.	Struct.	Overall
DINOv3 + MAE	0.818	**0.861**	**0.567**	**0.722**
DINOv3 + I-JEPA	**0.820**	0.857	0.548	0.713
MAE + I-JEPA	0.673	0.819	0.518	0.638
DINOv3 + MAE + I-JEPA	0.818	0.856	0.563	0.719

5.3 Encoder Composition

Table 4 reports HIFA retrained from scratch under the same frozen θ^* on each backbone pair. DINO+MAE reaches 0.722 overall (0.567 Structured), numerically competitive with three-encoder HIFA (0.719 / 0.563). DINO+I-JEPA is close (0.713 / 0.548), whereas MAE+I-JEPA without DINOv3 is much weaker (0.638 / 0.518). A paired test of Acc(trio) − Acc(DINO+MAE) over the 19 tasks yields $\bar{\Delta}$=−0.0030, 95% CI [−0.0071, +0.0006], Wilcoxon p=0.352 (W/T/L $7/4/8$); the configurations are not statistically distinguishable under this protocol. Encoder composition therefore matters, and predictive performance need not grow monotonically with B. Pair retraining is distinct from test-time backbone-off masking (Section 6.2): MAE-off barely hurts the trio, yet DINO+MAE is the strongest retrained pair, because the two protocols answer different questions.

5.4 Query-Key Normalization

On the multi-seed development leaderboard, dynamic head-wise query-key normalization (HIFA-QK) attains mean validation accuracy 0.704, compared with 0.696 for HIFA without the overlay (Δ=+0.74 percentage points; Appendix D). On the 19 tasks (held-out test splits) the mean paired difference is $\bar{\Delta} = \mathrm{Acc}(\text{HIFA-QK}) - \mathrm{Acc}(\text{HIFA}) = +0.0055$. A 95% bootstrap CI over datasets is $[-0.0011, +0.0135]$ (includes 0); a two-sided Wilcoxon signed-rank test yields p=0.182 (win/tie/loss $10/3/6$ at a 0.001 tie threshold). A static-temperature QK control also reaches approximately 0.724 overall, limiting claims that the dynamic temperature is necessary. We therefore treat HIFA-QK as an exploratory overlay that is numerically higher but not statistically significant on the 19-task held-out test protocol. Per-task scores are in Appendix E.

5.5 Efficiency and Scalability

HIFA is a compact shared head over cached frozen activations. ComBo (Ramtoula et al., 2025), the closest multi-backbone intermediate-representation neighbor, retains 196 spatial vectors per layer from four ViT-B encoders; HIFA uses one global descriptor per layer and Q=1 summaries. Table 5 gives analytical fusion-input payload and fusion-head MACs (Appendix M): payload is raw

Table 5: Analytical fusion footprint under respective published multi-backbone configs (not matched accuracy; not peak GPU memory).

Metric	ComBo	HIFA
Representation granularity	196 spatial / layer	1 global / layer
Raw fusion-input FP32 payload	$\sim$28.90 MB/image	$\sim$0.774 MB/image
Fusion-head compute	$\sim$1.217 GMAC	$\sim$0.155 GMAC
Trainable fusion params	$\sim$5.93M + cls	7.34M + cls
Backbone setting	4$\times$ViT-B	DINOv3+MAE+I-JEPA

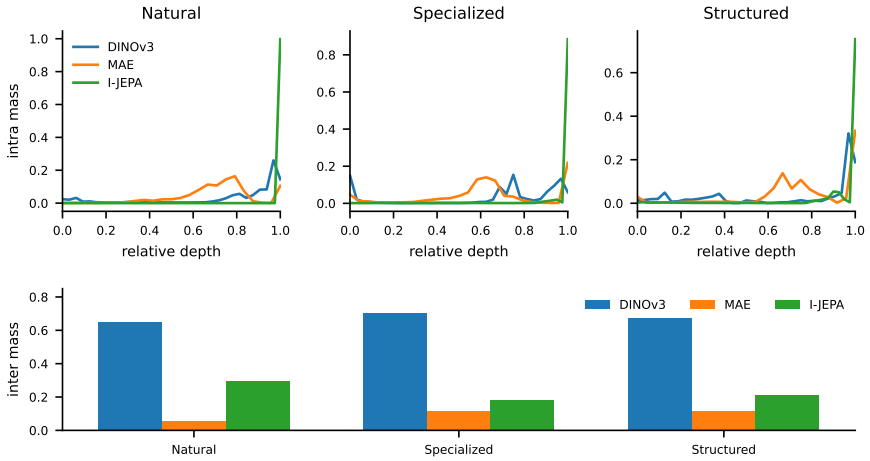


Figure 2: HIFA-QK attention on nine analysis tasks (mean over seeds). Top: intra mass vs. relative depth; bottom: inter mass over encoder summaries.

FP32 input footprint, not peak GPU memory; MACs exclude backbone forward and the classifier; encoder sets differ, so accuracy is unmatched. A matched ComBo adapter reproduction was not feasible under our accelerator-memory budget (Appendix O). Under these configurations, HIFA's raw fusion-input payload is $\sim$37.3$\times$ smaller with $\sim$7.9$\times$ fewer fusion MACs; parameter counts do not favor HIFA.

6 Understanding Hierarchical Fusion

Representation-utilization analyses use the selected HIFA-QK checkpoint on a nine-task subset and characterize that hierarchical instantiation (not vanilla HIFA by default). Geometry/CKA diagnostics are in Appendices J–K.

6.1 Layer and Backbone Utilization

Figure 2 shows intra mass versus relative depth and inter mass over $\{z_D, z_M, z_J\}$ (seeds $\{41, 42, 43\}$). Under this checkpoint, intra attention concentrates on late layers; inter attention favors DINOv3 (0.65–0.70), then I-JEPA, then MAE. We treat these weights as correlational preference, not causal necessity.

6.2 Test-time Masking Interventions

Masking attention logits with $-\infty$ before softmax yields $\Delta_c = \mathrm{Acc}_{\mathrm{intervention}} - \mathrm{Acc}_{\mathrm{full}}$ (Table 6). These are test-time interventions on a trained fusion head: large drops indicate functional dependence of that model on the masked sources, not that predictive information is uniquely localized there. Late-off collapses accuracy (Δ=-0.400); early/middle effects are small (-0.014/-0.002). DINOv3-off hurts most (-0.107); MAE-off is near zero (-0.004); I-JEPA-off is intermediate

Table 6: Test-time masking interventions on nine analysis tasks (mean Δ).

Intervention	Natural	Spec.	Struct.	Overall
Early off	-0.000	-0.011	-0.032	-0.014
Middle off	$+0.000$	$+0.000$	-0.007	-0.002
Late off	-0.523	-0.323	-0.355	-0.400
DINOv3 off	-0.128	-0.115	-0.078	-0.107
MAE off	-0.001	-0.001	-0.009	-0.004
I-JEPA off	-0.004	-0.004	-0.028	-0.012

(-0.012). High allocation coincides with large functional drops under masking; backbone-off dependence differs from pair retraining (Section 5.3).

7 Discussion and Limitations

Ladder: concat 0.660; packed-final 0.671; final-only / matched $0.702/0.704$; full HIFA 0.719. Matched $\approx$ final-only; remaining numerical lift tracks intermediate depth under fixed θ^*, unevenly across tasks. Gaps vs. weighted-sum / flat are numerical (Wilcoxon n.s. on 19 tasks); DINO+MAE is competitive with the trio; HIFA-QK remains exploratory.

Limitations. Frozen encoders; global descriptors; shared hyperparameters; six-$\mathcal{D}_{\text{dev}}$ sequential search; limited analysis coverage; weaker encoder-suite transfer; unmatched ComBo context; analytical (not measured) compute. Flat/static controls are not parameter- or search-budget-matched to HIFA, so numerical gaps do not isolate hierarchical factorization alone.

8 Conclusion

HIFA provides a compact two-level interface over frozen vision foundation models: intermediate layer descriptors $\rightarrow$ per-backbone summaries $\rightarrow$ cross-backbone fusion. Operating on global layer descriptors rather than dense spatial maps separates depth selection from cross-encoder combination under a shared hyperparameter structure that does not grow with B. Empirically, learned processing of strong final descriptors explains a substantial portion of the gain over ordinary concatenation: packed-final concat (0.671) adds little over conventional concat (0.660), while final-only / matched final-token HIFA ($0.702/0.704$) close most of the gap to full HIFA (0.719).

Under the architecture-matched comparison, intermediate depth adds a further but task-dependent numerical benefit; paired tests do not establish a uniform cross-task improvement on the full 19-task set. Encoder composition is non-monotonic, and attention plus test-time masking indicate highly non-uniform dependence on depth and encoder summaries. The global-descriptor design yields analytically smaller fusion-input footprint and fewer fusion MACs than dense multi-backbone spatial adapters under published configs (not measured runtime or peak GPU memory). Flat/static controls share the same optimization protocol but are not parameter- or search-budget-matched to HIFA. A useful direction is to better understand and design efficient fusion interfaces for heterogeneous frozen foundation models under clearer capacity and search matching.

AI use statement

Generative AI tools were used as research assistants for literature discovery and organization, scientific brainstorming, feedback on experimental design, manuscript organization and language refinement, and critical review and discussion of experimental results obtained by the authors. All research hypotheses, methodological decisions, experimental procedures, numerical results, analyses, interpretations, and conclusions were produced, reviewed, and validated by the authors. The authors take full responsibility for the content of this work.

Ethics statement

This work studies representation fusion on publicly available vision benchmarks and pretrained checkpoints. It does not involve human subjects, new data collection from individuals, or clinical decision-making claims about specialized medical datasets used only as transfer benchmarks. We are not aware of material ethics concerns beyond standard considerations for reuse of public datasets and foundation-model checkpoints.

Reproducibility statement

We provide the architecture definition, optimization protocol, model-selection procedure, random seeds, hyperparameter search spaces, selected configuration, and evaluation protocol in the main text and appendices to facilitate reproduction. Frozen encoder identities and dimensions, VTAB-1k train/validation/test discipline, and optimizer settings are specified therein. An anonymized implementation, training scripts, and experimental configurations are available at `https://github.com/anonymoussubmission231-stack/HIFA`.

References

Mahmoud Assran, Quentin Duval, Ishan Misra, Piotr Bojanowski, Pascal Vincent, Michael Rabbat, Yann LeCun, and Nicolas Ballas. Self-supervised learning from images with a joint-embedding predictive architecture. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. arXiv:2301.08243.

Arkabandhu Chowdhury, Mingqing Jiang, Swarat Chaudhuri, and Chris Jermaine. Few-shot image classification: Just use a library of pre-trained feature extractors and a simple classifier. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021.

Laure Ciernik, Marco Morik, Lukas Thede, Luca Eyring, Shinichi Nakajima, Zeynep Akata, and Lukas Muttenthaler. Attentive multi-layer fusion for vision transformers. In *Proceedings of the 43rd International Conference on Machine Learning (ICML)*, 2026. arXiv:2601.09322.

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 (ICLR)*, 2021.

Yonatan Dukler, Alessandro Achille, Hao Yang, Benjamin Bowman, Varsha Vivek, Luca Zancato, Avinash Ravichandran, Charless Fowlkes, Ashwin Swaminathan, and Stefano Soatto. Your representations are in the network: Composable and parallel adaptation for large scale models. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023. arXiv:2303.04105.

Utku Evci, Vincent Dumoulin, Hugo Larochelle, and Michael C. Mozer. Head2Toe: Utilizing intermediate representations for better transfer learning. In *Proceedings of the 39th International Conference on Machine Learning (ICML)*, 2022. arXiv:2201.03529.

Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. arXiv:2111.06377.

Alex Henry, Prudhvi Raj Dachapally, Shubham Shantaram Pawar, and Yuxuan Chen. Query-key normalization for transformers. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pp. 4246–4253, 2020. doi: 10.18653/v1/2020.findings-emnlp.379. URL `https://aclanthology.org/2020.findings-emnlp.379/`.

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 (ICML)*, 2019. arXiv:1905.00414.

Lu Liu, William Hamilton, Guodong Long, Jing Jiang, and Hugo Larochelle. A universal representation transformer layer for few-shot image classification. In *International Conference on Learning Representations (ICLR)*, 2021.

Michał Ostaszewski, Ehsan Imani, Razvan Pascanu, Piotr Miłoś, and Tomasz Trzciński. The tunnel effect: Building data representations in deep neural networks. *arXiv preprint arXiv:2305.19753*, 2023.

Benjamin Ramtoula, Pierre-Yves Lajoie, Paul Newman, and Daniele De Martini. Fantastic features and where to find them: A probing method to combine features from multiple foundation models. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2025. arXiv:2512.01405.

Oriane Siméoni, Huy V. Vo, Maximilian Seitzer, Federico Baldassarre, Maxime Oquab, et al. DINOv3. *arXiv preprint arXiv:2508.10104*, 2025.

Cheng-Hao Tu, Zheda Mai, and Wei-Lun Chao. Visual query tuning: Towards effective usage of intermediate representations for parameter efficient transfer learning. *arXiv preprint arXiv:2212.03220*, 2023.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2017.

Zhi-Fan Wu, Chaojie Mao, Xue Wang, Jianwen Jiang, Yiliang Lv, and Rong Jin. Structured model probing: Empowering efficient transfer learning by structured regularization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 16849–16858, 2024.

Xiaohua Zhai, Joan Puigcerver, Alexander Kolesnikov, Pierre Ruyssen, Carlos Riquelme, et al. A large-scale study of representation learning with the visual task adaptation benchmark. *arXiv preprint arXiv:1910.04867*, 2019.

A Learning-Rate Calibration

Grid over $\{10^{-5}, 10^{-4}, 10^{-3}, 10^{-2}, 10^{-1}\}$ on final-layer linear heads for (SUN397, DTD, SmallNORB azimuth) $\times$ (DINO, MAE, I-JEPA).

Table 7: Pooled LR ranking over nine (dataset, backbone) cells.

LR	Mean val. acc.	Std
$\mathbf{10^{-3}}$	**0.381**	0.224
10^{-2}	0.379	0.220
10^{-1}	0.341	0.188
10^{-4}	0.313	0.237
10^{-5}	0.134	0.127

Selected learning rate: 10^{-3}.

B Architecture Search

Coordinate search maximizes mean validation accuracy on $\mathcal{D}_{\text{dev}}$, requiring successful runs on all six development tasks. Axis order: intra attention, fusion dimension, vector policy, intra weighting (softmax; fixed), query design (Q=1 fixed), post-intra norm, inter attention (optional sanity recheck).

C Mechanism Screening

One-shot family screen of attention overlays on frozen hierarchical fusion. No screened family exceeds the hierarchical baseline on $\mathcal{D}_{\text{dev}}$.

Table 8: Selected shared architecture (mean val. 0.6983 on $\mathcal{D}_{\text{dev}}$).

Hyperparameter	Value
Intra attention	cross, 1 layer, 4 heads
Fusion dimension d	512
Query design	independent
Number of queries Q	1
Vector policy	all intermediate streams
Intra weighting	softmax
Pre-attention / post-intra norm	none
Inter attention	cross, 1 layer, 1 head

Table 9: Per-axis development winners (mean validation accuracy).

Axis	Selected	Score	Next best
Intra attention	cross $\times$ 1 $\times$ 4	**0.6983**	0.6925
Fusion dim	512	**0.6983**	0.6894
Vector policy	all intermediate	**0.6983**	0.6789
Query design	independent	**0.6983**	0.6833
Post-intra norm	none	**0.6983**	0.6892
Inter attention	1 $\times$ 1	**0.6983**	0.6881

D Query-Key Normalization Grid

We search temperature granularities (head-/layer-wise, static/dynamic) and radial magnitude terms (log-norm, z-score, raw, or none).

$$\Delta = 0.7036 - 0.6962 = 0.0074 \quad (+0.74\,\text{pp}). \tag{8}$$

Selection uses single-seed hierarchical search at 0.6983 for architecture choice and the multi-seed leaderboard above for the QK overlay; we do not mix the two baselines in a single Δ without stating the source. Final training also includes a static head-wise temperature control. A static-τ control trained under the same full protocol reaches overall test accuracy 0.724 (Natural 0.818, Specialized 0.854, Structured 0.576), statistically indistinguishable from dynamic HIFA-QK at the reported precision.

E Per-Task Test Comparison

Table 13 lists mean test accuracy of HIFA and HIFA-QK on each of the 19 VTAB-1k tasks (seeds $\{41, 42, 43\}$). $\Delta = \text{Acc}(\text{HIFA-QK}) - \text{Acc}(\text{HIFA})$. The paired summary in Section 5.4 is computed from these 19 differences.

F Paired Ablation Tests

Table 14 applies the same paired protocol as Section 5.4 to structural controls: $\Delta_d = \text{Acc}(\text{HIFA}) - \text{Acc}(\text{control})$ over the 19 tasks, 10,000-sample bootstrap CI, two-sided Wilcoxon signed-rank, and win/tie/loss at a 0.001 tie threshold. Artefacts: `results/paired_tests/ablations/`. The bootstrap interval targets the mean paired difference, whereas the Wilcoxon signed-rank test assesses paired signed ranks and therefore need not yield the same decision: a CI for the mean that excludes zero does not imply Wilcoxon $p<0.05$. We conservatively describe an ablation gap as statistically established only when the non-parametric paired test also reaches $p<0.05$.

Table 15 repeats the same paired protocol on the thirteen task identities not used during architecture selection. Weighted-sum and flat reach nominal two-sided Wilcoxon $p<0.05$ on this post-hoc thirteen-task subset, but these findings do not remain significant after Holm or Benjamini–Hochberg

Table 10: Design-choice ablations on VTAB-dev (mean validation accuracy).

Design choice	Variant	Dev
Intra aggregation	Cross-attention	**0.698**
	Self-attention	0.691
Query design	independent	**0.698**
	shared	0.683
Layer memory	All intermediate	**0.698**
	Mean stream only	0.679
Fusion dimension	512	**0.698**
	256	0.689

Table 11: Best variant per family vs. hierarchical baseline. $\Delta = 100 \times (\mathrm{Acc}_{\mathrm{family}} - \mathrm{Acc}_{\mathrm{base}})$.

Family (best variant)	Val mean	Δ (pp)
Hierarchical baseline	**0.6983**	—
Query-key geometry (screen)	0.6950	-0.33
Scalable softmax ($\gamma = 2.0$)	0.6900	-0.83
Hyper-connection	0.6892	-0.92
Group routing ($G = 16$)	0.6883	-1.00
Depth forgetting / selective	0.6775	-2.08
Post-attention gate	0.6633	-3.50

FDR correction across the structural-comparison family of eight HIFA-versus-control tests (Holm p=0.262 / 0.335; BH p=0.191 / 0.191; artefacts: `paired_nondev13_multiplicity.csv`). The remaining comparisons are numerical differences that are not established by the paired test even nominally. These statements are restricted to the thirteen non-development task identities and do not imply a broader claim about unseen datasets.

G HIFA vs. Matched Final-Token: Task-Level Pattern

Table 16 reports the paired difference $\Delta_{d,s} = \mathrm{Acc}(\text{HIFA}) - \mathrm{Acc}(\text{matched final-token})$ for every VTAB task and seed, using the completed three-seed runs (no new training). Tie convention for task-level mean signs is ± 0.001 as elsewhere. Artefacts: `results/statistics/hifa_vs_matched_final_token/`.

Mean task-level Δ is $+0.0144$ overall (median $+0.0011$; 10 positive / 2 tie / 7 negative). By group: Natural $+0.0371$ (median $+0.0083$; $5/1/1$), Specialized -0.0006 (median -0.0016; $1/0/3$), Structured $+0.0020$ (median $+0.0094$; $4/1/3$). The overall advantage is therefore *not* a uniform Structured-group lift: Structured mean accuracy moves only $0.561 \rightarrow 0.563$ because large positive and negative task deltas cancel (notably SmallNORB-ele $+0.115$ vs. KITTI -0.103). Natural's group mean is dominated by SVHN ($+0.249$), which alone accounts for roughly half of the summed positive task-level mass; excluding SVHN, the remaining Natural means are small. Largest positive / negative task-level means: SVHN $+0.249$, SmallNORB-ele $+0.115$; KITTI -0.103, SmallNORB-azi -0.046. This breakdown is descriptive only; it does not add a new hypothesis test beyond the paired protocol already reported in Table 14.

H Structural-Baseline Training and Search Protocol

All structural controls in Table 3 share the same global head optimization protocol documented in Appendix L (AdamW, LR 10^{-3}, weight decay 10^{-4}, 100 epochs, 10-epoch warmup, batch size 64, grad clip 5.0, seeds $\{41, 42, 43\}$, CrossEntropyLoss). Architecture axes not under intentional ablation are inherited from the selected θ^* (including d=512, Q=1, intra 1×4, inter 1×1, `all_available`, and default attention dropout 0.2). Only HIFA itself received an independent sequential architecture search on the six $\mathcal{D}_{\mathrm{dev}}$ tasks; weighted-sum, flat, linear-inter, final-only, and matched final-token were evaluated as post-θ^* controls with *no* independent development-set retuning of learning rate, epochs, or architecture width/depth. Thus the controls are optimization-protocol matched and θ^*-architecture inherited, but they

Table 12: Combined QK leaderboard (top entries; multi-seed development). Rank 1 is HIFA-QK.

Rank	Entry	Val mean	Extra
1	head-dynamic × none	**0.7036**	6156
2	layer-static × none	0.6986	396
3	layer-static × z-score	0.6975	792
4	HIFA (no overlay)	0.6962	0
5	head-static × none	0.6942	12

Table 13: Per-task test accuracy (HIFA vs. HIFA-QK).

Task	HIFA	HIFA-QK	Δ
Caltech101	0.908	**0.923**	+0.015
CIFAR-100	**0.775**	0.767	−0.007
CLEVR-count	0.873	**0.893**	+0.020
CLEVR-dist	0.608	**0.611**	+0.003
Retinopathy	0.749	**0.756**	+0.007
DMLab	**0.497**	**0.497**	+0.001
dSprites-loc	0.665	**0.696**	+0.031
dSprites-ori	**0.570**	0.560	−0.010
DTD	0.794	**0.796**	+0.001
EuroSAT	0.949	**0.957**	+0.009
KITTI-dist	0.608	**0.616**	+0.008
Flowers102	**0.996**	0.995	−0.001
Pets	**0.955**	0.951	−0.004
Camelyon	0.832	**0.838**	+0.006
RESISC45	0.895	**0.896**	+0.000
SmallNORB-azi	0.224	**0.279**	+0.054
SmallNORB-elev	**0.457**	0.451	−0.006
SUN397	**0.579**	0.556	−0.023
SVHN	**0.718**	**0.718**	+0.001

are *not* parameter- or search-budget-matched to HIFA (Limitation in Section 7). Artefact: `results/statistics/structural_baseline_training_audit.csv`.

I Development vs. Non-Development Task Identities

Table 18 reports mean overall accuracy on the six $\mathcal{D}_{\text{dev}}$ task identities used for architecture selection, on the thirteen task identities not used during selection, and on all 19. This split is post-hoc and descriptive: no architecture or checkpoint was selected using the thirteen-task subset. On the thirteen non-development identities, HIFA remains numerically ahead of concat, packed-final concat, final-only, and flat (0.727 vs. 0.668 / 0.679 / 0.704 / 0.707); paired tests for this subset are in Table 15. On the six development task identities, final-test differences among HIFA, weighted-sum, and flat are small.

J Representation Flow Diagnostics

How do intermediate representations evolve across depth within each frozen encoder? We analyze three representative VTAB tasks spanning Natural, Specialized, and Structured domains (Flowers102, Patch Camelyon, and dSprites orientation) with DINOv3, MAE, and I-JEPA as diagnostic proxies (not the full 19-task suite). For layer states $h_i \in \mathbb{R}^D$, let $r_i = \max(\|h_i\|_2, \varepsilon)$ and $\hat{h}_i = h_i / r_i$. We report directional similarity and log-radial trajectory relative to the first layer,

$$C_{ij} = \hat{h}_i^\top \hat{h}_j, \qquad \Delta_\ell^{\text{rad}} = \log(\|h_\ell\|_2 / \|h_1\|_2). \tag{9}$$

Figure 3 shows Δ_ℓ^{rad} across depth for each backbone; pairwise C_{ij} heatmaps appear in Figure 4.

Norm trajectories are strongly depth-, backbone-, and task-dependent, motivating learned aggregation across depth rather than relying on the final representation alone. Because Representation Flow uses encoder hidden states before projection and key maps, $\|h_\ell\|_2 \neq \|k_\ell\|_2$ in general: the diagnostic motivates geometry-aware fusion but does not measure attention-key scale directly. All

Table 14: Paired tests: HIFA versus structural controls (19 tasks). Positive $\bar{\Delta}$ means HIFA is higher. None of these comparisons reaches Wilcoxon $p<0.05$ on the full 19-task set.

Control	$\bar{\Delta}$	95% CI	Wilcoxon p	W/T/L
Final-only	$+0.0167$	$[-0.0063, +0.0442]$	0.196	11/3/5
Matched final-token	$+0.0144$	$[-0.0118, +0.0478]$	0.396	10/2/7
Weighted-sum	$+0.0146$	$[+0.0017, +0.0297]$	0.123	11/1/7
Flat	$+0.0134$	$[+0.0003, +0.0300]$	0.123	11/3/5
Inter-linear	$+0.0042$	$[-0.0087, +0.0183]$	0.709	10/1/8
Intra-only (DINOv3)	$+0.0100$	$[-0.0009, +0.0228]$	0.169	10/4/5
Concat + linear	$+0.0582$	$[+0.0093, +0.1152]$	0.182	11/1/7
Packed-final concat	$+0.0474$	$[+0.0020, +0.1012]$	0.441	8/1/10

Table 15: Paired tests: HIFA versus structural controls on the 13 non-development task identities (nominal Wilcoxon p). Positive $\bar{\Delta}$ means HIFA is higher. Multiplicity-corrected p-values for this family are discussed above; artefacts: `results/paired_tests/ablations/paired_nondev13.csv`.

Control	$\bar{\Delta}$	95% CI	Wilcoxon p	W/T/L
Final-only	$+0.0229$	$[-0.0091, +0.0607]$	0.244	7/3/3
Matched final-token	$+0.0216$	$[-0.0159, +0.0689]$	0.376	7/2/4
Weighted-sum	$+0.0233$	$[+0.0071, +0.0416]$	0.033	9/1/3
Flat	$+0.0205$	$[+0.0027, +0.0429]$	0.048	9/2/2
Inter-linear	$+0.0066$	$[-0.0097, +0.0249]$	0.685	7/1/5
Intra-only (DINOv3)	$+0.0160$	$[+0.0023, +0.0327]$	0.080	8/3/2
Concat + linear	$+0.0592$	$[+0.0038, +0.1181]$	0.216	7/1/5
Packed-final concat	$+0.0480$	$[-0.0017, +0.1012]$	0.305	6/1/6

quantities are means over the 1,000-image train+val pool. Euclidean displacement $D_{ij} = \|h_i - h_j\|_2$ is determined by the norms r_i, r_j and cosine C_{ij}, so we omit separate D and radial heatmaps.

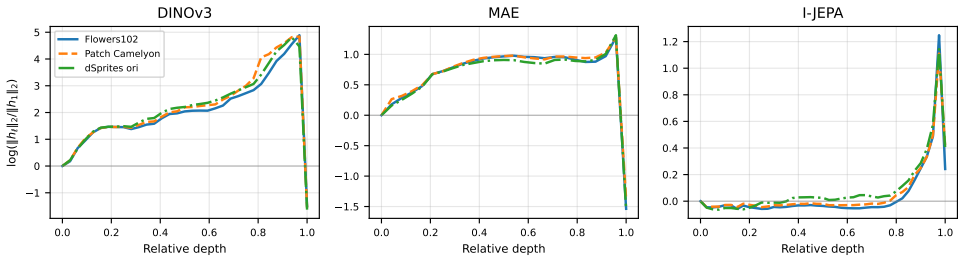


Figure 3: Representation geometry across depth. Log-radial trajectories $\Delta_\ell^{\mathrm{rad}} = \log(\|h_\ell\|_2/\|h_1\|_2)$ for DINOv3, MAE, and I-JEPA on representative Natural, Specialized, and Structured VTAB tasks. Each panel uses its own vertical scale so that within-backbone evolution is readable; absolute magnitudes are not compared directly across encoders.

K Cross-Backbone Representation Similarity

Do different encoders encode redundant information at corresponding depths? We measure linear CKA (Kornblith et al., 2019) between layer activations of DINOv3, MAE, and I-JEPA on the same three VTAB proxies (Figure 5; peak/mean values in Table 19).

$$\mathrm{CKA}(X, Y) = \frac{\|X_c^\top Y_c\|_F^2}{\|X_c^\top X_c\|_F \, \|Y_c^\top Y_c\|_F}, \tag{10}$$

with centered feature matrices X_c, Y_c.

Table 16: Per-seed paired Δ (HIFA $-$ matched final-token) and task-mean Δ. Positive means HIFA higher.

Task	Group	Seed 41	Seed 42	Seed 43	Mean Δ
CIFAR-100	Natural	$+0.004$	$+0.013$	$+0.009$	$+0.009$
Caltech101	Natural	-0.017	-0.016	-0.024	-0.019
DTD	Natural	$+0.005$	$+0.015$	$+0.014$	$+0.011$
Flowers102	Natural	$+0.001$	-0.001	$+0.001$	$+0.000$
Pets	Natural	$+0.003$	-0.001	$+0.001$	$+0.001$
SUN397	Natural	$+0.016$	-0.002	$+0.011$	$+0.008$
SVHN	Natural	$+0.247$	$+0.261$	$+0.239$	$+0.249$
Camelyon	Specialized	$+0.003$	$+0.000$	-0.008	-0.001
EuroSAT	Specialized	$+0.003$	$+0.002$	$+0.004$	$+0.003$
RESISC45	Specialized	-0.007	$+0.001$	-0.000	-0.002
Retinopathy	Specialized	-0.008	-0.015	$+0.017$	-0.002
CLEVR-count	Structured	-0.007	$+0.077$	-0.004	$+0.022$
CLEVR-dist	Structured	-0.039	$+0.011$	-0.030	-0.020
DMLab	Structured	$+0.002$	$+0.004$	-0.005	$+0.000$
KITTI	Structured	-0.087	-0.124	-0.097	-0.103
SmallNORB-azi	Structured	-0.031	-0.044	-0.063	-0.046
SmallNORB-ele	Structured	$+0.115$	$+0.124$	$+0.107$	$+0.115$
dSprites-loc	Structured	$+0.040$	$+0.057$	-0.042	$+0.019$
dSprites-ori	Structured	$+0.047$	$+0.041$	-0.005	$+0.028$

Table 17: Training/search treatment of structural baselines (descriptive audit of configs and expansion code; no new runs).

Baseline	Training protocol	Arch. from θ^*?	Independently tuned axes	Search budget
HIFA (θ^*)	Global AdamW protocol	Selected on $\mathcal{D}_{\text{dev}}$	All θ^* search axes	Full sequential $\mathcal{D}_{\text{dev}}$ search
Matched final-token	Same global protocol	Yes (full θ^*; memory length one)	None	None (post-θ^*)
Final-only	Same global protocol	Shared axes; `final_only` mode	None	None (post-θ^*)
Weighted-sum	Same global protocol	Yes; weighted-sum mode	None	None (post-θ^*)
Flat single-head	Same global protocol	Shared d/Q/depth; flat mode	None	None (post-θ^*)
Intra+linear inter	Same global protocol	Intra from θ^*; linear inter	None	None (post-θ^*)

Cross-backbone similarity is domain-dependent: Patch Camelyon shows high alignment (peaks ≈ 0.95), whereas Flowers102 is moderate and dSprites exhibits low mean alignment for DINO–MAE / DINO–I-JEPA while I-JEPA–MAE remains high early in depth. This depth structure is ignored by last-layer-only fusion. Low CKA does not by itself prove functional complementarity; utilization is interpreted via fusion accuracy and test-time masking dependence in the main text.

L Implementation Details

VTAB-1k tasks by group: Natural: CIFAR-100, Caltech101, DTD, Flowers102, Pets, SUN397, SVHN; Specialized: Patch Camelyon, EuroSAT, RESISC45, Retinopathy; Structured: CLEVR-count, CLEVR-dist, dSprites-loc, dSprites-ori, SmallNORB-azi, SmallNORB-elev, DMLab, KITTI-dist.

Frozen checkpoints and preprocessing. Cached features are extracted with Hugging Face Transformers processors bound to the checkpoint IDs in `config/config.yaml` (`AutoImageProcessor`+`AutoModel` for DINOv3/MAE; `AutoProcessor`+`AutoModel` for I-JEPA). Each extractor calls the official processor with `return_tensors="pt"` and applies *no* project-level resize/crop/normalize override. Table 20 lists the fields from each checkpoint's `preprocessor_config.json` (PIL `resample=2` $=$ bilinear). Center crop is not enabled (`do_center_crop`/`crop_size` absent or null). Images are resized to a square 224×224 canvas, rescaled by $1/255$, then normalized.

Table 18: Mean overall accuracy by task-identity scope (seeds $\{41, 42, 43\}$). $\mathcal{D}_{\text{dev}}$: n=6 selection identities; Non-dev: n=13 identities unused in architecture search.

Method	$\mathcal{D}_{\text{dev}}$ (n=6)	Non-dev (n=13)	All (n=19)
Concat + linear	0.644	0.668	0.660
Packed-final concat	0.654	0.679	0.671
Final-only HIFA	0.696	0.704	0.702
Matched final-token	0.701	0.706	0.704
Weighted-sum	0.704	0.704	0.704
Flat	0.702	0.707	0.705
Inter-linear	0.701	0.721	0.714
HIFA	0.700	0.727	0.719
HIFA-QK	**0.710**	**0.731**	**0.724**

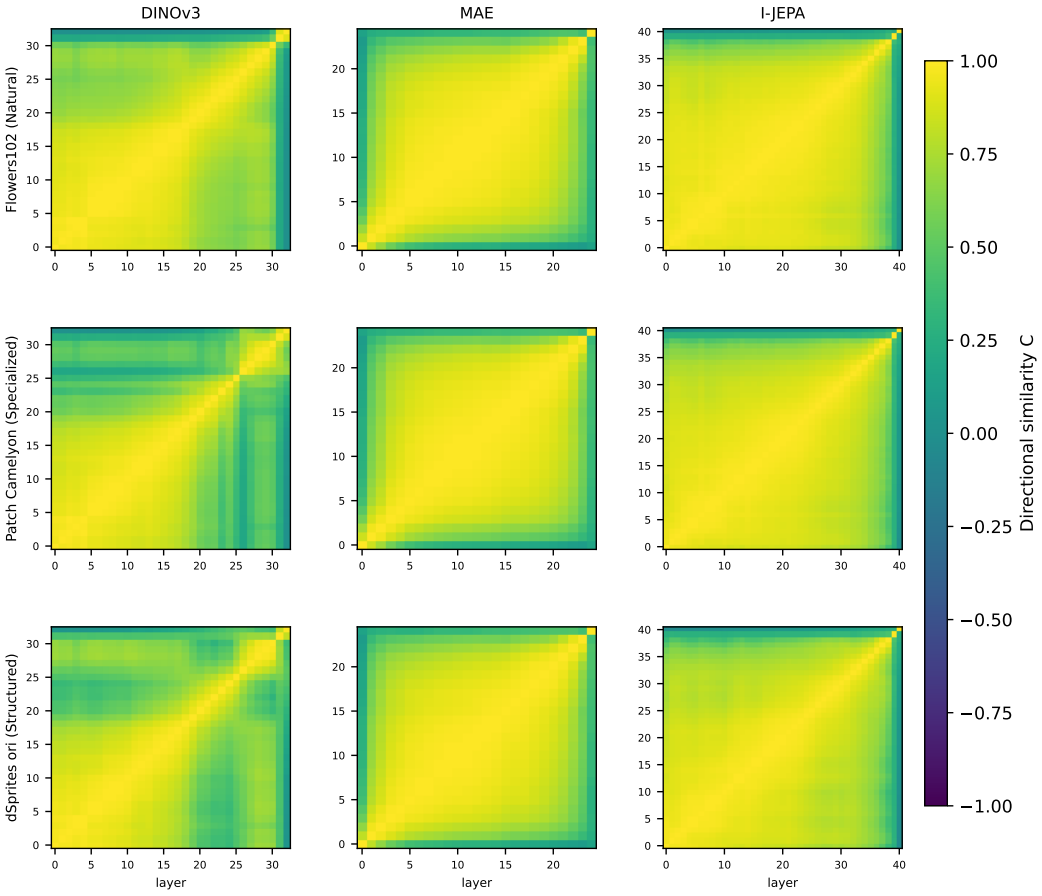


Figure 4: Mean directional similarity $C_{ij} = \hat{h}_i^\top \hat{h}_j$ across depth for each backbone and proxy task.

Extraction caches per-image NPZs with `layers_cls`/`layers_mean` and final `patch_cls`/`patch_mean` as available; HIFA training loads `vector_policy=all_available` (≡ `full_full_mean`): DINOv3/MAE pack $[\text{CLS} \,\|\, \text{mean}]$ along the feature axis per native layer, I-JEPA uses mean only, with `append_final=True`.

Selected configuration: intra cross-attention (1 layer / 4 heads), d=512, independent Q=1, `all_available`, softmax intra weighting, no post-intra norm, inter cross 1×1. HIFA-QK uses dynamic head-wise temperature on intra attention with radial term disabled. Training: head-only AdamW with weight decay 10^{-4}, cosine decay, 10-epoch warmup, 100 epochs, LR 10^{-3}, batch size 64, gradient clipping at max-norm 5.0, and seeds $\{41, 42, 43\}$. AdamW (β_1, β_2) and ε are

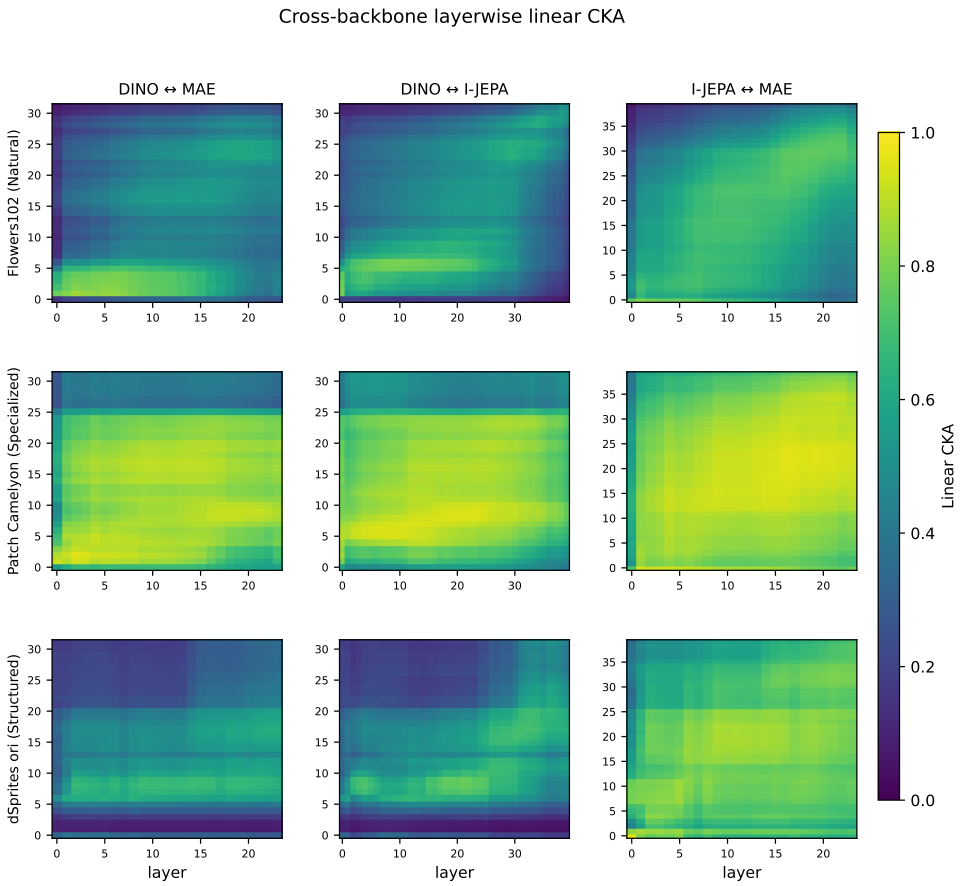


Figure 5: Layerwise linear CKA between encoder pairs on Natural (Flowers102), Specialized (Patch Camelyon), and Structured (dSprites orientation) proxies.

Table 19: Cross-backbone layerwise CKA: peak / mean over layer pairs.

Dataset	DINO–IJEPA	DINO–MAE	IJEPA–MAE
Flowers102	0.800 / 0.443	0.857 / 0.436	0.804 / 0.560
Camelyon	0.949 / 0.736	0.949 / 0.732	0.957 / 0.839
dSprites	0.772 / 0.398	0.686 / 0.382	0.966 / 0.726

the PyTorch defaults $(0.9, 0.999)$ and 10^{-8} (not overridden in config). The classification objective is `torch.nn.CrossEntropyLoss`. Eligibility for search requires successful validation on all six $\mathcal{D}_{\text{dev}}$ tasks.

Trainable parameter counts. Table 21 reports trainable parameters of the fusion head instantiated from θ^* under `all_available` widths, *excluding* the task classifier (class count varies across VTAB). Counts are from model instantiation only (no training). Concat+linear has no trainable fusion module before the classifier; its classifier parameters ($3712C+C$ for conventional concat; $6016C+C$ for packed-final concat) are excluded consistently with all other rows.

Table 20: Official Hugging Face preprocessor settings used for feature extraction (no project-level overrides).

Backbone	Checkpoint	Size	Resize / crop	Mean	Std
DINOv3	`facebook/dinov3-vith16plus-pretrain-lvd1689m`	224×224	resize; no center crop	$(0.485, 0.456, 0.406)$	$(0.229, 0.224, 0.225)$
MAE	`facebook/vit-mae-large`	224×224	resize; no center crop	$(0.485, 0.456, 0.406)$	$(0.229, 0.224, 0.225)$
I-JEPA	`facebook/ijepa_vitg16_22k`	224×224	resize; no center crop	$(0.5, 0.5, 0.5)$	$(0.5, 0.5, 0.5)$

Table 21: Trainable fusion-head parameters excluding the task classifier (d=512, Q=1). Concat variants have no fusion module before the classifier. Matched final-token instantiates the same modules as HIFA (identical count).

Variant	Params (excl. cls)
Concat + linear	0
Packed-final concat + linear	0
Final-only	4.13M
Weighted-sum	4.13M
Flat single-head	4.19M
DINOv3 intra-only	2.38M
Selected intra + linear inter	6.29M
Matched final-token	7.34M
HIFA	7.34M
HIFA-QK	7.35M

M Fusion Representation and Compute

Counts below are analytically estimated from the published ComBo adapter and from the executed HIFA configuration θ^* (`vector_policy=all_available`, `append_final=True`). They are not measured GPU memory or wall-clock runtime.

HIFA layout. DINOv3 and MAE pack CLS and mean by concatenating along the feature axis, yielding one token of dimension $2D_b$ per native layer; I-JEPA uses mean only. With the final representation appended, N_D=33 at D_D^{mem}=2560, N_M=25 at D_M^{mem}=2048, and N_J=41 at D_J^{mem}=1408. The raw fusion-input footprint is $N_{\text{HIFA}} = 33{\cdot}2560{+}25{\cdot}2048{+}41{\cdot}1408 = 193{,}408$ FP32 scalars (≈ 0.774 MB/image).

Projection MACs with d=512 are $\sum_b N_b D_b^{\text{mem}} d \approx 0.0990$ GMAC. Each intra/inter block is one multi-head attention layer with query/key/value/output projections; we count $2Nd^2{+}2Qd^2{+}2QNd$ and omit LayerNorm/dropout. With Q=1 this yields ≈ 0.0536 GMAC intra and ≈ 0.0021 GMAC inter, for ≈ 0.155 GMAC/image in total. Instantiating the head without training gives 7,342,592 trainable parameters excluding the task classifier.

ComBo layout. Four ViT-B encoders, 12 feature maps each, 196 spatial positions, native width 768: $N_{\text{ComBo}} = 4 \cdot 12 \cdot 196 \cdot 768 = 7{,}225{,}344$ scalars (≈ 28.90 MB/image). Adapter projection to d=128 costs $196 \cdot (48 \cdot 768) \cdot 128 \approx 0.925$ GMAC. A Transformer block with N=197 tokens is counted as $12Nd^2 + 2N^2d$; six blocks give ≈ 0.292 GMAC, hence ≈ 1.217 GMAC/image excluding backbone forward and the classifier. The ~ 5.93M adapter parameter figure follows the published ComBo specification, not a local measurement.

N Encoder-Suite Transfer

As a robustness check (not a comparison against ComBo), we freeze θ^* and retrain the hierarchical head on a different heterogeneous encoder suite (SAM, DFN-CLIP, SigLIP, DINOv2) without architecture re-search. This asks whether the HIFA architecture selected on DINOv3/MAE/I-JEPA transfers unchanged to another encoder set. Overall accuracy is 0.689 (Natural 0.805, Specialized 0.852, Structured 0.507), below the native-suite HIFA (0.719 / Structured 0.563). The hierarchi-

cal interface remains applicable without new search axes, but performance is sensitive to encoder composition and transfer of frozen θ^* across suites is not guaranteed.

O PUBLISHED VTAB-1K CONTEXT

Table 22 lists published overall VTAB-1k numbers from ComBo (Ramtoula et al., 2025) alongside our controlled scores. The encoder sets and probing protocols differ; the table is *not* a controlled ranking.

Table 22: Published VTAB-1k context (protocols are not matched). Head2Toe, SMP, and ComBo scores are from ComBo's re-evaluation / official protocol. HIFA / HIFA-QK are our Table 2 overall means.

Method	Backbone setting	Internal depth	Multi-bb	Overall
Head2Toe	ViT-B/16 IN21K	yes	no	70.9
SMP (Wu et al., 2024)	ViT-B/16 IN21K	yes	no	73.6
ComBo single	ViT-B/16 IN21K	yes	no	74.6
ComBo all-four	four FMs	yes	yes	78.2
ComBo top-two	top-2 FMs / task	yes	yes	78.6
HIFA	DINOv3+MAE+I-JEPA	yes	yes	71.9
HIFA-QK	DINOv3+MAE+I-JEPA	yes	yes	72.4