RARE BUT SUFFICIENT: EXTREMELY SPARSE NATIVE SUPPORTS IN FROZEN VISION REPRESENTATIONS

Anonymous authors
Paper under double-blind review

ABSTRACT

Frozen vision foundation models provide dense representations for downstream segmentation, yet these representations are typically used in full. Random-subset analyses reveal substantial diffuse redundancy. Their average-case results, however, leave open whether rare native supports can sustain unusually strong segmentation performance at extreme sparsity. To study this question without changing the representation basis, we define a support as a fixed dataset-level subset of the original spatial feature maps. We physically remove all other channels, discard the discovery model, and train a fresh decoder from scratch. Because exhaustive enumeration is infeasible, we use a coarse-to-fine backward search with fresh-decoder acceptance to identify candidate supports. On ACDC with frozen DINOv3 features, four of 384 native channels achieve a development Dice of 0.787, compared with 0.807 for the full representation. None of 512 uniformly sampled four-channel supports reaches this score. On unseen patients, the fixed support retains 95.2% of full-decoder Dice. Across ten additional discovery runs, independently selected supports contain only five to eight channels, each retaining at least 95% of its corresponding full-decoder test Dice. After excluding the four coordinates of the primary ACDC support from the search space, a separate search recovers a 13-channel support with comparable performance, demonstrating that high-performing native supports are not unique. The same qualitative sparse-support phenomenon recurs on COCO, ADE20K, M&Ms, MSD Liver, and G1020, spanning distinct anatomical targets, imaging settings, and visual domains. With a freshly trained decoder, the ACDC support also transfers directly to M&Ms, showing that sparse native-support structure can persist across related datasets. Together, these results indicate that redundancy in frozen dense representations is highly structured: broadly distributed task information can coexist with rare, non-unique, highly capable native supports. Code will be released on GitHub after the review process.

1 INTRODUCTION

Frozen vision encoders provide spatial features for correspondence, object discovery and segmentation (Caron et al., 2021; Amir et al., 2022; Oquab et al., 2024; Siméoni et al., 2025). Random-subset analyses offer a view of how task-relevant information is distributed within pretrained representations: uniformly sampled coordinate subsets recover much of the full-representation performance beyond a task- and pretraining-dependent width (Nanda et al., 2023). This pattern reveals substantial diffuse redundancy, with predictive information accessible from many different subsets of the representation. Diffuse redundancy, however, does not fully characterize the predictive capability of small coordinate subsets. The performance of a typical subset at a given width need not reflect that of a particular combination at the same width. Poor average performance at extreme sparsity can therefore coexist with rare, highly predictive subsets. Findings of sparse predictive coordinates in language models and frozen medical-image encoders further motivate examining these combinations in spatial representations (Durrani et al., 2023; Nooralahzadeh et al., 2026).

In this paper, we study native supports to characterize predictive structure beyond random-subset averages. A native support is a fixed dataset-level subset of the encoder's original spatial feature maps. Every image retains the same channel coordinates, and the representation basis remains unchanged. For each support, we remove all other maps, discard the discovery model, and train a fresh decoder from scratch using only the retained maps (Figure 1). The matched decoder using all channels,

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.

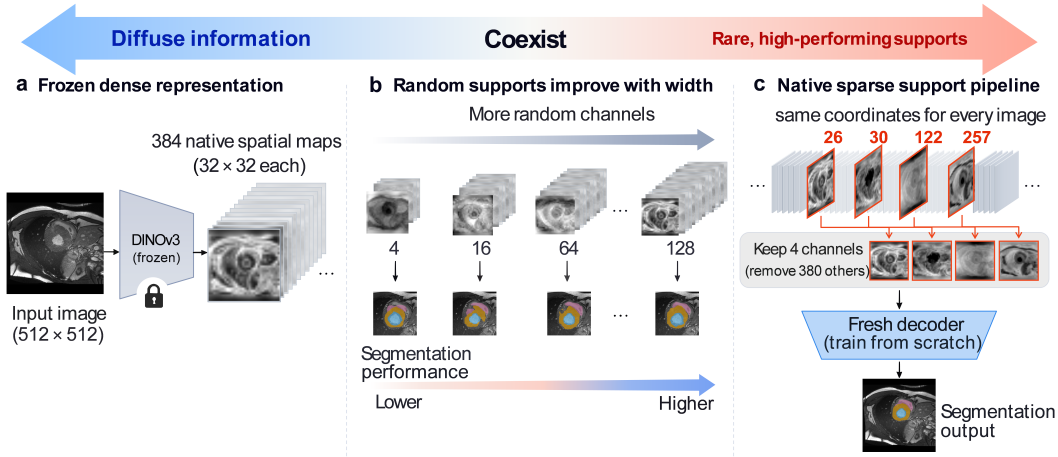

Figure 1: **Diffuse information and rare sparse supports are compatible. a**, A frozen encoder produces a dense collection of native spatial feature maps. **b**, Random-support performance typically improves as more native maps are retained. **c**, A native support fixes the same channel coordinates for every image; all other maps are removed and a fresh decoder is trained from scratch using only the retained maps. Our experiments test whether rare sparse supports can remain highly predictive despite the broad random-support trend.

denoted Full, provides the performance reference, while uniformly sampled same-width supports characterize how unusual the selected support is.

Our experiments show that broadly distributed task information coexists with rare, highly predictive native supports at extreme sparsity. On ACDC with frozen DINOv3 features, a four-channel support (4/384 native maps) reaches 0.772 test Dice versus 0.811 for the matched Full decoder, retaining 95.2%. In the development analysis used for discovery, the same support reaches 0.787 Dice versus 0.807 for Full, and none of 512 uniformly sampled four-channel supports reaches its score.

We identify candidate supports using Patience-Annealed Backward Feature Elimination (PA-BFE), a coarse-to-fine search that ranks retained maps by current-decoder knockout-loss increments and accepts deletions using fresh-decoder development performance. Ten additional searches recover 5–8-channel supports, each retaining at least 95% of its same-seed Full test Dice. High-performing supports are also non-unique: a separate search excluding the original four coordinates yields a disjoint 13-channel support with 0.773 test Dice versus 0.811 for Full, retaining 95.3%. Complete single-deletion audits reject every immediate child of both supports under the development acceptance criterion. Across M&Ms, MSD Liver, G1020, COCO and ADE20K, fixed sparse supports also retain most of their matched Full decoders' test Dice. Comparisons across datasets, frozen encoders and decoder architectures show that the support widths found depend strongly on the setting. The fixed ACDC coordinates also transfer directly to M&Ms with a freshly trained target decoder. Together, these results distinguish the performance of typical coordinate subsets from the predictive capability of rare native combinations, providing a more detailed account of how task-relevant information is organized in frozen visual representations.

2 Related Work

2.1 Redundancy and Sparse Predictive Structure

Our work builds on analyses of redundancy in pretrained representations. Representation-similarity studies provide a geometric perspective on redundancy, with Nguyen et al. (2021) examining how network width and depth affect hidden representations across architectures. At the coordinate level, Nanda et al. (2023) evaluate randomly sampled subsets of different widths using independently trained probes and representation-similarity measures. They find that the width needed to recover most of the full representation's predictive performance depends on the task and pretraining.

Coordinate subsets can also be selected for task relevance. Linguistic correlation analysis identifies predictive neurons in language models (Durrani et al., 2023). In frozen CT encoders, finding-specific coordinate scores and sparse mean-difference probes support radiological classification (Nooralahzadeh et al., 2026). We study this coordinate-level structure in segmentation, where each retained coordinate is a complete spatial feature map. We compare fixed selected supports with same-width random supports to distinguish typical subset performance from the predictive capability of particular combinations of native maps.

2.2 Dense visual representations and channel selection

Frozen patch features provide the spatial information needed to extend coordinate-level analysis to segmentation (Caron et al., 2021; Amir et al., 2022; Oquab et al., 2024; Siméoni et al., 2025). Register tokens and feature upsampling address the spatial organization and resolution of these features (Darcet et al., 2024; Fu et al., 2024). Segmentation methods exploit feature affinities, distill correspondences, or learn from masks derived from self-supervised features (Wang et al., 2022; Melas-Kyriazi et al., 2022; Hamilton et al., 2022; Wang et al., 2023). In-context methods use annotated examples to guide prediction (Cuttano et al., 2026).

Feature maps can also be adapted or combined for a task through learned adapters and projections, feature queries, and sparse feature fusion (Gao et al., 2025; Bi et al., 2025; Wu et al., 2024). OIParts explicitly selects DINOv2 and diffusion channels for each part using within-part variation in an annotated reference (Dai et al., 2024). Its selected channels define the feature similarities used to transfer part labels. In our setting, the same native channel indices are retained for every image, and the selected maps form the entire input to a freshly trained segmentation decoder. This evaluates their joint predictive capability.

2.3 Feature selection, model sparsity, basis changes and multiple supports

Feature selection provides a methodological starting point for native-support discovery. Stochastic Gates and LassoNet select input variables while learning a nonlinear predictor (Yamada et al., 2020; Lemhadri et al., 2021). These methods impose sparsity on the predictor's inputs. Weight pruning and sparse-network training act on model parameters and connectivity (Sanh et al., 2020; Lee et al., 2021; Evci et al., 2020), while structural pruning removes filters or coupled parameter groups (Chen et al., 2021; Sui et al., 2021; Fang et al., 2023). We use input selection to study the frozen representation: we discover subsets of its output maps and evaluate their predictive capability with fresh decoders.

The interpretation of input sparsity also depends on how those inputs are defined. A small set of concepts or learned factors can provide a compact representation without being a subset of the original channels. Concept bottleneck approaches use intermediate concept representations for prediction (Koh et al., 2020; Yuksekgonul et al., 2023), and CRAFT obtains concept directions by factorizing activations (Fel et al., 2023). Matryoshka Representation Learning trains a representation whose nested coordinate prefixes support prediction at different widths (Kusupati et al., 2022). These approaches construct or train the representation used for analysis or prediction. We examine sparsity within an existing native basis, keeping that basis fixed throughout selection and evaluation.

Once the basis is fixed, a further question is whether a predictive support is unique. Multiple predictive sets have been studied in Markov-boundary discovery and Bayesian families of near-optimal subsets (Statnikov et al., 2013; Kowal, 2022). We examine this multiplicity in segmentation by excluding all coordinates of the primary support and searching for a disjoint alternative with comparable predictive performance. Separately, complete single-deletion audits assess the local irreducibility of the resulting supports under the development acceptance criterion.

3 Probing Native Support Structure

3.1 Native Supports and Fresh-Decoder Evaluation

For an image x, a frozen encoder f produces a spatial representation $X(x) = f(x) \in \mathbb{R}^{D\times H\times W}$. Each native channel is a complete $H \times W$ spatial feature map. A dataset-level support $S \subseteq [D]$, where $[D] = \{0, \ldots, D-1\}$, selects the same channel coordinates for every image. The corresponding

representation is $X_S(x) = X(x)[S, :, :]$. Let $\mu \in \mathbb{R}^D$ be the channelwise mean computed from the training features. The mean-centered support input is $\widetilde{X}_S(x) = X_S(x) - \mu_S$, where μ_S is broadcast over spatial locations.

Support extraction gathers the retained maps from the frozen representation, without projection or channel mixing. Channels outside S are removed, leaving an input tensor of width $|S|$. We discard any discovery model and train a fresh decoder g_S from scratch on $\widetilde{X}_S$ under a fixed evaluation protocol $\mathcal{C}$. Within each comparison, $\mathcal{C}$ fixes the training, development and test splits, decoder architecture, optimization, training budget, checkpoint-selection rule, evaluation metric, and decoder seed.

For an evaluation split E, let $M^E_{\mathcal{C}}(S)$ denote the Dice of g_S on that split. The matched Full decoder uses all D native channels under the same protocol. We report the retained channel fraction $|S|/D$ and the Full-relative Dice retention:

$$R^E_{\mathcal{C}}(S) = \frac{M^E_{\mathcal{C}}(S)}{M^E_{\mathcal{C}}([D])}. \tag{1}$$

We use $E = \mathrm{dev}$ during discovery and $E = \mathrm{test}$ for final evaluation. Throughout the paper, *sufficient* describes supports that preserve most of the matched Full decoder's Dice.

3.2 Discovering Candidate Supports

We search for a compact support by first obtaining a suitable starting set and then progressively deleting channels. Candidate confirmation and deletion decisions use the same development-performance criterion. For a tolerance $\epsilon_{\mathrm{search}} > 0$, the acceptance floor is:

$$\tau_{\mathcal{C}} = M^{\mathrm{dev}}_{\mathcal{C}}([D]) - \epsilon_{\mathrm{search}}. \tag{2}$$

A nonempty support S passes when $M^{\mathrm{dev}}_{\mathcal{C}}(S) \geq \tau_{\mathcal{C}}$. The floor is fixed throughout a search.

Obtaining a starting support. We develop Dense-STG, a map-level adaptation of Stochastic Gates (STG) (Yamada et al., 2020), to propose compact supports from frozen dense representations. For each native channel $j \in [D]$, a learnable parameter $\alpha_j \in \mathbb{R}$ controls one stochastic scalar gate. The gate is shared across spatial locations and examples, so it acts on a complete feature map rather than individual spatial activations. With the encoder frozen, we jointly optimize the gate parameters and a segmentation decoder on the training data, adding a sparsity penalty summed over channels.

After training, we form deterministic gate values $\bar{z}_j = \mathrm{clip}(\alpha_j, 0, 1)$ and retain the channels whose values are at least 0.5: $S_{\mathrm{gate}} = \{j \in [D] : \bar{z}_j \geq 0.5\}$. This yields a candidate support. We discard the gates and jointly trained decoder, physically gather the selected maps, and evaluate the candidate under the fresh-decoder protocol defined above. A nonempty candidate that meets the criterion can serve as the starting support S_{init}. Searches initialized directly from Full instead use $S_{\mathrm{init}} = [D]$.

Ranking deletion proposals. We reduce S_{init} using Patience-Annealed Backward Feature Elimination (PA-BFE), which first tries coarse deletions and then performs local search. To choose which maps to try removing, PA-BFE measures how the current decoder's development loss changes when each map is separately set to zero. For a current support S and native index $i \in S$, let $\widetilde{X}_S^{(i \leftarrow 0)}$ denote the mean-centered input with map i zeroed at every spatial location. Let $\mathcal{L}_{\mathrm{dev}}(\widetilde{X}; g)$ be the segmentation loss of decoder g on the development set with input $\widetilde{X}$. With g_S fixed in evaluation mode and the input width unchanged, we compute:

$$d_i(S) = \mathcal{L}_{\mathrm{dev}}\left(\widetilde{X}_S^{(i \leftarrow 0)}; g_S\right) - \mathcal{L}_{\mathrm{dev}}\left(\widetilde{X}_S; g_S\right). \tag{3}$$

Channels with smaller $d_i(S)$ are tried first, with ties broken by native index. This score determines the trial order, not whether a deletion is accepted.

Evaluating proposed deletions. A proposed deletion removes a nonempty proper subset A of S, leaving $S' = S \setminus A$. We accept the deletion when the reduced support's fresh-decoder score satisfies $M^{\mathrm{dev}}_{\mathcal{C}}(S') \geq \tau_{\mathcal{C}}$. Acceptance updates the support to S' and triggers a new ranking using its decoder $g_{S'}$; rejection leaves the support and decoder unchanged. For both starting-support confirmation and deletion trials, a previously completed fresh-decoder evaluation may be reused when its full evaluation configuration matches. Otherwise, we train a new decoder from scratch on the physically gathered, mean-centered input.

Coarse and local search. PA-BFE first tests multi-channel deletions using prefixes of the current ranking, in descending block sizes. Rejection advances to the next smaller block on the same parent, whereas acceptance restarts the coarse schedule on the updated support. Once all scheduled proposals are rejected, the search switches permanently to the local stage.

Local search begins with single-channel trials in ranking order, skipping deletions already rejected for the current parent. Each accepted deletion clears this failure record for the updated support. When consecutive accepted steps reach the current patience threshold, the controller tries a wider deletion using a ranking prefix. The search stops when only one channel remains or when every single-channel deletion from the current support has been rejected; the latter certifies single-deletion local irreducibility under $\mathcal{C}$ with respect to the development acceptance floor. Supplement A.1 details the gate parameterization and evaluation reuse conditions. Supplement A.2 gives the exact coarse schedule, and Algorithm S1 specifies the local-stage patience and backoff rules.

4 Experiments

4.1 Experimental setup

ACDC cardiac MRI is the primary setting for our support-space analysis, repeated discovery and decoder-seed comparisons (Bernard et al., 2018). Additional experiments examine setting dependence using M&Ms cardiac MRI, G1020 retinal segmentation, MSD Liver CT, ADE20K scene parsing with 150 semantic categories and COCO with 80 object categories (Campello et al., 2021; Bajwa et al., 2020; Antonelli et al., 2022; Zhou et al., 2017; Lin et al., 2014). We use frozen DINOv3 ViT-S/16, DINOv2 ViT-S/14, and SAM3 features (Siméoni et al., 2025; Oquab et al., 2024; Carion et al., 2026), with a compact U-Net as the default decoder (Ronneberger et al., 2015). Primary experiments use 32×32 labels. The finer-label ACDC comparison uses 128×128 supervision and evaluation while retaining the 32×32 native feature grid. Support-evaluation decoders are trained for 20,000 steps, with checkpoints selected by development Dice. Development sets are used for support discovery and checkpoint selection, while held-out test sets are reserved for final evaluation. Exact split sizes and construction details are provided in Supplementary Table S1 and Supplement A.1.

4.2 Rare but Sufficient at Extreme Sparsity

Extreme sparsity does not necessarily imply a corresponding loss of predictive performance. On ACDC with frozen DINOv3, the support $S^{(1)} = \{26, 30, 122, 257\}$ retains only 4 of 384 native maps (1.04%), yet reaches 0.787 development Dice compared with 0.807 for Full, retaining 97.5% of Full's score. The striking aspect of this result is its position within the random-support landscape. As shown in Figure 2(a), 512 uniformly sampled four-channel supports achieve only 0.644 ± 0.030 development Dice (population standard deviation), and none reaches $S^{(1)}$. Indeed, the four-channel support performs comparably to the mean of random supports containing 48–64 channels. Thus, at extreme sparsity, support composition can produce an order-of-magnitude difference in the number of channels needed to reach similar predictive performance.

The same separation persists on the held-out test set. $S^{(1)}$ reaches 0.772 test Dice versus 0.811 for Full, retaining 95.2%, whereas 64 same-width random supports average 0.628 test Dice (Supplementary Table S2). Consistent with the development comparison, the observed 0/512 count places a one-sided 95% binomial-reference upper bound of 0.583% on the probability that a uniformly sampled four-channel support reaches the development score of $S^{(1)}$ (Supplement A.1). Together, these results identify a highly atypical region of the native support space in which most of Full's predictive performance is concentrated into only a few original feature maps.

4.3 Repeated Discovery and Tolerance Dependence

High predictive retention at extreme sparsity is reproducible across independent searches. As shown in Figure 2(b), ten additional ACDC searches at $\epsilon_{\text{search}} = 0.02$ begin from supports ranging from 6 to 24 channels, yet converge to a much narrower range of 5–8-channel endpoints. As shown in Figure 2(c), all ten endpoints retain at least 95% of their same-seed Full test Dice. The recovered supports also exhibit substantial coordinate recurrence across runs. Channels 122 and 257 appear in all ten endpoints, while channels 26 and 30 appear in nine and eight, respectively. Thus, repeated

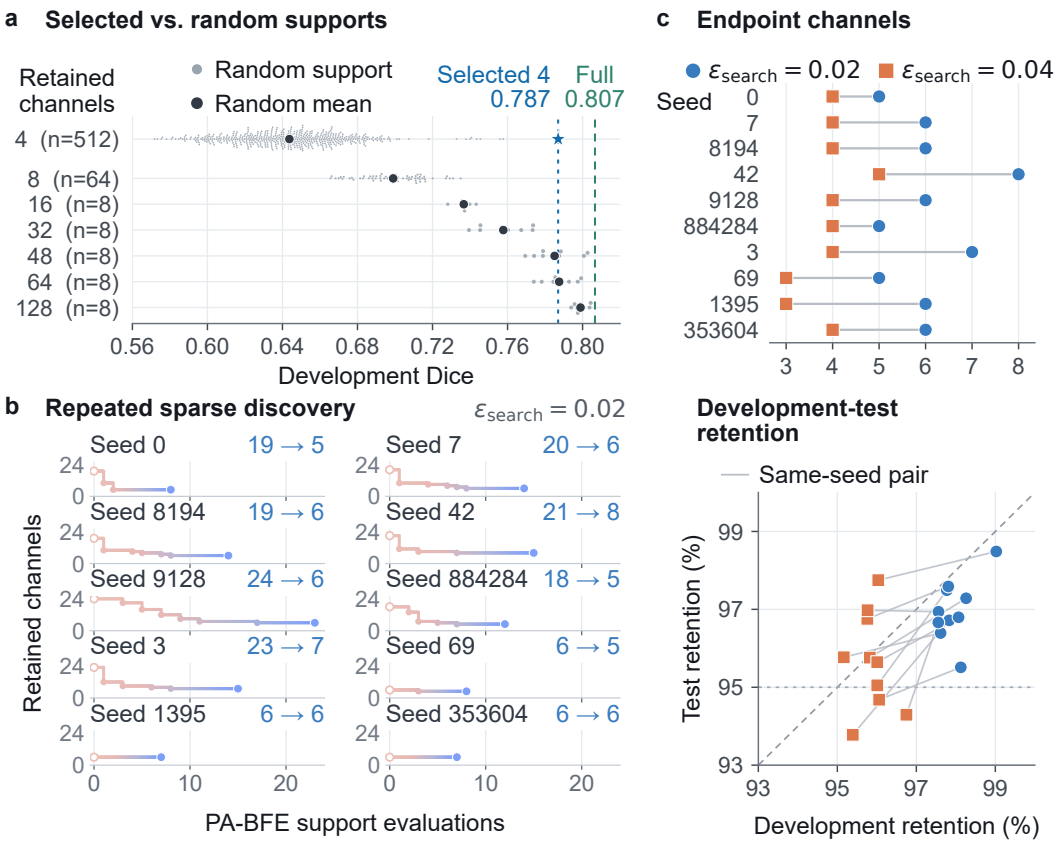

Figure 2: **Sparse-support performance, discovery, and tolerance dependence on ACDC. (a)**, Development Dice for random supports of different sizes, compared with the selected four-channel support and Full. **(b)**, Ten additional PA-BFE search trajectories at $\epsilon_{\text{search}} = 0.02$; steps indicate accepted deletions. **(c)**, Paired endpoint sizes (top) and development–test retention relative to same-seed Full (bottom) at $\epsilon_{\text{search}} = 0.02$ and 0.04. Gray lines connect same-seed pairs; dashed lines indicate equal retention and 95% test retention.

discovery does not simply reproduce the same support, but repeatedly identifies a compact family of high-retention supports with a strongly recurring core of native channels.

Figure 2(c) shows the tolerance trade-off. Increasing ϵ_{search} from 0.02 to 0.04 shifts the paired endpoints from 5–8 to 3–5 channels, while seven of ten looser-tolerance endpoints retain at least 95% of their same-seed Full test Dice. Thus, relaxing the development acceptance criterion yields smaller supports, but with less consistent retention on the test set. Supplement A.6 and Supplementary Table S4 provide the search details and individual endpoints.

4.4 Composition and Non-Uniqueness of Sparse Supports

Composition. The performance of $S^{(1)}$ does not arise from a single dominant channel. As shown in Figure 3(a), the strongest constituent, channel 257, achieves 0.717 development Dice, retaining 88.8% of Full's score, compared with 0.787 and 97.5% for $S^{(1)}$. Simple activation statistics also fail to recover the same performance: RMS Top-4 and variance Top-4 achieve 0.650 and 0.709 development Dice, respectively (test comparisons and paired intervals in Supplement A.5.1). The selected maps further carry image-dependent information that cannot be explained by a fixed spatial pattern. Replacing the four maps with their pixelwise training-set means reduces test Dice from 0.765–0.772 to 0.261–0.264 across decoder seeds 17, 23, and 41 (control details and the paired interval for seed 17 in Supplement A.5.2). Together, these controls show that the predictive capability

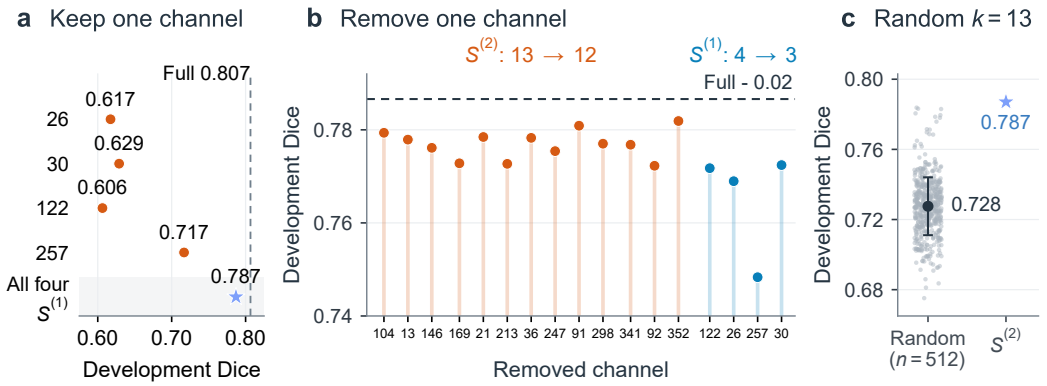

Figure 3: **Composition and non-uniqueness of sparse supports.** The supports satisfy $S^{(1)} \cap S^{(2)} = \varnothing$ by construction. **(a)**, Individual constituents versus the complete four-channel $S^{(1)}$ in development Dice; the dashed line marks Full. **(b)**, Single-channel deletions from $S^{(1)}$ (blue) and $S^{(2)}$ (orange). All children fall below the acceptance floor, Full $-\,0.02$. **(c)**, $S^{(2)}$ versus 512 uniformly sampled 13-channel supports from the full native basis in development Dice. The marker and bar show the random-support mean and population standard deviation; none reaches $S^{(2)}$.

of $S^{(1)}$ arises from the joint, image-dependent information carried by the selected maps, rather than from a single dominant channel, a simple activation ranking, or a fixed spatial template.

Non-Uniqueness. The repeated searches above show substantial coordinate recurrence, raising the question of whether the coordinates of $S^{(1)}$ are required for high sparse-support performance. To verify this, we excluded all four coordinates of $S^{(1)}$ from the search space and ran a separate search. This yielded $S^{(2)} = \{13, 21, 36, 91, 92, 104, 146, 169, 213, 247, 298, 341, 352\}$. Using 13 of 384 maps (3.39%), $S^{(2)}$ matches the 0.787 development Dice of $S^{(1)}$ and reaches 0.773 test Dice versus 0.811 for Full, retaining 95.3% (results across decoder seeds and test confidence intervals in Supplementary Table S3 and Supplement A.4.2).

The strong performance of $S^{(2)}$ is not explained simply by its larger width. As shown in Figure 3(c), 512 uniformly sampled 13-channel supports achieve a mean development Dice of 0.728 ± 0.016 (population standard deviation), and none reaches the 0.787 score of $S^{(2)}$ (additional random-support test comparisons in Supplement A.4.2). Thus, an entirely disjoint coordinate set can recover essentially the same predictive performance as $S^{(1)}$ while remaining highly atypical among same-width random supports.

Finally, both solutions are locally irreducible under single-channel deletion. As shown in Figure 3(b), complete deletion scans reject all four immediate children of $S^{(1)}$ and all thirteen of $S^{(2)}$ under Eq. (2). Thus, coordinate recurrence across searches does not imply a unique sparse solution: the frozen representation admits distinct disjoint high-performing supports, each locally irreducible under the development acceptance criterion.

4.5 Sparse Support Performance and Setting Dependence

Cross-setting test performance. The sparse-support phenomenon extends beyond the primary ACDC setting. Across M&Ms, MSD Liver, G1020, ADE20K and COCO, supports discovered with DINOv3 + U-Net retain 95.8–98.3% of their matched Full test Dice (Table 1), despite spanning cardiac MRI, liver CT, retinal imaging, and natural images. On ACDC, the same four-channel $S^{(1)}$ also retains 94.5% of Full test Dice when supervision and evaluation are increased from 32×32 to 128×128, while the native feature grid remains 32×32 (random controls and CIs in Supplement A.4.1).

Qualitative analysis. Figure 4 shows how the quantitative retention in Table 1 translates to the predicted masks. On ACDC, MSD Liver, and G1020, selected supports largely preserve the regions, shapes, and boundaries produced by Full, whereas same-width random supports recover the main

segmentation structure but exhibit larger boundary, shape, or regional errors. On COCO, selected supports similarly preserve object masks and category assignments, while random supports more often miss substantial regions or misassign categories. ADE20K exhibits the same qualitative pattern in complex scene parsing (Supplementary Figure S1). Thus, the retained performance of selected supports reflects preservation of Full's spatial predictions rather than merely similar aggregate Dice.

Table 1: **Test performance of sparse supports across datasets and evaluation settings.** Every row uses frozen DINOv3 features and a freshly trained U-Net. Retention compares the selected-support and matched Full-decoder Dice on the same test split.

Setting / support	k/D	Full	Selected	Retention (%)
ACDC, 32×32: $S^{(1)}$	4/384	0.811	0.772	95.2
ACDC, 32×32: $S^{(2)}$	13/384	0.811	0.773	95.3
ACDC, 128×128: $S^{(1)}$	4/384	0.842	0.796	94.5
M&Ms, native support	4/384	0.822	0.808	98.3
M&Ms, transferred ACDC $S^{(1)}$	4/384	0.822	0.797	97.0
MSD Liver	11/384	0.764	0.745	97.5
G1020	21/384	0.779	0.751	96.4
COCO	148/384	0.636	0.615	96.7
ADE20K	106/384	0.520	0.498	95.8

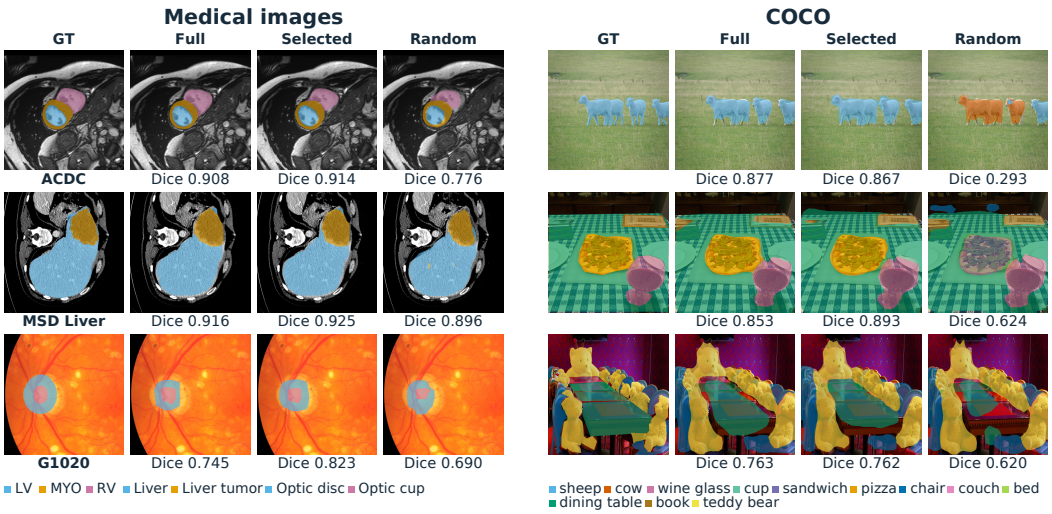

Figure 4: **Test-set segmentation examples from ACDC, MSD Liver, G1020, and COCO.** The left rows show ACDC, MSD Liver, and G1020; the right rows show COCO examples. Columns show ground truth, Full, selected, and a fixed same-width random support. Numbers report per-image Dice on the full original-image masks, averaged over all foreground classes for medical images and over foreground classes present in the ground truth for COCO.

Fixed-index transfer across datasets. The four-channel support discovered independently on M&Ms, $\{26, 30, 78, 257\}$, shares coordinates 26, 30, and 257 with the ACDC support $S^{(1)} = \{26, 30, 122, 257\}$. Keeping the complete ACDC support fixed and training a fresh decoder on M&Ms yields 0.797 test Dice, compared with 0.822 for Full (97.0% retention), as shown in Table 1. The support discovered directly on M&Ms reaches 0.808 test Dice (98.3% retention). Thus, a substantial part of the sparse native support is shared across the two cardiac datasets, and the complete ACDC support remains highly predictive after fixed-index transfer, although target-specific discovery achieves higher performance.

Support-width dependence. High-retention sparse supports recur across settings, but the degree of sparsity required to preserve performance varies substantially. Table 2 shows that selected supports remain close to their matched Full decoders even as the required support width changes by more

Table 2: **Support width and development performance across settings.** Random Dice is the mean over n same-width supports; the last column counts those matching or exceeding Selected.

Setting	Support k/D (%)	Full Dice	Selected Dice	Retention (%)	Random Dice (n)	Random $\geq$ selected
A. Dataset (DINOv3 + U-Net)						
ACDC	4/384 (1.04)	0.807	0.787	97.5	0.644 (512)	0/512
M&Ms	4/384 (1.04)	0.842	0.827	98.2	0.733 (64)	0/64
G1020	21/384 (5.47)	0.794	0.776	97.7	0.730 (64)	0/64
MSD Liver	11/384 (2.86)	0.788	0.769	97.6	0.636 (64)	0/64
COCO	148/384 (38.54)	0.618	0.603	97.6	0.584 (64)	0/64
ADE20K	106/384 (27.60)	0.528	0.509	96.4	0.484 (64)	0/64
B. Encoder (ACDC + U-Net)						
DINOv2	11/384 (2.86)	0.801	0.784	97.9	0.713 (64)	0/64
SAM3	4/256 (1.56)	0.834	0.819	98.2	0.780 (64)	0/64
C. Decoder (ACDC + DINOv3)						
Transformer	21/384 (5.47)	0.764	0.745	97.5	0.558 (64)	0/64
Pointwise MLP	67/384 (17.45)	0.784	0.764	97.4	0.692 (64)	0/64

than an order of magnitude. Across datasets, the narrowest supports are found in the two cardiac MRI settings, whereas broader supports are recovered for liver, retinal, and especially natural-image segmentation. This pattern is consistent with task-relevant information being less concentrated in the native representation for tasks with greater visual and semantic variation. The cardiac tasks involve a small set of recurring anatomical structures, whereas COCO and ADE20K span 80 object and 150 semantic categories, respectively, across highly heterogeneous scenes.

The encoder comparison suggests that this concentration is also representation-dependent. With the dataset and decoder fixed on ACDC, support widths vary across frozen encoders, indicating that pretrained representations organize the same task-relevant information differently across their native coordinates. The decoder comparison provides an even more controlled contrast, with both the task and frozen DINOv3 representation held fixed. Here, the pointwise MLP yields a substantially broader support than either U-Net or the Transformer. Unlike these decoders, it predicts independently at each spatial position and cannot aggregate information across neighboring or distant locations. Its broader support therefore indicates that decoder structure affects how efficiently information distributed across native channels can be combined for prediction. Support width is therefore not an intrinsic property of the frozen encoder alone: it reflects the interaction between the task, the organization of the frozen representation, and the downstream decoder. Every selected support in Table 2 also outperforms the evaluated same-width random supports in its own setting, showing that coordinate choice remains important across datasets, encoders, and decoders.

5 Conclusion

We study native supports in frozen vision representations and find that broad redundancy can coexist with rare, extremely sparse combinations that preserve most task-relevant predictive capability. Their advantage over same-width random subsets shows that particular native maps jointly carry image-dependent information that remains accessible to freshly trained decoders without changing the native basis. Recurring coordinates and disjoint high-performing alternatives reveal a structured but non-unique organization, with different channel combinations supporting comparable predictive capability. Across medical and natural-image segmentation, high-retention support width reflects interactions among task-specific information requirements, the frozen representation, and the downstream decoder. Our DINOv3 results further suggest that substantial task-relevant information may already be accessible through sparse combinations of unmodified native channels, motivating native-support analysis before or alongside learned adapters, projections, or alternative feature bases. Targeted support discovery complements random-subset analysis by characterizing highly predictive native subsets rather than only average robustness to channel removal. Understanding which supports remain effective across tasks and decoders may inform compact feature storage, reuse, and transfer.

AI USE STATEMENT

Generative AI tools were used to assist with and refine the writing. Experimental results and specific implementation details were provided by the authors. The authors are responsible for the final manuscript, references, and research conclusions.

REFERENCES

Shir Amir, Yossi Gandelsman, Shai Bagon, and Tali Dekel. Deep ViT features as dense visual descriptors. In *ECCV Workshop on What is Motion For?*, 2022. URL https://arxiv.org/abs/2112.05814v3.

Michela Antonelli, Annika Reinke, Spyridon Bakas, et al. The medical segmentation decathlon. *Nature Communications*, 13(1):4128, 2022.

Muhammad Naseer Bajwa, Gur Amrit Pal Singh, Wolfgang Neumeier, Muhammad Imran Malik, Andreas Dengel, and Sheraz Ahmed. G1020: A benchmark retinal fundus image dataset for computer-aided glaucoma detection. In *International Joint Conference on Neural Networks*, pp. 1–7, 2020.

Olivier Bernard, Alain Lalande, Clement Zotti, et al. Deep learning techniques for automatic MRI cardiac multi-structures segmentation and diagnosis: Is the problem solved? *IEEE Transactions on Medical Imaging*, 37(11):2514–2525, 2018. doi: 10.1109/TMI.2018.2837502.

Qi Bi, Jingjun Yi, Hao Zheng, Wei Ji, Yawen Huang, Yuexiang Li, and Yefeng Zheng. Learning generalized medical image representation by decoupled feature queries. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 47(12):11252–11269, 2025. doi: 10.1109/TPAMI.2025.3597364.

Victor M. Campello, Polyxeni Gkontra, Cristian Izquierdo, Carlos Martin-Isla, Alireza Sojoudi, Peter M. Full, Klaus Maier-Hein, Yao Zhang, Zhiqiang He, Jun Ma, Mario Parreno, Alberto Albiol, Fanwei Kong, Shawn C. Shadden, Jorge Corral Acero, Vaanathi Sundaresan, Mina Saber, Mustafa Elattar, Hongwei Li, Bjoern Menze, Firas Khader, Christoph Haarburger, Cian M. Scannell, Mitko Veta, Adam Carscadden, Kumaradevan Punithakumar, Xiao Liu, Sotirios A. Tsaftaris, Xiaoqiong Huang, Xin Yang, Lei Li, Xiahai Zhuang, David Vilades, Martin L. Descalzo, Andrea Guala, Lucia La Mura, Matthias G. Friedrich, Ria Garg, Julie Lebel, Filipe Henriques, Mahir Karakas, Ersin Cavus, Steffen E. Petersen, Sergio Escalera, Santi Segui, Jose F. Rodriguez-Palomares, and Karim Lekadir. Multi-centre, multi-vendor and multi-disease cardiac segmentation: The M&Ms challenge. *IEEE Transactions on Medical Imaging*, 40(12):3543–3554, 2021. doi: 10.1109/TMI.2021.3090082.

Nicolas Carion, Laura Gustafson, Yuan-Ting Hu, Shoubhik Debnath, Ronghang Hu, Didac Suris Coll-Vinent, Chaitanya Ryali, Kalyan Vasudev Alwala, Haitham Khedr, Andrew Huang, et al. SAM 3: Segment anything with concepts. In *International Conference on Learning Representations*, 2026.

Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021. URL https://arxiv.org/abs/2104.14294.

Tianyi Chen, Bo Ji, Tianyu Ding, Biyi Fang, Guanyi Wang, Zhihui Zhu, Luming Liang, Yixin Shi, Sheng Yi, and Xiao Tu. Only train once: A one-shot neural network training and pruning framework. In *Advances in Neural Information Processing Systems*, volume 34, 2021. URL https://proceedings.neurips.cc/paper/2021/hash/a376033f78e144f494bfc743c0be3330-Abstract.html.

Claudia Cuttano, Gabriele Trivigno, Christoph Reich, Daniel Cremers, Carlo Masone, and Stefan Roth. INSID3: Training-free in-context segmentation with DINOv3. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2026.

Zhenqi Dai, Ting Liu, Xingxing Zhang, Yunchao Wei, and Yanning Zhang. One-shot in-context part segmentation. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pp. 10966–10975, 2024. doi: 10.1145/3664647.3680989.

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

Nadir Durrani, Fahim Dalvi, and Hassan Sajjad. Discovering salient neurons in deep NLP models. *Journal of Machine Learning Research*, 24(362):1–40, 2023.

Utku Evci, Trevor Gale, Jacob Menick, Pablo Samuel Castro, and Erich Elsen. Rigging the lottery: Making all tickets winners. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 2943–2952. PMLR, 2020. URL https://proceedings.mlr.press/v119/evci20a.html.

Gongfan Fang, Xinyin Ma, Mingli Song, Michael Bi Mi, and Xinchao Wang. Depgraph: Towards any structural pruning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 16091–16101, 2023. URL https://openaccess.thecvf.com/content/CVPR2023/html/Fang_DepGraph_Towards_Any_Structural_Pruning_CVPR_2023_paper.html.

Thomas Fel, Agustin Picard, Louis Béthune, Thibaut Boissin, David Vigouroux, Julien Colin, Rémi Cadène, and Thomas Serre. CRAFT: Concept recursive activation FacTorization for explainability. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2711–2721, 2023. doi: 10.1109/CVPR52729.2023.00266. URL https://arxiv.org/abs/2211.10154.

Stephanie Fu, Mark Hamilton, Laura E. Brandt, Axel Feldmann, Zhoutong Zhang, and William T. Freeman. Featup: A model-agnostic framework for features at any resolution. In *International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=GkJiNn2QDF.

Yifan Gao, Haoyue Li, Feng Yuan, Xiaosong Wang, and Xin Gao. DINO U-Net: Exploiting high-fidelity dense features from foundation models for medical image segmentation. *arXiv preprint arXiv:2508.20909*, 2025. URL https://arxiv.org/abs/2508.20909v1.

Mark Hamilton, Zhoutong Zhang, Bharath Hariharan, Noah Snavely, and William T. Freeman. Unsupervised semantic segmentation by distilling feature correspondences. In *International Conference on Learning Representations*, 2022.

Pang Wei Koh, Thao Nguyen, Yew Siang Tang, Stephen Mussmann, Emma Pierson, Been Kim, and Percy Liang. Concept bottleneck models. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 5338–5348. PMLR, 2020. URL https://proceedings.mlr.press/v119/koh20a.html.

Daniel R. Kowal. Bayesian subset selection and variable importance for interpretable prediction and classification. *Journal of Machine Learning Research*, 23(108):1–38, 2022.

Aditya Kusupati, Gantavya Bhatt, Aniket Rege, Matthew Wallingford, Aditya Sinha, Vivek Ramanujan, William Howard-Snyder, Kaifeng Chen, Sham Kakade, Prateek Jain, and Ali Farhadi. Matryoshka representation learning. In *Advances in Neural Information Processing Systems*, volume 35, 2022.

Jaeho Lee, Sejun Park, Sangwoo Mo, Sungsoo Ahn, and Jinwoo Shin. Layer-adaptive sparsity for the magnitude-based pruning. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=H6ATjJ0TKdf.

Ismael Lemhadri, Feng Ruan, Louis Abraham, and Robert Tibshirani. Lassonet: A neural network with feature sparsity. *Journal of Machine Learning Research*, 22(127):1–29, 2021. URL https://www.jmlr.org/papers/v22/20-848.html.

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

Luke Melas-Kyriazi, Christian Rupprecht, Iro Laina, and Andrea Vedaldi. Deep spectral methods: A surprisingly strong baseline for unsupervised semantic segmentation and localization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022. URL https://arxiv.org/abs/2205.07839.

Vedant Nanda, Till Speicher, John P. Dickerson, Krishna P. Gummadi, Soheil Feizi, and Adrian Weller. Diffused redundancy in pre-trained representations. In *Advances in Neural Information Processing Systems*, volume 36, 2023.

Thao Nguyen, Maithra Raghu, and Simon Kornblith. Do wide and deep networks learn the same things? uncovering how neural network representations vary with width and depth. In *International Conference on Learning Representations*, 2021. URL https://arxiv.org/abs/2010.15327.

Farhad Nooralahzadeh, Lea Bogensperger, Christian Bluethgen, and Michael Krauthammer. Sparse concept channels in frozen 3d CT vision encoders. *arXiv preprint arXiv:2607.20993*, 2026.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, et al. DINOv2: Learning robust visual features without supervision. *Transactions on Machine Learning Research*, 2024.

Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-Net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention*, pp. 234–241, 2015.

Victor Sanh, Thomas Wolf, and Alexander M. Rush. Movement pruning: Adaptive sparsity by fine-tuning. In *Advances in Neural Information Processing Systems*, volume 33, pp. 20378–20389, 2020. URL https://proceedings.neurips.cc/paper/2020/hash/eae15aabaa768ae4a5993a8a4f4fa6e4-Abstract.html.

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

Alexander Statnikov, Nikita I. Lytkin, Jan Lemeire, and Constantin F. Aliferis. Algorithms for discovery of multiple Markov boundaries. *Journal of Machine Learning Research*, 14:499–566, 2013.

Yang Sui, Miao Yin, Yi Xie, Huy Phan, Saman Aliari Zonouz, and Bo Yuan. CHIP: Channel independence-based pruning for compact neural networks. In *Advances in Neural Information Processing Systems*, volume 34, pp. 24604–24616, 2021. URL https://proceedings.neurips.cc/paper/2021/hash/ce6babd060aa46c61a5777902cca78af-Abstract.html.

Xudong Wang, Rohit Girdhar, Stella X. Yu, and Ishan Misra. Cut and learn for unsupervised object detection and instance segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3124–3134, 2023. URL https://openaccess.thecvf.com/content/CVPR2023/html/Wang_Cut_and_Learn_for_Unsupervised_Object_Detection_and_Instance_Segmentation_CVPR_2023_paper.html.

Yangtao Wang, Xi Shen, Shell Xu Hu, Yuan Yuan, James L. Crowley, and Dominique Vaufreydaz. Self-supervised transformers for unsupervised object discovery using normalized cut. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022. URL https://arxiv.org/abs/2202.11539.

Boqian Wu, Qiao Xiao, Shiwei Liu, Lu Yin, Mykola Pechenizkiy, Decebal Constantin Mocanu, Maurice van Keulen, and Elena Mocanu. E2ENet: Dynamic sparse feature fusion for accurate and efficient 3d medical image segmentation. In *Advances in Neural Information Processing Systems*, volume 37, 2024.

Yutaro Yamada, Ofir Lindenbaum, Sahand Negahban, and Yuval Kluger. Feature selection using stochastic gates. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119, pp. 10648–10659, 2020.

Mert Yuksekgonul, Maggie Wang, and James Zou. Post-hoc concept bottleneck models. In *International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=nA5AZ8CEyow.

Bolei Zhou, Hang Zhao, Xavier Puig, Sanja Fidler, Adela Barriuso, and Antonio Torralba. Scene parsing through ADE20K dataset. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 633–641, 2017.

A SUPPLEMENTARY INFORMATION

This supplement provides implementation details, additional ACDC analyses, uncertainty estimates, and per-run discovery results.

A.1 IMPLEMENTATION DETAILS

Frozen encoders and feature extraction. For DINOv3 ViT-S/16 pretrained on LVD-1689M, we use the final-normalized patch features. Removing class and register tokens leaves 384 maps on a 32×32 grid. DINOv2 ViT-S/14 provides 384 final-normalized patch maps on a 37×37 grid after excluding the class token. For SAM3, we use the final 256-channel backbone feature-pyramid output after the visual neck. We exclude prompts and the transformer and mask decoders. All encoders remain frozen in evaluation mode. Feature maps are bilinearly resampled to 32×32.

Images, maps and labels. Medical images are represented as two-dimensional PNG slices. Grayscale inputs are replicated across three channels to form RGB images. We bilinearly resize each image while preserving its aspect ratio, then centre it on a black square canvas. The canvas side is 512, 518 or 1008 for DINOv3, DINOv2 or SAM3, respectively. Pixel intensities are divided by 255. DINO inputs use channel means and standard deviations. SAM3 inputs use $(x - 0.5)/0.5$ in each channel. Features are extracted once for each encoder and split and stored in FP32.

Except for the ACDC 128×128 experiment (Supplement A.4.1), segmentation labels are letterboxed to 64×64 with nearest-neighbour interpolation and then reduced to 32×32 by non-overlapping 2×2 mode pooling. Ties select the smallest class ID. Decoders predict on this 32×32 grid.

Data splits. Table S1 lists the split sizes used in this study. The original 100 ACDC training patients are deterministically partitioned into training and development sets, with all slices and cardiac phases from each patient kept in the same partition. Final evaluation uses the separate official test set. Background-only masks are retained in all splits.

For M&Ms (Campello et al., 2021), all training patients are supervised, including the 25 cases originally designated unlabeled whose ground truth is now public. We retain all short-axis slices from both annotated frames, including background-only slices.

Table S1: **Dataset splits used in our experiments.** Counts refer to images and subjects included in each split. Subject counts are not applicable to COCO or ADE20K and are not reported for G1020. NR: not reported; N/A: not applicable.

Dataset	Train		Development		Test	
	Images	Subjects	Images	Subjects	Images	Subjects
ACDC	1,091	80	366	20	996	50
M&Ms	3,860	175	806	34	3,242	136
MSD Liver	1,091	60	366	20	1,506	24
G1020	612	NR	204	NR	204	NR
COCO	20,000	N/A	5,000	N/A	5,000	N/A
ADE20K	16,165	N/A	2,000	N/A	4,045	N/A

For COCO, images are selected from train2017 by cycling through foreground categories in increasing ID order. Each category has a fixed ordering of images containing positive-area, non-crowd annotations, and each cycle takes the next unused image for that category. Development uses val2017. The test set is drawn from unused train2017 images using the category-round-robin procedure. Sorted category IDs map to consecutive foreground labels, with background 0. Non-crowd masks are painted from larger to smaller area, with ties broken by annotation ID, so smaller instances take precedence.

Loss and optimization. The decoder loss is the sum of weighted pixelwise cross-entropy and one minus soft Dice. Soft-Dice intersections and denominators are pooled over valid pixels within each mini-batch, averaged over the evaluated semantic classes, and smoothed by 10^{-6} in both numerator and denominator. Pixels carrying an ignore or void label are excluded from both loss terms and provide no background supervision. If n_c denotes the number of valid training-grid pixels assigned to supervised class c, cross-entropy weights are proportional to $1/\sqrt{\max(n_c, 1)}$ and normalized to

have mean one over the supervised classes, including background when background is itself a valid supervised class.

Fresh decoders are trained for 20,000 steps with AdamW using learning rate 10^{-3}, weight decay 10^{-4}, and $(\beta_1, \beta_2) = (0.9, 0.999)$. Training uses batch size 8, a 200-step warm-up, cosine decay to 10^{-6}, global gradient clipping at 1.0, and FP32. Mini-batches sample training examples uniformly with replacement. Development evaluation is performed every 200 steps; we retain the best development checkpoint and choose the earlier checkpoint on ties. The primary ACDC discovery and detailed analysis use decoder seed 17. Fixed-support decoder replications additionally use seeds 23 and 41.

Decoder architectures. The primary U-Net has a 1×1 Conv–GroupNorm–GELU stem mapping $|S|$ inputs to width 64. This is followed by two 3×3 blocks at width 64, stride-two stages at widths 128 and 256, and a two-block 8×8 bottleneck. Bilinear upsampling with skip connections restores the 32×32 grid. Blocks use GroupNorm and GELU without dropout, followed by a 1×1 classifier. Only input width varies across supports. The Transformer decoder projects memory to dimension 256. It uses 16 learned queries, six pre-normalized decoder layers, eight heads, a 2,048-dimensional feed-forward sublayer and zero dropout. Its 1×1 memory/mask projections produce class and mask logits. The pointwise MLP uses four width-400 Linear–LayerNorm–GELU–Dropout(0.1) layers and a class projection, independently at each spatial position.

Exact Dice convention. For the medical-image settings, let N be the number of evaluated images, C_{fg} the number of foreground classes, and d_{ic} the hard-mask Dice for image i and foreground class c. Both-empty prediction/target pairs receive one, and pairs with exactly one empty mask receive zero. We compute:

$$M_{\mathcal{C}}^{E}(S) = \frac{1}{N} \sum_{i=1}^{N} \frac{1}{C_{\text{fg}}} \sum_{c=1}^{C_{\text{fg}}} d_{ic}. \tag{4}$$

For COCO and ADE20K, intersections and predicted/target counts are pooled over the evaluation set for each evaluated semantic category before macro-averaging across categories. The two datasets differ only in the treatment of label 0: in ADE20K, label 0 denotes void pixels and is excluded entirely from metric computation; in COCO, label 0 denotes background, so foreground predictions on background pixels contribute false positives to the corresponding foreground categories.

Dense-STG initialization. Gate parameters are initialized as $\alpha_j = 0.5$. At each optimization step, independent noise $\xi_j \sim \mathcal{N}(0, 0.5^2)$ is sampled once per channel and the stochastic gate is $z_j = \text{clip}(\alpha_j + \xi_j, 0, 1)$. The training objective adds $\lambda \sum_j \Phi(\alpha_j / 0.5)$ to the segmentation loss, where Φ is the standard normal cumulative distribution function. Gate and decoder parameters use the AdamW schedule specified above. For all Dense-STG initializations, we search $\lambda \in \{10^{-5}, 3 \times 10^{-5}, 10^{-4}, 3 \times 10^{-4}, 10^{-3}, 3 \times 10^{-3}, 10^{-2}\}$.

Random-support sampling. At cardinality k, we sample supports uniformly from the native channel set, with k distinct indices per support. Duplicate supports and the corresponding selected support are excluded. The M&Ms native and ACDC-transfer comparisons share the same 64 four-channel random supports, sampled with NumPy seed 20260901 and excluding both fixed supports. For ACDC, the development panel contains 512 random four-channel supports, 64 of which are also evaluated on the test set. The remaining supports are evaluated only on the development set.

For a fixed selected score, observing 0 successes among 512 sampled supports gives a one-sided 95% binomial-reference upper bound of $1 - 0.05^{1/512} \approx 0.583\%$. This bounds the probability that a uniformly sampled support matches or exceeds the selected score. Because the number of possible supports is much larger than 512, the binomial reference closely approximates sampling without replacement.

Search and evaluation cost. Training a fresh decoder under the common budget counts as one oracle call. A previous fresh-decoder evaluation may be reused only when the support, feature cache, decoder seed, and complete decoder-training configuration are identical. Knockout rankings remain specific to the parent support and checkpoint. Reuse adds no training call; selector training and knockout ranking are counted separately.

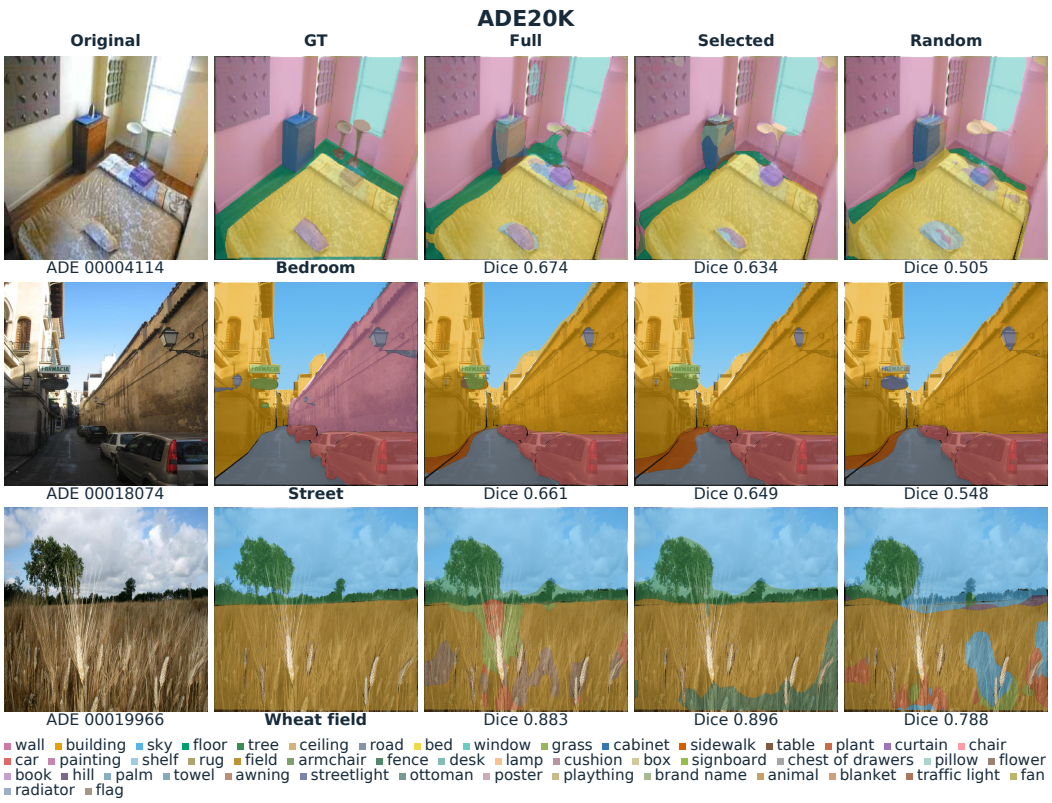

Figure S1: **ADE20K test-set segmentation examples.** Columns show the image, ground truth, Full (384 channels), Selected (106 channels), and a fixed same-width Random support. The same Random support is used in all rows. Numbers report per-image Dice on the original-image masks, averaged over ground-truth semantic classes; label 0 is void and excluded.

A.2 Exact PA-BFE controller

For a current support of size k, the coarse-stage deletion widths are:

$$\mathcal{B}(k) = \mathrm{unique}_{\downarrow}\left\{ \left\lfloor \frac{k}{2^j} \right\rfloor : j \geq 1,\ 2 \leq \left\lfloor \frac{k}{2^j} \right\rfloor < k \right\}, \tag{5}$$

where $\mathrm{unique}_{\downarrow}$ removes duplicates and orders the remaining widths from largest to smallest. The widths are tried in this order; after an accepted deletion, $\mathcal{B}(|S|)$ is recomputed for the updated support. If all coarse widths are rejected, control passes to the local stage.

Let $\mathrm{Rank}(S, g_S)$ denote the native-channel ordering induced by increasing $d_i(S)$ in Eq. (3), with ties broken by native index. Let $\mathrm{Eval}_{\mathcal{C}}(S')$ return the development Dice $M_{\mathcal{C}}^{\mathrm{dev}}(S')$ and the corresponding decoder $g_{S'}$ under the fresh-decoder protocol.

A.3 Random-support frontier and test uncertainty

A.3.1 Test random-support frontier

Table S2 reports test scores and support counts across ACDC widths; Figure 2 shows the development distributions. Sampling is specified in Supplement A.1.

A.3.2 Test uncertainty

For every ACDC interval, each paired bootstrap replicate samples 50 patients with replacement, retains all slices of each sampled patient and recomputes image-mean foreground Dice. We use 10,000

Algorithm S1 Exact local-stage controller for PA-BFE

Require: current passing support S, decoder g_S, fixed floor $\tau_{\mathcal{C}}$

$b \leftarrow 1$, $c \leftarrow 0$, $F \leftarrow \varnothing$; $P_1 \leftarrow 4$, $P_j \leftarrow 2$ for $j \geq 2$
$r \leftarrow \mathrm{Rank}(S, g_S)$
while $|S| > 1$ and $F \neq S$ **do**
  if $b > |S| - 1$ **then**
    $b \leftarrow |S| - 1$; $c \leftarrow 0$
  end if
  $e \leftarrow [b + 1 \leq |S| - 1 \text{ and } c \geq P_b]$
  if e **then**
    $t \leftarrow b + 1$; $A \leftarrow$ first t channels of r
  else if $b = 1$ **then**
    $t \leftarrow 1$; $A \leftarrow \{\text{first channel of } r \text{ outside } F\}$
  else
    $t \leftarrow b$; $A \leftarrow$ first t channels of r
  end if
  $S' \leftarrow S \setminus A$
  $(m, g_{S'}) \leftarrow \mathrm{Eval}_{\mathcal{C}}(S')$
  if $m \geq \tau_{\mathcal{C}}$ **then**
    $(S, g_S) \leftarrow (S', g_{S'})$
    $F \leftarrow \varnothing$
    $r \leftarrow \mathrm{Rank}(S, g_S)$
    if e **then**
      $b \leftarrow t$; $c \leftarrow 1$
    else
      $b \leftarrow t$; $c \leftarrow c + 1$
    end if
  else
    $c \leftarrow 0$
    if $t = 1$ **then**
      $F \leftarrow F \cup A$; $b \leftarrow 1$
    else if e **then**
      $P_b \leftarrow 2P_b$
    else
      $P_{t-1} \leftarrow 2P_{t-1}$; $b \leftarrow \max(1, \lfloor t/2 \rfloor)$
    end if
  end if
end while
return S, F

Table S2: **ACDC random-support frontier on test data.** Random Dice entries are means across n evaluated supports. Full and $S^{(1)}$ are included as references.

Support	k	Test n	Test Dice
Full	384	1	0.811
Selected $S^{(1)}$	4	1	0.772
Uniform random	4	64	0.628
Uniform random	8	64	0.680
Uniform random	16	8	0.723
Uniform random	32	8	0.748
Uniform random	48	8	0.770
Uniform random	64	8	0.775
Uniform random	128	8	0.789

replicates and the 2.5th/97.5th percentiles, recomputing retention as a ratio within each replicate. This respects patient clustering while preserving the image-weighted estimand. The intervals quantify

patient-sampling uncertainty for each fixed trained decoder. Table S3(b) reports the fixed-support results and paired intervals.

A.4 FIXED-SUPPORT ACDC FOLLOW-UPS

These experiments keep the discovered supports and decoder family fixed, using the ACDC splits, features and training protocol in Supplement A.1. All selected and same-width random supports are fixed before these additional decoders are trained and tested.

A.4.1 FINER LABEL SUPERVISION AND EVALUATION

Original masks are resized directly using nearest-neighbour interpolation while preserving the aspect ratio, and are then centred on a 128×128 canvas. Feature maps, training-only feature means and support indices remain fixed. Bilinear interpolation maps 32×32 U-Net logits to 128×128 before loss and hard-mask scoring. We train ten fresh seed-17 decoders for 20,000 steps: Full, $S^{(1)}$ and eight prespecified random four-channel supports, using the training-only class weights and checkpoint-selection rule.

Full and $S^{(1)}$ achieve 0.836 and 0.794 development Dice (95.0% retention). Same-width random supports achieve 0.672 ± 0.031 development Dice and 0.665 ± 0.025 test Dice (mean $\pm$ population SD). Table 1 reports the Full and selected test scores. Under the bootstrap in Supplement A.3.2, test support-minus-Full Dice is -0.046 (95% interval $[-0.059, -0.033]$), and retention is 94.5% (92.9–96.1%). The selected support remains above the random mean. Each support is compared with Full on the same label grid.

A.4.2 DECODER INITIALIZATION AND FIXED-SUPPORT TEST UNCERTAINTY

With $S^{(1)}$ and $S^{(2)}$ fixed, decoders use seeds 17, 23 and 41 under the original 32×32 protocol, each paired with the same-seed Full decoder. Table S3(a,b) preserves their development and test results. Eight fixed random 13-channel decoders average 0.703 ± 0.015 test Dice (population SD). These experiments vary decoder initialization. Across the six fixed-support test comparisons, retention ranges from 93.7% to 96.5% (Table S3). For seed 17, the displayed test Dice of $S^{(1)}$ is 0.039 lower than that of Full. Absolute-Dice 95% intervals are $[0.802, 0.819]$ for Full and $[0.756, 0.787]$ for $S^{(1)}$. Bootstrap intervals use the underlying predictions.

A.5 ACDC SUPPORT CONTROLS

A.5.1 TRAINING-STATISTIC CONTROLS

We rank native channels by RMS activation or variance over uncentered training features, before the channelwise centering used for decoder training. The Top-4 supports are $\{112, 116, 341, 350\}$ for RMS and $\{21, 26, 112, 341\}$ for variance; each uses the seed-17 U-Net protocol. The selected support differs from both: channel 30 ranks 350th by RMS and 321st by variance.

On the test set, Full, $S^{(1)}$, RMS Top-4 and variance Top-4 score 0.811, 0.772, 0.655 and 0.707 Dice. The selected support exceeds RMS Top-4 by 0.117 (95% interval $[0.101, 0.131]$) and variance Top-4 by 0.065 ($[0.047, 0.083]$), using the paired patient bootstrap in Supplement A.3.2.

A.5.2 IMAGE-DEPENDENT VARIATION BEYOND A FIXED SPATIAL FIELD

Replacing each selected field with its pixelwise average over the 1,091 centered training examples preserves four 32×32 maps while removing image-specific variation. Fresh U-Nets at seeds 17, 23 and 41 use the original 20,000-step protocol and development-selected checkpoints. For seed 17, real features exceed the fixed mean field by 0.508 test Dice (paired patient-bootstrap 95% interval $[0.466, 0.549]$). Thus, image-dependent variation contributes beyond the retained spatial template.

Table S3: **Decoder-seed variation for the fixed ACDC supports.** All comparisons use same-seed Full decoders and 32×32 labels.

(a) Development performance

		$S^{(1)}$: 4 channels			$S^{(2)}$: 13 channels	
Seed	Full Dev. Dice	Selected Dev. Dice	Δ_{dev}	Dev. retention (%)	Selected Dev. Dice	Dev. retention (%)
17	0.807	0.787	−0.020	97.5	0.787	97.5
23	0.808	0.776	−0.032	96.0	0.782	96.8
41	0.815	0.774	−0.041	95.0	0.783	96.1

(b) Test performance and paired patient-bootstrap intervals

Support	Seed	Full test Dice	Selected test Dice	Test retention (%)	Test retention 95% CI (%)	Test Dice gap 95% CI
$S^{(1)}$	17	0.811	0.772	95.2	[93.5, 96.9]	$[-0.053, -0.025]$
$S^{(2)}$	17	0.811	0.773	95.3	[93.9, 96.8]	$[-0.050, -0.026]$
$S^{(1)}$	23	0.811	0.765	94.3	[91.8, 96.5]	$[-0.067, -0.028]$
$S^{(2)}$	23	0.811	0.760	93.7	[91.9, 95.3]	$[-0.066, -0.038]$
$S^{(1)}$	41	0.793	0.765	96.5	[94.9, 98.0]	$[-0.040, -0.015]$
$S^{(2)}$	41	0.793	0.765	96.5	[94.9, 97.9]	$[-0.041, -0.016]$

Dev. denotes development; Δ_{dev} is $S^{(1)}$−Full development Dice. Test uses 996 slices from 50 official ACDC test patients; Dice gaps are support minus Full. CIs use 10,000 paired patient-bootstrap replicates to quantify patient-sampling uncertainty for each trained decoder. Displayed retention and point differences use the reported three-decimal Dice; bootstrap intervals use the underlying predictions.

A.6 Sensitivity to the acceptance tolerance

We repeat discovery with ten additional random seeds using the same ACDC split, frozen DINOv3 features, and U-Net evaluation protocol, with 20,000 steps per model. Each run uses one seed for Full, Dense-STG, physical confirmation, and PA-BFE. For each seed, we use the Dense-STG grid specified in Supplement A.1 and physically confirm every distinct nonempty final-gate support with a fresh decoder. A candidate passes if its development Dice is at least the same-seed, unrounded Full Dice minus 0.02. After all confirmations, we select the smallest passing support.

The detailed $S^{(1)}$ case uses gate seed 1701 and decoder seed 17. Its 17-channel candidate is the smallest passing support in the seven-point grid, and PA-BFE reduces it through $17 \to 9 \to 5 \to 4$. The ten additional runs use the same controller and seed assignment.

Follow-up PA-BFE searches relax ϵ_{search} from 0.02 to 0.04 in Eq. (2). Each follow-up shares the same Full reference as its paired 0.02 run and either restarts from the confirmed pre-search support or continues from the 0.02 endpoint. Completed fresh-decoder evaluations are reused under the matching conditions specified in Supplement A.1; unevaluated supports receive fresh decoders. Search stops when a complete singleton scan rejects every candidate.

Table S4 reports all ten paired endpoints in the original run order. All endpoints are evaluated on the same 996 test slices from 50 ACDC patients. These ten pairs exclude the original $S^{(1)}$ discovery.

Table S4: **Paired ACDC discovery at two acceptance tolerances.** Acceptance uses development Dice; endpoint scores are reported on the test set. Retention is computed from unrounded scores.

Seed	Confirmed start k	ϵ_{search}	Final k	Final support	Test Dice	Test retention (%)
0	19	0.02	5	$\{26, 30, 122, 247, 257\}$	0.784	98.49
		0.04	4	$\{26, 30, 122, 257\}$	0.778	97.75
7	20	0.02	6	$\{26, 30, 122, 213, 247, 257\}$	0.781	97.49
		0.04	4	$\{26, 122, 247, 257\}$	0.762	95.05
8194	19	0.02	6	$\{26, 30, 122, 213, 247, 257\}$	0.785	97.28
		0.04	4	$\{26, 30, 122, 257\}$	0.772	95.76
42	21	0.02	8	$\{13, 26, 78, 122, 213, 247, 257, 323\}$	0.784	97.59
		0.04	5	$\{13, 26, 122, 257, 323\}$	0.777	96.76
9128	24	0.02	6	$\{13, 30, 78, 122, 257, 309\}$	0.778	96.39
		0.04	4	$\{13, 30, 122, 257\}$	0.773	95.77
884284	18	0.02	5	$\{4, 26, 30, 122, 257\}$	0.770	95.51
		0.04	4	$\{26, 30, 122, 257\}$	0.764	94.68
3	23	0.02	7	$\{21, 26, 122, 213, 247, 257, 323\}$	0.783	96.94
		0.04	4	$\{21, 26, 122, 257\}$	0.784	96.98
69	6	0.02	5	$\{26, 30, 78, 122, 257\}$	0.782	96.72
		0.04	3	$\{26, 122, 257\}$	0.758	93.78
1395	6	0.02	6	$\{26, 30, 99, 122, 213, 257\}$	0.773	96.80
		0.04	3	$\{30, 122, 257\}$	0.764	95.64
353604	6	0.02	6	$\{26, 30, 122, 213, 257, 304\}$	0.782	96.66
		0.04	4	$\{26, 30, 122, 257\}$	0.763	94.30