Cut-ViT: Task-Specific Model Pruning via Gram Anchoring Subspace Consistency

Jianjian Yin[1,4,*], Liulei Li[2,*], Tao Chen[1,4], Yi Chen[3], Yazhou Yao[1,4(✉)], and Wenguan Wang[2]

[1]Nanjing University of Science and Technology, Nanjing, China
[2]Zhejiang University, Hangzhou, China. [3]Nanjing Normal University, Nanjing, China
[4]State Key Laboratory of Intelligent Manufacturing of Advanced Construction Machinery, Nanjing, China
[*]Equal contribution [✉]Corresponding author
https://github.com/NUST-Machine-Intelligence-Laboratory/Cut-ViT

Abstract. Pruning visual foundation models has attracted considerable attention. However, existing methods focus on rigid point-to-point token alignment on a single dataset for pruning, suffering from two limitations: **i**) robustness degradation, and **ii**) task-specificity deficiency. To address these limitations, we propose a task-specific pruning pipeline, named Cut-ViT. Specifically, we first construct gram anchoring matrices from both spatial and semantic perspectives, and perform the subspace decomposition to extract the corresponding subspace bases. Basis-agnostic and residual constraints are then adopted to align the gram subspaces between the native and pruned DINOv3 models along spatial and channel dimensions, enabling subnetworks to inherit robust feature representations of native DINOv3. Furthermore, we design spectral entropy adaptation, which quantifies the information density of feature manifolds along spatial and channel dimensions, thereby adapting the pruning objective to specific downstream tasks. Experiments show that Cut-ViT requires approximately **one** minute on a single A100 GPU to obtain subnetworks at various sparsity levels, using only **20.9%** of the time and **45.5%** of the GPU memory compared with previous methods, while achieving SOTA performance on **six** tasks across **nine** datasets.

Keywords: Model pruning · Gram anchoring · Subspace constraint

1 Introduction

The emergence of visual foundation models (VFMs) [5, 22, 38, 49], notably the DINO series [5,38,49], has pushed the boundaries of visual representation learning [4,29,37,52,67,72,73]. However, despite VFMs achieving competitive performance across a wide range of downstream tasks [3,15,30,64,68,75], their massive computational overhead poses significant challenges for deployment on resource-constrained edge devices. Consequently, the efficient compression of VFMs without compromising performance has become a research focus [16, 31, 71, 76, 77]. Among various model compression paradigms, training-free one-shot structured

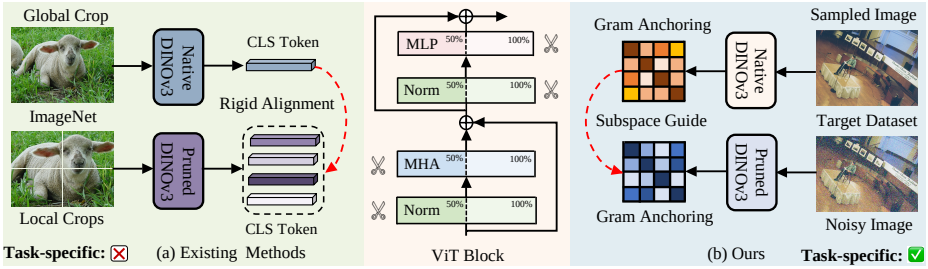

Fig. 1: Comparison with prior work. (a) To ensure feature consistency, existing methods enforce rigid alignment between CLS tokens; (b) CUT-VIT aligns the subspace induced by the inherent gram anchoring within DINOv3, thereby enabling the subnetworks to inherit the robust feature representations of the original network.

pruning (OSP) [23,25,35,50] has recently emerged as a highly promising solution. By dynamically ranking parameter importance using loss gradients, OSP enables rapid generation of lightweight subnetworks, making it particularly suitable for large-scale VFMs where retraining is computationally prohibitive.

However, existing OSP methods face two bottlenecks that hinder their applications in real-world scenarios. ❶ **Robustness degradation**. Current OSP methods rely on rigid point-to-point alignment between the tokens of native and pruned models. This localized strategy encourages the subnetwork to memorize specific numerical features rather than preserving the global semantic topology. By prioritizing exact token matching over structural consistency, these methods fail to inherit the robust manifold structure of VFMs, leading to representation degradation. ❷ **Task-specificity deficiency**. Prior training-free methods typically follow a rigid pipeline, that optimizes on a single generic dataset (*e.g.*, ImageNet) to deploy a universal subnetwork across all downstream tasks (Fig. 1(a)). While seemingly elegant, this paradigm overlooks that for edge computing, model capacity has already been sacrificed to achieve faster inference and lower memory usage. Emphasizing a shared, task-agnostic subnetwork inevitably leads to a further compromise in precision and results in a sub-optimal trade-off between precision and efficiency for specific downstream applications.

To address these challenges, this work proposes CUT-VIT, a task-specific OSP algorithm that can generate subnetworks at various sparsity levels in approximately ***one minute*** for DINOv3 [49] using a single A100 GPU. To alleviate ❶, we rethink feature alignment from the perspective of manifold consistency. Previous studies [49] found that the robustness of DINOv3 stems from the rich global topology encoded in its gram anchoring, which acts as a dense regularization constraint. However, rigid point-to-point alignment fails to preserve this structure and is sensitive to noise. Inspired by this, a gram anchoring subspace decomposition strategy (Fig. 1(b)) is proposed. Rather than matching raw features, CUT-VIT constructs spatial and channel gram matrices and decomposes them to extract orthonormal bases. These bases span the directions of maximal joint variation, thereby decoupling the core semantic topology from high-frequency noise, which enables the design of a basis-invariant subspace con-

sistency strategy. By imposing basis-agnostic and residual constraints, Cut-ViT forces the feature subspace of subnetworks to geometrically align with that of the teacher. This allows the subnetwork to inherit the robust spatial and semantic manifold of DINOv3 without overfitting to specific token values. To answer ❷, we design a pruning pipeline that adapts to both data distributions and task properties. **First**, unlike generic pruning, we perform optimization directly on the target dataset. By encoding the target data distribution into the gram anchoring and propagating it via subspace consistency, the subnetwork implicitly adapts to domain shifts without requiring ground-truth annotations. **Second**, Cut-ViT introduces spectral entropy to quantify the information density of feature manifolds. Since different tasks require different information granularities (*e.g.*, classification relies on global semantics, while segmentation demands high spatial frequency) [18,57], spectral entropy measures this complexity along spatial and channel dimensions. This serves to dynamically weight the pruning objective to tailor the model architecture for specific downstream tasks.

Cut-ViT demonstrates superior performance over existing training-free OSP methods [23,35,50] across **nine** datasets covering **six** downstream tasks: semantic segmentation, object detection, video object segmentation, semantic matching, depth estimation, and image classification. Beyond standard benchmarks, our out-of-distribution experiments further confirm that Cut-ViT can preserve the robustness of the native DINOv3 [49] model. In terms of *efficiency*, Cut-ViT offers a compelling advantage. Specifically, it achieves a **3.4%** improvement in $(\mathcal{J}\&\mathcal{F})_m$ for video object segmentation on DAVIS-2017 [39], **5.1%** improvement in PCK@0.05 for semantic matching on FG3DCar [33], **2.8%** improvement in mAP for object detection on MS COCO [32] compared with prior state-of-the-art OSP approaches, while requiring only **20.9%** of the inference time and **45.5%** of the GPU memory. Moreover, our approach remains competitive even against training-based model pruning alternatives [16, 77], while consuming only **0.44%** of the pruning time and **28.2%** of the GPU memory.

2 Related Work

Visual Foundation Model. The emergence of large-scale pretraining [58,59,63] and the success of visual encoders [10, 34, 53] have catalyzed the development of VFMs, such as CLIP [42], SAM [22], and DINO [5, 38, 49]. Trained on vast datasets, these models learn robust and generalized visual representations that are transferable across a wide array of vision tasks [3,28,40,41,61,65,66,69]. However, their considerable computational overhead creates significant challenges for deployment on edge devices. Several studies have aimed to lightweight VFMs, including Efficientsam [60], Theia [46], and Tinysam [47]. These methods primarily rely on knowledge distillation to transfer the capabilities of computationally intensive encoders to more efficient, lightweight encoders. In contrast, our approach selects parameters directly from the VFM's encoder, generating networks with varying sparsities in a single round of inference.

Model Compression. The limitation of computational resources has sparked increased interest in model compression research [20, 26, 31, 76]. In recent years, quantization [13, 19, 43], token optimization [2, 7, 21, 27, 36, 70], and structured pruning [16, 24] have become the primary research directions in this field. Quantization [54, 55] aims to reduce the precision of model weights, thus striving to minimize computational overhead while preserving model performance. Token optimization [1, 6] pruns redundant tokens and merges tokens with high similarity, leading to acceleration in model computation. Structured pruning [45, 51], which eliminates network parameters, has inspired a series of influential works. HydraViT [16] generates multiple sub-networks by stacking attention during training, progressively modifying the embedding dimension and the head number of attention layer. Scala [71] employs scale-aware coordination during training to ensure that each pruned network learns robust objectives. EA-ViT [77] adopts a two-stage training paradigm and incorporates a lightweight router to select sub-networks that meet budget constraints. However, these methods require intensive training, leading to high computational cost.

One-shot Pruning without Retraining. One-shot structured pruning [12, 24, 62] eliminates the need for training, requiring only a single inference pass, during which parameter importance is dynamically ranked based on gradients derived from the loss function. This allows for the generation of networks with varying sparsities in a single pass. Specifically, LAMP [25] introduces a layer-adaptive pruning score that enables pruning without requiring hyperparameter tuning, while SNIP [23] employs a magnitude-based criterion. SNOWS [35] and SnapViT [50] are notable OSP methods designed for post-processing pretrained models. The former introduces a second-order optimization framework combining Hessian-free optimization with a customized conjugate gradient method for layer-wise pruning, effectively addressing second-order approximation issues. The latter jointly evaluates parameter importance for pruning using a self-supervised objective and a genetic algorithm. The aforementioned methods focus on rigid point-to-point alignment between tokens for pruning, which undermines the robustness of the subnetworks. Conversely, our work encourages alignment of gram anchoring in subspace, ensuring that the subnetworks inherit the robust feature representations of the native DINOv3.

Departing from prior SVD-based work [8, 56] that rely on *static* subspace construction, Cut-ViT pioneers rotation-invariant *dynamic* optimization to align the subspaces of the teacher and pruned models during pruning, ensuring the inheritance of robust topological structures.

3 Methodology

3.1 Preliminary

Standard training-free OSP methods [12,23,25] aim to identify an optimal binary mask $\boldsymbol{m} \in \{0,1\}^{|\boldsymbol{W}|}$ for the weights $\boldsymbol{W}$ of a pre-trained VFM (*e.g.,* DINOv3 [49]) to minimize the performance gap between the pruned student and the native

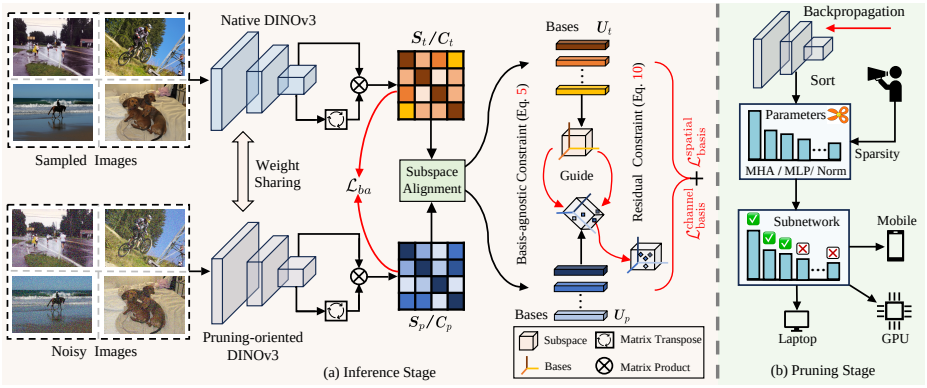

Fig. 2: The pipeline of our Cut-ViT. Pairs of images are input into the native and pruning-guided DINOv3 to obtain feature embeddings. $\mathcal{L}_{ba}$ denotes the dense feature matching loss. Subspace decomposition (§3.2) is applied to the spatial and channel gram matrices ($\boldsymbol{S}_t$, $\boldsymbol{S}_p$, $\boldsymbol{C}_t$, $\boldsymbol{C}_p$) to derive subspace bases ($\boldsymbol{U}_t$, $\boldsymbol{U}_p$). Subspace consistency is enforced with basis-agnostic and residual constraints (§3.3), yielding spatial and channel gram losses, $\mathcal{L}_{\text{basis}}^{\text{spatial}}$ and $\mathcal{L}_{\text{basis}}^{\text{channel}}$. During pruning, backpropagation computes the gradients of each parameter, which are then ranked in descending order of magnitude to select parameters and generate subnetworks with varying sparsity levels.

teacher. Typically, this is formulated as a reconstruction problem:

$$H \approx \frac{1}{N} \sum_{n=1}^{N} \|\nabla_{\boldsymbol{m}} \mathcal{L}_n\|^2, \qquad \mathcal{L} = \Upsilon(\boldsymbol{O}_s, \boldsymbol{O}_t). \tag{1}$$

The parameter importance score H is approximated using the empirical Fisher matrix over N calibration samples. Based on these scores, search algorithms such as xNES [14] are employed to locate the optimal subnetwork. $\boldsymbol{O}_s$ and $\boldsymbol{O}_t$ denote the feature outputs of the student (local crops) and teacher (global crops), respectively. $\Upsilon(\cdot, \cdot)$ represents the distillation objective, traditionally implemented as a rigid element-wise comparison (*e.g.*, MSE or KL divergence).

3.2 Gram Anchoring Subspace Decomposition

Motivation. The strong generalization ability of DINOv3 [49] arises from the rich structural and semantic relationships encoded in its dense feature representations. While prior work [35, 50] enforce rigid point-wise alignment of tokens, such exact numerical reconstruction imposes overly strict constraints that can encourage overfitting to redundant correlations and high-frequency noise. Therefore, we propose decomposing the gram matrix into subspaces spanned by dense features using singular value decomposition (SVD), and then regularizing the subspace consistency for gram anchoring. The detailed pipeline of Cut-ViT is shown in Fig. 2. By injecting noise into input images, the derived bases are encouraged to capture a more robust and noise-invariant semantic topology.

.

Gram Formulation. Let $\boldsymbol{F} \in \mathbb{R}^{L \times D}$ denote a feature embedding, where L is the token number and D is the channel dimension. We define a generalized gram matrix $\boldsymbol{G}$ to model the second-order correlation of the features along a specific dimension, and then extract the manifold structure by performing SVD on $\boldsymbol{G}$:

$$\boldsymbol{G} = \boldsymbol{X} \cdot \boldsymbol{X}^\top \approx \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{U}^\top, \tag{2}$$

where $\boldsymbol{X}$ represents the feature matrix unfolded along the target dimension, $\boldsymbol{\Sigma}$ is the diagonal matrix of singular values, and $\boldsymbol{U}$ contains the orthonormal basis vectors spanning the dominant subspace. This formulation allows us to capture correlations from two complementary perspectives:

• **Spatial Gram Anchoring.** To capture the topological relationships between tokens (*e.g.*, the spatial layout of objects), we compute the spatial gram matrix $\boldsymbol{S}$ by aggregating features across the channel dimension, yielding:

$$\boldsymbol{S} = \boldsymbol{F} \cdot \boldsymbol{F}^\top \in \mathbb{R}^{L \times L}. \tag{3}$$

After applying Eq. 2 to $\boldsymbol{S}$, the spatial basis matrix $\boldsymbol{U}^{\boldsymbol{S}}$ can be obtained, which identifies the dominant modes of inter-token correlation and can group spatially coherent patches based on their feature similarity ("Where").

• **Channel Gram Anchoring.** Complementary to spatial structure, channel correlations encode semantic attributes (*e.g.*, texture, style, and class). We compute the channel gram matrix $\boldsymbol{C}$ by treating tokens as samples, resulting in:

$$\boldsymbol{C} = \boldsymbol{F}^\top \cdot \boldsymbol{F} \in \mathbb{R}^{D \times D}. \tag{4}$$

Similarly, decomposing $\boldsymbol{C}$ yields the channel basis $\boldsymbol{U}^{\boldsymbol{C}}$, which decouples abstract concepts from spatial instantiations. Consequently, $\boldsymbol{U}^{\boldsymbol{C}}$ serves as a compact descriptor of the global semantics, and answers the question of "What" is depicted. Detailed process is provided in Fig. 3. By extracting $\boldsymbol{U}^{\boldsymbol{S}}$ and $\boldsymbol{U}^{\boldsymbol{C}}$, the alignment between teacher and student models is converted into a subspace consistency problem, so as to preserve the robust structural and semantic topology of DINOv3 [49].

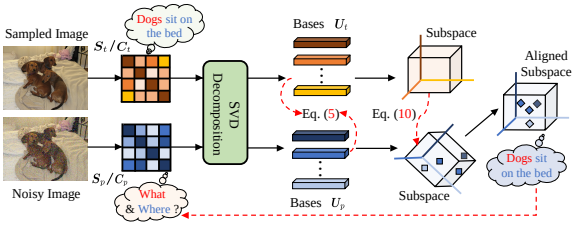

Fig. 3: The process of spatial and channel gram anchoring subspace consistency.

3.3 Basis-Invariant Subspace Consistency

This section addresses the challenge of aligning feature subspaces of pruned and native models. Since the singular vectors (bases) derived by SVD are not unique, directly aligning the basis vectors element-by-element via MSE loss is ill-posed. To tackle this, we propose a basis-agnostic constraint that aligns the subspaces independent of the specific basis orientation. A residual constraint is further proposed to suppress noise outside the aligned subspace.

Basis-agnostic Constraint. Let $\boldsymbol{U}_p \in \mathbb{R}^{L\times K}$ and $\boldsymbol{U}_t \in \mathbb{R}^{L\times K}$ denote the truncated Top-K basis matrices for the pruned and native models. The affinity matrix is defined as $\boldsymbol{M} = \boldsymbol{U}_p^\top \boldsymbol{U}_t$. The subspace consistency is enforced by maximizing the subspace overlap (*i.e.*, correlation energy) between two sets of bases:

$$\mathcal{L}_{\text{basis}} = 1 - ||\boldsymbol{M}||_F^2 = 1 - \sum\nolimits_{i=1}^{K}\sum\nolimits_{j=1}^{K} M_{ij}^2, \tag{5}$$

where $||\cdot||_F^2$ denotes the Frobenius norm, and K is the number of selected components. $\mathcal{L}_{\text{basis}}$ is robust to the non-uniqueness of SVD, as established below.

Theorem. ***(Basis Invariance).*** The loss function $\mathcal{L}_{\text{basis}}$ depends solely on the subspace spanned by the bases and is invariant to arbitrary orthogonal transformations of the basis vectors $\boldsymbol{U}_p$ and $\boldsymbol{U}_t$.

Proof. Let $\boldsymbol{Q}_p, \boldsymbol{Q}_t \in \mathbb{R}^{K\times K}$ be arbitrary orthogonal matrices (*i.e.*, $\boldsymbol{Q}_p^\top \boldsymbol{Q}_p = \boldsymbol{I}$). The non-uniqueness of SVD implies that valid bases can be transformed versions $\tilde{\boldsymbol{U}}_p = \boldsymbol{U}_p\boldsymbol{Q}_p$ and $\tilde{\boldsymbol{U}}_t = \boldsymbol{U}_t\boldsymbol{Q}_t$. We examine the invariance of $\mathcal{L}_{\text{basis}}$ in two cases:
Case 1: Simultaneous Transformation. Consider the scenario where both the native and pruned bases undergo arbitrary rotations. Substituting the transformed bases into the Frobenius norm term yields:

$$||\tilde{\boldsymbol{U}}_p^\top \tilde{\boldsymbol{U}}_t||_F^2 = ||(\boldsymbol{U}_p\boldsymbol{Q}_p)^\top(\boldsymbol{U}_t\boldsymbol{Q}_t)||_F^2 = ||\boldsymbol{Q}_p^\top(\boldsymbol{U}_p^\top\boldsymbol{U}_t)\boldsymbol{Q}_t||_F^2. \tag{6}$$

Recalling the unitary invariance property of the Frobenius norm (*i.e.*, $||\boldsymbol{UAV}||_F = ||\boldsymbol{A}||_F$ for orthogonal $\boldsymbol{U}, \boldsymbol{V}$), we have:

$$||\boldsymbol{Q}_p^\top \boldsymbol{M}\boldsymbol{Q}_t||_F^2 = ||\boldsymbol{M}||_F^2. \tag{7}$$

Case 2: Unilateral Transformation. In our task-specific pruning setting, the native DINOv3 serves as a frozen teacher, whose basis $\boldsymbol{U}_t$ is fixed ($\boldsymbol{Q}_t = \boldsymbol{I}$). The basis $\boldsymbol{U}_p$ of the pruned network evolves during optimization and may rotate arbitrarily. Therefore, the objective becomes:

$$||\tilde{\boldsymbol{U}}_p^\top \boldsymbol{U}_t||_F^2 = ||\boldsymbol{Q}_p^\top(\boldsymbol{U}_p^\top\boldsymbol{U}_t)||_F^2 = ||\boldsymbol{Q}_p^\top\boldsymbol{M}||_F^2. \tag{8}$$

Since $\boldsymbol{Q}_p^\top$ is orthogonal, the multiplication represents an isometric rotation which preserves the Frobenius norm, *i.e.*, $||\boldsymbol{Q}_p^\top\boldsymbol{M}||_F = ||\boldsymbol{M}||_F$. Details are provided in *supplementary materials*. Thus, $\mathcal{L}_{\text{basis}}$ consistently aligns the subspace of the pruned model to the teacher, regardless of how pruned bases are oriented. □

Residual Constraint. While $\mathcal{L}_{\text{basis}}$ ensures the alignment of principal directions, the pruned feature embedding $\boldsymbol{F}_p$ inevitably contains redundant information orthogonal to the target subspace. To explicitly filter this noise, we employ a geometric projection. Let $\boldsymbol{P}_t = \boldsymbol{U}_t\boldsymbol{U}_t^\top$ be the orthogonal projection matrix onto the native DINOv3 subspace. Any feature embedding $\boldsymbol{F}_p$ from the pruned network can be uniquely decomposed into two orthogonal components:

$$\boldsymbol{F}_p = \underbrace{\left(\boldsymbol{U}_t\boldsymbol{U}_t^\top\boldsymbol{F}_p\right)}_{\text{Aligned Component}} + \underbrace{\left((\boldsymbol{I}-\boldsymbol{U}_t\boldsymbol{U}_t^\top)\boldsymbol{F}_p\right)}_{\text{Orthogonal Residual}}. \tag{9}$$

The first term represents features successfully projected onto the robust teacher subspace. The second term corresponds to the residual error or noise lying outside the target manifold. We minimize the energy of irrelevant information:

$$\mathcal{L}_{\text{residual}} = ||(\boldsymbol{I} - \boldsymbol{U}_t \boldsymbol{U}_t^\top)\boldsymbol{F}_p||_F^2. \tag{10}$$

$\mathcal{L}_{\text{residual}}$ suppresses activation that does not structurally align with the gram anchoring of the teacher model, thereby enhancing feature purity.

3.4 Task-specific Pruning via Spectral Entropy Adaptation

In this section, we inject tailored properties for each task into the training-free pruning pipeline. **First**, rather than performing universal pruning on a generic dataset (*e.g.,* ImageNet [9]), which ignores the distinct data distributions, our pruning is conducted on a minimal subset of N samples from the target dataset. This allows the subnetwork to capture the underlying data manifold of the target domain without requiring annotations. It is noteworthy that, benefited from our efficient design (*i.e.*, 61s *v.s.* 292s of SnapViT [50]), CUT-VIT maintains lower latency than prior methods even when performing multiple searches. **Second**, we introduce spectral entropy to adaptively weight the pruning objective based on the information density of feature manifolds across spatial and channel axes. Let $\sigma_i^{\boldsymbol{S}}$ and $\sigma_j^{\boldsymbol{C}}$ denote the singular values of the native spatial and channel gram matrices. We first normalize these singular values to obtain the energy probability distributions within the spatial and channel subspaces:

$$p_i^{\boldsymbol{S}} = \frac{\sigma_i^{\boldsymbol{S}}}{\sum_k \sigma_k^{\boldsymbol{S}}}, \quad p_j^{\boldsymbol{C}} = \frac{\sigma_j^{\boldsymbol{C}}}{\sum_k \sigma_k^{\boldsymbol{C}}}. \tag{11}$$

The complexity of spatial and channel representations can then be measured as:

$$\mathbb{H}(\boldsymbol{S}_t) = -\textstyle\sum_i p_i^{\boldsymbol{S}} \log p_i^{\boldsymbol{S}}, \quad \mathbb{H}(\boldsymbol{C}_t) = -\textstyle\sum_j p_j^{\boldsymbol{C}} \log p_j^{\boldsymbol{C}} \tag{12}$$

Theoretically, the eigenspectrum decay of the feature covariance matrices reflects the degree of representation collapse within the token space. For instance, classification tasks optimize for shift-invariant global semantics, which drives spatial tokens toward uniformity. This concentrates the spatial Gram matrix energy into a few **dominant principal components** (lower $\mathbb{H}(\boldsymbol{S}_t)$). Conversely, dense prediction tasks (*e.g.,* semantic segmentation, correspondence matching) should preserve high-frequency topological boundaries and pixel-level semantics. This motivates a penalty in representation collapse, and enforces a much **flatter singular value distribution** along the spatial axis (higher $\mathbb{H}(\boldsymbol{S}_t)$). On this basis, we remodel the basis consistency loss to be task-aware:

$$\mathcal{L}_{\text{basis}}^{\text{all}} = w^{\text{spatial}} \cdot \mathcal{L}_{\text{basis}}^{\text{spatial}} + w^{\text{channel}} \cdot \mathcal{L}_{\text{basis}}^{\text{channel}}, \tag{13}$$

where the dynamic weights are defined as:

$$w^{\text{spatial}} = \frac{\mathbb{H}(\boldsymbol{S}_t)}{\mathbb{H}(\boldsymbol{S}_t) + \mathbb{H}(\boldsymbol{C}_t)}, \quad w^{\text{channel}} = \frac{\mathbb{H}(\boldsymbol{C}_t)}{\mathbb{H}(\boldsymbol{S}_t) + \mathbb{H}(\boldsymbol{C}_t)}. \tag{14}$$

This ensures that our pruning algorithm adapts to given downstream tasks *w.r.t.* the underlying representational demands.

4 Experiments

This section presents a comprehensive evaluation of Cut-ViT across multiple visual tasks at sparsity levels ranging from 10% to 30%. The datasets used for each task are listed below.

Datasets. We use the following datasets for various tasks: ADE20K [74] and PASCAL VOC 2012 [11] for semantic segmentation, COCO [32] for object detection, NYUv2 [48] for depth estimation, DAVIS-2017 [39] for video object segmentation, and FG3DCAR [33], JODS [44], and SBD [17] for semantic matching, as well as ImageNet [9] for image classification. ADE20K is a large semantic segmentation dataset containing 150 categories, with 20,210/2,000/3,000 images allocated for train/val/test. PASCAL VOC 2012 consists of 20 semantic classes and one background class, with 1,464/1,449/1,456 images used for train/val/test. COCO has 80 object categories for object detection, with 118,287 training images, 5,000 validation images, and 40,000 test images. NYUv2 is split into 795 images for training and 654 images for testing. DAVIS-2017 contains 60 training videos and 30 validation videos. The three datasets for semantic matching together comprise 400 pairs of test images with dense correspondence annotations. ImageNet consists of 1,000 categories, with 1.2 million images allocated for training, 50,000 for validation, and 100,000 for testing.

Model Architecture. We adopt DINOv3 [49] (ViT-B/16) as the encoder for all tasks. Detailed parameter counts and FLOPs at different sparsity levels are provided in the supplementary material. Object detection uses a standard deformable attention decoder [78]. For classification, semantic segmentation, and depth estimation, we employ linear probing. Video object segmentation and semantic matching are evaluated via feature matching, without any decoder.

Evaluation Metrics. The performance of classification, semantic segmentation, object detection, and semantic matching is evaluated using Top-1 Accuracy, mIoU, mAP with IoU thresholds in the range $[0.5 : 0.05 : 0.95]$, and PCK@0.05, respectively. Depth estimation is assessed using ARel (lower is better) and δ_1 (higher is better), while video object segmentation performance is measured by mean region similarity $\mathcal{J}_m$, mean contour-based accuracy $\mathcal{F}_m$, and $(\mathcal{J}\&\mathcal{F})_m$.

Implementation Details. The model is implemented using the PyTorch framework. Pruning is performed on a single GPU, while all other experiments are conducted on four NVIDIA A100 GPUs. For semantic segmentation, the learning rate is set to 1e-3, with training conducted over 100 epochs. For object detection, the learning rate is 5e-4, and the model is trained for 72 epochs. For depth estimation, the learning rate is 1e-6, with training spanning 60 epochs. The batch size is set to 16 for all tasks. For image classification, the learning rate is 5e-3, the batch size is 1024, and the model is trained for 20 epochs. Video object segmentation and semantic matching tasks do not require training; models pruned at various sparsity levels are directly validated on the datasets. We set $N = 1000$

Table 1: Quantitative results (§4.1) on DAVIS-2017 [39] video object segmentation using $(\mathcal{J}\&\mathcal{F})_m$ / $\mathcal{J}_m$ / $\mathcal{F}_m$.

Method	Dataset	Different Sparsity Levels of DINOv3		
		30%	20%	10%
Training-based (0% : 69.7 / 67.00 / 72.5)				
HydraViT [NeurIPS24] [16]	DAVIS	58.6 / 58.1 / 59.1	64.5 / 63.1 / 65.9	68.4 / 66.1 / 70.7
EA-ViT [ICCV25] [77]	DAVIS	60.3 / 59.4 / 61.2	65.4 / 63.9 / 66.9	69.3 / 66.8 / 71.8
Training-free (0% : 69.7 / 67.0 / 72.5)				
LAMP [ICLR21] [25]	ImageNet	35.0 / 32.9 / 37.0	40.4 / 38.1 / 42.8	60.2 / 58.4 / 62.1
NVIT [CVPR23] [62]	ImageNet	36.3 / 34.2 / 38.4	48.3 / 46.9 / 49.7	61.2 / 59.7 / 62.7
SNIP [ICCV23] [23]	ImageNet	36.5 / 33.8 / 39.2	51.0 / 49.2 / 52.9	61.9 / 60.3 / 64.4
SNOWS [ICLR25] [35]	ImageNet	52.9 / 50.7 / 55.1	59.4 / 57.8 / 61.0	63.8 / 61.3 / 66.3
SnapViT [NeurIPS25] [50]	ImageNet	56.0 / 54.3 / 57.7	60.8 / 58.8 / 62.8	65.0 / 63.4 / 66.7
Cut-ViT (Ours)	DAVIS	**59.4** / **58.0** / **60.8** ↑3.4 / ↑3.7 / ↑3.1	**65.0** / **63.2** / **66.8** ↑4.2 / ↑4.4 / ↑4.0	**68.7** / **66.3** / **71.2** ↑3.7 / ↑2.9 / ↑4.5

Table 2: Quantitative results (§4.1) of semantic matching on the FG3DCar / JODS / SBD (FJS) dataset [17, 33, 44].

Method	Dataset	Different Sparsity Levels of DINOv3		
		30%	20%	10%
Training-based (0% : 92.0 / 77.1 / 59.7)				
HydraViT [NeurIPS24] [16]	FJS	69.9 / 55.7 / 37.6	90.7 / 72.5 / 51.9	91.2 / 76.3 / 59.1
EA-ViT [ICCV25] [77]	FJS	72.0 / 57.9 / 38.5	91.4 / 74.2 / 54.6	92.0 / 77.0 / 59.5
Training-free (0% : 92.0 / 77.1 / 59.7)				
LAMP [ICLR21] [25]	ImageNet	55.2 / 38.8 / 29.9	79.5 / 56.9 / 38.9	84.2 / 59.2 / 50.4
NVIT [CVPR23] [62]	ImageNet	56.3 / 38.2 / 29.8	80.2 / 56.4 / 39.2	86.3 / 60.7 / 51.1
SNIP [ICCV23] [23]	ImageNet	59.6 / 39.1 / 30.4	82.3 / 58.9 / 40.9	88.4 / 63.6 / 53.2
SNOWS [ICLR25] [35]	ImageNet	64.8 / 44.2 / 32.1	85.3 / 62.9 / 50.6	90.3 / 71.7 / 54.9
SnapViT [NeurIPS25] [50]	ImageNet	65.7 / 43.3 / 32.4	86.0 / 64.0 / 50.8	90.2 / 75.1 / 56.0
Cut-ViT (Ours)	FJS	**70.8** / **56.3** / **37.9** ↑5.1 / ↑12.0 / ↑5.5	**90.6** / **73.0** / **53.2** ↑4.6 / ↑9.0 / ↑2.4	**92.0** / **76.9** / **58.9** ↑1.7 / ↑1.8 / ↑2.9

and the number of principal components $K = 192$ in the low-rank SVD. Except for image classification, which relies solely on the CLS token for basis-invariant subspace consistency, all other downstream tasks incorporate basis-agnostic and residual constraints from the perspectives of spatial and channel gram anchoring.

4.1 Comparison with SOTA Methods

Video Object Segmentation. We compare Cut-ViT with other state-of-the-art methods on the DAVIS-2017 dataset [39] for video object segmentation, as shown in Table 1. The proposed method significantly outperforms the training-free method on the $(\mathcal{J}\&\mathcal{F})_m$, $\mathcal{J}_m$, and $\mathcal{F}_m$ metrics at various pruning sparsity levels. Notably, at 20% sparsity, Cut-ViT improves the corresponding metrics by over **4.2%**, **4.4%**, and **4.0%**, respectively. When compared to training-based pruning methods, our approach also outperforms HydraViT [16], although it slightly trails behind EA-ViT [77] in performance. These results highlight the superiority of the backbone obtained using our pruning approach.

Semantic Matching. Table 2 reports the semantic matching results on the FJS dataset [17,33,44]. Across all datasets and pruning sparsity levels, Cut-ViT con-

Table 3: Quantitative results (§4.1) for object detection on COCO [32].

Method	Dataset	Different Sparsities		
		30%	20%	10%
Training-based (0% : 57.8)				
HydraViT[NeurIPS24] [16]	COCO	47.6	52.3	56.1
EA-ViT[ICCV25] [77]	COCO	49.4	54.1	56.9
Training-free (0% : 57.8)				
LAMP[ICLR21] [25]	ImageNet	42.3	43.8	50.6
NVIT[CVPR23] [62]	ImageNet	42.7	44.6	51.4
SNIP[ICCV23] [23]	ImageNet	43.8	45.2	48.4
SNOWS[ICLR25] [35]	ImageNet	45.1	50.6	54.4
SnapViT[NeurIPS25] [50]	ImageNet	45.8	51.0	54.1
Cut-ViT (Ours)	COCO	**48.6** ↑2.8	**53.0** ↑2.0	**56.5** ↑2.1

Table 4: Quantitative results (§4.1) for semantic segmentation on ADE20K [74].

Method	Dataset	Different Sparsities		
		30%	20%	10%
Training-based (0% : 51.8)				
HydraViT [NeurIPS24] [16]	ADE20K	46.4	48.9	49.6
EA-ViT [ICCV25] [77]	ADE20K	48.2	50.9	51.4
Training-free (0% : 51.8)				
LAMP [ICLR21] [25]	ImageNet	43.7	45.4	47.3
NVIT [CVPR23] [62]	ImageNet	43.5	46.1	47.9
SNIP [ICCV23] [23]	ImageNet	44.0	47.6	48.0
SNOWS [ICLR25] [35]	ImageNet	45.8	47.8	50.0
SnapViT [NeurIPS25] [50]	ImageNet	46.0	48.3	50.2
Cut-ViT (Ours)	ADE20K	**47.2** ↑1.2	**50.0** ↑1.7	**51.0** ↑0.8

sistently outperforms existing training-free methods as well as the training-based HydraViT [16]. For instance, at 30% sparsity, Cut-ViT improves performance over the training-free methods by **5.1%**, **12.1%**, and **5.5%** on FG3DCar, JODS, and PASCAL, respectively, and exceeds HydraViT [16] by 0.9%, 0.6%, and 0.3%. These results indicate that the proposed pruning strategy effectively preserves the fine-grained feature matching capability of DINOv3 [49].

Object Detection. Table 3 compares the performance of Cut-ViT with state-of-the-art methods on the COCO dataset [32]. Cut-ViT outperforms training-free approaches and achieves competitive performance relative to the training-based EA-ViT [77]. For instance, at 30% sparsity, Cut-ViT improves mAP by **2.8%** over the former, while slightly trailing behind the latter by 0.4%. These results further demonstrate that our approach more effectively preserves the detection capabilities of DINOv3 [49].

Semantic Segmentation. Table 4 presents experimental results comparing our method with SOTA approaches on the ADE20K dataset [74]. At sparsity levels of 30%, 20%, and 10%, Cut-ViT outperforms training-free methods by **1.2%**, **1.7%**, and **0.8%** in mIoU, respectively, and significantly surpasses the training-based HydraViT [16]. These results show that Cut-ViT preserves robust semantic information in subnetworks pruned from DINOv3 [49].

Depth Estimation. Table 5 presents the depth estimation results on the NYUv2 dataset [48], comparing the proposed method with state-of-the-art approaches. Consistent with its performance on other dense prediction tasks, Cut-ViT outperforms training-free pruning methods in both ARel and δ_1, while achieving competitive performance compared to training-based methods. Furthermore, when pruning an equal number of parameters, our approach yields models whose performance is closer to the original DINOv3 [49] than those produced by training-free pruning methods. These results across multiple dense prediction tasks demonstrate that Cut-ViT effectively preserves the robust dense feature representations of the native DINOv3 [49] during pruning.

Image Classification. Table 7 presents the image classification results of Cut-ViT on the ImageNet dataset [9], compared with state-of-the-art methods. Even without gram anchoring, the method outperforms training-free pruning approaches by relying solely on basis-agnostic and residual constraints. At sparsity levels of

Table 5: Quantitative results (§4.1) for depth estimation on NYUv2 [48] using ARel / δ_1.

Method	Dataset	Different Sparsities		
		30%	20%	10%
Training-based (0% : 3.0 / 99.9)				
HydraViT [NeurIPS24] [16]	NYUv2	5.2 / 98.4	4.9 / 99.0	4.1 / 99.7
EA-ViT [ICCV25] [77]	NYUv2	3.9 / 99.3	3.4 / 99.7	3.1 / 99.9
Training-free (0% : 3.0 / 99.9)				
LAMP [ICLR21] [25]	ImageNet	15.8 / 91.8	14.0 / 93.4	11.3 / 96.4
NVIT [CVPR23] [62]	ImageNet	14.4 / 93.3	12.7 / 95.1	10.3 / 96.1
SNIP [ICCV23] [23]	ImageNet	11.1 / 95.0	10.4 / 96.3	9.2 / 96.8
SNOWS [ICLR25] [35]	ImageNet	8.3 / 97.6	5.7 / 98.2	5.5 / 98.1
SnapViT [NeurIPS25] [50]	ImageNet	7.7 / 97.6	5.6 / 98.5	5.0 / 99.2
CUT-VIT (Ours)	NYUv2	**4.6** / **98.7** ↑3.1 / ↑1.1	**3.8** / **99.3** ↑1.8 / ↑0.8	**3.3** / **99.9** ↑1.7 / ↑0.7

Table 6: Quantitative results (§4.2) for out of distribution validation.

Method	Dataset	ADE20K → VOC	
		20%	10%
Training-based (0% : 76.0)			
HydraViT [NeurIPS24] [16]	ADE20K	73.6	74.8
EA-ViT [ICCV25] [77]	ADE20K	74.9	75.7
Training-free (0% : 76.0)			
LAMP [ICLR21] [25]	ImageNet	70.8	74.2
NVIT [CVPR23] [62]	ImageNet	70.7	73.9
SNIP [ICCV23] [23]	ImageNet	71.7	74.5
SNOWS [ICLR25] [35]	ImageNet	72.9	74.1
SnapViT [NeurIPS25] [50]	ImageNet	72.6	73.6
CUT-VIT (Ours)	ADE20K	**74.7** ↑1.8	**75.2** ↑1.1

Table 7: Quantitative results (§4.1) for image classification on ImageNet [9].

Method	Dataset	Different Sparsities		
		10%	20%	30%
Training-based (0% : 89.30)				
HydraViT [NeurIPS24] [16]	ImageNet	87.8	78.6	55.9
EA-ViT [ICCV25] [77]	ImageNet	89.1	81.7	60.3
Training-free (0% : 89.30)				
NVIT [CVPR23] [62]	ImageNet	71.3	61.9	54.7
SNIP [ICCV23] [23]	ImageNet	72.1	63.4	55.2
SNOWS [ICLR25] [35]	ImageNet	86.2	76.7	56.1
SnapViT [NeurIPS25] [50]	ImageNet	86.8	78.2	56.4
CUT-VIT (Ours)	ImageNet	**88.6** ↑1.8	**79.4** ↑1.2	**57.2** ↑0.8

Table 8: Complexity analysis (§4.3) on a single A100 GPU under the same settings.

Method	Pruning Time (Second)	GPU (GB)	mIoU
Training-based			
EA-ViT [ICCV25] [77]	13920	37.9	50.9
Training-free			
SNIP [ICCV23] [23]	220	27.0	47.6
SNOWS [ICLR25] [35]	6121	21.3	47.8
SnapViT [NeurIPS25] [50]	292	23.5	48.3
CUT-VIT (Ours)	**61**	**10.7**	**50.0**

10%, 20%, and 30%, it achieves gains of **1.8%**, **1.2%**, and **0.8%**, respectively. Furthermore, consistent with dense prediction tasks, CUT-VIT achieves competitive performance relative to training-based pruning methods. These results indicate that basis-invariant subspace consistency, even when applied solely to CLS features, effectively captures rich global information.

4.2 Out of Distribution Generalization

The generalization of CUT-VIT is evaluated by performing inference on the PASCAL VOC 2012 dataset [11] using models trained on ADE20K [74]. Table 6 compares its performance with state-of-the-art methods. CUT-VIT consistently outperforms training-free pruning approaches across all sparsity levels, achieving a notable gain of **1.5%** mIoU at 20% sparsity. These results provide strong evidence that the proposed pruning strategy enables the subnetworks to effectively inherit the generalization ability of the native DINOv3 [49].

4.3 Ablation Study

Complexity Analysis. Table 8 presents a comparison of the algorithmic complexity of CUT-VIT with state-of-the-art methods. Compared with both training-free and training-based methods, CUT-VIT achieves promising performance while

Table 9: Component analysis (§4.3) of basis-agnostic, residual constraints and task-specific on ADE20K [74] with 20% sparsity.

Baseline	Basis-agnostic Constraint	Residual Constraint	Task-specific	mIoU
✓				48.4
✓	✓			49.1
✓		✓		49.3
✓	✓	✓	✓	**49.7**

Table 10: Component analysis (§4.3) of spatial and channel gram anchoring on ADE20K [74] with 20% sparsity.

Baseline	Spatial Gram Anchoring	Channel Gram Anchoring	mIoU
✓			48.4
✓	✓		49.7
✓		✓	49.5
✓	✓	✓	**50.0**

Table 11: Parameter analysis (§4.3) of principal component number K on ADE20K [74] at 20% sparsity. CEVR denotes the cumulative explained variance ratio.

K	96	128	144	192	256	300	500
CEVR	91.8	93.9	96.6	98.9	99.2	99.4	99.6
Time	0.43	0.46	0.5	0.59	0.72	0.80	1.12
Classification	79.1	79.2	79.2	79.4	79.4	79.4	79.5
Segmentation	49.6	49.8	49.9	50.0	50.1	50.1	50.2

Table 12: Performance analysis (§4.3) between static weighting and spectral entropy weighting across different tasks at 30% sparsity.

Setup	Static Weighting	Spectral Entropy Weighting
Detection	47.9	**48.6**
Segmentation	58.8 / 57.2 / 60.3	**59.4** / **58.0** / **60.8**
Matching	70.2 / 55.5 / 37.3	**70.8** / **56.3** / **37.9**

significantly reducing pruning time and GPU memory usage. Specifically, it requires only **20.9%** of the time and **45.5%** of the GPU memory required by SnapViT [50], while delivering a **1.7%** improvement in mIoU. Although the method trails training-based pruning approaches, such as EA-ViT [77], by 0.9% in performance, it requires only **0.44%** of the time and **28.2%** of the GPU memory. These results demonstrate that competitive performance can be achieved with minimal computational overhead.

Component Analysis. The results for each component and the different gram matrices are reported in Tables 9 and 10, respectively. Local dense feature alignment $\mathcal{L}_{ba}$ is adopted as the baseline. The basis-agnostic constraint, residual constraint, and task-specific adaptation mutually complement each other, enabling the model to achieve excellent experimental performance. Furthermore, incorporating spatial and channel gram matrices encourages the pruned model to learn spatial contextual information and channel-wise semantic representations aligned with those of DINOv3 [49]. The complementary effects of the two gram matrices allow the model to accurately capture semantic and spatial contextual information, thereby achieving optimal performance.

Parameter & Weighting Analysis. Tables 11 and 12 present the performance analysis of the number of principal components K and different loss weightings, respectively. For the former, we analyze it from the following three aspects: **i)** Information preservation. We introduce the *cumulative explained variance ratio* (CEVR) to quantify the information retained by the selected Top-K principal components. **ii)** Efficiency. We measure the time to obtain the corresponding principal components from a single matrix via SVD. **iii)** Cross-task robustness. The results below confirm that $K = 192$ achieves an optimal trade-off among information preservation, computational efficiency, and performance. For the

Table 13: Ablation analysis (§4.3) on different pruning datasets at 30% sparsity.

Method	Dataset	Video Object Segmentation	Depth Estimation	Semantic Matching
		$(\mathcal{J}\&\mathcal{F})_m$ / $\mathcal{J}_m$ / $\mathcal{F}_m$	ARel ↓ / δ_1 ↑	FG3DCar / JODS / SBD
SNOWS [ICLR25] [35]	ImageNet	52.9 / 50.7 / 55.1	8.3 / 97.6	64.8 / 44.2 / 32.1
	Target	53.4 / 50.9 / 56.0	8.0 / 97.6	65.4 / 44.6 / 32.6
SnapViT [NeurIPS25] [50]	ImageNet	56.0 / 54.3 / 57.7	7.7 / 97.6	65.7 / 43.3 /32.4
	Target	56.4 / 54.6 / 58.1	7.3 / 97.7	66.4 / 43.5 / 32.8
CUT-VIT (Ours)	Target	**59.4** / **58.0** / **60.8** ↑3.0 / ↑3.4 / ↑2.7	**4.6** / **98.7** ↑2.7 / ↑1.0	**70.8** / **56.3** / **37.9** ↑4.4 / ↑11.7 / ↑5.1

Table 14: Robustness analysis (§4.3) of different architectures on their corresponding tasks at 30% sparsity.

Method	SAM	DeiT	CLIP
	Promptable Seg.	Video Object Seg.	Open-vocabulary Seg.
SNOWS [35]	67.9	40.1 / 42.3 /37.8	45.7
SnapViT [50]	68.6	40.6 / 43.0 / 38.1	46.9
CUT-VIT	**71.3** ↑2.7	**44.6** ↑4.0 / **46.7** ↑3.7 / **42.5** ↑4.4	**51.1** ↑3.2

Table 15: Ablation analysis (§4.3) between MSE and BI.

Baseline	MSE	BI	mAP
✓			54.3
✓	✓		54.7
✓		✓	**55.7**

latter, the performance on object detection, video object segmentation, and semantic matching highlights the superiority of spectral entropy weighting.

Pruning Dataset Analysis. We conduct an ablation analysis on pruning with different datasets across various tasks, as shown in Table 13. The experimental results on the three tasks show that even when pruning is performed on the corresponding target datasets, the state-of-the-art SNOWS [35] and SnapViT [50] achieve only marginal performance gains. This directly demonstrates that the performance improvements stem from the superiority of our method itself, rather than from simply replacing the pruning dataset.

Architecture Robustness Analysis. In addition to DINOv3 [49], we further apply our method to SAM [22], DeiT [53], and CLIP [42], and compare it against state-of-the-art methods to validate its robustness across architectures, as reported in Table 14. Their performance on prompt segmentation, video object segmentation, and open-vocabulary segmentation is evaluated using 0.1% of SA-1B, DAVIS-2017, and PASCAL-Context, respectively. The results demonstrate that our method generalizes robustly across diverse architectures.

Subspace Alignment Analysis. Table 15 and Fig. 4 provide an extensive analysis of subspace alignment strategies from quantitative and qualitative perspectives, respectively. Taken together, these results show that naively aligning and pruning models with standard distance metrics (*e.g.*, MSE) penalizes meaningless rotational differences, leading to ill-posed optimization. In contrast, our basis invariance (BI) proves that $\mathcal{L}_{basis}$ depends exclusively on the geometric overlap between subspaces and is invariant to basis rotations, thereby providing a robust foundation for subspace alignment. Compared with the diffused representations induced by MSE, BI preserves the precise object boundaries of DINOv3 [49] representations and achieves superior performance.

Retention & **Overlap Analysis.** Fig. 5 shows the parameter retention and overlap ratios for both training-free OSP and our task-specific methods across different blocks. As illustrated in (**Left**), the parameter retention rate of the generic method fluctuates significantly from the 3rd to the 10th layer, while our

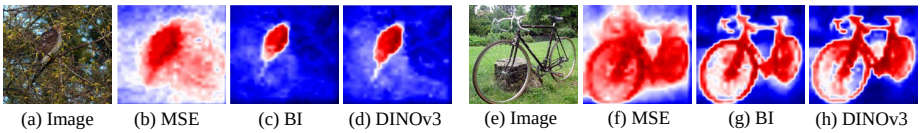

Fig. 4: Qualitative comparison (§4.3) of different subspace alignment strategies.

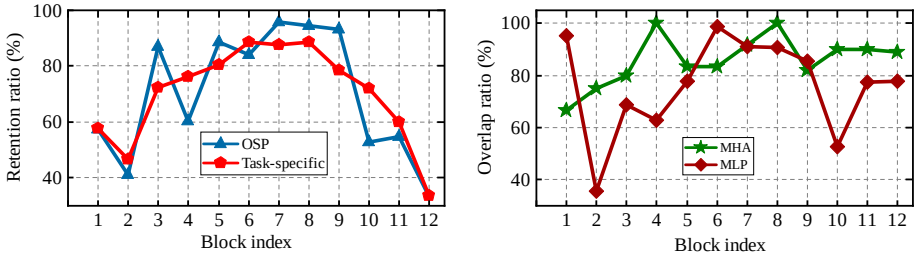

Fig. 5: The parameter retention and overlap analysis (§4.3) for each block in DINOv3 [49]. (**Left**) shows the parameter retention rate for each block in both training-free OSP and our task-specific methods. (**Right**) illustrates the overlap ratio of the retained parameters in the MLP and MHA layers of each block for the two types of methods.

task-specific method exhibits more stable changes, indicating that our approach better preserves the fine-grained texture features of DINOv3 [49]. The curve in (**Right**) reveals that both pruning paradigms focus more on the information within the attention layers across several intermediate blocks.

5 Conclusion

In this paper, we propose Cut-ViT, a task-specific model pruning paradigm to mitigate robustness degradation and task-specificity deficiency present in existing work. Specifically, we construct gram anchoring matrices from spatial and channel perspectives and apply subspace decomposition to extract the corresponding bases in latent space. Basis-agnostic and residual constraints are designed to enforce subspace consistency between the original and pruning-oriented DINOv3 models across both domains. This enables the subnetwork to inherit the robust feature representations of foundation models such as DINOv3. Moreover, we introduce spectral entropy adaptation, which quantifies the information density of feature manifolds along spatial and channel dimensions, thereby adapting the pruning objective to specific downstream tasks. Extensive experiments on **six** tasks across **nine** datasets show that Cut-ViT achieves state-of-the-art performance with minimal computational overhead compared to prior approaches.

Acknowledgments. This work was supported by the National Defense Science and Technology Industry Bureau Technology Infrastructure Project (JSZL2024606C001).

References

1. Bergner, B., Lippert, C., Mahendran, A.: Token cropr: Faster vits for quite a few tasks. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 9740–9750 (2025)
2. Bolya, D., Fu, C.Y., Dai, X., Zhang, P., Feichtenhofer, C., Hoffman, J.: Token merging: Your vit but faster. arXiv preprint arXiv:2210.09461 (2022)
3. Cai, X., Li, L., Pei, G., Chen, T., Pan, J., Yao, Y., Wang, W.: Unbiased object detection beyond frequency with visually prompted image synthesis. In: Int. Conf. Learn. Represent. (2026)
4. Cai, X., Pei, G., Sun, Z., Yao, Y., Shen, F., Wang, W.: Iris: Bringing real-world priors into diffusion model for monocular depth estimation. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 26909–26919 (2026)
5. Caron, M., Touvron, H., Misra, I., Jégou, H., Mairal, J., Bojanowski, P., Joulin, A.: Emerging properties in self-supervised vision transformers. In: Int. Conf. Comput. Vis. pp. 9650–9660 (2021)
6. Chen, J., Ye, L., He, J., Wang, Z.Y., Khashabi, D., Yuille, A.: Efficient large multimodal models via visual context compression. Adv. Neural Inform. Process. Syst. **37**, 73986–74007 (2024)
7. Chen, L., Zhao, H., Liu, T., Bai, S., Lin, J., Zhou, C., Chang, B.: An image is worth 1/2 tokens after layer 2: Plug-and-play inference acceleration for large vision-language models. In: Eur. Conf. Comput. Vis. pp. 19–35 (2024)
8. Chen, Z., Li, K., Jia, Y., Ye, L., Ma, Y.: Accelerating diffusion transformer via increment-calibrated caching with channel-aware singular value decomposition. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 18011–18020 (2025)
9. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: Imagenet: A large-scale hierarchical image database. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 248–255 (2009)
10. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D.: An image is worth 16x16 words: Transformers for image recognition at scale. In: Int. Conf. Learn. Represent. (2021)
11. Everingham, M., Eslami, S.A., Van Gool, L., Williams, C.K., Winn, J., Zisserman, A.: The pascal visual object classes challenge: A retrospective. Int. J. Comput. Vis. **111**(1), 98–136 (2015)
12. Frantar, E., Alistarh, D.: Sparsegpt: Massive language models can be accurately pruned in one-shot. In: Int. Conf. Mach. Learn. pp. 10323–10337 (2023)
13. Gholami, M., Akbari, M., Cannons, K., Zhang, Y.: Casp: Compression of large multimodal models based on attention sparsity. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 9372–9381 (2025)
14. Glasmachers, T., Schaul, T., Yi, S., Wierstra, D., Schmidhuber, J.: Exponential natural evolution strategies. In: Proceedings of the 12th annual conference on Genetic and evolutionary computation. pp. 393–400 (2010)
15. Gu, H., Pei, G., Sun, Z., Ren, M., Shu, X., Yao, Y., Shen, F.: Medfg-vqa: Low-frequency memory and graph attention for lightweight medical vqa. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 42755–42764 (2026)
16. Haberer, J., Hojjat, A., Landsiedel, O.: Hydravit: Stacking heads for a scalable vit. Adv. Neural Inform. Process. Syst. **37**, 40254–40277 (2024)
17. Hariharan, B., Arbeláez, P., Bourdev, L., Maji, S., Malik, J.: Semantic contours from inverse detectors. In: Int. Conf. Comput. Vis. pp. 991–998 (2011)
18. He, K., Chen, X., Xie, S., Li, Y., Dollár, P., Girshick, R.: Masked autoencoders are scalable vision learners. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 16000–16009 (2022)

19. He, Y., Liu, L., Liu, J., Wu, W., Zhou, H., Zhuang, B.: Ptqd: Accurate post-training quantization for diffusion models. Adv. Neural Inform. Process. Syst. **36**, 13237–13249 (2023)
20. Kim, D., Park, S., Han, G., Kim, S.W., Seo, P.H.: Random conditioning with distillation for data-efficient diffusion model compression. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 18607–18618 (2025)
21. Kim, K., Park, J., Kim, J., Kwon, H., Sohn, K.: Faster parameter-efficient tuning with token redundancy reduction. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 30189–30198 (2025)
22. Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A.C., Lo, W.Y., et al.: Segment anything. In: Int. Conf. Comput. Vis. pp. 4015–4026 (2023)
23. Kohama, H., Minoura, H., Hirakawa, T., Yamashita, T., Fujiyoshi, H.: Single-shot pruning for pre-trained models: Rethinking the importance of magnitude pruning. In: Int. Conf. Comput. Vis. pp. 1433–1442 (2023)
24. Kwon, W., Kim, S., Mahoney, M.W., Hassoun, J., Keutzer, K., Gholami, A.: A fast post-training pruning framework for transformers. Adv. Neural Inform. Process. Syst. **35**, 24101–24116 (2022)
25. Lee, J., Park, S., Mo, S., Ahn, S., Shin, J.: Layer-adaptive sparsity for the magnitude-based pruning. In: Int. Conf. Learn. Represent. (2021)
26. Lee, J., Das, D., Hayat, M., Choi, S., Hwang, K., Porikli, F.: Customkd: Customizing large vision foundation for edge model improvement via knowledge distillation. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 25176–25186 (2025)
27. Lee, S., Choi, J., Kim, H.J.: Multi-criteria token fusion with one-step-ahead attention for efficient vision transformers. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 15741–15750 (2024)
28. Li, L., Wang, W., Yang, Y.: Human-object interaction detection collaborated with large relation-driven diffusion models. Adv. Neural Inform. Process. Syst. **37**, 23655–23678 (2024)
29. Li, L., Wang, W., Zhou, T., Quan, R., Yang, Y.: Semantic hierarchy-aware segmentation. IEEE Trans. Pattern Anal. Mach. Intell. **46**(4), 2123–2138 (2023)
30. Li, L., Yang, Y., Wang, W.: Reconciling visual perception and generation in diffusion models. In: Int. Conf. Learn. Represent. (2026)
31. Li, Y., Yang, C., Zeng, H., Dong, Z., An, Z., Xu, Y., Tian, Y., Wu, H.: Frequency-aligned knowledge distillation for lightweight spatiotemporal forecasting. In: Int. Conf. Comput. Vis. pp. 7262–7272 (2025)
32. Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft coco: Common objects in context. In: Eur. Conf. Comput. Vis. pp. 740–755 (2014)
33. Lin, Y.L., Morariu, V.I., Hsu, W., Davis, L.S.: Jointly optimizing 3d model fitting and fine-grained classification. In: Eur. Conf. Comput. Vis. pp. 466–480 (2014)
34. Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., Guo, B.: Swin transformer: Hierarchical vision transformer using shifted windows. In: Int. Conf. Comput. Vis. pp. 10012–10022 (2021)
35. Lucas, R., Mazumder, R.: Preserving deep representations in one-shot pruning: A hessian-free second-order optimization framework. In: Int. Conf. Learn. Represent. (2025)
36. Luo, S., Yang, H., Xin, Y., Yi, M., Wu, G., Zhai, G., Liu, X.: Tr-pts: Task-relevant parameter and token selection for efficient tuning. In: Int. Conf. Comput. Vis. pp. 4360–4369 (2025)

37. Ma, C., Xiao, X., Wang, T., Wang, X., Shen, Y.: Learning straight flows: Variational flow matching for efficient generation. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 38154–38164 (2026)
38. Oquab, M., Darcet, T., Moutakanni, T., Vo, H., Szafraniec, M., Khalidov, V., Fernandez, P., Haziza, D., Massa, F., El-Nouby, A., et al.: Dinov2: Learning robust visual features without supervision. Trans. Mach. Learn Res. (2024)
39. Pont-Tuset, J., Perazzi, F., Caelles, S., Arbeláez, P., Sorkine-Hornung, A., Van Gool, L.: The 2017 davis challenge on video object segmentation. arXiv preprint arXiv:1704.00675 (2017)
40. Qu, W., Mei, G., Wang, J., Wu, Y., Huang, X., Xiao, L.: Robust single-stage fully sparse 3d object detection via detachable latent diffusion. In: AAAI. vol. 40, pp. 8668–8676 (2026)
41. Qu, W., Mei, G., Wu, Y., Gong, Y., Huang, X., Xiao, L.: A self-conditioned representation guided diffusion model for realistic text-to-lidar scene generation. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 9434–9444 (2026)
42. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: Int. Conf. Mach. Learn. pp. 8748–8763 (2021)
43. Rastegari, M., Ordonez, V., Redmon, J., Farhadi, A.: Xnor-net: Imagenet classification using binary convolutional neural networks. In: Eur. Conf. Comput. Vis. pp. 525–542 (2016)
44. Rubinstein, M., Joulin, A., Kopf, J., Liu, C.: Unsupervised joint object discovery and segmentation in internet images. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 1939–1946 (2013)
45. Sanh, V., Wolf, T., Rush, A.: Movement pruning: Adaptive sparsity by fine-tuning. Adv. Neural Inform. Process. Syst. **33**, 20378–20389 (2020)
46. Shang, J., Schmeckpeper, K., May, B.B., Minniti, M.V., Kelestemur, T., Watkins, D., Herlant, L.: Theia: Distilling diverse vision foundation models for robot learning. In: CoRL. pp. 724–748 (2025)
47. Shu, H., Li, W., Tang, Y., Zhang, Y., Chen, Y., Li, H., Wang, Y., Chen, X.: Tinysam: Pushing the envelope for efficient segment anything model. In: AAAI. vol. 39, pp. 20470–20478 (2025)
48. Silberman, N., Hoiem, D., Kohli, P., Fergus, R.: Indoor segmentation and support inference from rgbd images. In: Eur. Conf. Comput. Vis. pp. 746–760 (2012)
49. Siméoni, O., Vo, H.V., Seitzer, M., Baldassarre, F., Oquab, M., Jose, C., Khalidov, V., Szafraniec, M., Yi, S., Ramamonjisoa, M., et al.: Dinov3. arXiv preprint arXiv:2508.10104 (2025)
50. Simoncini, W., Dorkenwald, M., Blankevoort, T., Snoek, C.G., Asano, Y.M.: Elastic vits from pretrained models without retraining. Adv. Neural Inform. Process. Syst. pp. 2123–2138 (2025)
51. Sun, M., Liu, Z., Bair, A., Kolter, J.Z.: A simple and effective pruning approach for large language models. In: Int. Conf. Learn. Represent. (2024)
52. Sun, Z., Yao, Y., Liu, T., Li, Z., Shen, F., Tang, J.: Jo-snc: Combating noisy labels through fostering self- and neighbor-consistency. IEEE Trans. Pattern Anal. Mach. Intell. **48**(4), 4708–4725 (2026)
53. Touvron, H., Cord, M., Douze, M., Massa, F., Sablayrolles, A., Jégou, H.: Training data-efficient image transformers & distillation through attention. In: Int. Conf. Mach. Learn. pp. 10347–10357 (2021)
54. Wang, C., Wang, Z., Xu, X., Tang, Y., Zhou, J., Lu, J.: Q-vlm: Post-training quantization for large vision-language models. Adv. Neural Inform. Process. Syst. **37**, 114553–114573 (2024)

55. Wang, H., Shang, Y., Yuan, Z., Wu, J., Yan, J., Yan, Y.: Quest: Low-bit diffusion model quantization via efficient selective finetuning. In: Int. Conf. Comput. Vis. pp. 15542–15551 (2025)
56. Wang, X., Zheng, Y., Wan, Z., Zhang, M.: Svd-llm: Truncation-aware singular value decomposition for large language model compression. In: Int. Conf. Learn. Represent. vol. 2025, pp. 19299–19319 (2025)
57. Woo, S., Debnath, S., Hu, R., Chen, X., Liu, Z., Kweon, I.S., Xie, S.: Convnext v2: Co-designing and scaling convnets with masked autoencoders. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 16133–16142 (2023)
58. Xiao, X., Zhang, Y., Li, X., Wang, T., Wang, X., Wei, Y., Hamm, J., Xu, M.: Visual instance-aware prompt tuning. In: ACM Int. Conf. Multimedia. pp. 2880–2889 (2025)
59. Xiao, X., Zhang, Y., Zhao, L., Liu, Y., Liao, X., Mai, Z., Li, X., Wang, X., Xu, H., Hamm, J., et al.: Prompt-based adaptation in large-scale vision models: A survey. arXiv preprint arXiv:2510.13219 (2025)
60. Xiong, Y., Varadarajan, B., Wu, L., Xiang, X., Xiao, F., Zhu, C., Dai, X., Wang, D., Sun, F., Iandola, F., et al.: Efficientsam: Leveraged masked image pretraining for efficient segment anything. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 16111–16121 (2024)
61. Xu, J., Pei, G., Liu, H., Yao, Y.: Gsv2x: Geometry-aware uncertainty modeling and orthogonal fusion for robust roadside perception. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 21409–21419 (2026)
62. Yang, H., Yin, H., Shen, M., Molchanov, P., Li, H., Kautz, J.: Global vision transformer pruning with hessian-aware saliency. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 18547–18557 (2023)
63. Yang, L., Kang, B., Huang, Z., Xu, X., Feng, J., Zhao, H.: Depth anything: Unleashing the power of large-scale unlabeled data. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 10371–10381 (2024)
64. Yang, Z., Pei, G., Chen, T., Yuan, X., Zhang, H., Shu, X., Yao, Y.: Beyond quadratic: Linear-time change detection with rwkv. In: AAAI. vol. 40, pp. 11811–11819 (2026)
65. Yang, Z., Pei, G., Chen, T., Zhou, Y., Zhou, T., Yao, Y., Shen, F.: Efficiency follows global-local decoupling. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 25524–25535 (2026)
66. Yang, Z., Pei, G., Yao, Y., Zhou, T., Ding, L., Shen, F.: ChangeTitans: Toward remote sensing change detection with neural memory. IEEE Trans. Geosci. Remote Sens. **63**, 1–14 (2025)
67. Yin, J., Chen, T., Chen, Y., Pei, G., Shu, X., Yao, Y., Shen, F.: Pca-seg: Revisiting cost aggregation for open-vocabulary semantic and part segmentation. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 27633–27643 (2026)
68. Yin, J., Chen, T., Pei, G., Liu, H., Yao, Y., Nie, L., Hua, X.: Semi-supervised semantic segmentation with multi-constraint consistency learning. IEEE Trans. Multimedia **27**, 6449–6461 (2025)
69. Yin, J., Jiang, X., Chen, T., Pei, G., Yao, Y., Shen, F., Shen, H.T.: Depmatch: Boosting semi-supervised semantic segmentation by exploring depth difference knowledge. IEEE Trans. Image Process. **35**, 3256–3270 (2026)
70. Zeng, F., Yu, D., Kong, Z., Tang, H.: Token transforming: A unified and training-free token compression framework for vision transformer acceleration. arXiv preprint arXiv:2506.05709 (2025)

71. Zhang, Y., Coskun, H., Ma, X., Wang, H., Ma, K., Chen, X., Hu, D.H., Fu, Y.: Slicing vision transformer for flexible inference. Adv. Neural Inform. Process. Syst. **37**, 42649–42671 (2024)
72. Zhou, B., Lai, Q., Sun, Z., Shu, X., Yao, Y., Wang, W.: Learning 3d representations for spatial intelligence from unposed multi-view images. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 22550–22560 (2026)
73. Zhou, B., Li, L., Wang, Y., Liu, H., Yao, Y., Wang, W.: Unialign: Scaling multimodal alignment within one unified model. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 29644–29655 (2025)
74. Zhou, B., Zhao, H., Puig, X., Fidler, S., Barriuso, A., Torralba, A.: Scene parsing through ade20k dataset. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 633–641 (2017)
75. Zhou, T., Wang, W.: Prototype-based semantic segmentation. IEEE Trans. Pattern Anal. Mach. Intell. **46**(10), 6858–6872 (2024)
76. Zhou, Y., Gao, X., Chen, Z., Huang, H.: Attention distillation: A unified approach to visual characteristics transfer. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 18270–18280 (2025)
77. Zhu, C., Zhao, W., Zhang, H., Zhou, Y., Tang, W., Wang, S., Yuan, Z., Shang, Y., Peng, X., Wang, K., et al.: Ea-vit: Efficient adaptation for elastic vision transformer. In: Int. Conf. Comput. Vis. pp. 1038–1047 (2025)
78. Zhu, X., Su, W., Lu, L., Li, B., Wang, X., Dai, J.: Deformable detr: Deformable transformers for end-to-end object detection. arXiv preprint arXiv:2010.04159 (2020)