Steering Sparse Autoencoder Latents to Control Dynamic Head Pruning in Vision Transformers (Student Abstract)

Yousung Lee[1], **Dongsoo Har**[1]

[1]Korea Advanced Institute of Science and Technology, Daejeon 34051, Korea
yslee410@kaist.ac.kr, dshar@kaist.ac.kr

Abstract

Dynamic head pruning in Vision Transformers (ViTs) improves efficiency by removing redundant attention heads, but existing pruning policies are often difficult to interpret and control. In this work, we propose a novel framework by integrating Sparse Autoencoders (SAEs) with dynamic pruning, leveraging their ability to disentangle dense embeddings into interpretable and controllable sparse latents. Specifically, we train an SAE on the final-layer residual embedding of the ViT and amplify the sparse latents with different strategies to alter pruning decisions. Among them, per-class steering reveals compact, class-specific head subsets that preserve accuracy. For example, *bowl* improves accuracy (76%→82%) while reducing head usage (0.72→0.33) via heads h_2 and h_5. These results show that sparse latent features enable class-specific control of dynamic pruning, effectively bridging pruning efficiency and mechanistic interpretability in ViTs.

Introduction

Vision Transformers (ViTs) leverage multi-head self-attention to capture diverse token interactions. However, many attention heads are redundant, increasing computation without proportional performance gain. To address this, adaptive frameworks such as AdaViT (Meng et al. 2022) have been introduced, which use auxiliary networks to select which heads to prune. This input-dependent pruning substantially reduces computation while preserving accuracy.

However, a key limitation remains: since these pruning policies rely on residual embeddings, the decision process is often opaque and difficult to control at the latent level. As a result, while dynamic head pruning improves efficiency, its mechanism lacks interpretability. If pruning decisions could be explained or controlled at the latent level, head selection would become both interpretable and controllable.

Sparse Autoencoders (SAEs) offer a natural tool for this purpose, as they disentangle dense, polysemantic embeddings into sparse latent features that tend to encode more monosemantic and interpretable concepts in transformer representations (Cunningham et al. 2023; Lim et al. 2025). Recent studies have shown that this disentanglement enables steering specific SAE latent dimensions to control model behavior in a desired direction (Kang, Wang, and Xiong 2024; Chatzoudis et al. 2025).

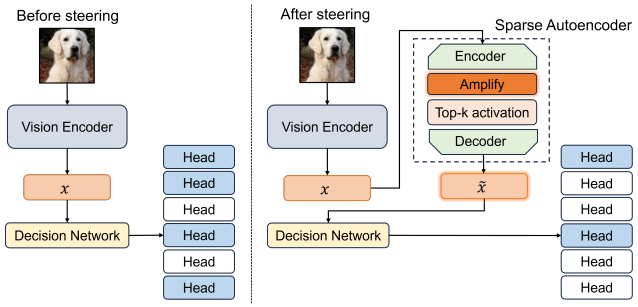

Figure 1: Comparison of dynamic head pruning before (left) and after (right) SAE latent steering, where x denotes the CLS token from the final-layer residual stream input, and $\tilde{x}$ denotes the steered embedding.

In this work, we propose a novel framework that integrates k-sparse autoencoders (Makhzani and Frey 2013) with dynamic head pruning in ViTs to make pruning decisions controllable at the latent level. This framework also enhances interpretability by revealing class-specific head subsets. As illustrated in Figure 1, we amplify selected SAE latent dimensions, reconstruct the steered embedding, and feed it into the decision network to observe how pruning behavior changes. Our experiments show that per-class latent steering is particularly effective, reducing head usage while largely maintaining accuracy. Overall, these results suggest that sparse latents provide an effective way to interpret and control dynamic pruning, bridging the gap between efficiency and mechanistic interpretability in ViTs.

Method

In this work, we adopt AdaViT as the baseline dynamic pruning framework, with a particular focus on head pruning. We first train a Vision Transformer with layer-wise decision networks. In this setup, each lightweight network receives the class (CLS) token from the residual stream input and outputs head importance logits $a_\ell \in \mathbb{R}^H$, where ℓ denotes the layer index and H the number of attention heads. We use the CLS token as input for the decision network since it encodes global context relevant for classifica-

tion. Binary masks $M_{\ell,i} \in \{0,1\}$ are obtained via Gumbel-Sigmoid sampling from $a_{\ell,i}$, where i denotes the head index. The masked attention is computed as follows:

$$h_{\ell,i} = M_{\ell,i}\,\mathrm{Attn}(Q,K,V)_{\ell,i}. \tag{1}$$

The Vision Transformer and the decision network are trained jointly to preserve accuracy while enforcing head sparsity toward a target head usage ratio. After training, we extract $x \in \mathbb{R}^d$, the `CLS` token from the final layer's residual input and use it to train a k-sparse autoencoder. Formally, the encoder and decoder are given by

$$z = \mathrm{TopK}(W_{\mathrm{enc}}(x - b_{\mathrm{dec}})), \quad W_{\mathrm{enc}} \in \mathbb{R}^{n\times d}, \tag{2}$$

$$\hat{x} = W_{\mathrm{dec}} z + b_{\mathrm{dec}}, \quad W_{\mathrm{dec}} \in \mathbb{R}^{d\times n}, \tag{3}$$

where z is the sparse latent representation of x, $\hat{x}$ is its reconstruction, and b_{dec} denotes the decoder bias. The parameters n and d represent the SAE latent and input embedding dimensions, respectively. Sparsity is enforced through a top-k activation, which preserves only the k largest dimensions. The SAE is trained with the following mean squared error (MSE) reconstruction objective:

$$\mathcal{L}_{\mathrm{rec}} = \|x - \hat{x}\|_2^2, \tag{4}$$

which encourages $\hat{x}$ to remain close to x. After training the SAE, we amplify the selected latent dimensions S by

$$z'_i = \begin{cases} z_i + \alpha, & i \in S, \\ z_i, & i \notin S, \end{cases} \qquad \tilde{z} = \mathrm{TopK}(z'), \tag{5}$$

where α is the amplification strength and $\tilde{z}$ denotes the steered sparse latent vector obtained after amplification and top-k activation. The steered embedding is then reconstructed as $\tilde{x} = W_{\mathrm{dec}}\tilde{z} + b_{\mathrm{dec}}$, which is finally fed into the decision network to obtain the steered pruning mask.

Experiments

Dynamic pruning baseline. We fine-tune an ImageNet-pretrained ViT-Small (12 layers, 6 heads per layer, 384-d embeddings) on CIFAR-100, reaching 91.27% accuracy. Each layer employs a single decision network for head selection, as described in the Method section. The final model prunes 30% of heads while maintaining 89.79% accuracy.

Sparse autoencoder. The SAE expands the 384-d residual embedding into the 3072-d latent space ($8\times$) with top-k activation ($k = 64$). It is trained for 100 epochs, achieving an MSE loss of 0.0228. Replacing the original embedding with its reconstruction yields only minor differences in accuracy (-0.12%) and head usage ratio ($+0.025$), showing that the SAE preserves essential information for effective pruning.

Steering dynamic pruning. Inspired by the top-k masking experiments in PatchSAE (Lim et al. 2025), we adapt this idea to dynamic head pruning using the latent steering defined in Eq. (5), which amplifies selected SAE latent dimensions. For each sample, we evaluate three strategies for defining the index set S based on training-set activation frequency statistics: (1) **Per-class frequent** — top-k most frequently activated latents within each class, (2) **Global frequent** — top-k most frequently activated latents across all classes, and (3) **Random**. Figure 2 shows that per-class steering reduces head usage while largely preserving accuracy, whereas global and random strategies lead to larger accuracy drops as α increases. The low overlap between global and per-class top-k frequent latent dimensions (0.1641) indicates that the SAE captures class-discriminative concepts. Figure 3 further illustrates head usage patterns in the final layer under per-class steering. For *bowl*, accuracy rises ($76\% \rightarrow 82\%$) while head usage falls ($0.72 \rightarrow 0.33$), mainly relying on h_2 and h_5. *pine_tree* shows a similar pattern ($79\% \rightarrow 84\%$, $0.93 \rightarrow 0.35$), relying on h_2 and h_3. Interestingly, semantically related classes such as *bowl* and *plate* share similar head subsets h_2 and h_5, indicating that per-class steering reveals class-level semantic relationships among heads. These examples suggest that amplifying per-class top-k frequent activations enriches class-specific signals in the decision network input, leading to class-specific pruning behaviors. Overall, these results demonstrate that the Sparse Autoencoder provides an effective way to interpret and control dynamic head pruning in ViTs.

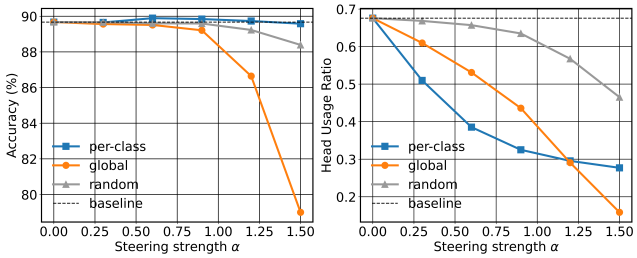

Figure 2: Accuracy (%) and head usage ratio in the final layer under different strategies as α increases from 0 to 1.5.

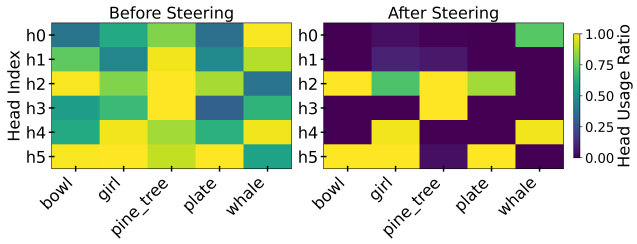

Figure 3: Effect of per-class steering ($\alpha = 1.2$) on head activation patterns. Columns indicate the top-5 classes ranked by accuracy gain, and rows correspond to heads (h_0–h_5).

Conclusion

This paper introduces a novel Sparse Autoencoder-based framework that makes dynamic head pruning interpretable and controllable at the latent level in Vision Transformers. By amplifying per-class frequent activations, we reveal class-specific pruning behaviors that are both efficient and interpretable. Our current work focuses on the final layer and small datasets, and future work will extend the framework to earlier layers and foundation models.

Acknowledgments

This work was supported by the Technology Innovation Program (RS-2025-02613131) funded by the Ministry of Trade, Industry & Energy (MOTIE, Korea).

References

Chatzoudis, G.; Li, Z.; Moran, G. E.; Wang, H.; and Metaxas, D. N. 2025. Visual Sparse Steering: Improving Zero-shot Image Classification with Sparsity Guided Steering Vectors. *arXiv preprint arXiv:2506.01247*.

Cunningham, H.; Ewart, A.; Riggs, L.; Huben, R.; and Sharkey, L. 2023. Sparse autoencoders find highly interpretable features in language models. *arXiv preprint arXiv:2309.08600*.

Kang, H.; Wang, T.; and Xiong, C. 2024. Interpret and control dense retrieval with sparse latent features. *arXiv preprint arXiv:2411.00786*.

Lim, H.; Choi, J.; Choo, J.; and Schneider, S. 2025. Sparse Autoencoders Reveal Selective Remapping of Visual Concepts during Adaptation. In *Proceedings of the 13th International Conference on Learning Representations (ICLR)*.

Makhzani, A.; and Frey, B. 2013. K-sparse autoencoders. *arXiv preprint arXiv:1312.5663*.

Meng, L.; Li, H.; Chen, B.-C.; Lan, S.; Wu, Z.; Jiang, Y.-G.; and Lim, S.-N. 2022. Adavit: Adaptive vision transformers for efficient image recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 12309–12318.