Enhancing Mixture-of-Experts Specialization via Cluster-Aware Upcycling

Supplementary Material

A. Additional Quantitative Results

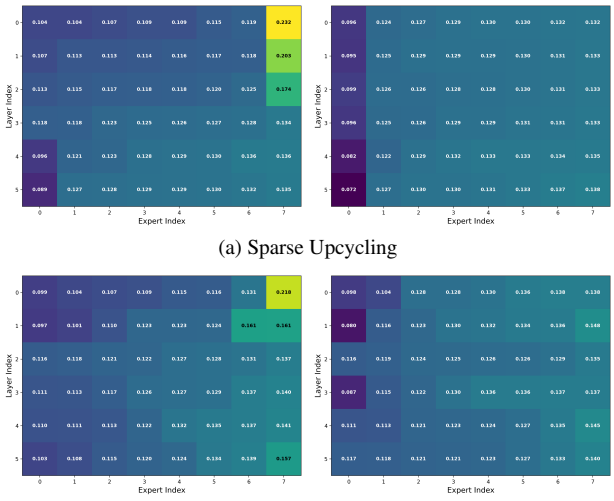
To complement the comparisons presented in the main paper, we provide additional quantitative results on the upcycled ViT-B/32 in Table A.1. We first revisit Drop-Upcycling [4], which reinitializes a randomly selected fraction r of channels using Gaussian noise estimated from the pretrained statistics. While the original work reports results with $r = 0.5$, we observe that smaller perturbation ratios generally lead to better performance across most benchmarks. This suggests that aggressively perturbing pretrained weights can degrade the underlying representation learned during pretraining. However, even under its best-performing configuration ($r = 0.25$), Drop-Upcycling generally underperforms Sparse Upcycling [2].

To explore a more structured alternative to random perturbation, we introduce a simple variant termed Drop-SVD. Instead of randomly perturbing channels, Drop-SVD applies singular value decomposition to each weight matrix and reinitializes only the singular vectors corresponding to the lowest 25% of the spectrum. This preserves the dominant semantic subspace encoded in the pretrained weights while introducing diversity through controlled perturbations in the least informative directions. In contrast to Drop-Upcycling, Drop-SVD achieves performance comparable to Sparse Upcycling and even surpasses it on several benchmarks. These results indicate that respecting the structure of the pretrained representations is important when designing initialization strategies for MoE upcycling.

Nevertheless, perturbation-based approaches do not explicitly leverage the semantic structure of the pretrained representation space. In contrast, Cluster-aware Upcycling directly leverages activation-space clustering to initialize experts according to the structure of the pretrained representation space.

B. Additional Analysis

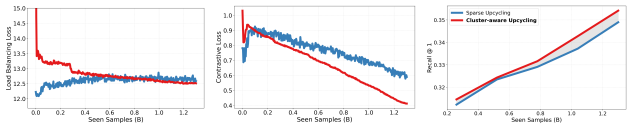
Expert utilization We extend the layer-wise expert utilization analysis of the vision tower shown in Figure 5 to both modalities. As shown in Figure B.1, Cluster-aware Upcycling exhibits slightly more diverse expert utilization patterns than Sparse Upcycling, while maintaining balanced expert loads. Such moderate non-uniformity is expected as specialization emerges, since experts naturally evolve to process different subsets of tokens. Consistent with these patterns, Figure B.2a shows that the load-balancing loss converges to a value comparable to or lower than that of Sparse Upcycling, confirming well-balanced expert utilization during training.



(a) Sparse Upcycling

(b) Cluster-aware Upcycling

Figure B.1. Expert utilization across mixture-of-experts layers for Sparse Upcycling and Cluster-aware Upcycling. Left: Vision towers. Right: Text towers. Best viewed in color.



(a) Load-balancing loss (b) Contrastive loss (c) MSCOCO I→T

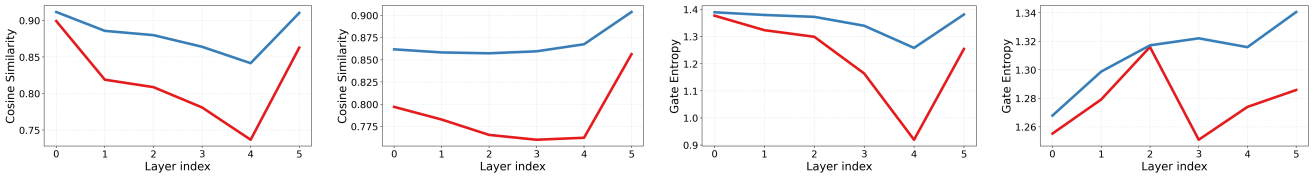
Figure B.2. Training dynamics and epoch-wise performance.

Scaling behavior We further analyze the scaling behavior and training dynamics of Cluster-aware Upcycling. As model capacity increases (e.g., ViT-B/32 → ViT-B/16), the effectiveness of the proposed method becomes more pronounced, resulting in larger performance gaps between our method and the baselines reported in Table 1 of the main paper. Moreover, both the task loss gap and the benchmark performance gap widen as the number of seen samples increases, as illustrated in Figure B.2b and Figure B.2c, respectively. Our experiments use 5.3B seen samples (4B for dense pretraining and 1.3B for MoE upcycling), whereas typical CLIP pre-training scales to around 13B samples. These observations suggest that the proposed method could further benefit from training at larger scales.

Specialization across depth To understand how expert specialization evolves across network depth, we illustrate expert similarity and routing entropy across layers in both vision and text towers, in Figure B.3. Across both modalities, expert similarity exhibits a consistent depth-wise pat-

Table A.1. Additional comparison of upcycling methods for ViT-B/32. Cluster-aware Upcycling achieves the strongest overall performance across most benchmarks.

Model	MSCOCO			ImageNet-1k							VTAB
MoE Init	I→T	T→I	Avg.	Val	V2	A	R	Sketch	ObjNet	Avg.	Natural
Drop-Upcycling (r=0.75) [4]	29.1	45.6	37.4	55.7	47.3	12.4	62.3	40.1	33.0	41.8	58.6
Drop-Upcycling (r=0.50) [4]	29.7	46.5	38.1	56.0	47.7	12.9	63.4	40.8	34.3	42.5	57.8
Drop-Upcycling (r=0.25) [4]	30.6	47.6	39.1	56.7	48.4	13.1	63.5	41.0	35.7	43.1	57.7
Sparse Upcycling [2]	30.8	48.0	39.4	57.1	49.1	13.8	64.3	41.8	36.0	43.7	58.0
CLIP-MoE [5]	29.5	46.8	38.2	56.6	48.1	**14.3**	64.2	41.4	35.7	43.4	58.8
DeRS-LM [1]	**31.0**	47.7	39.4	56.8	48.6	13.9	64.2	41.1	36.4	43.5	58.1
Drop-SVD	30.7	47.9	39.3	**57.4**	48.8	14.1	64.6	42.0	36.1	43.8	58.6
Cluster-aware Upcycling	**31.0**	**48.2**	**39.6**	57.3	**49.2**	14.0	**65.2**	**42.3**	**36.5**	**44.1**	**59.1**

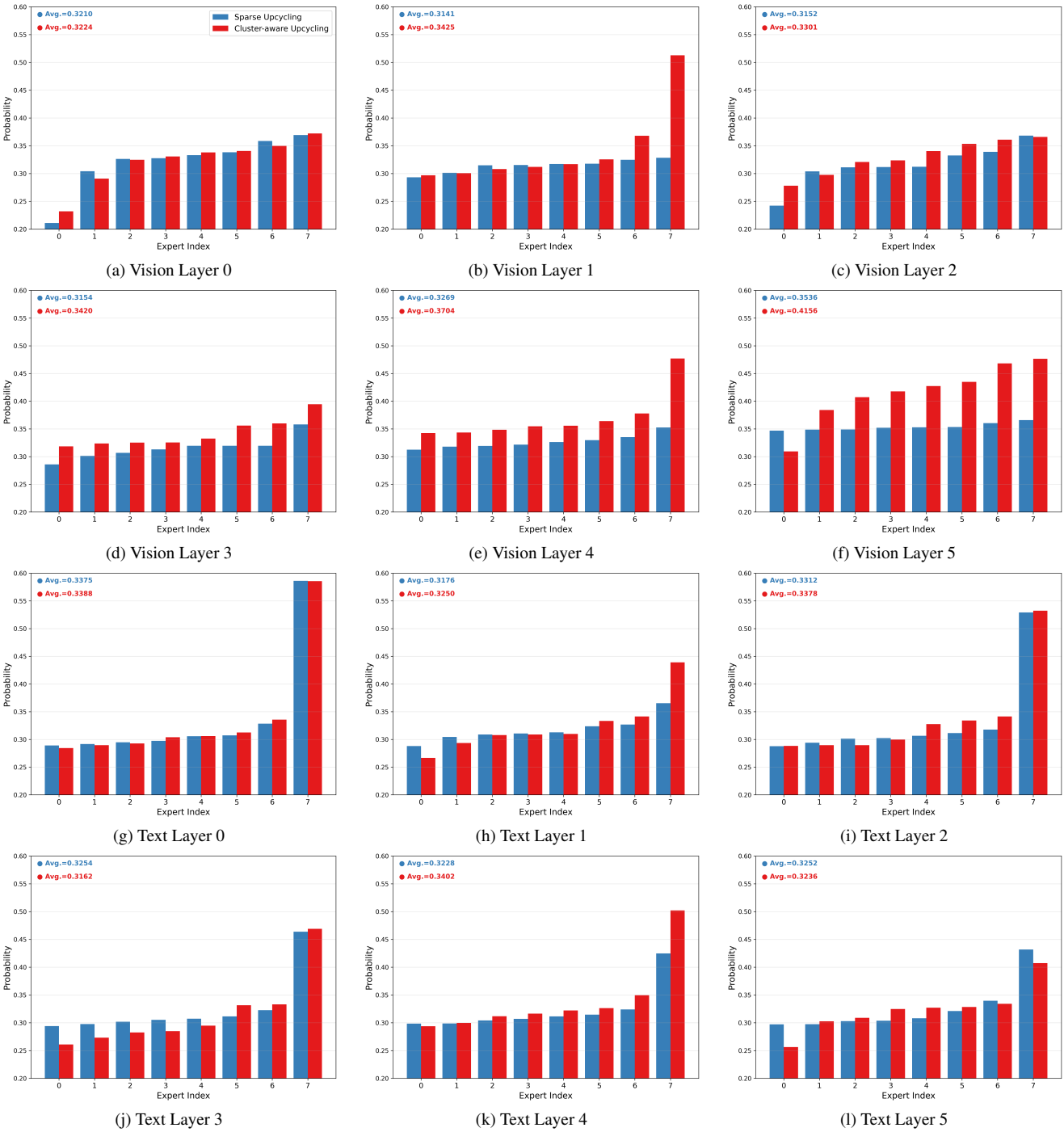


(a) Expert similarity (vision) (b) Expert similarity (text) (c) Routing entropy (vision) (d) Routing entropy (text)

Figure B.3. Layer-wise expert similarity and routing entropy across MoE layers in the vision and text towers of Sparse Upcycling (blue) and Cluster-aware Upcycling (red).

tern: it decreases through most layers, then increases in the final layer, consistent with observations in language MoE models [3]. Routing entropy shows a similar trend, with higher entropy in layers where experts are more similar. However, this pattern is more pronounced in the vision tower, where early layers exhibit higher similarity, possibly due to shared low-level visual features. The text tower shows a milder pattern, with less variation across layers. Despite these modality differences, Cluster-aware Upcycling consistently produces lower expert similarity and routing entropy across all depths, indicating stronger expert differentiation.

Expert-level routing probability To better understand routing behavior, we analyze the average routing probability for each expert, computed over the tokens assigned to that expert. This provides an expert-level perspective that complements the token-level routing entropy analysis presented in Section 4.5 of the main paper. As shown in Figure B.4, Sparse Upcycling exhibits broadly similar assignment probabilities across experts with top-2 routing, whereas Cluster-aware Upcycling shows noticeable variability. Such variability reflects emerging differences in activation strength across experts and corresponds to a more heterogeneous expert distribution.



(a) Vision Layer 0 (b) Vision Layer 1 (c) Vision Layer 2

(d) Vision Layer 3 (e) Vision Layer 4 (f) Vision Layer 5

(g) Text Layer 0 (h) Text Layer 1 (i) Text Layer 2

(j) Text Layer 3 (k) Text Layer 4 (l) Text Layer 5

Figure B.4. Expert routing probabilities for Sparse Upcycling and Cluster-aware Upcycling.

References

[1] Yongqi Huang, Peng Ye, Chenyu Huang, Jianjian Cao, Lin Zhang, Baopu Li, Gang Yu, and Tao Chen. DeRS: Towards extremely efficient upcycled mixture-of-experts models. In *CVPR*, 2025. 2

[2] Aran Komatsuzaki, Joan Puigcerver, James Lee-Thorp, Carlos Riquelme Ruiz, Basil Mustafa, Joshua Ainslie, Yi Tay, Mostafa Dehghani, and Neil Houlsby. Sparse Upcycling: Training mixture-of-experts from dense checkpoints. In *ICLR*, 2023. 1, 2

[3] Ka Man Lo, Zeyu Huang, Zihan Qiu, Zili Wang, and Jie Fu. A closer look into mixture-of-experts in large language models. In *NAACL Findings*, 2025. 2

[4] Taishi Nakamura, Takuya Akiba, Kazuki Fujii, Yusuke Oda, Rio Yokota, and Jun Suzuki. Drop-upcycling: Training sparse mixture of experts with partial re-initialization. In *ICLR*, 2025. 1, 2

[5] Jihai Zhang, Xiaoye Qu, Tong Zhu, and Yu Cheng. CLIP-MoE: Towards building mixture of experts for clip with diversified multiplet upcycling. In *EMNLP*, 2025. 2