SAGE: Shape-Adapting Gated Experts for Adaptive Histopathology Image Segmentation

Supplementary Material

A. Semantic Affinity Routing Analysis

Quantitative analysis of the utilization of Mixture of Experts for heterogeneous inputs reveals a fair utilization of Mixture of Experts. Analysis of the heatmap for activation of the model and attention for the expert map reveals that Semantic Affinity Routing (SAR) achieves a structured allocation of tokens to the experts, ensuring that there is no routing collapse, which is a common failure mode for sparse Mixture of Experts.

All visualizations for the routing were produced using the $GlaS$ test A set [3] with the SAGE-ConvNeXt+ViT-UNet model and $K = 4$.

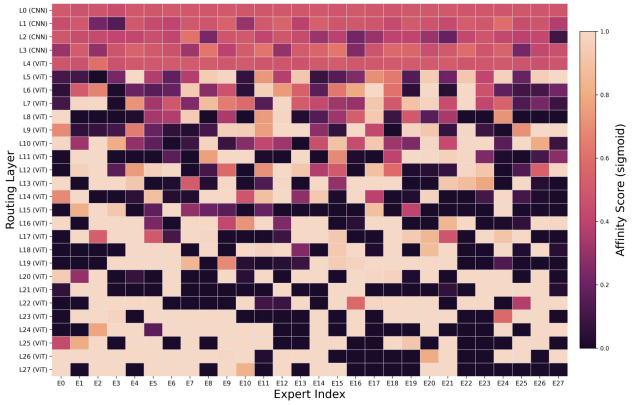


Figure 1. **Normalized affinity score heatmap.** The visualization shows the normalized affinity scores, which are the gating probabilities per expert per layer, with a color map ranging from red (low affinity) to dark green (high affinity). The rows in this heatmap represent the model's layer, where $L0$ to $L3$ represent the CNN layers, followed by $L4$ to $L27$, which represent the Vision Transformer layers. The columns represent the 28 experts in the expert pool, where $E0$ to $E27$ represent each individual expert.

Figure 1 shows a non-uniform affinity landscape, indicating layer-dependent expert specialization rather than routing collapse. No expert remains consistently dominant across all layers, and most experts exhibit alternating high/low affinity bands over depth. Stronger contrast appears in the early CNN-to-shallow ViT transition (approximately $L1$–$L4$), while many middle Transformer layers are closer to moderate values, with localized high-affinity reactivation in later layers. Overall, the pattern supports depth-dependent specialization rather than uniform expert usage.

In addition to the global balance observation, Figure 1 reveals clear layer-wise variation. The early CNN and shallow ViT layers show sharper affinity contrast across experts,

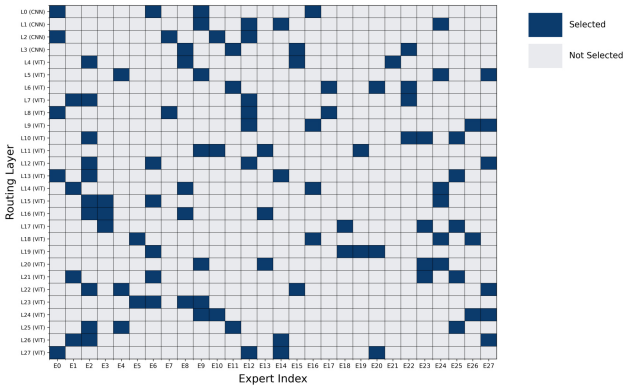


Figure 2. **Top-K activation map.** The binary heatmap illustrates the routing decisions, where K is equal to 4, across the 28 layers in the model (rows: $L0$ to $L3$, which represent the CNN layers; $L4$ to $L27$, which represent the Vision Transformer layers) and the 28 experts in the shared pool (columns: $E0$ to $E27$). The activation of an expert in the top K, at least for one token in the layer after processing the entire batch, is represented by blue, while empty boxes indicate no activation.

whereas many middle Transformer layers appear more centered around moderate affinity values. Selected late layers still present localized high-affinity experts, suggesting that specialization is redistributed across depth rather than monotonically increasing toward deeper layers.

Figure 2 complements this view by showing activation sparsity patterns per layer at $K = 4$. Early CNN layers activate fewer experts per batch, whereas deeper Transformer layers trigger broader expert sets, matching their higher semantic complexity.

Taken together, the two diagnostics indicate that SAR does not behave uniformly across depth: specialization increases in later stages, while early stages remain more shared. This depth-dependent behavior is the key supplementary finding from the routing visualizations. Figure 3 provides a complementary decomposition view, showing how CNN self-affinity, Transformer self-affinity, and cross-architectural affinity jointly contribute to the final Top-K routing decisions.

B. Group-Level Gate Analysis

To supplement the mechanistic evidence presented in the main paper, this section visualizes the evolution of the group-level gate g_s on GlaS [3] during stage-2 training (Figure 4). The hierarchical gate's formulation is identical to that in the main paper; the present analysis is restricted to

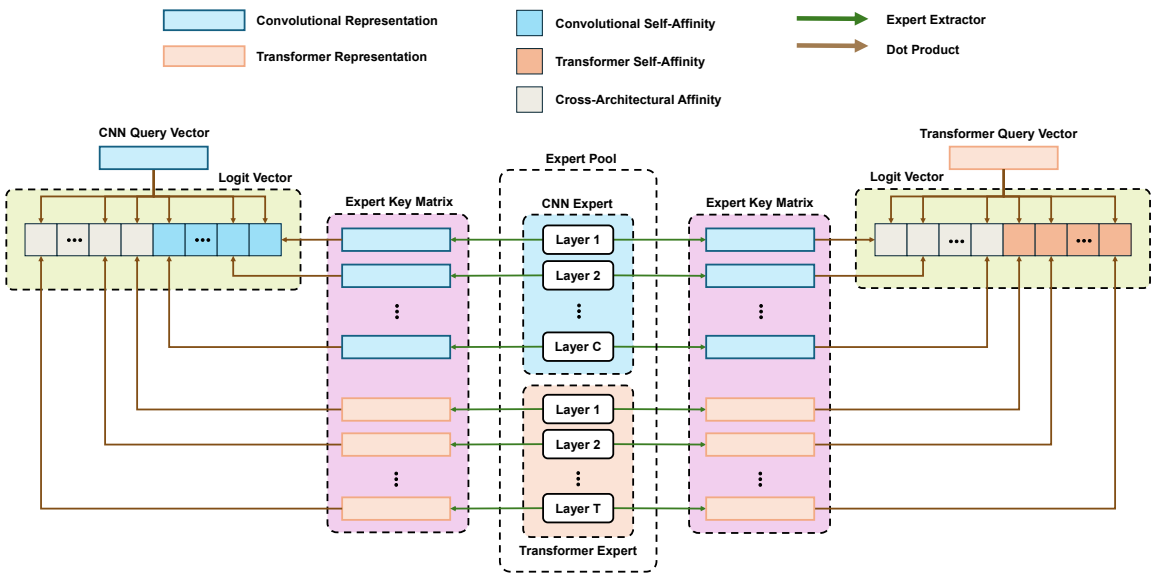


Figure 3. **Diagnostic view of SAR affinity decomposition.** Given CNN and Transformer query vectors, SAR computes per-expert routing logits by dot-product matching against expert keys extracted from both CNN and Transformer experts. The colored components separate CNN self-affinity, Transformer self-affinity, and cross-architectural affinity contributions, which are aggregated before group-level logit modulation and Top-K selection.

its empirical behavior.

Distribution Shift Across Training. Figure 4(a) demonstrates that the final distribution is broader and more polarized than the initial distribution, with substantial mass in both low-g_s and high-g_s regions. This observation indicates that training enhances routing selectivity at the group level, rather than converging to a narrow unimodal regime.

Dynamic Load Balancing. Figure 4(b) shows that the mean g_s oscillates around the neutral value (0.5): it decreases below 0.5 during early and mid epochs, increases above 0.5 in later epochs, and stabilizes near 0.5 at the end of training. The broad standard deviation band reflects considerable sample-wise heterogeneity throughout training, consistent with input-dependent switching between shared and fine-grained experts.

Architectural Role Specialization. Figure 4(c) reveals a consistent architectural gap, with CNN layers exhibiting higher g_s values than Transformer layers.

- **CNN Layers:** Higher g_s indicates stronger reliance on shared experts, which aligns with the extraction of low-level structural features.
- **Transformer Layers:** Lower and more variable g_s values suggest a more frequent preference for fine-grained experts, with periodic shifts back toward neutrality as training advances.

These results support the intended division of labor within the hybrid encoder. CNN layers are more oriented toward shared experts, whereas Transformer layers exhibit a tendency toward mixed routing.

C. Shape-Adapting Hub Analysis

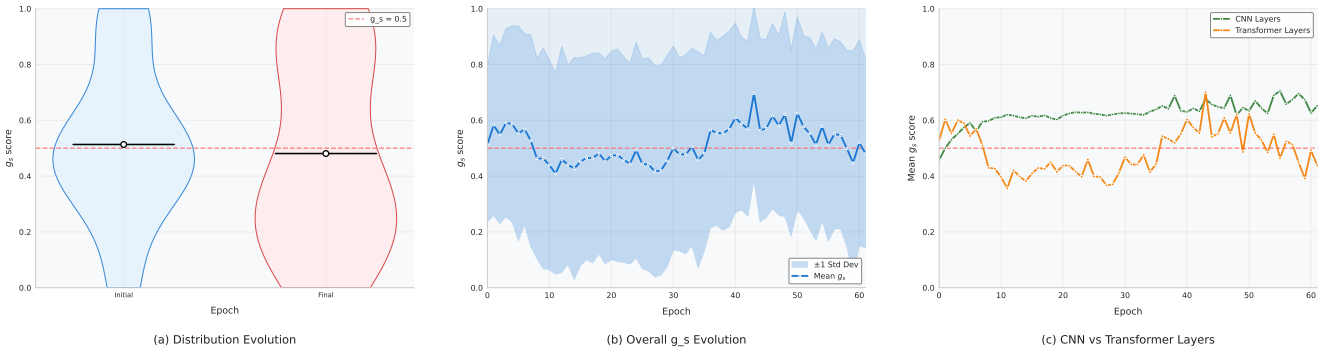
Figure 5 provides an implementation-level perspective of the SA-Hub and highlights three operational details that are essential for interpreting routing behavior.

First, shape adaptation is performed for each activated expert. Each selected expert receives a dedicated input-side reshape and output-side realignment, enabling heterogeneous expert blocks to operate within the same routed layer.

Second, the diagram distinguishes between expert selection and expert contribution. Top-K routing identifies the active experts, while subsequent weighting and fusion stages determine the extent to which each active expert influences the final representation.

Third, the dashed metadata paths indicate that adaptation is conditioned on shape or interface information rather than a single fixed reshape rule. This perspective clarifies why runtime increases with the number of activated experts, even when parameter growth remains modest.

Overall, the SA-Hub perspective illustrated in Figure 5 elucidates the integration of heterogeneous experts into a shape-consistent representation pathway and offers a mechanistic explanation for the observed trade-off between routing flexibility and inference cost.



(a) Distribution Evolution (b) Overall g_s Evolution (c) CNN vs Transformer Layers

Figure 4. **Group-level gate analysis during training.** The figure illustrates the evolution of g_s, where higher values indicate a preference for shared experts and lower values indicate a preference for fine-grained experts. **(a)** Distribution of g_s at the start and end of training. **(b)** Mean g_s with standard deviation across 60 epochs, with a neutral reference at 0.5. **(c)** Mean g_s by architecture type (CNN versus Transformer), showing that CNN layers remain higher while Transformer layers stay closer to the neutral region.

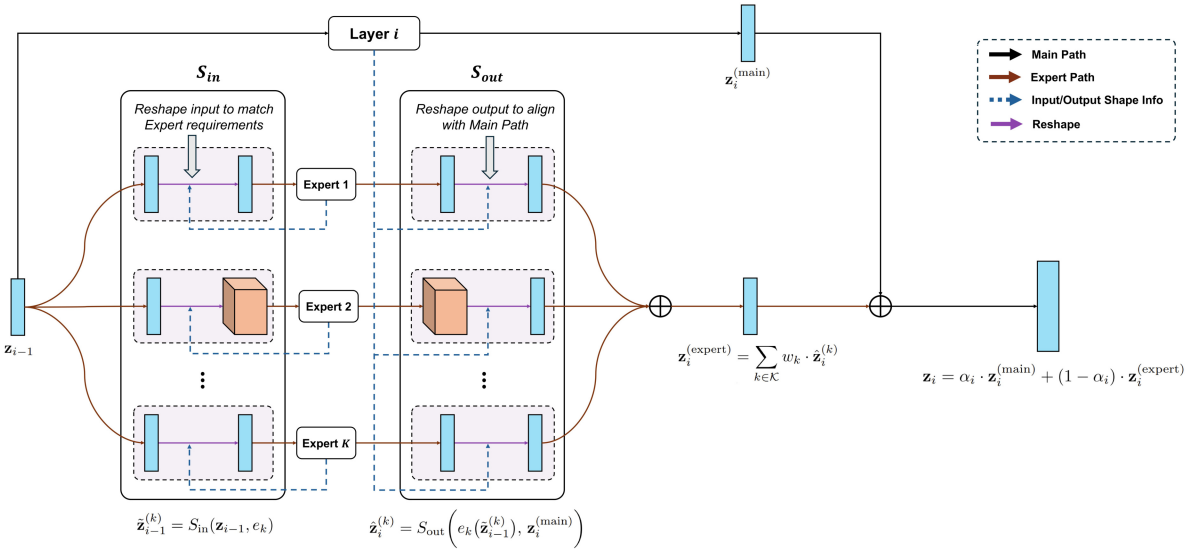


Figure 5. **Detailed execution view of the Shape-Adapting Hub (SA-Hub).** The diagram highlights the adaptation path for each expert, which includes input reshaping, expert execution, and output realignment. This process is followed by weighted aggregation across activated experts and gated fusion with the main branch. Blue dashed arrows indicate the flow of shape metadata, while purple arrows represent reshape operations.

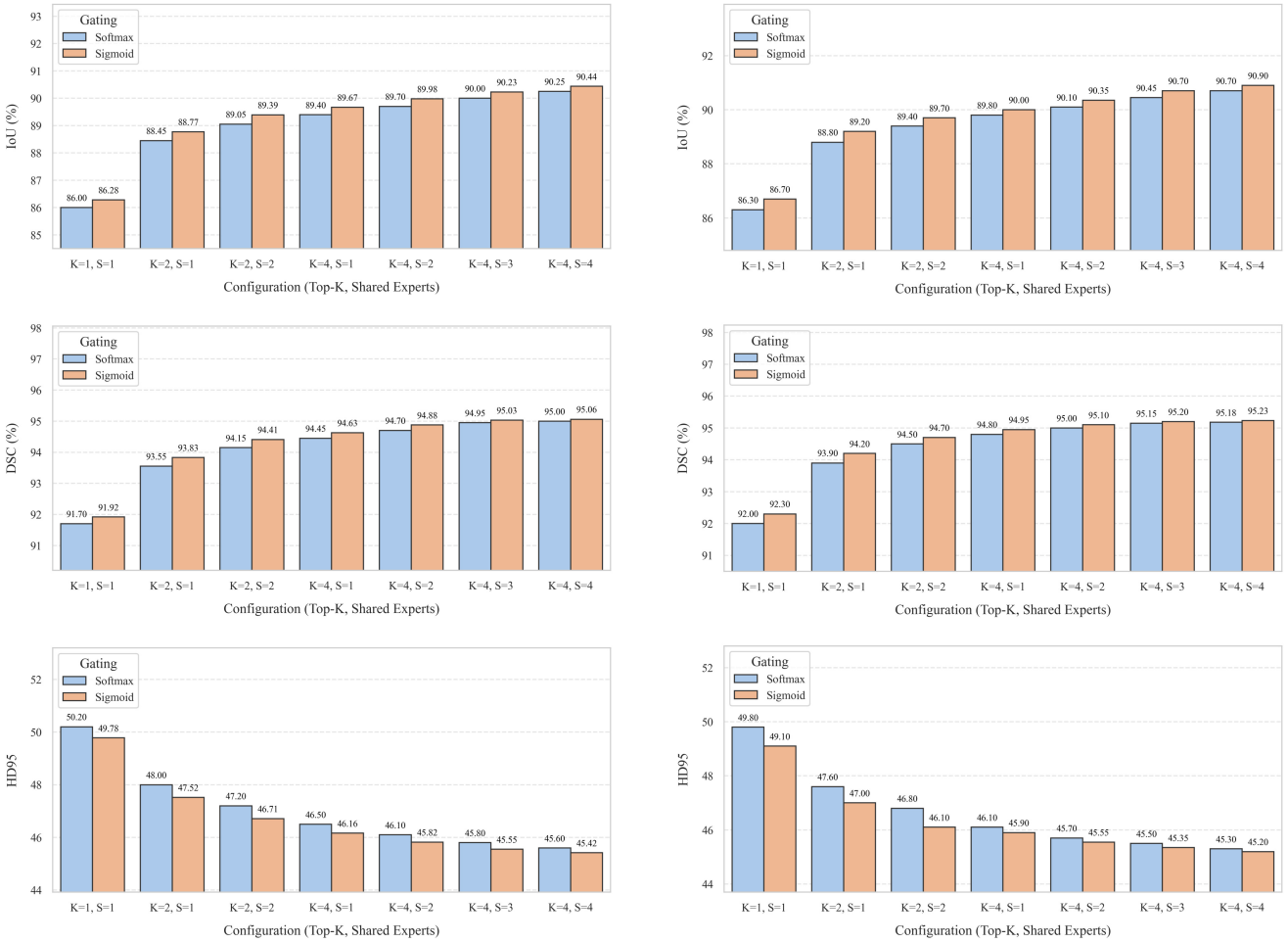
D. Ablation Studies

We investigate the impact of gating strategies and expert capacity on the EBHI dataset [2]. Figure 6 summarizes the trend, and Table 1 reports the exact scores.

Unified Comparison (Without vs. With Logit Modulation). Table 1 is structured to facilitate row-wise comparisons under identical (gating, K, S) settings. In all seven configuration pairs, logit modulation consistently increases Acc, IoU, DSC, and BF1, while reducing HD95 for both sigmoid and softmax gating.

Gating Mechanism. The selection of the gating function is critical for the router's capacity to address complex tissue morphologies. As shown in Table 1, sigmoid gating consistently outperforms softmax gating in both regimes (with and without logit modulation) across all evaluated settings. In the most expansive configuration ($K = 4$, $S = 4$), the sigmoid router achieves a DSC of 95.06% without logit modulation and 95.23% with modulation, while the softmax router attains 95.00% and 95.18%, respectively. This trend aligns with previous findings in sigmoid-based routing and attention mechanisms [1, 4]. The non-competitive nature of sigmoid gating is particularly beneficial for histopathology image segmentation, where the integration of multiple distinct feature extractors is often required due to the presence of diverse and overlapping tissues.

Effect of Logit Modulation. The performance improve-



(a) Without Logit Modification

(b) With Logit Modification

Figure 6. **Ablation of gating strategies and expert capacity.** Comparison of sigmoid vs softmax gating with and without logit modification across varying Top-K (K) and shared experts (S).

ments from logit modulation are systematic and increase additively with greater routing capacity. For example, in the optimal sigmoid configuration ($K = 4$, $S = 4$), enabling logit modulation raises the DSC from 95.06% to 95.23% ($+0.17\%$) and the boundary-aware BF1 score from 57.63 to 58.10 ($+0.47$). It also reduces the HD95 distance from 45.42 to 45.20 (-0.22), indicating enhanced structural boundary precision.

Expert Capacity (K and S). In both modulation regimes, increasing the number of selected experts (K) yields the largest performance improvements. After achieving a high level of specialization ($K = 4$), further increasing the shared expert pool (S) results in smaller but consistent incremental gains. These findings indicate that K primarily enhances the model's ability to process diverse inputs, while S serves as a stable common-knowledge anchor that mitigates feature fragmentation at higher routing capacities.

Based on these comprehensive results, the $K = 4$ and $S = 4$ architecture with sigmoid gating and logit modulation is adopted as the reference configuration, as it offers the optimal balance between specialized routing and stable feature aggregation.

References

[1] Huy Nguyen, Nhat Ho, and Alessandro Rinaldo. Sigmoid gating is more sample efficient than softmax gating in mixture of experts, 2024. 3

[2] Liyu Shi, Xiaoyan Li, Weiming Hu, Haoyuan Chen, Jing Chen, Zizhen Fan, Minghe Gao, Yujie Jing, Guotao Lu, Deguo Ma, Zhiyu Ma, Qingtao Meng, Dechao Tang, Hongzan Sun, Marcin Grzegorzek, Shouliang Qi, Yueyang Teng, and Chen Li. Ebhi-seg: A novel enteroscope biopsy histopathological hematoxylin and eosin image dataset for image segmentation tasks. *Frontiers in Medicine*, Volume 10 - 2023, 2023. 3

Table 1. **Unified ablation on EBHI with and without logit modulation.** This table compares matched routing configurations under both settings. Higher values indicate better performance for Acc, IoU, DSC, and BF1, while lower values are preferable for HD95. **Best** and Second denote the best and second-best results within each modulation regime. The rightmost gain column reports (With − Without) to explicitly quantify the benefit of modulation.

Config			Without Logit Modulation					With Logit Modulation					Gain (With-Without)				
Gating	K	S	**Acc** ↑	**IoU** ↑	**DSC** ↑	**HD95** ↓	**BF1** ↑	**Acc** ↑	**IoU** ↑	**DSC** ↑	**HD95** ↓	**BF1** ↑	**ΔAcc** ↑	**ΔIoU** ↑	**ΔDSC** ↑	**ΔHD95** ↓	**ΔBF1** ↑
Softmax	1	1	88.85	86.00	91.70	50.20	49.70	89.10	86.30	92.00	49.80	50.20	+0.25	+0.30	+0.30	-0.40	+0.50
Sigmoid	1	1	89.07	86.28	91.92	49.78	50.18	89.40	86.70	92.30	49.10	51.00	+0.33	+0.42	+0.38	-0.68	+0.82
Softmax	2	1	91.95	88.45	93.55	48.00	53.60	92.20	88.80	93.90	47.60	54.10	+0.25	+0.35	+0.35	-0.40	+0.50
Sigmoid	2	1	92.21	88.77	93.83	47.52	54.12	92.55	89.20	94.20	47.00	55.10	+0.34	+0.43	+0.37	-0.52	+0.98
Softmax	2	2	92.60	89.05	94.15	47.20	55.40	92.90	89.40	94.50	46.80	56.00	+0.30	+0.35	+0.35	-0.40	+0.60
Sigmoid	2	2	92.88	89.39	94.41	46.71	55.92	93.25	89.70	94.70	46.10	56.90	+0.37	+0.31	+0.29	-0.61	+0.98
Softmax	4	1	93.10	89.40	94.45	46.50	56.60	93.45	89.80	94.80	46.10	57.10	+0.35	+0.40	+0.35	-0.40	+0.50
Sigmoid	4	1	93.34	89.67	94.63	46.16	56.97	93.70	90.00	94.95	45.90	57.60	+0.36	+0.33	+0.32	-0.26	+0.63
Softmax	4	2	93.30	89.70	94.70	46.10	57.20	93.62	90.10	95.00	45.70	57.80	+0.32	+0.40	+0.30	-0.40	+0.60
Sigmoid	4	2	93.58	89.98	94.88	45.82	57.48	93.85	90.35	95.10	45.55	58.00	+0.27	+0.37	+0.22	-0.27	+0.52
Softmax	4	3	93.45	90.00	94.95	45.80	57.40	93.78	90.45	95.15	45.50	58.00	+0.33	+0.45	+0.20	-0.30	+0.60
Sigmoid	4	3	93.66	90.23	95.03	45.55	57.56	93.96	90.70	95.20	45.35	58.05	+0.30	+0.47	+0.17	-0.20	+0.49
Softmax	4	4	93.55	90.25	95.00	45.60	57.50	93.88	90.70	95.18	45.30	58.05	+0.33	+0.45	+0.18	-0.30	+0.55
Sigmoid	4	4	93.74	90.44	95.06	45.42	57.63	94.03	90.90	95.23	45.20	58.10	+0.29	+0.46	+0.17	-0.22	+0.47

[3] Korsuk Sirinukunwattana, Josien P. W. Pluim, Hao Chen, Xiaojuan Qi, Pheng-Ann Heng, Yun Bo Guo, Li Yang Wang, Bogdan J. Matuszewski, Elia Bruni, Urko Sanchez, Anton Böhm, Olaf Ronneberger, Bassem Ben Cheikh, Daniel Racoceanu, Philipp Kainz, Michael Pfeiffer, Martin Urschler, David R. J. Snead, and Nasir M. Rajpoot. Gland segmentation in colon histology images: The glas challenge contest, 2016. 1

[4] Fanqi Yan, Huy Nguyen, Pedram Akbarian, Nhat Ho, and Alessandro Rinaldo. Sigmoid self-attention is better than softmax self-attention: A mixture-of-experts perspective. *arXiv preprintarXiv: 2502.00281*, 2025. 3