Outlier-Suppressed Transformer for Efficient 3D Medical Image Segmentation

Zhihong Liu, Lei Zhang, Zeyu Fu, and Xujiong Ye*

Department of Computer Science, University of Exeter, UK
X.Ye2@exeter.ac.uk

Abstract. Transformers have been widely adopted in medical image segmentation due to their superior capacity for capturing long-term dependencies. However, their high computational cost and memory footprint pose challenges for deployment in resource-constrained clinical scenarios. To bridge this gap, post-training quantization has emerged as a standard paradigm for model compression due to its data-efficient and training-free nature. However, its efficacy is severely hindered by activation outliers—extreme values that severely skew the activation distribution and induce prohibitive quantization errors. Existing approaches primarily rely on post-hoc compensation or complex quantization schemes, ignoring the architectural factors that lead to outlier formation in model representations. In this work, we propose to suppress activation outliers at the architectural level to foster a quantization-friendly model. We introduce a lightweight, plug-and-play attention-gating block that constrains the dynamic range of activations during forward propagation by modulating the attention outputs. This facilitates reliable low-bit deployment under standard post-training quantization protocols, eliminating the need for complex or specialized quantization strategies. Experiments on three 3D abdominal CT benchmarks and two transformer-based medical image segmentation backbones demonstrate that our method effectively suppresses activation outliers, leading to more stable quantization behavior. Our approach provides a viable pathway for the high-precision deployment of medical segmentation models on resource-constrained hardware. The code is available at: Outlier-Suppressed Transformer

Keywords: Medical Image Segmentation · Post-Training Quantization · Outliers in Transformers.

1 Introduction

Image segmentation plays a fundamental role in medical image analysis, as it is typically the first step for identifying and analysing anatomical structures [21]. Inspired by the remarkable success of Transformer models in natural language processing, researchers have introduced Transformers into medical image segmentation [27,4,11,10,30]. Although Transformer-based segmentation models have achieved state-of-the-art performance across a wide range of medical imaging tasks, the rapidly increasing model size, together with the growing

scale of medical imaging datasets [28,23], poses significant computational challenges. In practical clinical scenarios, hardware resources are often limited. The high computation cost of Transformer-based segmentation models leads to inference latency, excessive memory cost, and increased power consumption [8,26]. Therefore, it is required to compress these models to enable their deployment in real-world clinical environments.

Quantization, which converts floating-point numbers into low-bit fixed-point numbers to reduce storage and computational overhead, has emerged as a promising model compression technique [9]. In particular, post-training quantization (PTQ), which calibrates quantization parameters using a small set of unlabeled data without retraining, is considered an efficient and practical solution [1,14]. Although PTQ has demonstrated great success on convolutional neural networks, directly applying PTQ to Transformer-based models often results in performance degradation, especially in aggressive ultra-low-bit settings [19,29]. The primary bottleneck lies in the emergence of outliers, as their extreme magnitudes significantly expand the dynamic range of activations, leading to severe precision loss in low-bit matrix operations [3,6].

To alleviate this issue, recent studies for vision Transformers focus on post-hoc compensation at the quantization stage. They identify components with highly skewed activation distributions and design finer-grained PTQ schemes to improve quantization performance [17,29,16]. These approaches treat outliers as an immutable byproduct of the model. However, recent evidence suggests that these outliers are not inherent necessities, but are closely tied to architectural behaviors. He et al. [12] investigated the impact of architectural choices on outlier emergence, revealing that modifying standard normalization layers can significantly reduce activation magnitudes. Furthermore, several studies have established a direct correlation between the attention mechanism and outlier formation. Darcet et al. [5] observed that attention maps in modern vision Transformers often exhibit visible artifacts, which are closely associated with high-norm outlier tokens. Sun et al. [25] further demonstrated that excessively large activations in large Transformer models tend to cause attention to concentrate on a small subset of tokens.

In this paper, instead of pursuing model efficiency through post-hoc compensation, we adopt a proactive architectural strategy to explicitly mitigate activation outliers at the representation level. We introduce the Outlier-Suppressed Transformer, featuring an Attention-Gating (AG) block designed to recalibrate activation distributions, as shown in Fig. 1. Our approach is inspired by the observation that gating mechanisms can effectively mitigate localized concentration in attention maps [22]. Specifically, our pipeline incorporates the proposed AG block into the segmentation network during full-precision training, while strictly following the standard PTQ protocol during deployment. By modulating attention outputs, the AG block ensures that the learned representations remain compact and statistically stable. This architectural refinement effectively constrains the activation dynamic range, enabling high-performance deployment via standard, hardware-friendly PTQ schemes.

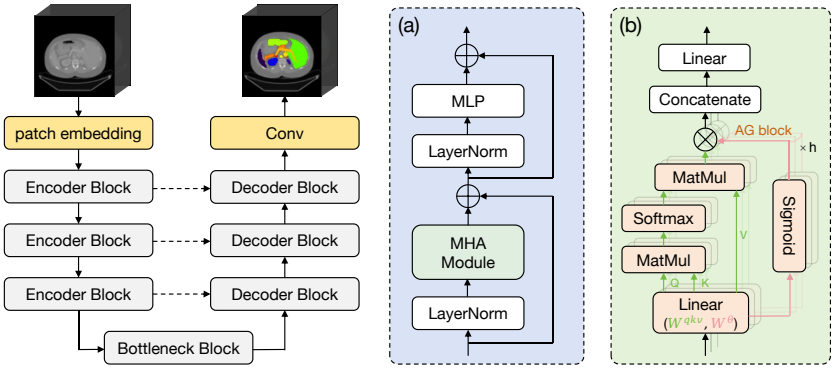

Fig. 1. The overall architecture of the outlier-suppressed transformer for medical image segmentation. (a) The architecture of a single Transformer layer, highlighting the strategic placement of the Attention mechanism. (b) Detailed view of the Multi-head Attention (MHA) module integrated with Attention-Gating (AG) block to adaptively modulate activation distributions.

Our main contributions can be summarized as follows: 1) We systematically explore the characteristics of outliers in Transformer-based segmentation models and quantify their impact on segmentation performance. 2) We introduce a lightweight Attention-Gating (AG) block that suppresses outliers at the representation level by adaptively modulating attention outputs. 3) Extensive evaluations across the BTCV, AMOS, and AbdomenCT-1K benchmarks demonstrate that our method improves segmentation accuracy under aggressive low-bit settings while preserving competitive full-precision performance.

2 Method

2.1 Preliminary

Quantization maps floating-point values to low-bit fixed-point numbers, reducing model size and improving computational efficiency. The uniform quantizer is widely adopted due to its simplicity and strong hardware support, and can be formulated as follows:

$$Q(x) = \left(\text{clip}\left(\left\lfloor \frac{x}{s} \right\rceil + zp; 0,\ 2^b - 1\right) - zp\right) \cdot s, \tag{1}$$

where s and zp denote the quantization scale and zero point, respectively, and b represents the bit-width. The scale factor s is typically determined by the dynamic range of the input x as $s = \frac{\max(x) - \min(x)}{2^b - 1}$. The quantization performance is highly sensitive to the distribution of x and the quantization error is theoretically minimized when the input x follows a uniform distribution.

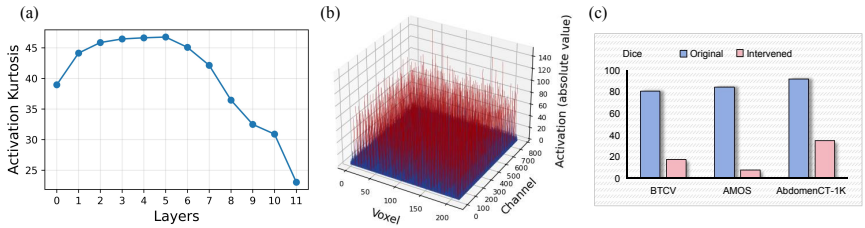

Fig. 2. Analysis of outliers in the UNETR model. (a) Activation kurtosis of the residual stream across layers in a UNETR model trained on the BTCV dataset. (b) Activation behavior of the UNETR model, with outliers highlighted in red. (c) Outlier intervention in UNETR. The identified outliers are replaced with the median of the corresponding activation tensor, which leads to a substantial decline in Dice scores.

2.2 Outliers in Transformer-based medical segmentation models

We first investigate the activation characteristics of well-established Transformer-based 3D medical image segmentation models, such as UNETR [11] and Swin-UNETR [10]. While these models excel at capturing long-range dependencies, our analysis reveals the persistent emergence of activation outliers with extreme values. To characterize the distribution of activations, we employ Kurtosis as the metric [12,7]. For an activation matrix $\mathbf{X} \in \mathbb{R}^{n \times d}$, the Kurtosis is defined as the ratio of the fourth moment m_4 to the squared second moment m_2:

$$\mathrm{Kurt}(\mathbf{X}) = \frac{m_4(\mathbf{X})}{m_2(\mathbf{X})^2} \overset{\text{def}}{=} \frac{\frac{1}{d}\sum_{j=1}^{d} s_j^4}{\left(\frac{1}{d}\sum_{j=1}^{d} s_j^2\right)^2}, \tag{2}$$

where s_j represents the root-mean-square (RMS) activation for the j-th channel across n inputs, and d represents the feature dimension. This metric effectively measures how much the distribution is dominated by extreme values. A Kurtosis value of 1 signifies an even distribution, while a value close to d indicates that the entire range is overwhelmed by a few extreme spikes. As illustrated in Fig. 2(a), the observed Kurtosis values consistently exceed 1 across layers, indicating the presence of extreme outliers throughout the network.

To further evaluate the impact of outliers on model performance, we conduct sensitivity tests using a standard intervention strategy. Following previous studies [3], outliers are identified as values exceeding six standard deviations from the mean of the corresponding activation tensor, as shown in Fig. 2(b). Across the BTCV dataset [15], such outliers account for approximately 0.4% of all activation values in the UNETR model. Substituting these values with the corresponding median activation results in a severe performance degradation, as shown in Fig. 2(c). This observation further indicates that a narrow quantization scale fails to capture extreme outliers, leading to irreversible clipping errors that compromise segmentation accuracy. Conversely, increasing the quantization scale to accommodate these outliers results in a coarse representation for the majority of activations, thereby sacrificing the numerical precision of most features.

This reflects a fundamental trade-off between dynamic range coverage and quantization resolution, which poses a central challenge for low-bit quantization of Transformer-based medical segmentation models.

2.3 Outlier-suppressed transformer

To resolve the aforementioned conflict between clipping errors and quantization resolution, we propose the Outlier-Suppressed Transformer. Our framework prioritizes the cultivation of quantization-friendly representations through architectural refinement.

The architecture of the Outlier-Suppressed Transformer is illustrated in Fig. 1, featuring a lightweight Attention-Gating (AG) block to modulate activation distributions and constrain the emergence of extreme outliers. Within each Transformer layer, the AG block is integrated into the Multi-Head Attention (MHA) module. Specifically, for each attention head, the gating mechanism generates dynamic, input-dependent gating coefficients to modulate the output as follows:

$$Y_i' = g(Y_i, X, W_i^\theta, \sigma) = Y_i \odot \sigma(XW_i^\theta), \tag{3}$$

$$Y_i = \text{softmax}\left(\frac{Q_i K_i^T}{\sqrt{d_k}}\right) V_i. \tag{4}$$

Here, Y_i represents the standard output of the i-th attention head, σ denotes the Sigmoid activation function, and W_i^θ refers to the learnable weight parameters of the gate. The gating scores $\sigma(XW_i^\theta) \in (0, 1)$ are applied element-wise as a scaling factor, which enables the network to modulate attention responses adaptively.

Following the standard MHA formulation, the modulated outputs of all attention heads are concatenated and projected to get the output as follows:

$$\text{MultiHead}(Q, K, V) = \text{Concat}(Y_1', ..., Y_h')W^O, \tag{5}$$

where W^O is the projection matrix. Crucially, the AG block introduces non-linearity between the attention outputs and the projection matrix of the MHA. This non-linearity prevents large attention responses from being directly propagated and further amplified by the subsequent linear mixing. As a result, extreme activation values are effectively suppressed, leading to a reduced effective dynamic range and improved quantization accuracy under low-bit quantization.

3 Experiments

3.1 Experimental Setup

Evaluation settings We use three publicly available datasets, BTCV [15], AMOS [13], and AbdomenCT-1K [20], for evaluation. These datasets are strictly divided into training and validation sets. We evaluate our method on two popular Transformer-based medical image segmentation models: UNETR [11] and Swin UNETR [10]. These models both follow the successful "U-shaped" network [24]

Table 1. Quantization results across baselines and datasets. "8/8" means weights are quantized to 8-bit and activations are quantized to 8-bit.

Model	BTCV				AMOS				AbdomenCT-1K			
	32/32	8/8	6/6	4/4	32/32	8/8	6/6	4/4	32/32	8/8	6/6	4/4
UNETR	80.1	80.0	78.3	7.8	84.0	84.0	82.0	6.3	91.4	91.3	90.6	25.2
UNETR-AG	81.6	**81.4**	**80.0**	**12.8**	84.2	**84.1**	**83.1**	**6.7**	91.6	**91.6**	**91.2**	**49.2**
SwinUNETR	83.7	82.7	31.3	13.7	86.0	84.8	25.1	14.3	93.7	**93.6**	50.6	40.9
SwinUNETR-AG	83.6	**83.5**	**55.8**	**22.4**	86.2	**86.1**	**73.2**	**51.9**	93.7	**93.6**	**66.0**	**45.0**

design for the encoder and decoder. The UNETR utilizes a transformer as the encoder to learn sequence representations of the input volume [27]. The Swin UNETR introduces a swin transformer encoder to extract features at different resolutions by utilizing shifted windows for computing self-attention [18].

Implementation Details We run all experiments based on PyTorch and MONAI framework. The AG block is trained jointly with the segmentation model from scratch using one RTX 3090 GPU. The default optimizer is AdamW with a learning rate 1e-4 and a weight decay 1e-5. The input images are cropped to a size of $96 \times 96 \times 96$. After training, standard post-training quantization is applied. We randomly select 4 sample from the training set of a dataset to calibrate the quantization parameters. For the calibration strategy, we apply the standard MinMax method, with layer-wise quantization for weights and activations in the inference process. For evaluation, the widely used medical image segmentation metric, DICE [2], is employed.

3.2 Results and Discussion

Main Results Table 1 reports the quantization results across different backbones on the BTCV, AMOS and AbdomenCT-1K datasets under various bit-width settings. We compare FP32, W8/A8, W6/A6 and W4/A4. We observe that most models maintain the comparable performance to their full-precision baselines under 8-bit quantization. However, they experience varying degrees of performance drop at lower bit-widths. Under the 4-bit setting, all models suffer from a performance collapse. In the 6-bit configuration, UNETR experiences a Dice drop of over 1.5% on both BTCV and AMOS datasets. The integration of the AG block effectively mitigates the loss, yielding a substantial boost in quantization robustness. This improvement is even more pronounced for SwinUNETR, which exhibits extreme sensitivity to low-bit quantization. Unlike UNETR, the vanilla SwinUNETR shows a noticeable performance decline (over 1%) even at 8-bit on BTCV and AMOS. This vulnerability escalates dramatically under the 6-bit setting, where the model fails significantly. For instance, its Dice score on the AMOS dataset drops from 86.0% (32/32) to only 25.1% (6/6). By contrast, adding the AG block successfully restores its performance to 73.2%, effectively bridging the gap to the high-precision baseline. These results demonstrate that while Transformer-based segmentation models exhibit varying degrees of quantization resilience, the proposed AG block effectively alleviates this vulnerability.

Table 2. Quantitative analysis of activation statistics across different backbones and datasets. "s-Act" denotes the average standard deviation of activations across layers, and "M-Act" represents the Maximum activation magnitude observed within the hidden states. Dice scores are reported under 6-bit quantization.

Model	BTCV			AMOS			AbdomenCT-1K		
	s-Act	M-Act	6/6	s-Act	M-Act	6/6	s-Act	M-Act	6/6
UNETR	12	309	78.3	13	369	82.0	19	464	90.6
UNETR-AG	8	227	80.0	9	288	83.1	12	244	91.2
SwinUNETR	4	68	31.3	4	68	25.1	5	83	50.6
SwinUNETR-AG	3	58	55.8	3	46	73.2	4	83	66.0

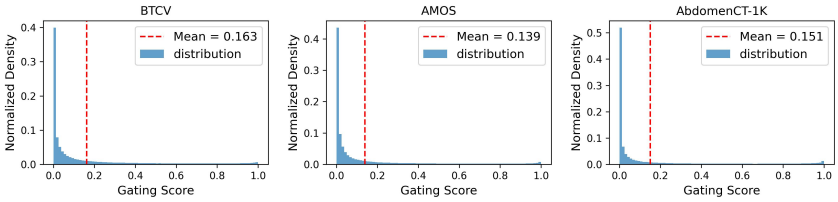

Fig. 3. Gating score means and distributions for UNETR-AG across three benchmarks: BTCV (left), AMOS (middle), and AbdomenCT-1K (right). Most gating scores are less than 0.5, indicating that the gating scores are sparse.

Analysis of hidden states To investigate the underlying causes of the disparate quantization resilience, we conduct an analysis of the internal hidden states, with the results reported in Table 2. For UNETR, we observe that the standard deviation and the maximum activation are reduced, demonstrating that the AG block acts as a statistical regularizer to foster a compact, quantization-resilient feature space. As evidenced by Fig. 3, the gating scores for UNETR concentrate near zero, indicating that the module adaptively attenuates responses to optimize the hidden state distribution. Intriguingly, while SwinUNETR exhibits lower maximum activations, it suffers a severe performance drop. This suggests that its quantization bottleneck arises not only from large activation magnitudes but also from the sensitivity of its hierarchical attention structure to numerical perturbations. Nevertheless, integrating the AG block consistently stabilizes activations and restores performance, highlighting its effectiveness across different architectures.

Qualitative Analysis The qualitative segmentation results across varying precision levels are illustrated in Fig. 4. It is observed that 8-bit quantization serves as a robust baseline, maintaining structural fidelity with minimal visible degradation. However, as the bit-width decreases to 6-bit and 4-bit, the baseline model exhibits notable performance degradation and structural artifacts. In contrast, models integrated with the AG block consistently produce superior segmentation quality, effectively preserving structural integrity of organs. These results demonstrate the AG block's capability to enhance the quantization robustness of Transformer-based segmentation models in resource-constrained settings.

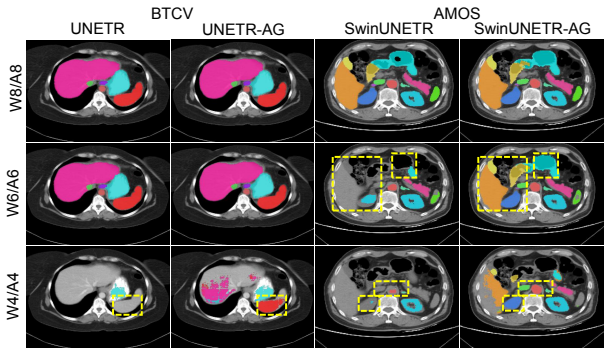

Fig. 4. Qualitative comparison of segmentation results on BTCV and AMOS. While 8-bit results maintain high fidelity, baselines degrade significantly at lower bit-widths. In contrast, our proposed method preserves the structural integrity of organs (yellow boxes) under low-bit settings.

Table 3. Comparison of model complexity and computational efficiency.

Model	Params (M)	FLOPs (G)	Latency (s)	Dice
UNETR	92.62	82.71	0.022	80.1
UNETR-AG	99.70	84.23	0.022	81.6
SwinUNETR	61.99	329.84	0.076	83.7
SwinUNETR-AG	62.38	331.05	0.077	83.6

Complexity and efficiency Analysis We evaluate the architectural complexity and computational efficiency of our proposed method on the BTCV dataset, with FLOPs and latency measured using a standard input volume of $96 \times 96 \times 96$. As summarized in Table 3, the integration of the AG block incurs minimal computational overhead. For the UNETR backbone, the AG block introduces a modest increase of 7.6% in parameters and only 1.8% in FLOPs. Notably, the real-world inference latency remains nearly identical to the baseline, indicating that the proposed block is lightweight and can be seamlessly integrated into existing Transformer-based segmentation architectures.

4 Conclusion

In this paper, we have conducted a systematic analysis of outliers in Transformer-based medical image segmentation models. Our findings demonstrate that while these outliers are essential for maintaining the segmentation performance, their vast dynamic range inherently presents a significant barrier to post-training quantization. To address this, we introduce the Outlier-Suppressed Transformer, featuring a simple yet effective Attention-Gating (AG) block designed to proactively modulate activation distributions. Extensive experiments demonstrate that our method preserves competitive full-precision performance while significantly enhancing model robustness in low-bit settings.

5 Limitations

Our work has several limitations. First, the proposed method requires jointly training the AG block with the segmentation network before quantization. Second, the generalization of the proposed method to other architectures and imaging modalities remains to be explored. Third, while our method significantly improves robustness under 6-bit quantization, extremely low-bit settings (e.g., W4/A4) remain challenging and require further investigation.

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Banner, R., Nahshan, Y., Soudry, D.: Post training 4-bit quantization of convolutional networks for rapid-deployment. Advances in neural information processing systems **32** (2019)
2. Bertels, J., Eelbode, T., Berman, M., Vandermeulen, D., Maes, F., Bisschops, R., Blaschko, M.B.: Optimizing the dice score and jaccard index for medical image segmentation: Theory and practice. In: International conference on medical image computing and computer-assisted intervention. pp. 92–100. Springer (2019)
3. Bondarenko, Y., Nagel, M., Blankevoort, T.: Understanding and overcoming the challenges of efficient transformer quantization. In: Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing. pp. 7947–7969 (2021)
4. Chen, J., Lu, Y., Yu, Q., Luo, X., Adeli, E., Wang, Y., Lu, L., Yuille, A.L., Zhou, Y.: Transunet: Transformers make strong encoders for medical image segmentation (2021), `https://arxiv.org/abs/2102.04306`
5. Darcet, T., Oquab, M., Mairal, J., Bojanowski, P.: Vision transformers need registers. arXiv preprint arXiv:2309.16588 (2023)
6. Dettmers, T., Lewis, M., Belkada, Y., Zettlemoyer, L.: Gpt3. int8 (): 8-bit matrix multiplication for transformers at scale. Advances in neural information processing systems **35**, 30318–30332 (2022)
7. Elhage, N., Lasenby, R., Olah, C.: Privileged bases in the transformer residual stream. Transformer Circuits Thread **24** (2023)
8. Gao, Y., Zhou, M., Liu, D., Yan, Z., Zhang, S., Metaxas, D.N.: A data-scalable transformer for medical image segmentation: architecture, model efficiency, and benchmark. arXiv preprint arXiv:2203.00131 (2022)
9. Gholami, A., Kim, S., Dong, Z., Yao, Z., Mahoney, M.W., Keutzer, K.: A survey of quantization methods for efficient neural network inference. In: Low-power computer vision, pp. 291–326. Chapman and Hall/CRC (2022)
10. Hatamizadeh, A., Nath, V., Tang, Y., Yang, D., Roth, H.R., Xu, D.: Swin unetr: Swin transformers for semantic segmentation of brain tumors in mri images. In: International MICCAI brainlesion workshop. pp. 272–284. Springer (2021)
11. Hatamizadeh, A., Tang, Y., Nath, V., Yang, D., Myronenko, A., Landman, B., Roth, H.R., Xu, D.: Unetr: Transformers for 3d medical image segmentation. In: Proceedings of the IEEE/CVF winter conference on applications of computer vision. pp. 574–584 (2022)

12. He, B., Noci, L., Paliotta, D., Schlag, I., Hofmann, T.: Understanding and minimising outlier features in transformer training. Advances in Neural Information Processing Systems **37**, 83786–83846 (2024)
13. Ji, Y., Bai, H., Ge, C., Yang, J., Zhu, Y., Zhang, R., Li, Z., Zhanng, L., Ma, W., Wan, X., et al.: Amos: A large-scale abdominal multi-organ benchmark for versatile medical image segmentation. Advances in neural information processing systems **35**, 36722–36732 (2022)
14. Kravchik, E., Yang, F., Kisilev, P., Choukroun, Y.: Low-bit quantization of neural networks for efficient inference. In: Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops. pp. 0–0 (2019)
15. Landman, B., Xu, Z., Igelsias, J., Styner, M., Langerak, T., Klein, A.: Miccai multi-atlas labeling beyond the cranial vault–workshop and challenge. In: Proc. MICCAI multi-atlas labeling beyond cranial vault—workshop challenge. vol. 5, p. 12. Munich, Germany (2015)
16. Li, Z., Xiao, J., Yang, L., Gu, Q.: Repq-vit: Scale reparameterization for post-training quantization of vision transformers. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 17227–17236 (2023)
17. Lin, Y., Zhang, T., Sun, P., Li, Z., Zhou, S.: Fq-vit: Post-training quantization for fully quantized vision transformer. arXiv preprint arXiv:2111.13824 (2021)
18. 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: Proceedings of the IEEE/CVF international conference on computer vision. pp. 10012–10022 (2021)
19. Liu, Z., Wang, Y., Han, K., Zhang, W., Ma, S., Gao, W.: Post-training quantization for vision transformer. Advances in Neural Information Processing Systems **34**, 28092–28103 (2021)
20. Ma, J., Zhang, Y., Gu, S., Zhu, C., Ge, C., Zhang, Y., An, X., Wang, C., Wang, Q., Liu, X., et al.: Abdomenct-1k: Is abdominal organ segmentation a solved problem? IEEE Transactions on Pattern Analysis and Machine Intelligence **44**(10), 6695–6714 (2021)
21. Monteiro, M., Newcombe, V.F.J., Mathieu, F., Adatia, K., Kamnitsas, K., Ferrante, E., et al.: Multiclass semantic segmentation and quantification of traumatic brain injury lesions on head ct using deep learning: an algorithm development and multicentre validation study. The Lancet Digital Health **2**(6), e314–e322 (2020)
22. Qiu, Z., Wang, Z., Zheng, B., Huang, Z., Wen, K., Yang, S., Men, R., Yu, L., Huang, F., Huang, S., et al.: Gated attention for large language models: Non-linearity, sparsity, and attention-sink-free. arXiv preprint arXiv:2505.06708 (2025)
23. Qu, C., Zhang, T., Qiao, H., Tang, Y., Yuille, A.L., Zhou, Z., et al.: Abdomenatlas-8k: Annotating 8,000 ct volumes for multi-organ segmentation in three weeks. Advances in Neural Information Processing Systems **36**, 36620–36636 (2023)
24. Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. In: International Conference on Medical image computing and computer-assisted intervention. pp. 234–241. Springer (2015)
25. Sun, M., Chen, X., Kolter, J.Z., Liu, Z.: Massive activations in large language models. arXiv preprint arXiv:2402.17762 (2024)
26. Tang, Y., Yang, D., Li, W., Roth, H.R., Landman, B., Xu, D., Nath, V., Hatamizadeh, A.: Self-supervised pre-training of swin transformers for 3d medical image analysis. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 20730–20740 (2022)

27. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. Advances in neural information processing systems **30** (2017)
28. Wasserthal, J., Breit, H.C., Meyer, M.T., Pradella, M., Hinck, D., Sauter, A.W., Heye, T., Boll, D.T., Cyriac, J., Yang, S., et al.: Totalsegmentator: robust segmentation of 104 anatomic structures in ct images. Radiology: Artificial Intelligence **5**(5), e230024 (2023)
29. Yuan, Z., Xue, C., Chen, Y., Wu, Q., Sun, G.: Ptq4vit: Post-training quantization for vision transformers with twin uniform quantization. In: European conference on computer vision. pp. 191–207. Springer (2022)
30. Zhou, H.Y., Guo, J., Zhang, Y., Han, X., Yu, L., Wang, L., Yu, Y.: nnformer: Volumetric medical image segmentation via a 3d transformer. IEEE transactions on image processing **32**, 4036–4045 (2023)