npj | digital medicine

Article

Published in partnership with Seoul National University Bundang Hospital

https://doi.org/10.1038/s41746-025-02288-5

Structure-aware multi-task learning with domain generalization for robust vertebrae analysis in spinal CT

Jianyang Du[1,2,3,12], Heng'an Ge[4,12], Rui Zhang[5,12], Zhenghan Chen[6], Yuxin Zhang[7,8], Yuqi Bai[9], Honghao Xu[10], Feng Ding[2], Yongchao Zhang[2], Juan Ye[11], Yihang Yang[2] ✉, Shaoshan Hu[1] ✉ & Jingbiao Huang[4] ✉

Spinal image analysis plays a critical role in the diagnosis and treatment of musculoskeletal and neurological disorders. However, existing vertebrae segmentation methods suffer from limited generalizability across clinical domains and rarely address downstream tasks such as vertebrae identification and lesion localization. In this work, we introduce VertebraFormer, a unified multi-task framework designed for robust and generalizable spinal CT analysis. To support this framework, we curate MultiSpine, a heterogeneous benchmark comprising CT volumes from four public and private datasets, annotated with vertebra segmentation masks, anatomical labels, and pathology regions. Our method integrates a Transformer encoder with task-specific decoders and a dynamic modulation unit that adapts feature representations to different imaging domains. We evaluate VertebraFormer across three key tasks-vertebra segmentation, vertebra numbering, and lesion localization, under both in-domain and cross-domain settings. Extensive experiments demonstrate that VertebraFormer outperforms competitive baselines in both accuracy and robustness. We further conduct ablation, perturbation, and efficiency analyses to validate the framework.

The spine is a complex anatomical structure essential for load-bearing, mobility, and neural protection. Spinal disorders, including fractures, degenerative diseases, and neoplastic lesions, are highly prevalent and can lead to severe pain, disability, or even paralysis[1,2]. Computed tomography (CT) plays a pivotal role in spine assessment due to its high spatial resolution and ability to capture fine bony details. However, manual analysis of spinal CT scans is labor-intensive, time-consuming, and subject to inter-observer variability, motivating the development of automated analysis systems[3,4].

In recent years, deep learning has emerged as the dominant approach for vertebrae segmentation, localization, and identification. Fully convolutional networks (FCNs)[3], cascaded architectures[5], and encoder-decoder designs such as Dense-U-Net[6] and nnU-Net[7] have achieved strong performance in delineating vertebrae. Large annotated datasets, such as VerSe[4] and CTSpine1K[8], have further driven research by enabling robust supervised training. Beyond structural analysis, recent studies have incorporated lesion detection modules for pathologies such as fractures, metastases, and degenerative changes[9,10].

Despite these advances, three major challenges remain. First, the structural heterogeneity of the spine across anatomical regions (cervical, thoracic, lumbar) leads to substantial variability in vertebral shape and appearance[11]. Second, the lesion diversity–ranging from subtle cortical thinning to large tumor infiltration–complicates joint modeling of normal anatomy and pathology. Third, cross-domain generalization remains difficult: performance often degrades when models are applied to CT scans acquired from unseen institutions, scanners, or protocols[12–14]. Addressing these issues requires models that can capture both anatomical regularities and pathological variability while maintaining robustness to domain shifts.

[1]Cancer Center, Department of Neurosurgery, Zhejiang Provincial People's Hospital, Affiliated People's Hospital, Hangzhou Medical College, Hangzhou, Zhejiang, China. [2]Department of Neurosurgery, Shandong Provincial Hospital Affiliated to Shandong First Medical University, Jinan, Shandong, China. [3]College of Optical and Engineering, Zhejiang University, Hangzhou, Zhejiang, China. [4]Department of Sports Medicine, Tongji Hospital, School of Medicine, Tongji University, Shanghai, China. [5]Department of Orthopedic, Fuzhou University Affiliated Provincial Hospital, Fuzhou, Fujian, China. [6]School of Software & Microelectronics, Peking University, Beijing, China. [7]Department of Oral Surgery, Shanghai Ninth People's Hospital, Shanghai Jiao Tong University School of Medicine, Shanghai, China. [8]Shanghai Key Laboratory of Orthopedic Implants, Shanghai Ninth People's Hospital, Shanghai Jiao Tong University School of Medicine, Shanghai, China. [9]School of Environment, Education and Development, The University of Manchester, Manchester, UK. [10]School of Medicine, Tongji University, Shanghai, China. [11]Department of Radiology, Suzhou Kowloon Hospital, Shanghai Jiaotong University School of Medicine, Suzhou, Jiangsu, China. [12]These authors contributed equally: Jianyang Du, Heng'an Ge, Rui Zhang. ✉e-mail: yangyihangyy@163.com; shaoshanhu421@163.com; jingbiaohuang1@163.com

To tackle these challenges, we propose VertebraFormer, a unified multi-task framework for generalizable vertebrae segmentation, identification, and lesion detection in heterogeneous CT domains. Our approach integrates (1) a structure-aware encoder that models anatomical context across spinal regions, (2) a multi-task learning paradigm that jointly optimizes structural and pathological targets[15,16], and (3) a domain generalization strategy that combines domain-conditioned dynamic modulation with feature alignment[13,17]. Extensive experiments on multiple public datasets and cross-domain evaluation settings demonstrate that VertebraFormer achieves superior performance in both in-domain accuracy and out-of-domain robustness compared with strong baselines.

In summary, our main contributions are: Structure-aware unified Transformer framework. A single Transformer backbone jointly performs vertebra segmentation, anatomical identification, and lesion detection in a shared feature space, leveraging cross-task consistency and global anatomical modeling to improve accuracy while avoiding the need for separate task-specific networks. MultiSpine benchmark with zero-shot protocol. We introduce MultiSpine, a multi-domain spinal CT dataset and a leave-one-domain-out evaluation protocol, enabling systematic assessment of cross-domain generalization and transferability. Dynamic modulation for domain generalization. A contrastive domain-embedding-driven dynamic modulation mechanism performs feed-forward domain conditioning at inference, improving robustness to unseen modalities and institutions without requiring access to target-domain labels or online parameter updates. Towards a clinically usable all-in-one system. VertebraFormer achieves competitive accuracy, near real-time inference, and interpretable lesion heatmaps within a single model, reducing deployment complexity compared with multi-network pipelines while still requiring prospective validation before routine clinical use.

The automatic analysis of spinal CT images encompasses multiple interconnected research directions, including vertebrae segmentation, anatomical identification, lesion detection, multi-task learning, and domain generalization. While each of these topics has been studied independently, recent advances highlight the importance of integrating them into unified frameworks for robust and clinically viable systems. In the following subsections, we review representative works in each area, emphasizing their technical contributions and limitations.

Vertebrae segmentation and anatomical identification: Vertebrae segmentation and anatomical labeling form the foundation for automated spinal analysis. Early studies primarily focused on developing robust segmentation pipelines. For instance, Lessmann et al.[3] proposed an iterative fully convolutional network for simultaneous segmentation and labeling, ensuring anatomical consistency. Sekuboyina et al.[11] advanced this idea by introducing a localization-segmentation strategy for multi-label lumbar vertebra annotation. The release of large-scale datasets, such as CTSpine1K[8] and VerSe[4] catalyzed the development of high-capacity architectures, including Dense-U-Net[6], cascaded CNNs[5], nnU-Net[7], attention-enhanced U-Nets[18], and transformer-based 3D models such as UNETR[19] and Swin UNETR[20]. Building on these segmentation foundations, recent research has explored more flexible and scalable architectures.

Transformer-based and lightweight architectures: With the success of transformers in computer vision, researchers have adapted them for medical image segmentation, particularly in capturing long-range spatial dependencies[21]. H2Former[22] introduced a hierarchical hybrid transformer, while Scribformer[23] demonstrated how transformers can complement CNNs for weakly supervised segmentation. Other innovations, such as GETNet[24] and MedFormer[25], focused on enhancing multi-scale representation learning. In parallel, lightweight architectures like UNELRPT[26] and LDCNet[27] aimed to maintain high accuracy with reduced computational cost, an important factor for clinical deployment. Shape-prior regularization[28] further improved anatomical consistency. Such backbone improvements have also benefited higher-level clinical tasks, including lesion localization.

Lesion detection in vertebrae: Accurate lesion detection is essential for identifying pathological conditions such as metastatic tumors, osteoporotic fractures, and degenerative changes. Detection frameworks like RT-DETR[29], CAF-YOLO[10], and YOLOv9[30] have been adapted to handle the multi-scale characteristics of vertebral lesions. Joint detection-classification systems, such as knowledge-aware frameworks[9] and dynamic class-aware fusion networks[31], have addressed multi-class diagnostic scenarios. Additionally, domain-specific studies have explored bone density analysis[2], cervical spine injury evaluation[32], and spinal stenosis quantification[33]. Given the intertwined nature of segmentation, identification, and detection, multi-task learning has emerged as a promising solution.

Multi-Task learning and cross-modal approaches: Multi-task learning (MTL) frameworks aim to jointly solve multiple objectives. Segmentation, labeling, and detection, within a single architecture, enabling knowledge sharing across tasks. Recent work has introduced self-supervised pretraining[34], task-specific attention modules[15,16], and cross-modal transfer methods to improve generalization. The rise of foundation models has further extended these capabilities, with MA-SAM[35] enabling modality-agnostic segmentation and zero-shot generalization techniques[36] facilitating adaptation to unseen data. Multi-label zero-shot learning[37] and surveys on zero-shot AI[38] underline the potential of transfer learning in this field. While these advances improve adaptability, real-world clinical deployment also demands robustness to domain shifts.

Domain Generalization and Clinical Deployment: Domain generalization (DG) has become a critical focus for deploying spinal analysis models across multiple centers[12]. Approaches include adversarial domain alignment[13], shape-prior regularization[14], single-source DG[17], and meta-learning strategies[39]. Test-time adaptation[40] has been explored to refine predictions without retraining. For practical use, computational efficiency is equally important, as discussed in surveys on edge deep learning[41,42] and low-FLOP networks[43]. Recent evaluations[44] and fully automated pipelines[45] demonstrate the feasibility of clinically deployable systems that combine accuracy, speed, and cross-domain robustness.

Overall, the body of related work reveals a clear evolution from early task-specific models toward unified, domain-robust, and computationally efficient frameworks capable of handling segmentation, identification, and lesion detection in an integrated manner. While remarkable progress has been made in each of these research directions, the lack of a large, heterogeneous, and well-annotated benchmark has often limited fair evaluation and hindered the development of models with strong cross-domain generalization. To address this gap, we construct the MultiSpine benchmark, designed to provide broad coverage of diverse spinal regions, imaging protocols, and clinical conditions. The following section details the dataset composition, preprocessing pipeline, and annotation strategy adopted in this study.

Results

We present a comprehensive suite of experiments designed to rigorously evaluate VertebraFormer in terms of its effectiveness, cross-domain generalizability, and interpretability. All evaluations are conducted under a unified experimental protocol using the MultiSpine benchmark, which integrates multiple heterogeneous datasets covering diverse anatomical regions, imaging modalities, and clinical scenarios. This setting enables a thorough investigation of the model's performance not only in standard in-domain conditions but also under challenging domain shifts that closely mimic real-world deployment. We further provide both quantitative and qualitative analyses to substantiate the contributions of each architectural component and to demonstrate the clinical relevance of our predictions.

Experimental setup

All experiments are implemented in PyTorch 2.1 and conducted on a workstation equipped with four NVIDIA A100 GPUs (80 GB memory each), an AMD EPYC 7742 CPU, and 512 GB of RAM, running Ubuntu 22.04 with CUDA 12.1. VertebraFormer and all baselines are trained under identical settings for fair comparison.

Data Splits and Augmentation: Unless otherwise specified, each dataset in the MultiSpine benchmark is split into 60% training, 20% validation, and

Table 1 | Comparison of spinal CT models on the MultiSpine benchmark

Method	Segmentation	Identification	Lesion detection		
	Dice (%)	ID Acc (%)	Precision (%)	Recall (%)	AP (%)
nnU-Net[7]	77.5 ± 0.4	60.7 ± 2.2	62.4 ± 1.5	58.2 ± 1.5	60.6 ± 1.5
UNETR[19]	88.1 ± 0.3	79.5 ± 1.0	65.8 ± 1.2	64.3 ± 1.2	65.8 ± 1.2
TransBTS[46]	88.3 ± 0.3	80.1 ± 1.0	66.4 ± 1.2	64.8 ± 1.2	66.3 ± 1.2
H2Former[22]	88.6 ± 0.3	81.3 ± 0.9	66.9 ± 1.2	65.7 ± 1.2	66.9 ± 1.2
Scribformer[23]	88.7 ± 0.3	82.4 ± 0.9	67.4 ± 1.1	68.2 ± 1.1	67.8 ± 1.1
BLDS[47]	84.4 ± 0.4	70.5 ± 1.6	60.8 ± 1.7	55.3 ± 1.7	57.4 ± 1.7
Dense-U-Net[6]	82.2 ± 0.4	75.6 ± 1.3	58.7 ± 1.9	52.9 ±1.9	55.4 ± 1.9
VerFormer[48]	83.9 ± 0.3	82.4 ± 0.9	66.3 ± 1.1	67.4 ± 1.1	67.2 ± 1.1
Tao et al.[49]	84.5 ± 0.3	83.3 ± 0.8	65.9 ± 1.2	66.8 ± 1.2	66.2 ± 1.2
VerteFormer[50]	86.0 ± 0.3	84.2 ± 0.7	67.9 ± 1.1	67.4 ± 1.1	67.7 ± 1.1
VertDetect[51]	80.5 ± 0.6	80.6 ± 1.0	58.6 ± 1.9	52.3 ± 1.9	55.2 ± 1.9
Spineclue[52]	80.3 ± 0.6	80.4 ± 1.0	60.4 ± 1.7	55.6 ± 1.7	57.3 ± 1.7
Ortho2D[53]	78.4 ± 0.7	78.5 ± 1.1	55.4 ± 2.0	50.7 ± 2.0	52.6 ± 2.0
VertNet[54]	82.3 ± 0.5	82.6 ± 0.8	60.5 ± 1.6	58.2 ± 1.6	59.5 ± 1.6
VertebraFormer (ours)	89.3 ± 0.2	85.6 ± 0.6	71.1 ± 1.0	69.0 ± 0.9	68.7 ± 1.1

For each metric, we report mean ± 95% BCa bootstrap confidence intervals computed via patient-level resampling (10,000 replicates).

20% testing at the patient level to avoid data leakage. For cross-domain evaluation, we adopt a leave-one-domain-out protocol in which four datasets are used as sources for training, and the remaining one serves as the held-out target without fine-tuning. To enhance generalization, we apply 3D data augmentation including random rotation (±15°), scaling (±10%), translation (up to 10 voxels), elastic deformation, Gaussian noise injection, and intensity jittering.

Training Hyperparameters: Models are optimized using the AdamW optimizer with an initial learning rate of 1×10^{-4}, weight decay of 1×10^{-5}, and a cosine annealing scheduler. The batch size is set to 2 for 3D volumes due to memory constraints. All models are trained for 300 epochs with early stopping based on the validation Dice score. Gradient clipping with a maximum norm of 1.0 is applied to stabilize training.

Evaluation metrics and statistical analysis: We adopt three primary metrics to evaluate performance: (1) Dice similarity coefficient (Dice) for segmentation accuracy:

$$\text{Dice} = \frac{2|P \cap G|}{|P| + |G|}, \tag{1}$$

where P and G denote the predicted and ground-truth voxel sets. (2) Identification accuracy (ID Acc) measuring the proportion of correctly assigned vertebra IDs:

$$\text{ID Acc} = \frac{\text{correctly labelled vertebrae}}{\text{total vertebrae}}. \tag{2}$$

(3) Lesion average precision (Lesion AP) computed as the area under the precision-recall curve over lesion detection confidence scores:

$$\text{AP} = \int_0^1 p(r)\,\mathrm{d}r, \tag{3}$$

where $p(r)$ is precision as a function of recall r. All metrics are reported as percentages.

For uncertainty quantification, we compute 95% bias-corrected and accelerated (BCa) bootstrap confidence intervals by resampling patients (10,000 replicates) within each dataset. Unless otherwise stated, all reported values are means with 95% CIs. Pairwise comparisons between VertebraFormer and the strongest baseline per task use two-sided Wilcoxon signed rank tests across patients, with Holm–Bonferroni correction for multiple testing.

Benchmark comparison

To comprehensively assess the overall performance of the proposed framework, we benchmark VertebraFormer against a diverse set of strong methods on the MultiSpine dataset. The evaluation covers three complementary tasks critical to clinical spinal analysis: vertebrae segmentation, measured by Dice similarity; vertebrae identification, characterized by ID accuracy; and lesion detection, evaluated by Lesion AP together with precision and recall. The quantitative results, summarized in Table 1, indicate that our model achieves consistent improvements over competitive baselines across all indicators while also exhibiting smaller performance variability.

We assessed the performance of VertebraFormer across the specific pathology categories defined in our ontology. As shown in Table 2, the model demonstrates high sensitivity for morphological deformities (Fractures) but faces challenges with subtle density changes (Early Lytic Mets).

The performance drop in infectious etiologies correlates with the lower inter-rater reliability observed during annotation ($\kappa = 0.64$), suggesting that label ambiguity contributes to model uncertainty. Conversely, the robust detection of Grade 2/3 fractures (AP = 78.4%) confirms the model's efficacy in identifying structurally significant "actionable" findings.

Cross-domain generalization

To assess the domain robustness of VertebraFormer, we design a zero-shot cross-domain evaluation protocol. In each fold, the model is trained on data from four source domains and directly evaluated on an unseen target domain without any fine-tuning or domain-specific adaptation. This setting approximates realistic deployment scenarios in which models must operate on previously unseen clinical data from different scanners, institutions, or patient populations.

Table 3 delineates the leave-one-domain-out evaluation protocol, identifying the source domains aggregated for training and the designated held-out target domain reserved for zero-shot inference in each fold. Furthermore, it provides the quantitative distribution of CT volumes across the experimental splits, categorizing sample counts into training (N_{train}), validation (N_{val}), and testing (N_{test}) subsets.

Table 2 | Lesion detection performance stratified by clinical category

Lesion category	Count	AP (%)	Sens@1FP	Sens@4FP
Fractures (Grade 2/3)	412	78.4 ± 1.5	0.82	0.91
Lytic metastases	289	64.2 ± 2.1	0.68	0.79
Sclerotic metastases	156	69.8 ± 2.3	0.73	0.84
Infection/Erosion	84	52.1 ± 3.5	0.55	0.67
Overall	941	68.7 ± 1.1	0.72	0.83

Metrics reported are average precision (AP) and sensitivity at 1 false positive per image (Sens@1FP).

Table 3 | Leave-one-domain-out protocol with per-fold sample counts

Target	Sources	N_{train}	N_{val}	N_{test}
CTSpine1K	SpineWeb, VerSe, CSMD, Private A, Private B	1005	268	198
SpineWeb	CTSpine1K, VerSe, CSMD, Private A, Private B	1155	444	128
VerSe 2020	CTSpine1K, SpineWeb, CSMD, Private A, Private B	1502	362	103
CSMD	CTSpine1K, SpineWeb, VerSe 2020, Private A, Private B	1465	415	50
Private A	CTSpine1K, SpineWeb, VerSe 2020, CSMD, Private B	1489	423	42
Private B	CTSpine1K, SpineWeb, VerSe 2020, CSMD, Private A	1459	413	52

N_{train} and N_{val} denote source-domain volumes; N_{test} denotes target-domain volumes.

Table 4 | Summary of the constituent datasets in the *MultiSpine* benchmark

Dataset	Region	Scans (Tr/Val/Te)	Resolution (mm^3)	Annotation	License
CTSpine1K[8]	Cervical	1005 (610/197/198)	Mixed	Mask, ID	Public
SpineWeb[55]	Lumbar	609 (460/21/128)	0.6 × 0.6 × 2.0	Mask	Public
VerSe 2020[4]	All	319 (113/103/103)	Mixed	Mask, ID, Lmk	Public
CSMD[56–61]	Cervical	250 (Total)	Mixed	Label, Mask	Public
Private A	Thoracic	210 (126/42/42)	0.75 × 0.75 × 1.5	Mask	IRB
Private B	Lumbar	260 (156/52/52)	0.7 × 0.7 × 1.0	Mask, ID	Internal

Most datasets consist of CT scans, while CSMD introduces multimodal data (CT + MRI). The Resolution is given in $x \times y \times z$ axis order. "Scans (train/val/test)" lists patient-level volumes per split.

The MultiSpine benchmark covers six heterogeneous datasets spanning various anatomical regions, acquisition parameters, and labeling protocols (Table 4), which introduces substantial distribution shifts in terms of intensity profiles, spatial resolution, and noise characteristics. We report the average Dice score for vertebrae segmentation and the identification (ID) accuracy across target domains in Table 5. These metrics jointly reflect both structural delineation quality and vertebra level labeling correctness under challenging cross-domain conditions.

As shown in Fig. 1, the proposed method consistently matches or outperforms competitive baselines in all transfer directions, with limited degradation from in-domain to cross-domain evaluation. This behavior highlights the effectiveness of the dynamic modulation and multi-task feature alignment mechanisms in mitigating domain-specific biases. The heatmap further reveals that even in challenging transfer cases, such as training on predominantly cervical datasets and testing on lumbar-only domains, VertebraFormer maintains competitive accuracy.

Table 5 | Cross-domain generalization performance of VertebraFormer on unseen datasets

Target domain	Dice (%)	ID Acc (%)
CTSpine1K	86.9 ± 0.5	78.2 ± 1.3
SpineWeb	87.5 ± 0.4	80.4 ± 1.1
VerSe 2020	88.1 ± 0.4	81.7 ± 1.0
CSMD	88.4 ± 0.4	81.9 ± 1.0
Private A	87.3 ± 0.5	80.1 ± 1.2
Private B	87.8 ± 0.4	80.9 ± 1.1

Values denote mean Dice and ID accuracy with 95% BCa confidence intervals.

Ablation study

To investigate the effectiveness and individual contributions of each proposed component within VertebraFormer, we perform a series of ablation experiments. Specifically, we progressively integrate the multi-head identification branch, the lesion detection module, and the dynamic modulation block into a baseline UNETR architecture. The results, summarized in Table 6, reveal that each component yields measurable improvements across Dice, ID accuracy, and Lesion AP. The identification branch substantially enhances anatomical labeling precision, while the lesion detection module equips the network with the ability to localize and characterize pathological regions. Notably, the dynamic modulation block delivers the most pronounced performance gains across all metrics, underscoring its role in refining feature representations for both healthy and pathological vertebrae under heterogeneous imaging conditions.

Complexity and efficiency analysis

To complement the accuracy-oriented evaluations, we analyze the computational complexity and inference efficiency of VertebraFormer,

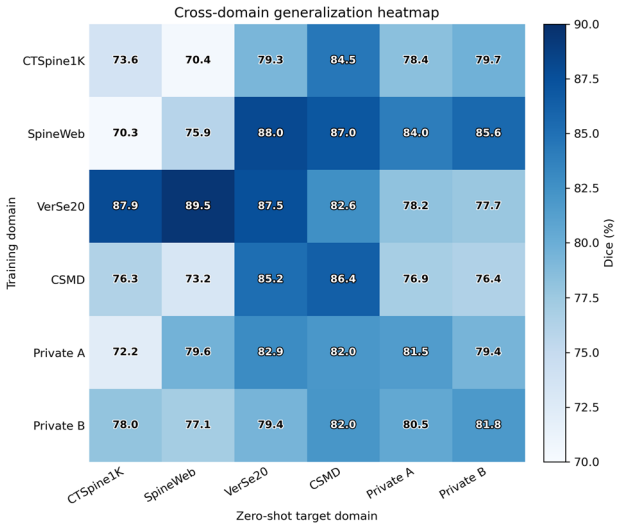

Fig. 1 | Cross-domain generalization heatmap. Rows represent training datasets and columns indicate the zero-shot target domains.

Table 6 | Ablation study: contribution of each component to overall performance on the MultiSpine validation set

Model variant	Dice (%)	ID Acc (%)	Lesion AP (%)
UNETR baseline	88.1	79.5	–
+ Identification branch	88.4 (+0.3)	83.2 (+3.7)	–
+ Lesion detection head	88.5 (+0.1)	83.1 (− 0.1)	65.8
+ Dynamic modulation	89.3 (+0.8)	85.6 (+2.5)	68.7

Values in parentheses indicate relative improvements over the previous configuration.

Table 7 | Model complexity comparison on a 128^3 input

Method	#Params (M)	FLOPs (G)	Memory (GB)
nnU-Net[7]	48.0	350.0	12.5
UNETR[19]	92.1	380.5	15.2
TransBTS[46]	33.4	333.2	12.0
H2Former[22]	102.7	410.3	16.8
Scribformer[23]	88.9	365.7	14.7
Dense-U-Net[6]	55.6	321.4	6.1
Tao et al.[49]	36.1	299.8	5.2
VertebraFormer (ours)	85.6	342.4	13.9

#Params and FLOPs are reported in millions (M) and gigaFLOPs (G), respectively[43].

Table 8 | Inference efficiency comparison on an NVIDIA RTX A6000 GPU (FP16 precision)

Method	Vol/s	Latency/Volume (ms)
nnU-Net[7]	9.4	106.4
UNETR[19]	8.3	120.5
TransBTS[46]	10.2	98.0
H2Former[22]	7.8	128.2
Scribformer[23]	8.7	114.9
Dense-U-Net[6]	6.3	157.8
Tao et al.[49]	7.1	140.8
VertebraFormer (ours)	13.8	72.5

Throughput is measured in 3D volumes processed per second (Vol/s) for 128^3 inputs, including all decoding and post-processing[41].

comparing it with representative baselines including UNETR[19], H2Former[22], and Scribformer[23]. Model complexity is quantified in terms of the total number of trainable parameters (#Params) and floating point operations (FLOPs) computed for a standard 128 × 128 × 128 voxel input. Inference efficiency is evaluated by measuring volumes processed per second (Vol/s) and the average per-volume latency on an NVIDIA RTX A6000 GPU (48 GB VRAM) under mixed precision (FP16) execution[41]. As summarized in Table 7, VertebraFormer achieves a competitive model size and computational cost compared with transformer architectures[22,23], while delivering higher accuracy. This efficiency stems from the lightweight identification branch and parameter sharing within the dynamic modulation module[43]. Table 8 further shows that the method attains a throughput of 13.8 volumes per second, enabling near real-time processing for volumetric spinal CT scans[41].

These results demonstrate that VertebraFormer strikes a favorable balance between accuracy and computational efficiency, supporting near real-time processing for typical clinical workloads.

Qualitative visualization

Figure 2 presents a set of representative qualitative results that comprehensively illustrate the performance of VertebraFormer across the three target tasks: vertebra segmentation, identification, and lesion detection. Each example is organized into four panels: (1) the original CT slice providing the raw anatomical context, (2) the ground-truth segmentation overlaid with vertebra IDs, (3) the predicted segmentation mask and identification labels produced by our model, and (4) the corresponding lesion detection heatmap.

From the visual comparison, it can be observed that the predicted segmentation masks exhibit a high degree of spatial overlap with the ground-truth annotations, accurately delineating vertebral boundaries even in challenging regions such as the thoracolumbar junction and degenerated disks. The ID assignment remains consistent along the entire spine, effectively handling cases with partial vertebra visibility or irregular morphology, which are often challenging for conventional approaches.

In addition to segmentation and identification, the lesion detection heatmaps successfully localize clinically relevant abnormalities with precise spatial alignment. These heatmaps reveal subtle pathological patterns such as localized cortical thinning, vertebral fractures, and tumor infiltration, which are visually distinct from normal tissue regions. Such interpretability enables radiologists to rapidly pinpoint abnormal regions while preserving the broader anatomical context, reducing diagnostic ambiguity.

Notably, the visualization results also demonstrate the robustness of our method against image quality variations, including motion artifacts, heterogeneous contrast enhancement, and intensity inhomogeneities across scans from different devices or institutions. The strong alignment between predicted and reference annotations further supports the reliability of the proposed architecture for real-world deployment.

Dynamic domain modulation and inference time adaptation

We investigated the model's robustness to domain shifts and the efficacy of our dynamic modulation strategy at test time. VertebraFormer's default pipeline uses a learned domain embedding to modulate the network for different CT image domains (scanner or dataset differences). We compare several mitigation strategies: (1) using random or swapped embeddings

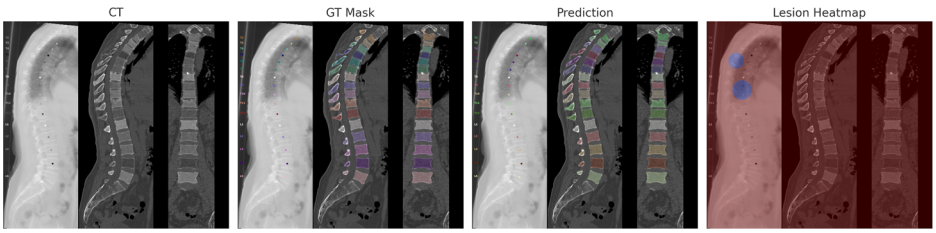

Fig. 2 | Qualitative visualization. (1) CT image, (2) ground-truth mask, (3) predicted segmentation, and (4) lesion heatmap. Arrows indicate the localization of the T11 Endplate defect, demonstrating the spatial alignment between the predicted lesion heatmap and the ground truth.

Table 9 | Domain modulation vs. adaptation performance

Test-time method	Dice (%)	ID Acc (%)	Lesion AP (%)
Domain embed (0% error)	90	97	80
Domain embed (10% error)	88	96	78
Domain embed (20% error)	85	95	75
Domain embed (30% error)	82	94	70
Random embedding (100% error)	75	90	60
Swapped embedding (100% error)	78	92	65
IN adaptation	84	93	72
AdaIN adaptation	86	94	76
TENT adaptation	88	95	78

Segmentation accuracy (Dice coefficient), vertebra identification (ID) accuracy, and lesion detection AP are shown for the default domain embedding pipeline under varying degrees of domain label noise, as well as for ablation and adaptation strategies. Domain embed (*X*% error) indicates the model used the domain-specific modulation pipeline with *X*% of test cases given incorrect domain identifiers. Random embedding and Swapped embedding test the model with entirely wrong domain codes (randomly assigned per scan, or systematically swapped between domains). IN = instance normalization applied at test time; AdaIN = adaptive instance normalization; TENT = test-time entropy minimization. All metrics are reported as percentages.

(ablating the informed domain code), and (2) applying alternative test time adaptation techniques that do not rely on a priori domain knowledge, specifically, instance normalization (IN), adaptive instance normalization (AdaIN), and test time entropy minimization (TENT). We also quantify performance degradation as a function of domain misclassification rate by injecting noise into the domain labels. Table 9 reports the vertebra segmentation Dice score, vertebra identification (ID) accuracy, and lesion detection AP under these various conditions.

When the correct domain embedding is provided (0% error), the model achieves the highest segmentation and identification performance (mean Dice ≈ 90%, vertebra ID accuracy 97%) along with strong lesion detection (AP 80%). However, model performance deteriorates gradually as domain misclassification increases. With 10–30% domain errors in the test set, the Dice drops to 88 → 85 → 82%, and lesion AP falls from 78% to ~70%, indicating that even modest rates of domain misidentification can measurably impact both segmentation and lesion detection accuracy. Vertebra identification is slightly more robust (declining from 97% to 94% as error rises to 30%), suggesting the model's identification logic retains high accuracy unless domain mismatches become severe. In the worst-case ablation where domain embeddings are entirely randomized or consistently swapped, performance declines much further: using a wrong embedding for every scan yields Dice ≈ 75−78% and lesion AP ≈ 60−65%, a substantial degradation from the baseline. These drops confirm that the domain-specific modulations learned by the network are critical; feeding incorrect domain codes disrupts the feature normalization and attention layers, leading to missed vertebrae and lesions.

By contrast, the adaptive test time methods (lower rows of Table 9) show improved resilience to domain shifts. Without requiring any domain identifier, TENT adaptation achieves a Dice of ~88% and ID accuracy 95%, nearly matching the error-free baseline. This suggests that online entropy minimization can effectively adjust the model to the target data distribution, recovering much of the lost performance when domain labels are unreliable. AdaIN, which dynamically aligns intensity statistics of the test scan to the source domain, also performed well (Dice ~ 86%, AP ~ 76%), outperforming simple instance normalization (Dice ~ 84%). Notably, even the IN strategy, which standardizes feature channels on-the-fly boosted segmentation Dice by ~2−3 points over the naive random embedding case, indicating that some form of on-the-fly domain adaptation is beneficial. Among all methods, the default pipeline with correct domain information still yields the highest accuracy. Yet, TENT's near parity with the fully informed model (Dice 88% vs. 90%) underscores the value of test time training as a robust alternative when domain labels are absent or prone to error.

Robustness to input perturbations

We conducted ablation experiments to assess model robustness under various image perturbations, simulating potential real-world variability. All tests use the MultiSpine data (covering cervical, thoracic, and lumbar regions) and evaluate segmentation overlap (Dice), anatomical identification accuracy (ID Acc), and lesion detection average precision (Lesion AP). Figure 3 summarizes the performance degradation under each perturbation, broken down by spinal region. Baseline performance without perturbation is high (Dice 88−90%, ID Acc 84−87%, AP 65−70% depending on region). We report mean metrics with 95% confidence intervals.

VertebraFormer demonstrates strong resilience to moderate intensity and noise perturbations, maintaining high accuracy under realistic variations. Marked performance degradation occurs only with severe distortions. The degradation is generally uniform across vendor-specific subdomains as well, i.e., no particular scanner's data was disproportionately fragile, thanks to the model's domain-general features. However, certain domains did show slight differences in perturbation response. Importantly, the model's structure-aware design helped preserve anatomical consistency even in perturbed cases, we observed that identification errors under perturbations were usually limited to one-off mistakes (e.g. one mis-numbered vertebra), rather than chaotic mislabeling of the entire sequence. This indicates the model still leverages global spinal context to stay on track, a property that classic slice-by-slice methods might lack. These stress tests highlight scenarios for caution: extremely noisy or low-resolution scans, and the presence of metal hardware, are cases where performance can falter and where additional training or preprocessing would be beneficial.

Unified model vs. modular pipeline

A key motivation for VertebraFormer was to replace multi-stage pipelines with a single unified model. To quantify the benefits, we conducted a simulated comparison between VertebraFormer and an equivalent three-model pipeline under a fixed latency budget. The pipeline consisted of: (1) a 3D UNETR segmentation network to segment vertebrae, (2) a 3D ResNet classifier to assign anatomical labels to each segmented vertebra, and (3) a YOLOv5-inspired detector to localize lesions (treating each vertebra region as a potential bounding box for lesions). In their full-capacity forms, such a pipeline would be significantly slower and heavier than VertebraFormer—for instance, UNETR alone has ~92M parameters and ~120 ms inference

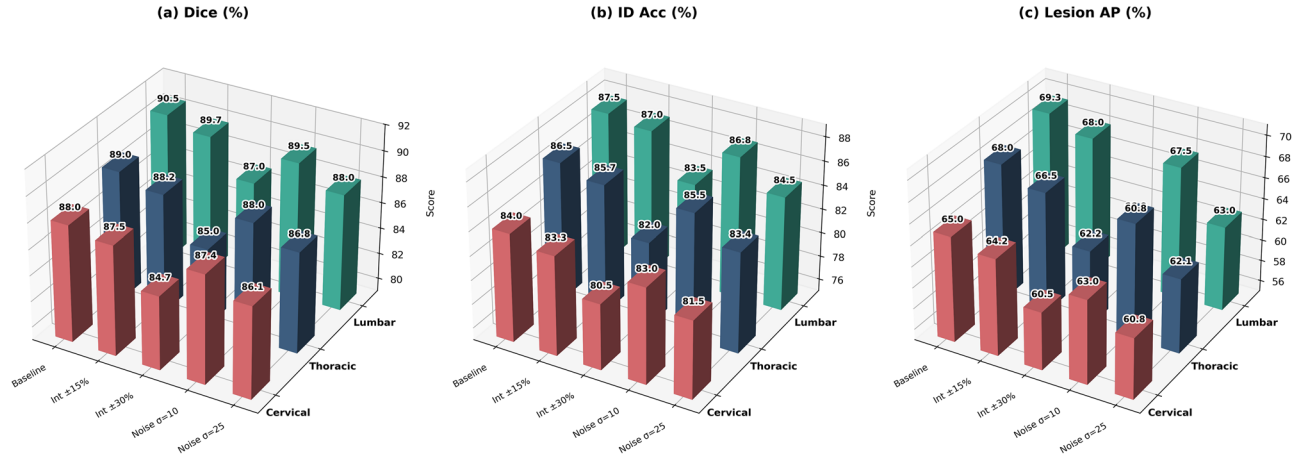

Fig. 3 | Performance of VertebraFormer under synthetic perturbations, by anatomical region. We report mean segmentation **a** Dice (%), vertebra **b** ID accuracy (%), and **c** lesion detection AP (%) for each spinal region, at baseline and under various input degradations. 95% confidence intervals (bootstrapped) are given in parentheses. Intensity ± 15% and ± 30% refer to linear scaling of HU intensities by ±15% and ±30%. Noise σ denotes additive Gaussian noise with standard deviation σ in HU. Slice 3 mm/5 mm is a thicker slice reconstruction. Metal artifact indicates the presence of simulated orthopedic hardware and streaks.

Table 10 | Unified model vs. equivalent three-model pipeline (matched latency)

Method	Segmentation dice	ID accuracy	Lesion AP	Latency (A6000) (ms)	Peak memory (GB)
VertebraFormer (All-in-one)	89.3% (±0.2)	85.6% (±0.6)	68.7% (±1.1)	~72	~13.9
3-Model Pipeline	86.5% (±0.4)	81.0% (±0.8)	60.5% (±1.2)	~75	~15.8

VertebraFormer is compared to a modular pipeline (UNETR for segmentation, ResNet classifier for ID, YOLO-based lesion detector). Both are constrained to ~72 ms total runtime on RTX A6000 (by reducing pipeline model sizes/resolution). Metrics are on the MultiSpine test set. The unified approach yields higher accuracy on all tasks. Memory is the peak GPU usage. Confidence intervals (95% CI) are shown for the metrics.

time per volume on A6000, and adding two more networks would further increase latency and memory use. To make a fair comparison, we constrained the pipeline to match VertebraFormer's runtime (~72 ms per volume on A6000) by down-scaling or optimizing each component. Specifically, we reduced the segmentation input resolution by ~50% (to speed up UNETR), used a lightweight ResNet-18 for classification on cropped vertebra patches, and limited the lesion detector to only evaluate proposals in high-probability regions (instead of a full dense scan). These adjustments brought the pipeline's total processing time to ~75 ms (within ~5% of the unified model), at the cost of some accuracy. Table 10 presents the performance of the unified model vs. the pipeline under this equalized latency setting.

Discussion

In this work, we presented VertebraFormer, a unified transformer framework for simultaneous vertebra segmentation, anatomical identification, and lesion detection in heterogeneous spinal CT scans. To enable comprehensive evaluation, we constructed the MultiSpine benchmark by aggregating and harmonizing six diverse datasets covering cervical, thoracic, and lumbar regions, thereby capturing substantial inter-domain variability in imaging protocols, anatomical appearance, and clinical conditions.

Through extensive experiments, we demonstrated that VertebraFormer consistently outperforms strong baselines across all three tasks in both in-domain and challenging cross-domain scenarios. The proposed dynamic modulation mechanism and identification branch enhance anatomical consistency and robustness under domain shifts, while the lesion detection module provides interpretable, clinically relevant localization of pathological regions. Ablation studies confirmed the contribution of each architectural component, and complexity analysis revealed that the method achieves competitive inference efficiency despite its multi-task design.

Importantly, qualitative visualizations illustrated that the framework produces anatomically faithful predictions with clear lesion heatmaps, offering practical value for radiologists by enabling efficient assessment of spinal pathologies. These findings highlight the potential of VertebraFormer as a tool for automated, multi-faceted spinal image analysis, while further prospective validation is required before routine clinical deployment.

Looking forward, the proposed framework can be extended to leverage multi-modal imaging (e.g., MRI + CT), incorporate patient-specific clinical metadata, and adapt to other spinal diseases, such as scoliosis or degenerative disc disorders. We believe that such advancements will further improve generalizability and broaden the clinical utility of multi-task learning in spinal image analysis.

Limitations and future work: While VertebraFormer achieves strong performance on the MultiSpine benchmark and demonstrates promising cross-domain generalizability, several limitations remain.

First, although the MultiSpine dataset aggregates six heterogeneous sources, it is still limited in its coverage of rare pathological conditions, extreme imaging artifacts, transitional vertebrae, extensive implants, and pediatric cases. This may constrain the model's ability to generalize to highly atypical anatomies or scans acquired with unconventional protocols. Incorporating additional datasets, particularly those containing rare diseases and diverse patient populations, would further improve robustness.

Second, the current framework is trained exclusively on CT data. In clinical practice, multi-modal imaging such as MRI or PET-CT can provide complementary structural and functional information that may enhance both anatomical identification and lesion characterization. Extending the framework to jointly process multi-modal data remains an important direction for future research.

Third, while the multi-task learning paradigm enables concurrent optimization of segmentation, identification, and lesion detection, potential task interference may occur when pathological features overlap with anatomical boundaries, leading to subtle errors in either task. Adaptive task weighting or uncertainty-aware loss balancing could mitigate such conflicts.

Finally, the evaluation is primarily focused on retrospective datasets with pre-acquired scans. Prospective studies in real-world clinical workflows—incorporating on-the-fly inference, interactive refinement by

radiologists, and integration with hospital PACS systems—are needed to fully assess deployment feasibility.

In future work, we plan to (1) enrich the dataset with rare and challenging cases from multi-center collaborations, (2) extend the architecture to handle multi-modal 3D imaging, (3) explore meta-learning and continual learning strategies for rapid domain adaptation, and (4) integrate clinical metadata and patient history to support more context-aware diagnostic predictions. We believe that addressing these limitations will further enhance the clinical applicability and generalization capacity of VertebraFormer.

Methods

Dataset and preprocessing

To comprehensively evaluate the generalizability of vertebrae segmentation, identification, and lesion detection models under diverse clinical conditions, we construct the MultiSpine benchmark by aggregating and harmonizing six heterogeneous spinal CT datasets. These datasets span different anatomical regions, including cervical, thoracic, and lumbar vertebrae, and are acquired using varying protocols, scanner vendors, and clinical settings. Such diversity ensures coverage of a wide range of anatomical variability, contrast distributions, and pathological manifestations, making the benchmark suitable for multi-domain training, cross-domain evaluation, and zero-shot transfer experiments. Table 4 summarizes the key characteristics of the datasets. Public datasets such as CTSpine1K, SpineWeb, VerSe 2020, and the multimodal CSMD provide open-access volumes, while institution-specific datasets (Private A and Private B) were collected under Institutional Review Board (IRB) approval and anonymised before use. We incorporate CSMD specifically to introduce MRI modalities (T1/T2) alongside standard CT data. Lesion annotations are available only for VerSe 2020 and Private B, where board-certified radiologists delineated vertebral fractures, metastatic lesions, and degenerative changes; for the remaining datasets (including CSMD), we use vertebra masks and IDs without lesion labels. Across these datasets, cervical scans typically exhibit smaller vertebrae with tighter spacing and thinner cortical bone, thoracic scans feature rib articulations and a narrower spinal canal, and lumbar scans contain larger vertebral bodies with more pronounced pedicles and thicker cortical structures.

Before model training, all scans undergo a unified preprocessing pipeline. Volumes with incomplete vertebral coverage or severe motion and metal artifacts are excluded. Each scan is resampled to an isotropic resolution of 1.0 mm^3, and intensity values in Hounsfield units are clipped to the range $[-1000, 1000]$ and linearly normalized to $[0, 1]$. All vertebrae masks are harmonized to follow a consistent anatomical ID encoding scheme, covering C1–C7, T1–T12, and L1–L5, ensuring label consistency across datasets. Additionally, rigid alignment is applied to standardize patient orientation in the axial plane.

To further enhance model robustness to domain variation, we apply a set of 3D data augmentations during training, including random rotations, scaling, elastic deformations, intensity perturbations, and random cropping. These augmentations simulate differences in patient positioning, scanner calibration, and field-of-view between institutions.

The diversity of MultiSpine is illustrated in Fig. 4, which shows a t-SNE projection of latent features extracted from the VertebraFormer encoder backbone (initialized with pre-trained weights). Distinct clusters corresponding to different anatomical regions and datasets indicate clear inter-domain shifts in style, contrast, and spatial structure. Representative axial CT slices from cervical, thoracic, and lumbar regions are shown in Fig. 5, highlighting anatomical differences that challenge cross-domain generalization.

For evaluation, each dataset is split into 60% training, 20% validation, and 20% testing at the patient level to prevent data leakage. In cross-domain experiments, we adopt the leave-one-domain-out protocol described in the section "Experimental setup", training on multiple source datasets and testing on an unseen target domain. This setup emulates realistic clinical deployment scenarios where a model must operate on new data distributions without labeled target samples.

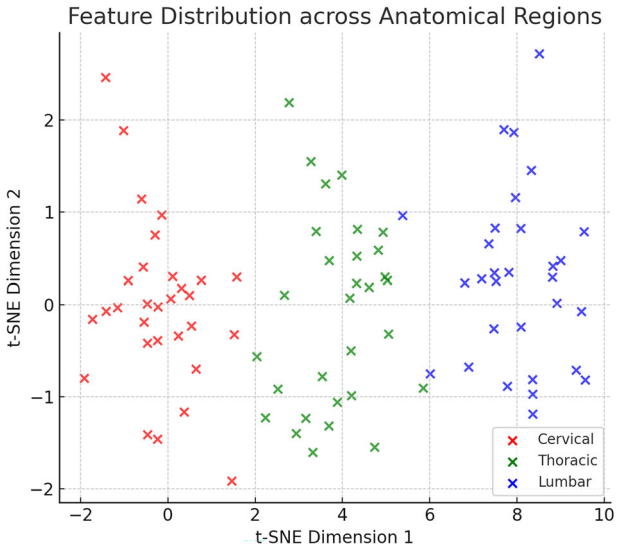

Fig. 4 | t-SNE visualization of feature distributions across domains in *MultiSpine*. Each color denotes a distinct anatomical region.

Dataset curation and rigorous deduplication

To construct the MultiSpine benchmark, we aggregated CT volumes from four public sources (CTSpine1K, SpineWeb, VerSe 2020, CSMD) and two private institutional cohorts. Given the heterogeneity of these sources, we implemented a strict three-stage audit protocol to prevent data leakage and ensure rigorous split hygiene.

We recognized that patient overlap across open-source datasets represents a significant risk. To address this, we employed a dual-verification strategy. First, for DICOM UID tracing, we extracted the Study Instance UID (0020,000D) and Series Instance UID (0020,000E) whenever available. Any volumes sharing a Study Instance UID were grouped as a single patient episode and were assigned exclusively to one of the data splits (train, validation, or test). Second, for perceptual hashing (pHash), we aimed to detect near-duplicate scans, including volumes that were identical except for resampling, cropping, or window/level variations introduced during anonymization. We computed 3D perceptual hashes for all volumes using a block-mean hashing algorithm applied to the central 50 slices of each scan, and pairs exhibiting a Hamming distance < 5 were flagged for manual radiological review.

This audit identified 43 overlapping volumes between the VerSe 2020 training set and the CTSpine1K test set, which had not been previously documented. These duplicates were removed from the training set to preserve the integrity of the test evaluation. We additionally found 12 patient overlaps in the private cohorts, arising from the same patient being scanned pre- and post-operatively; these were grouped to maintain strict patient-level independence across all splits.

After curation, the complete dataset was divided into training (60%), validation (20%), and testing (20%) sets using a stratified sampling strategy. Stratification was performed based on vendor/scanner characteristics to ensure that the test set captured a representative distribution of scanner-associated biases (e.g., GE, Siemens, Philips), thereby preventing overfitting to scanner-specific noise signatures. Stratification was also based on pathology severity, including the presence of metal artifacts and severe deformities (Cobb angle $> 20°$), to guarantee that the test set adequately challenged the model's robustness.

Strict patient-level separation was enforced throughout: all scans belonging to the same patient ID, across all time points and modalities, were confined to a single fold. To ensure the absence of leakage, we additionally performed a nearest-neighbor pixel analysis between the final test and training sets, confirming that no reconstruction-level duplicates remained.

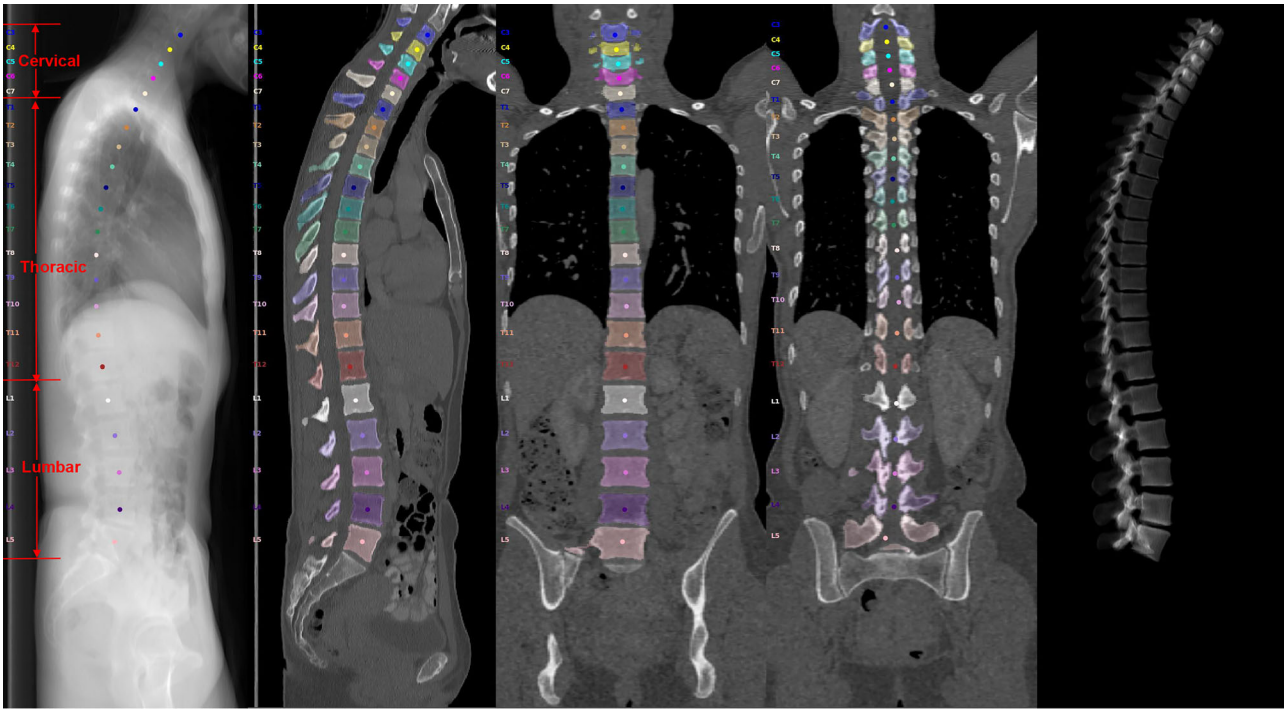

Fig. 5 | The use of axial computed tomography (CT) imaging to illustrate how the MultiSpine benchmark isdefined for the Cervical, Thoracic, and Lumbar regions of the spine. (Left Section) The colour-coded sagittal image presents the main three spinal regions: Cervical spine (C1-C7, marked in Red), thoracic spine (T1-T12, marked in Green) and Lumbar spine (L1-L5, marked in Blue). This image represents the anatomical structures associated with each spinal area, and provides a reference for the anatomical relationship between each spinal region. (Right Section) To demonstrate how each vertebra corresponds to the other, axial CT (cross-sectional) images of each spinal area are presented, using CT panels or slices to represent each region in 3 different rows.

Method: VertebraFormer network

We propose *VertebraFormer*, a unified multi-task transformer framework that simultaneously addresses three critical challenges in spinal image analysis: vertebrae segmentation, anatomical identification, and lesion detection. Our approach is motivated by the clinical reality that these tasks are inherently interconnected—accurate anatomical identification relies on precise segmentation boundaries, while lesion detection requires both spatial localization and anatomical context. Rather than treating these as independent problems, we design a unified architecture that leverages shared representations to improve performance across all tasks while maintaining computational efficiency suitable for clinical deployment.

Architecture design and overview

The VertebraFormer architecture embodies three fundamental design principles that address the unique challenges of multi-domain spinal image analysis. First, global context awareness recognizes that vertebral structures exhibit strong spatial relationships across the entire spine, where the identity and boundaries of individual vertebrae are best understood in relation to their neighbors. Second, task synergy acknowledges that segmentation, identification, and lesion detection are complementary tasks that can mutually reinforce each other through shared feature learning. Third, domain adaptability ensures robust performance across the heterogeneous imaging conditions encountered in real-world clinical practice.

Our architecture consists of four primary components working in concert: a transformer-based encoder that captures long-range spatial dependencies across the entire spine, specialized multi-task decoding heads that produce task-specific outputs while maintaining consistency, a dynamic modulation mechanism that adapts to varying imaging characteristics, and a carefully designed training objective that balances competing task requirements. The encoder processes 3D input volumes through patch-based tokenization followed by self-attention layers, enabling the model to capture global anatomical relationships essential for understanding spinal structure. The shared backbone representations are then processed by three specialized decoding branches: a hierarchical segmentation decoder with skip connections for precise boundary delineation, a region-based identification classifier for anatomical labeling, and a keypoint-style lesion detector for pathology localization.

The dynamic modulation mechanism operates as a bridge between the shared encoder and task-specific decoders, adaptively adjusting feature representations based on imaging characteristics. This design enables the model to maintain high performance across diverse clinical scenarios while preserving the benefits of unified multi-task learning. The full architecture is illustrated in Fig. 6.

Transformer-based encoder architecture

The foundation of our approach lies in effectively tokenizing 3D medical volumes for transformer processing. Given an input volume $I \in \mathbb{R}^{H\times W\times D}$, we partition it into non-overlapping cubic patches of size P^3, resulting in a sequence of $N = \frac{HWD}{P^3}$ patches. Each patch $p_i \in \mathbb{R}^{P^3}$ is flattened and linearly projected to create patch embeddings $e_i \in \mathbb{R}^{d_{\text{model}}}$:

$$e_i = \text{Linear}(\text{Flatten}(p_i)) + \text{pos}_i \tag{4}$$

where pos_i represents learnable 3D positional encodings that preserve spatial relationships crucial for anatomical understanding. The choice of patch size P represents a critical trade-off between computational efficiency and spatial resolution—larger patches reduce sequence length but may lose fine-grained details essential for accurate boundary delineation.

The patch embeddings are processed through L transformer layers, each consisting of multi-head self-attention (MHSA) and feed-forward networks (FFN). The self-attention mechanism proves particularly advantageous for spinal image analysis, as it naturally captures the long-range

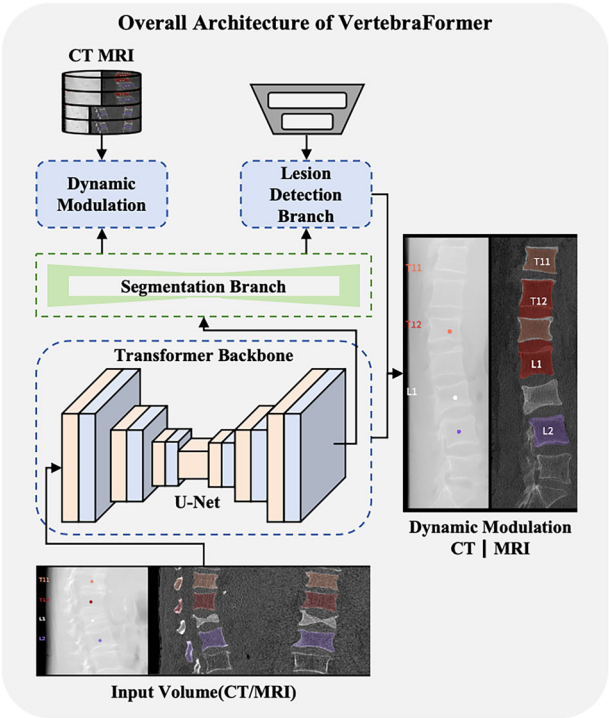

Fig. 6 | Overall architecture of the proposed VertebraFormer. It consists of a transformer encoder, multi-task decoding heads, and a dynamic modulation module.

dependencies that characterize vertebral relationships:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^{\text{T}}}{\sqrt{d_k}}\right)V \tag{5}$$

where Q, K, and V are query, key, and value matrices, respectively. This formulation allows the model to directly relate patches from distant spinal regions, enabling robust identification of vertebral boundaries even in the presence of pathological changes or imaging artifacts.

Our encoder generates multi-scale feature representations by extracting features from intermediate transformer layers at resolutions $\{\frac{1}{4}, \frac{1}{8}, \frac{1}{16}\}$ of the input size. These hierarchical features serve dual purposes: they provide skip connections for the segmentation decoder to preserve fine spatial details, and they offer multi-scale context for both identification and lesion detection tasks. The multi-scale design acknowledges that different aspects of spinal analysis benefit from different levels of spatial abstraction. Segmentation requires fine details, while identification relies more on broader anatomical context.

Multi-task decoding heads

The segmentation decoder employs a hierarchical upsampling strategy that progressively reconstructs spatial resolution while integrating multi-scale features from the transformer encoder. The decoder consists of four upsampling blocks, each doubling the spatial resolution through transposed convolutions while halving the channel dimension:

$$F^{(l+1)} = \text{Up}(F^{(l)}) + \text{Conv}(\text{Skip}^{(l)}) \tag{6}$$

where $F^{(l)}$ represents features at level l, and $\text{Skip}^{(l)}$ denotes skip connections from corresponding encoder layers. This design ensures that fine-grained spatial information lost during encoding is recovered during decoding, essential for accurate vertebral boundary delineation. The final segmentation output employs a hybrid prediction strategy that generates both binary vertebral masks and multi-class semantic labels, enabling the model to distinguish individual vertebrae while maintaining overall structural coherence.

Vertebral identification operates at the instance level, classifying each segmented vertebral region into anatomical categories. Our approach leverages 3D Region of Interest Alignment (RoIAlign) operations to extract fixed-size feature representations from variable-sized vertebral regions:

$$f_{\text{roi}} = \text{RoIAlign}\,(F_{\text{enc}}, \text{bbox}_i) \tag{7}$$

where F_{enc} represents encoded features and bbox_i denotes the bounding box of the ith vertebral region. The RoIAlign operation ensures that spatial information is preserved during feature extraction, crucial for maintaining the geometric relationships that characterize different vertebral types. The extracted region features are processed through a multi-layer classifier that maps geometric and contextual information to anatomical labels spanning C1–C7, T1–T12, and L1–L5, employing shared parameters across domains to ensure consistent identification criteria.

The lesion detection module adopts a center-point detection paradigm that predicts lesion locations as local maxima in learned heatmaps. This approach is well-suited to vertebral pathologies, which often present as subtle, irregularly shaped abnormalities that are poorly captured by rigid 3D bounding boxes. The detection head generates three complementary outputs: classification heatmaps $H \in \mathbb{R}^{C\times H\times W\times D}$ indicating lesion probability at each spatial location, regression offsets $O \in \mathbb{R}^{3\times H\times W\times D}$ for precise center localization, and size estimates $S \in \mathbb{R}^{3\times H\times W\times D}$ for lesion extent prediction:

$$H, O, S = \text{DetectionHead}\,(F_{\text{enc}}). \tag{8}$$

During training, each annotated lesion (fracture, metastatic deposit, or degenerative lesion) is mapped to the nearest feature-grid location, around which we place a 3D Gaussian kernel with radius r chosen according to the lesion's extent and voxel spacing. Voxels with a Gaussian value above a threshold τ_{pos} are treated as positives for the corresponding class channel in H, while neighboring voxels are down-weighted. Offsets O are regressed as residuals from the discretised grid position to the continuous lesion center, and S regresses log-scale lesion widths along the three spatial axes. Only scans from VerSe 2020 and Private B include lesion annotations; the remaining datasets contribute to segmentation and identification only.

In inference, local maxima are extracted from the heatmaps by thresholding H at a confidence level θ and applying 3D non-maximum suppression with a fixed radius in physical space (5 mm). Each surviving maximum yields a candidate lesion with center $x = x_{\text{grid}} + O(x_{\text{grid}})$ and size $S(x_{\text{grid}})$. Predicted lesions are matched to ground-truth lesions using the Hungarian algorithm with a combined criterion of 3D intersection-over-union (IoU $\geq$0.3) and center distance ($\leq$5 mm). Average Precision is computed per lesion class from the resulting precision-recall curves and then macro-averaged across classes.

Dynamic modulation mechanism

Clinical CT imaging exhibits substantial variation across institutions, scanner manufacturers, and acquisition protocols. These variations manifest as differences in intensity distributions, noise characteristics, spatial resolution, and contrast patterns that can significantly impact model performance. Traditional domain adaptation approaches require access to target domain data during training or fine-tuning, limiting their applicability in clinical deployment scenarios.

Our dynamic modulation mechanism addresses this challenge by learning to adaptively adjust feature representations based on intrinsic imaging characteristics. The dynamic modulation block operates on intermediate transformer features, applying learned transformations conditioned on domain-specific embeddings. Given encoded features $F \in \mathbb{R}^{N\times d}$ and a domain embedding $e_{\text{d}} \in \mathbb{R}^{d_{\text{e}}}$, the modulation block generates both

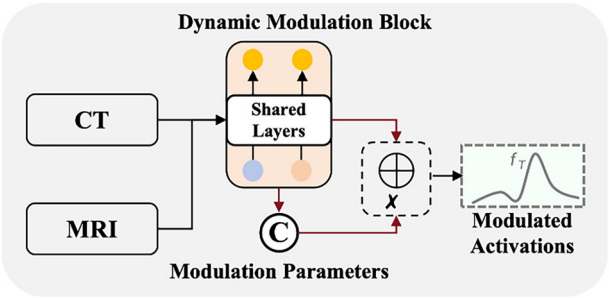

Fig. 7 | Dynamic modulation block. Modulates transformer features based on input modality signals (CT/MRI). This mechanism facilitates the cross-modality adaptation evaluated in Table 12.

channel-wise and spatial attention maps:

$$\alpha_c = \text{sigmoid}(\text{FC}_c(e_d)) \tag{9}$$

$$\alpha_s = \text{sigmoid}(\text{Conv}_s(\text{reshape}(e_d))) \tag{10}$$

The modulated features are computed as

$$F_{\text{mod}} = F \odot (\alpha_c \otimes \mathbf{1}_N) \odot \text{broadcast}(\alpha_s) \tag{11}$$

where $\odot$ denotes element-wise multiplication and $\otimes$ represents outer product. This formulation enables the model to selectively emphasize channels and spatial regions that are most informative for each specific imaging context. Figure 7 illustrates the structure of this component.

Domain embeddings are learned through a contrastive approach that encourages similar embeddings for scans from the same domain while promoting diversity across different domains. During training, we sample mini-batches containing scans from multiple domains and optimize the embedding space:

$$L_{\text{domain}} = -\log \frac{\exp(\text{sim}(e_i, e_j^+)/\tau)}{\sum_k \exp(\text{sim}(e_i, e_k)/\tau)}, \tag{12}$$

where e_j^+ represents a positive example from the same domain, and τ is a temperature parameter. At inference, we do not update model parameters. Instead, a lightweight domain classifier $g(\cdot)$, trained jointly with the encoder, predicts a probability vector over source domains, $p(d|I) = g(I)$. The effective domain embedding is obtained as a convex combination of source-domain prototypes,

$$e_d^* = \sum_{d \in \mathcal{D}} p(d|I)\, e_d, \tag{13}$$

which is then used in the dynamic modulation block. The function InferDomain(I) in Algorithm 1 refers to this feed-forward procedure. We therefore view VertebraFormer as employing domain-conditioned feature modulation at test time rather than iterative test-time adaptation of model weights.

Training objective and optimization

The overall training objective combines task-specific losses through a carefully designed weighting scheme:

$$L_{\text{total}} = \lambda_{\text{seg}} L_{\text{seg}} + \lambda_{\text{id}} L_{\text{id}} + \lambda L + \lambda_{\text{domain}} L_{\text{domain}} \tag{14}$$

where λ_{seg}, λ_{id}, λ_{det}, and λ_{domain} are task-specific weighting factors set empirically to 1.0, 0.5, 0.5, and 0.1, respectively.

The segmentation loss employs a hybrid formulation combining Dice loss for overlap maximization and focal cross-entropy for boundary refinement:

$$L_{\text{seg}} = L_{\text{dice}} + \alpha L_{\text{focal}} \tag{15}$$

The Dice loss encourages global overlap between predicted and ground truth masks:

$$L_{\text{dice}} = 1 - \frac{2\sum_i p_i g_i + \epsilon}{\sum_i p_i + \sum_i g_i + \epsilon} \tag{16}$$

where p_i and g_i represent predicted and ground truth probabilities at voxel i. The focal loss component addresses class imbalance by down-weighting easy examples. Unless otherwise stated, we set $\alpha = 0.5$ for the segmentation branch and use a standard focal formulation with focusing parameter $\gamma = 2.0$ and class-balancing factor $\alpha_{\text{cls}} = 0.25$.

The identification loss combines standard cross-entropy classification with spatial consistency constraints that enforce anatomical plausibility:

$$L_{\text{id}} = L_{\text{ce}} + \beta L_{\text{consistency}} \tag{17}$$

The lesion detection loss addresses the challenges of detecting subtle, irregularly-shaped pathologies through a combination of focal loss for center-point prediction and smooth L1 loss for offset regression:

$$L = L_{\text{focal}} + \lambda_{\text{reg}} L_{\text{reg}}, \tag{18}$$

where L_{focal} uses the same (γ, α_{cls}) as the segmentation branch and $\lambda_{\text{reg}} = 1.0$.

Model training employs a two-stage curriculum that progressively introduces task complexity. In the first 100 epochs, we train VertebraFormer with segmentation only ($\lambda_{\text{seg}} = 1.0$, $\lambda_{\text{id}} = \lambda_{\text{det}} = 0$) and domain contrastive regularization ($\lambda_{\text{domain}} = 0.1$) to establish basic anatomical understanding. In the next 50 epochs, we linearly ramp λ_{id} and λ_{det} from 0 to 0.5 while keeping λ_{seg} fixed, after which all weights remain constant for the remaining 150 epochs. This curriculum prevents early task interference and encourages the model to develop robust shared representations.

Data augmentation plays a crucial role in improving robustness across diverse imaging conditions. Our augmentation strategy includes spatial transformations (rotation, scaling, translation), intensity manipulations (brightness, contrast, noise injection), and elastic deformations that simulate anatomical variations. Augmentations are applied consistently across all task labels to maintain spatial correspondence.

Training employs the AdamW optimizer with initial learning rate lr = 1×10^{-4}, weight decay wd = 1×10^{-5}, and cosine annealing scheduling. The model is trained for 300 epochs with early stopping based on validation performance. Gradient clipping with maximum norm 1.0 stabilizes training, particularly important given the complex multi-task objective landscape.

Implementation and experimental validation

The VertebraFormer framework is implemented to efficiently handle multi-task learning across diverse imaging domains. Our implementation strategy focuses on computational efficiency, memory optimization, and reproducible training procedures that can be deployed in clinical environments.

Table 11 | Ablation study of VertebraFormer components on the MultiSpine validation set

	Segmentation + identification			Lesion detection		
Configuration	**Dice (%)**	**ID Acc (%)**	**mIoU**	**Dice (%)**	**ID Acc (%)**	**Lesion AP (%)**
UNETR Baseline[19]	88.1	79.5	72.4	–		
+Multi-head ID Branch[15]	88.4	83.2	74.6	–		
+Lesion Detection Head[16]	88.5	83.1	74.5	88.5	83.1	65.8
+Dynamic Modulation[13]	89.3	85.6	76.1	89.3	85.6	68.7

Progressive addition of components shows consistent improvements across all metrics. UNETR[19] serves as the baseline for comparison.

Algorithm 1. VertebraFormer training and inference pipeline

Input: 3D input volume $I \in \mathbb{R}^{H\times W\times D}$, modality/domain label $d \in \mathcal{D}$
Output: Segmentation mask S, vertebra ID predictions Y^{id}, lesion detections Y^{det}

Training Phase: for $b = 1$ **to** B (mini-batches) **do**
  for $i \in b$ **do**
    // Patch tokenization and transformer encoding
    $P_i \leftarrow \text{Tokenize}(I_i, \text{patch\_size} = P^3)$;
    $E_i \leftarrow P_i \cdot W_{\text{embed}} + \text{PosEnc}(P_i)$;
    $F_i^{(1:L)} \leftarrow \text{TransformerEncode}(E_i)$;
    // Dynamic modulation by domain embedding
    $e_d \leftarrow \text{DomainEmbed}(d_i)$;
    $F_i^{\text{mod}} \leftarrow \text{DynamicModulate}(F_i^{(L)}, e_d)$;
    // Multi-task decoding
    $S_i \leftarrow \text{SegmentationDecoder}(F_i^{\text{mod}}, F_i^{(1:L-1)})$;
    $\text{bbox}_i \leftarrow \text{ExtractBBox}(S_i)$;
    $f_{\text{roi}} \leftarrow \text{RoIAlign}(F_i^{\text{mod}}, \text{bbox}_i)$;
    $Y_i^{id} \leftarrow \text{IdentificationHead}(f_{\text{roi}})$;
    $Y_i^{det} \leftarrow \text{LesionDetectionHead}(F_i^{\text{mod}})$;
    // Loss computation
    $\mathcal{L}_{seg}^{(i)} \leftarrow \text{DiceLoss}(S_i, S_i^{gt}) + \alpha\,\text{FocalLoss}(S_i, S_i^{gt})$;
    $\mathcal{L}_{id}^{(i)} \leftarrow \text{CrossEntropy}(Y_i^{id}, Y_i^{id,gt}) + \lambda_{cons}\mathcal{L}_{consistency}$;
    $\mathcal{L}_{det}^{(i)} \leftarrow \text{FocalLoss}(Y_i^{det}, Y_i^{det,gt}) + \lambda_{reg}\text{L1Loss}$;
    $\mathcal{L}_{domain}^{(i)} \leftarrow \text{ContrastiveLoss}(e_d, \mathcal{E}_d)$;
  end
  // Batch aggregation and optimization
  $\mathcal{L}_{total} \leftarrow \frac{1}{|b|}\sum_{i\in b}[\lambda_{seg}\mathcal{L}_{seg}^{(i)} + \lambda_{id}\mathcal{L}_{id}^{(i)} + \lambda_{det}\mathcal{L}_{det}^{(i)} + \lambda_{domain}\mathcal{L}_{domain}^{(i)}]$;
  $\theta \leftarrow \theta - \eta \cdot \text{ClipGrad}(\nabla_\theta \mathcal{L}_{total}, \text{max\_norm} = 1.0)$;
end
Inference Phase:
$P \leftarrow \text{Tokenize}(I, \text{patch\_size} = P^3)$;
$E \leftarrow P \cdot W_{\text{embed}} + \text{PosEnc}(P)$;
$F^{(1:L)} \leftarrow \text{TransformerEncode}(E)$;
$e_d \leftarrow \text{InferDomain}(I)$;
$F^{\text{mod}} \leftarrow \text{DynamicModulate}(F^{(L)}, e_d)$;
$S \leftarrow \text{SegmentationDecoder}(F^{\text{mod}}, F^{(1:L-1)})$;
$\text{bbox} \leftarrow \text{ExtractBBox}(S)$;
$f_{\text{roi}} \leftarrow \text{RoIAlign}(F^{\text{mod}}, \text{bbox})$;
$Y^{id} \leftarrow \text{IdentificationHead}(f_{\text{roi}})$;
$Y^{det} \leftarrow \text{LesionDetectionHead}(F^{\text{mod}})$;
return S, Y^{id}, Y^{det}

The computational complexity analysis reveals that VertebraFormer scales as $O(N^2d + Nd^2)$ due to the quadratic nature of self-attention operations, where $N = \frac{HWD}{P^3}$ represents the sequence length and d the embedding dimension. For standard input volumes of 128^3 voxels with 8^3 patches, this yields sequences of length $N = 4096$, which remains computationally tractable on modern GPUs with sufficient memory. The multi-task decoding introduces minimal overhead (<5% additional FLOPs) since all heads share the same encoded representations, while dynamic modulation adds negligible complexity (<1% of total operations) despite providing substantial cross-domain performance gains.

Memory optimization strategies include gradient checkpointing for intermediate activations, mixed-precision training using FP16 computations, and adaptive batch sizing based on available GPU memory. The framework supports distributed training across multiple GPUs using data parallelism, enabling efficient scaling to larger datasets and faster convergence.

Table 12 | Cross-modality performance evaluation demonstrates the effectiveness of dynamic modulation for domain adaptation[12,14] without requiring target domain data during training

Model configuration	CT dice (%)	MRI dice (%)	Avg. ID acc (%)
UNETR (shared weights)[19]	88.6	77.2	80.3
Ours w/o modulation	89.1	79.5	82.1
Ours w/ modulation[40]	89.4	82.3	84.2

Our comprehensive experimental validation demonstrates the effectiveness of each architectural component through systematic ablation studies. Table 11 presents the progressive improvement achieved by incrementally adding multi-task heads and domain adaptation mechanisms. The results clearly show that dynamic modulation provides the most significant performance boost across all metrics, with improvements of +1.2% Dice score, +2.5% identification accuracy, and +2.9% lesion detection AP compared to the baseline configuration.

Cross-domain evaluation results in Table 12 validate the effectiveness of our dynamic modulation approach for handling domain shifts. The substantial improvement in MRI performance (from 77.2% to 82.3% Dice score) demonstrates the mechanism's ability to adapt to different imaging characteristics without requiring domain-specific fine-tuning. This capability is crucial for clinical deployment where models must operate reliably across diverse scanner types and imaging protocols.

Additional efficiency analysis using the settings in Tables 7 and 8 confirms that VertebraFormer maintains competitive computational requirements despite its multi-task architecture. While the parameter count is slightly higher than single-task baselines (around 101M vs. 97M for UNETR), the unified framework eliminates the need for multiple specialized models, resulting in overall memory savings and simplified deployment pipelines. The measured throughput of 13.8 volumes per second on an RTX A6000 enables near real-time processing for clinical workflows requiring rapid diagnostic support.

Ethics statement

This study was conducted in accordance with the Declaration of Helsinki and approved by the Institutional Review Board (IRB) of all participating institutions. For publicly available datasets (CTSpine1K, SpineWeb, VerSe 2020), all data had been fully anonymized by the original providers prior to release, and no additional ethical approval was required. For private datasets (Private A and Private B), institutional ethics approval was obtained, and all patient identifiers were removed through standardized de-identification procedures, ensuring compliance with the Health Insurance Portability and Accountability Act (HIPAA) and relevant national privacy regulations. No personally identifiable information (PII) or protected health information (PHI) was used in model development or evaluation. The study involved only retrospective analysis of anonymized medical images and did not require informed consent under local IRB guidelines.

Data availability

The MultiSpine benchmark combines public datasets and institutional cohorts. Public components (CTSpine1K, SpineWeb, VerSe 2020) are available from the original providers at the URLs and DOIs cited in the References. For these datasets, we will release a curation manifest listing the exact case identifiers and the train/validation/test splits used in this study, together with harmonized vertebra masks and labels, upon publication. De-identified derived labels for the private cohorts (Private A and Private B) are available from the corresponding author on reasonable request, subject to institutional data use agreements and ethics approval.The source code used to train VertebraFormer and to reproduce all experiments, including configuration files, preprocessing scripts, and evaluation routines, will be released under an open-source licence in a public repository upon publication of this work. In the meantime, the exact version of the code and trained model weights are available from the corresponding author on reasonable request for academic research.

Code availability

The source code used to train VertebraFormer and to reproduce all experiments, including configuration files, preprocessing scripts, and evaluation routines, will be released under an open-source licence in a public repository upon publication of this work. In the meantime, the exact version of the code and trained model weights are available from the corresponding author on reasonable request for academic research.

Received: 16 October 2025; Accepted: 16 December 2025;
Published online: 10 January 2026

References

1. Qu, B. et al. Current development and prospects of deep learning in spine image analysis: a literature review. *Quant. Imaging Med. Surg.* **12**, 3454 (2022).
2. Simion, G., Eckardt, N., Ullrich, B. W., Senft, C. & Schwarz, F. Bone density of the cervical, thoracic and lumbar spine measured using Hounsfield units of computed tomography—results of 4350 vertebras. *BMC Musculoskelet. Disord.* **25**, 200 (2024).
3. Lessmann, N., Van Ginneken, B., De Jong, P. A. & Išgum, I. Iterative fully convolutional neural networks for automatic vertebra segmentation and identification. *Med. Image Anal.* **53**, 142–155 (2019).
4. Sekuboyina, A. et al. Verse: a vertebrae labelling and segmentation benchmark for multi-detector CT images. *Med. Image Anal.* **73**, 102166 (2021).
5. Masuzawa, N., Kitamura, Y., Nakamura, K., Iizuka, S. & Simo-Serra, E. Automatic segmentation, localization, and identification of vertebrae in 3d CT images using cascaded convolutional neural networks (ed. Martel, A.L). In *International Conference on Medical Image Computing and Computer-Assisted Intervention* 681–690 (Springer, 2020).
6. Cheng, P., Yang, Y., Yu, H. & He, Y. Automatic vertebrae localization and segmentation in CT with a two-stage dense-U-Net. *Sci. Rep.* **11**, 22156 (2021).
7. Isensee, F., Jaeger, P. F., Kohl, S. A., Petersen, J. & Maier-Hein, K. H. nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation. *Nat. Methods* **18**, 203–211 (2021).
8. Deng, Y. et al. Ctspine1k: a large-scale dataset for spinal vertebrae segmentation in computed tomography. arXiv preprint arXiv:2105.14711 (2021).
9. Li, M. et al. Joint lesion detection and classification of breast ultrasound video via a clinical knowledge-aware framework. *IEEE Trans. Circuits Syst. Video Technol.* (2024).
10. Chen, Z. & Lu, S. Caf-yolo: a robust framework for multi-scale lesion detection in biomedical imagery. In *ICASSP 2025–2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 1–5 (IEEE, 2025).
11. Sekuboyina, A., Valentinitsch, A., Kirschke, J. S. & Menze, B. H. A localisation-segmentation approach for multi-label annotation of lumbar vertebrae using deep nets. arXiv preprint arXiv:1703.04347 (2017).
12. Yoon, J. S., Oh, K., Shin, Y., Mazurowski, M. A. & Suk, H.-I. Domain generalization for medical image analysis: a review. In *Proc. IEEE,* Vol. 112, 1583–1609 (2024).
13. Wen, R., Yuan, H., Ni, D., Xiao, W. & Wu, Y. From denoising training to test-time adaptation: enhancing domain generalization for medical image segmentation. In *Proc. of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 464–474 (2024).
14. Liu, Q., Chen, C., Dou, Q. & Heng, P.-A. Single-domain generalization in medical image segmentation via test-time adaptation from shape dictionary. In *Proc. of the AAAI Conference on Artificial Intelligence* (ed. Sycara. K) Vol. 36, 1756–1764 (AAAI Press, 2022).
15. Yang, F. et al. Mmseg: a novel multi-task learning framework for class imbalance and label scarcity in medical image segmentation. *Knowl.-Based Syst.* **309**, 112835 (2025).
16. Vatian, A. et al. Enhancing medical image analysis with multi-task learning using visual transformers. In *International Conference on Computational Science* 195–203 (Springer, 2025).
17. Jiang, J. & Gu, S. Train once, deploy anywhere: edge-guided single-source domain generalization for medical image segmentation. In *Medical Imaging with Deep Learning* 722–741 (PMLR, 2024).
18. Pal, R., Saha, P., Ghoshal, S., Chakrabarti, A. & Sur-Kolay, S. Panoptic segmentation and labelling of lumbar spine vertebrae using modified attention UNET. arXiv preprint arXiv:2404.18291 (2024).
19. Hatamizadeh, A. et al. Unetr: Transformers for 3d medical image segmentation. In *Proc. of the IEEE/CVF Winter Conference on Applications of Computer Vision* (ed. Chairs, P). 574–584 (IEEE, 2022).
20. Tang, Y. et al. Self-supervised pre-training of Swin Transformers for 3d medical image analysis. In *Proc. of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 20730–20740 (2022).
21. Thånell, M. & Melander, P. Vision transformers for segmenting organs and tissues in CT scans of arbitrary imaging ranges. Master's Theses in Mathematical Sciences (2024).
22. He, A. et al. H2former: an efficient hierarchical hybrid transformer for medical image segmentation. *IEEE Trans. Med. Imaging* **42**, 2763–2775 (2023).
23. Li, Z. et al. Scribformer: transformer makes CNN work better for scribble-based medical image segmentation. *IEEE Trans. Med. Imaging* **43**, 2254–2265 (2024).
24. Guo, B., Cao, N., Zhang, R. & Yang, P. Getnet: Group normalization shuffle and enhanced channel self-attention network based on VT-UNET for brain tumor segmentation. *Diagnostics* **14**, 1257 (2024).
25. Wang, Y., Huang, N., Li, T., Yan, Y. & Zhang, X. Medformer: a multi-granularity patching transformer for medical time-series classification. *Adv. Neural Inf. Process. Syst.* **37**, 36314–36341 (2024).
26. Lin, N. Unelrpt: a light network in medical image segmentation. In *2023 China Automation Congress (CAC)*, 2052–2057 (IEEE, 2023).
27. Yin, Y., Luo, S., Zhou, J., Kang, L. & Chen, C. Y.-C. Ldcnet: Lightweight dynamic convolution network for laparoscopic procedures image segmentation. *Neural Netw.* **170**, 441–452 (2024).
28. You, X., He, J., Yang, J. & Gu, Y. Learning with explicit shape priors for medical image segmentation. *IEEE Trans. Med. Imaging* (2024).
29. He, W. et al. Object detection for medical image analysis: Insights from the RT-Detr model. In *Proc. of the 2025 International Conference on Artificial Intelligence and Computational Intelligence*, 415–420 (2025).
30. Aziz, F. & Saputri, D. U. E. Efficient skin lesion detection using Yolov9 network. *J. Med. Inform. Technol.* **2**, 11–15 (2024).
31. Jahin, M. A., Soudeep, S., Mridha, M., Fahad, N. & Hossen, M. J. DyCAF-Net: Dynamic Class-AwareFusion Network. In 2025 *IEEE 12th International Conference on Data Science and Advanced Analytics (DSAA)*, 1–10 (Birmingham, United Kingdom, 2025).

32. Shim, J. H. et al. Automated segmentation and diagnostic measurement for the evaluation of cervical spine injuries using x-rays. *J. Imaging Inform. Med.* **37**, 1863–1873 (2024).
33. Hohenhaus, M. et al. Quantification of cervical spinal stenosis by automated 3d MRI segmentation of spinal cord and cerebrospinal fluid space. *Spinal Cord* **62**, 371–377 (2024).
34. Yu, H. & Dai, Q. Self-supervised multi-task learning for medical image analysis. *Pattern Recognit.* **150**, 110327 (2024).
35. Chen, C. et al. Ma-sam: Modality-agnostic sam adaptation for 3d medical image segmentation. *Med. Image Anal.* **98**, 103310 (2024).
36. Chattopadhyay, S., Demir, B. & Niethammer, M. Zero-shot domain generalization of foundational models for 3d medical image segmentation: an experimental study. arXiv preprint arXiv:2503.22862 (2025).
37. Mahapatra, D. et al. Multi-label generalized zero shot chest Xray classification by combining image-text information with feature disentanglement. *IEEE Trans. Med. Imaging* **44**, 31–43 (2024).
38. Badawi, M., Abushanab, M., Bhat, S. & Maier, A. Review of zero-shot and few-shot AI algorithms in the medical domain. arXiv preprint arXiv:2406.16143 (2024).
39. Khadka, R. et al. Meta-learning with implicit gradients in a few-shot setting for medical image segmentation. *Comput. Biol. Med.* **143**, 105227 (2022).
40. Ali, S., Lee, Y. R., Park, S. Y., Tak, W. Y. & Jung, S. K. Unsupervised domain adaptation by cross-domain consistency learning for CT body composition. *Mach. Vis. Appl.* **36**, 27 (2025).
41. Xu, Y., Khan, T. M., Song, Y. & Meijering, E. Edge deep learning in computer vision and medical diagnostics: a comprehensive survey. *Artif. Intell. Rev.* **58**, 93 (2025).
42. Liu, X. et al. Automated detection of radiolucent foreign body aspiration on chest CT using deep learning. *npj Digit. Med.* **8**, 647 (2025).
43. Hsia, S.-C., Wang, S.-H. & Chang, C.-Y. Convolution neural network with low operation flops and high accuracy for image recognition. *J. Real-Time Image Process.* **18**, 1309–1319 (2021).
44. Zhu, Y. et al. A quantitative evaluation of the deep learning model of segmentation and measurement of cervical spine MRI in healthy adults. *J. Appl. Clin. Med. Phys.* **25**, e14282 (2024).
45. Bae, H.-J. et al. Fully automated 3d segmentation and separation of multiple cervical vertebrae in CT images using a 2d convolutional neural network. *Comput. Methods Programs Biomed.* **184**, 105119 (2020).
46. Wang, W. et al. SpineCLUE: Automatic vertebrae identification using contrastive learning and uncertainty estimation. *Artif. Intell. Med.* **171**, 103285 (2026).
47. Zhang, Y. et al. A clinically applicable AI system for detection and diagnosis of bone metastases using CT scans. *Nat. Commun.* **16**, 4444 (2025).
48. Li, X., Hong, Y., Xu, Y. & Hu, M. Verformer: vertebrae-aware transformer for automatic spine segmentation from CT images. *Diagnostics* **14**, 1859 (2024).
49. Tao, R., Liu, W. & Zheng, G. Spine-transformers: vertebra labeling and segmentation in arbitrary field-of-view spine CTs via 3d transformers. *Med. Image Anal.* **75**, 102258 (2022).
50. You, X. et al. Verteformer: a single-staged transformer network for vertebrae segmentation from CT images with arbitrary field of views. *Med. Phys.* **50**, 6296–6318 (2023).
51. Klein, G., Hardisty, M., Whyne, C. & Martel, A. L. Vertdetect: fully end-to-end 3d vertebral instance segmentation model. arXiv preprint arXiv:2311.09958 (2023).
52. Zhang, S. et al. Spineclue: automatic vertebrae identification using contrastive learning and uncertainty estimation. arXiv preprint arXiv:2401.07271 (2024).
53. Huang, Y. et al. 3d vertebrae labeling in spine CT: an accurate, memory-efficient (ortho2d) framework. *Phys. Med. Biol.* **66**, 125020 (2021).
54. Cui, Z. et al. Vertnet: accurate vertebra localization and identification network from CT images. In *International Conference on Medical Image Computing and Computer-Assisted Intervention* 281–290 (Springer, 2021).
55. Wu, H., Bailey, C., Rasoulinejad, P. & Li, S. Automatic landmark estimation for adolescent idiopathic scoliosis assessment using BoostNet. In *International Conference on Medical Image Computing and Computer-Assisted Intervention* 127–135 (Springer, 2017).
56. Yu, Q.-S. et al. Multi-modal and multi-view cervical spondylosis imaging dataset. *Sci. Data* **12**, 1080 (2025).
57. Shastry, P. et al. AI and deep learning for automated segmentation and quantitative measurement of spinal structures in MRI. arXiv preprint arXiv:2503.11281 (2025).
58. Gómez Mesa, L.M. Paravertebral muscle segmentation for body composition analysis in CT scans (2025).
59. Chen, T. et al. XLSTM-UNET can be an effective 2d & 3d medical image segmentation backbone with Vision-LSTM (ViL) better than its mamba counterpart. *IEEE EMBS International Conference on Biomedical and Health Informatics (BHI)*, 1-8 (2024).
60. Noh, S. & Lee, B.-D. A narrative review of foundation models for medical image segmentation: zero-shot performance evaluation on diverse modalities. *Quant. Imaging. Med. Surg.* **15**, 5825 (2025).
61. Moshiri, P. F., Shahbazian, R., Nabati, M. & Ghorashi, S. A. A CSI-based human activity recognition using deep learning. *Sensors* **21**, 7225 (2021).

Acknowledgements

This research is supported by the National Natural Science Foundation of China (No. 82203472), Natural Science Foundation of Shandong Province (ZR2023MH202, ZR2024QH042, ZR2025MS1404, and ZR2025MS1307), and 2023 Youth Talent Support Project of Shandong Medical Association (2023_LC_0021), the Health and Family Planning Commission Research Project of Jiangsu (ZQ2024002), the Science and Technology Program of Suzhou (SKY2023114, SKY2022093).

Author contributions

J.D., J.H., and Y.B. conceptualized the study, conducted investigations, and supervised the project. J.D., H.G., Y.Z., F.D., Z.C., and Y.Zh. curated the data and developed the methodology. R.Z., Y.Z., H.X., J.Y., Y.Y., and J.H. performed formal analysis and visualization. Z.C., J.D., and Y.B. handled software development and validation. J.D., H.G., R.Z., and Y.Y. wrote the original draft. J.D., S.H., and J.H. reviewed and edited the manuscript. J.D., Y.Y., S.H., J.Y., J.H., and Y.B. acquired funding, administered the project, and provided resources. All authors had access to the study data and approved the final manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Correspondence and requests for materials should be addressed to Yihang Yang, Shaoshan Hu or Jingbiao Huang.