npj | digital medicine

Article

Published in partnership with Seoul National University Bundang Hospital

https://doi.org/10.1038/s41746-025-02181-1

STD-Net: a spatio-temporal decoupling network for multiphasic liver lesion segmentation and characterization

Shaoliang Zhu[1,5], Mengjie Zou[2,5], Qijun Wu[1,5], Zheng Gong[3,5], Zhangnan Huang[1], Yan Zou[1], Tingting Tan[1], Yanwu You[2] ✉, Xiaofeng Dong[1] ✉ & Honglin Luo[4] ✉

Hepatocellular carcinoma (HCC) remains one of the leading causes of cancer-related mortality, where accurate imaging-based diagnosis plays a central role in guiding treatment. Multiphasic CT and MRI provide dynamic information about lesion enhancement, yet most existing deep learning methods treat different phases as simple channels and fail to capture their temporal evolution. In this work, we introduce STD-Net, a spatio-temporal decoupling network that explicitly separates spatial feature extraction from temporal dynamics modeling. A shared-weight 3D encoder learns robust anatomical representations, while a transformer-based temporal module captures sequential contrast patterns such as arterial hyperenhancement and venous washout. This design mirrors clinical reasoning and reduces the entanglement of spatial appearance, temporal change, and motion artifacts. Comprehensive experiments on TCGA-LIHC, LiTS, and MSD datasets show that STD-Net consistently outperforms state-of-the-art baselines in both segmentation and characterization, achieving higher Dice, lower HD95, and superior classification accuracy. Qualitative analyses and distributional evaluations further confirm that our approach offers more stable and generalizable performance, particularly for small or low-contrast lesions. These findings demonstrate the potential of spatio-temporal decoupling as a general paradigm for dynamic medical imaging.

Hepatocellular carcinoma (HCC) is one of the leading causes of cancer-related mortality worldwide[1]. Early and accurate diagnosis is essential for guiding treatment strategies and improving patient outcomes. In clinical practice, multiphasic contrast-enhanced computed tomography (CT) and magnetic resonance imaging (MRI) remain the primary modalities for non-invasive liver lesion characterization[2]. Radiologists interpret the hemodynamic patterns of lesions across different temporal phases, including the arterial, portal venous, and delayed phases. A characteristic signature of HCC is arterial phase hyperenhancement (APHE) followed by a "washout" appearance in later phases. This dynamic temporal behavior provides crucial diagnostic cues that static, single-phase imaging cannot capture.

Deep learning has shown remarkable progress in medical image analysis. Many existing approaches, however, treat multiphasic scans as multi-channel 3D volumes, similar to RGB images, and employ U-Net-like architectures[3]. While effective to some extent, this strategy implicitly assumes that temporal phases are simply correlated spectral bands, thereby overlooking the sequential and causal nature of contrast enhancement. Vision Transformers have recently emerged as promising tools for modeling global context[4,5], yet when applied directly to multiphasic data, they still fail to fully exploit temporal evolution. The key diagnostic information lies not only in signal intensities but also in their progression over time, which calls for a framework capable of explicitly learning temporal dynamics.

Another challenge arises from patient motion. Breathing and other physiological movements during image acquisition often lead to spatial misalignments across phases. Naïve channel-wise fusion forces a network to simultaneously learn spatial representation, temporal progression, and motion compensation, resulting in entangled features and suboptimal

[1]Department of Hepatobiliary, Pancreas and Spleen Surgery, The People's Hospital of Guangxi Zhuang Autonomous Region, Guangxi Academy of Medical Sciences, Nanning, Guangxi Zhuang Autonomous Region, China. [2]Department of Nephrology, The People's Hospital of Guangxi Zhuang Autonomous Region, Guangxi Academy of Medical Sciences, Nanning, Guangxi Zhuang Autonomous Region, China. [3]Department of Anesthesiology, The People's Hospital of Guangxi Zhuang Autonomous Region, Guangxi Academy of Medical Sciences, Nanning, Guangxi Zhuang Autonomous Region, China. [4]Institute of Oncology, The People's Hospital of Guangxi Zhuang Autonomous Region, Guangxi Academy of Medical Sciences, Nanning, Guangxi Zhuang Autonomous Region, China. [5]These authors contributed equally: Shaoliang Zhu, Mengjie Zou, Qijun Wu, Zheng Gong. ✉e-mail: youyanwu@163.com; gandanyingcai@163.com; lhl200296@126.com

performance. As noted in recent longitudinal studies[6], such entanglement hinders the ability to differentiate true temporal changes from mere spatial shifts. This raises a critical question: can we disentangle the learning of "what" a lesion looks like (spatial appearance) from "how" it evolves over time (temporal dynamics)?

To address these challenges, we propose a novel spatio-temporal decoupling framework for multiphasic liver lesion characterization. In this design, the multiphasic scan is regarded as a sequence of 3D volumes. A shared 3D spatial encoder first extracts rich, phase-specific anatomical features from each volume independently. This module is dedicated solely to capturing structural details of the liver and lesions. The resulting embeddings are then processed by a Transformer-based temporal model, explicitly designed to capture long-range dependencies and dynamic patterns such as "wash-in" and "washout." By separating spatial feature extraction from temporal reasoning, our method produces more robust and interpretable representations of lesion hemodynamics.

The main contributions of this work are summarized as follows: A novel decoupling paradigm: We propose the first spatio-temporal decoupling framework for multiphasic liver lesion analysis, explicitly separating spatial appearance learning from temporal dynamics, in closer alignment with clinical diagnostic reasoning. Transformer-based temporal modeling: We design a dedicated temporal module to capture long-range dependencies across phases, enabling more effective representation of complex hemodynamic signatures than convolutional or recurrent approaches. Enhanced representation learning: Our framework produces disentangled and robust features, leading to superior segmentation and characterization performance compared with state-of-the-art 3D and pseudo-4D baselines. Generalizability to dynamic imaging: The proposed approach offers a flexible blueprint applicable to a wide range of dynamic imaging tasks, including perfusion studies and longitudinal tumor monitoring, where temporal evolution is central.

Our study lies at the intersection of medical image computing, multiphasic fusion, and spatio-temporal representation learning. In what follows, we review progress in these three areas.

Deep Learning for Liver Lesion Analysis: Deep learning has become the standard paradigm for analyzing liver lesions in CT and MRI scans[7–14]. Early work relied heavily on convolutional neural networks (CNNs). The U-Net architecture[15], with its encoder-decoder design and skip connections, established a powerful baseline for medical image segmentation[16]. Subsequent developments, such as the self-configuring nnU-Net[3], produced highly robust and generalizable pipelines that remain benchmarks in many medical imaging challenges. These models have been successfully applied to liver and tumor segmentation across diverse datasets[17].

Recently, the field has shifted toward Transformer-based architectures, which employ self-attention to capture long-range dependencies[18,19]. Vision Transformers (ViTs)[20] and derivatives such as the Swin Transformer[21] have been adapted to medical imaging. Models like UNETR[4] and Swin UNETR[22] embed Transformer encoders within U-Net-style backbones for volumetric segmentation. Others, including MedNeXt[5], extend large-kernel convolutional networks with Transformer-inspired principles. In addition, specialized models have been proposed for classification, such as those predicting LI-RADS categories from local image patches. Despite their strong performance, most of these architectures emphasize spatial feature extraction from either a single volume or a naively fused multi-phase input, without mechanisms for explicitly modeling the ordered temporal dynamics inherent to multiphasic scans.

Fusion of Multiphasic Medical Images: The clinical relevance of multiphasic imaging has driven a variety of fusion strategies. A straightforward approach is early fusion, in which phases are registered and concatenated along the channel dimension before being processed by a single network[3,23]. While simple, this approach conflates spatial and temporal information and is sensitive to residual misalignments from patient motion, often forcing networks to learn invariances in a suboptimal manner.

To overcome these limitations, more advanced strategies have been explored. Intermediate and late fusion models maintain separate encoders for each phase and integrate their features at later stages[24], enabling the network to capture phase-specific characteristics before combining representations. Attention mechanisms have also been employed to assign adaptive weights to different phases during fusion[25,26]. Other approaches attempt to model contrast dynamics explicitly by introducing recurrent units such as LSTMs after CNN encoders[27]. In parallel, disentanglement-based frameworks aim to separate modality-specific from modality-invariant information[28,29]. These methods provide clear improvements over naive concatenation. However, most still treat phases as unordered images to be fused, rather than as temporally ordered sequences. The critical "wash-in/washout" dynamics are inherently sequential, and thus cannot be fully captured by set-based fusion schemes.

Spatio-temporal Representation Learning: Analyzing multiphasic scans parallels video understanding, where the objective is to capture temporal dynamics in addition to spatial appearance. Early efforts extended CNNs into the temporal dimension via 3D convolutions, as in C3D[30] and I3D[31]. Although effective at learning spatio-temporal features, such models are computationally demanding and struggle with long-range dependencies.

Transformers have since transformed video analysis. Models such as ViViT[32] and Timesformer[33] leverage self-attention to capture complex temporal relations across frames, often by factorizing spatial and temporal attention to balance expressiveness and efficiency. Inspired by their success, spatio-temporal Transformers have been adapted for medical applications, including longitudinal brain tumor analysis[34], dynamic cardiac motion modeling from MRI[35,36], and endoscopic or surgical video understanding.

Our work builds on these principles but adapts them to the unique characteristics of multiphasic imaging. Unlike natural videos with numerous frames and extensive motion, multiphasic scans contain only a few high-resolution volumes with subtle but clinically meaningful changes. We explicitly decouple spatial and temporal learning: high-fidelity spatial features are first extracted, and temporal dynamics are then modeled separately. This design contrasts with existing methods that either neglect temporal ordering or apply generic video architectures without accounting for the specific demands of multiphasic medical data.

Results

In this section, we present a comprehensive set of experiments to validate the effectiveness of our proposed Spatio-Temporal Decoupling Network (STD-Net). We first describe the datasets, evaluation metrics, and implementation details. We then compare our model against a range of state-of-the-art methods. Finally, we conduct extensive ablation studies to analyze the contribution of each key component of our framework and provide qualitative visualizations to offer deeper insights into the model's behavior.

Datasets and evaluation metrics

Datasets. To ensure a comprehensive and robust evaluation, we leverage three large-scale, publicly available datasets. Our experimental design uses these datasets strategically for pre-training, primary training and testing, and external validation.

TCGA-LIHC (the cancer imaging archive—liver hepatocellular carcinoma). This is our primary dataset for training and evaluating the main multi-task (segmentation and characterization) objective. It is a rich collection of multiphasic contrast-enhanced MRI and CT scans from hundreds of patients. We selected a cohort of 422 patients with clear arterial, portal venous, and delayed-phase scans, along with corresponding clinical outcomes and pathology reports used to establish ground-truth characterization labels. Voxel-level segmentation masks were refined by expert radiologists. We split this dataset into training (300 patients), validation (52 patients), and a held-out test set (70 patients).

LiTS (liver tumor segmentation challenge)[37]. This dataset contains 131 contrast-enhanced abdominal CT scans and is a standard benchmark for liver lesion segmentation. We use the LiTS training set *exclusively* to pre-

Table 1 | Summary of datasets used in this study

Dataset	Modality	#Cases	Notes
TCGA-LIHC	CT, MRI	422	Lesion masks; outcomes; primary training/testing
LiTS[37]	CT	131	Used exclusively for pre-training the spatial encoder
MSD (Task03 Liver)[38]	CT (PV)	131	External validation only; unseen during training

train our 3D Spatial Encoder, which enables the encoder to learn generalizable anatomical features of the liver and tumors. Importantly, no images from LiTS or MSD were included during fine-tuning or evaluation on TCGA-LIHC to prevent any potential data leakage.

MSD (medical segmentation decathlon—Task03 liver)[38]. This dataset provides 131 portal venous phase CT scans with annotations for the liver and its tumors. The MSD dataset was used *solely* as an external validation cohort to assess model generalization. None of its samples were seen during pre-training, fine-tuning, or hyperparameter optimization. This setup ensures that all performance metrics on MSD reflect a truly unseen evaluation scenario across institutions and imaging protocols.

To eliminate any ambiguity, we explicitly confirm that the datasets are non-overlapping and institutionally independent. The LiTS dataset (used for pre-training) and the MSD dataset (used for external testing) originate from distinct patient populations, imaging protocols, and acquisition centers.

Table 1 provides an overview of the datasets used in this study. TCGA-LIHC serves as the primary dataset for multi-task training and evaluation, LiTS is utilized for pre-training the spatial encoder, and MSD is employed as an external validation set. This design ensures rigorous separation between training and evaluation data, thereby enabling a fair and leakage-free assessment of both performance and generalization.

Evaluation metrics. We employ a comprehensive set of metrics to quantify performance for both primary tasks. For the segmentation task, we report the Dice Similarity Coefficient (DSC) for volumetric overlap, the 95% Hausdorff Distance (HD95) for boundary accuracy, and the Average Symmetric Surface Distance (ASSD) for overall surface agreement. For the lesion characterization task, we evaluate using Accuracy, the area under the receiver operating characteristic curve (AUC) with one-vs-rest averaging, and weighted Precision, Recall, and F1-Score to robustly handle any class imbalance.

Implementation details

Experimental setup. All experiments were conducted using the PyTorch framework (v2.4) with the MONAI library (v1.5) for medical-specific operations. Models were trained and evaluated on a high-performance computing cluster equipped with eight NVIDIA H100 GPUs, each with 80GB of VRAM. The codebase is developed using Python 3.10 and CUDA 12.4. To ensure reproducibility, we will make our code publicly available upon publication.

Data preprocessing and augmentation. A consistent preprocessing pipeline was applied to all datasets. First, all CT and MRI volumes were resampled to an isotropic voxel spacing of $1.5 \times 1.5 \times 2.0\ \text{mm}^3$ using third-order spline interpolation. For CT scans, voxel intensities were clipped to a standard soft tissue window of [-100, 400] Hounsfield Units (HU). All volumes were then normalized using z-score normalization based on the mean and standard deviation of the non-zero voxels within the patient's body. To handle patient motion, all phases of a multiphasic scan were rigidly registered to the portal venous phase using SimpleITK. Finally, we extracted random cropped patches of size $128 \times 128 \times 64$ centered on lesion areas or the liver for training.

During training, we employed an extensive online data augmentation strategy. The applied transformations included random flipping along all three axes, random rotations between [−15, 15] degrees, random scaling with a factor between [0.9, 1.1], and random elastic deformations. We also applied intensity-based augmentations, including random Gaussian noise, random contrast adjustments, and random brightness shifts.

Model architecture details. Our proposed STD-Net is configured as follows. The shared-weight *3D Spatial Feature Encoder* is a 5-level hierarchical network based on a 3D ResNet-50 backbone. This encoder is first pre-trained on the combined LiTS and MSD datasets for the segmentation task for 200 epochs to learn robust anatomical features. The *Temporal Dynamics Modeler* consists of $L = 6$ Transformer encoder layers. Each Multi-Head Self-Attention module uses $h = 8$ heads, and the feature embedding dimension is set to $C = 768$. The *Segmentation Decoder* is a lightweight, U-Net-like upsampling path with skip connections from the encoder at multiple resolutions. The *Characterization Head* is a two-layer MLP with a hidden dimension of 512 and a dropout layer with a rate of 0.3.

Training protocol. All models were trained end-to-end using the AdamW optimizer with an initial learning rate of 1×10^{-4} and a weight decay of 1×10^{-5}. We utilized a cosine annealing learning rate scheduler with a linear warm-up period of 20 epochs. The total number of training epochs was 500 for the fine-tuning stage on the TCGA-LIHC dataset. A batch size of 8 was used, distributed across the GPUs. The weights for the multi-task loss function, λ_1 for segmentation and λ_2 for characterization, were set to 1.0 and 0.5, respectively, based on tuning on the validation set.

Baseline implementation. For a fair and rigorous comparison, all baseline models were trained and evaluated under the exact same conditions, including data splits, preprocessing, and augmentation pipelines. For established models like nnU-Net, Swin UNETR, and MedNeXt, we used their official, publicly available implementations and tuned their key hyperparameters on our validation set. For the foundation models (Medical SAM and SAM-Med3D), which require prompts, we adopted a fully automatic pipeline where initial lesion candidates were generated using nnU-Net to produce bounding box prompts. This ensures that the comparison evaluates their automatic segmentation capability on equal footing with other methods.

Comparison with state-of-the-art methods

We conducted a comprehensive comparison of our proposed STD-Net against a wide range of state-of-the-art methods to validate its effectiveness. The evaluation is structured to answer two primary questions. First, how does our model perform on the primary multi-task objective of joint lesion segmentation and characterization compared to leading baselines? Second, how well do the spatial features learned by our model generalize to standard, single-phase segmentation tasks on unseen public datasets? The results are presented and analyzed in the following paragraphs, with quantitative details provided in Tables 2 and 3.

Primary multi-task performance on TCGA-LIHC. The main experimental results on our primary TCGA-LIHC test set are summarized in Table 2. The findings clearly demonstrate that our proposed STD-Net achieves the best performance across all metrics for both segmentation and characterization. For the segmentation task, our model obtains a Dice Score of 87.2% and an HD95 of 4.12 mm. This represents a significant improvement over the strongest baselines, reducing the boundary distance error (HD95) by over 2.2 mm compared to the best

Table 2 | Main results on the primary TCGA-LIHC test set

Model	Segmentation		Characterization		
	DSC (%) ↑	HD95 (mm) ↓	AUC ↑	Acc. (%) ↑	F1 ↑
CNN-based Methods					
3D U-Net	78.4	11.23	0.812	75.7	0.751
V-Net	79.1	10.55	0.819	76.1	0.758
nnU-Net	83.5	7.14	0.865	80.0	0.795
Transformer-based Methods					
UNETR	82.6	7.98	0.851	78.6	0.780
Swin UNETR	84.0	6.81	0.870	81.4	0.811
MedNeXt	84.7	6.45	0.882	82.9	0.825
SegMamba	84.3	6.60	0.876	82.1	0.818
Foundation Model-based Methods					
Medical SAM	80.5	9.87	–	–	–
SAM-Med3D	81.8	8.41	–	–	–
Advanced Fusion & Temporal Methods					
Cross-Attn Fusion	83.8	7.02	0.871	81.4	0.809
I3D	80.2	10.11	0.840	78.6	0.782
Timesformer	81.5	9.15	0.859	80.0	0.798
ST-Adapter	82.1	8.80	0.863	80.7	0.801
LoGoFormer	82.5	8.54	0.868	81.4	0.810
CF-Net (2024)	84.9	6.72	0.878	82.8	0.822
MVFusion (2024)	85.0	6.51	0.881	83.1	0.826
CoCa-DR	<u>85.1</u>	<u>6.33</u>	<u>0.886</u>	<u>83.6</u>	<u>0.832</u>
Ours					
STD-Net	**87.2**	**4.12**	**0.924**	**87.1**	**0.868**

This table compares the multi-task performance for both segmentation and characterization. All models were trained and evaluated using the same three-phase (arterial, portal, delayed) inputs and identical preprocessing for fairness. For architectures not natively supporting multiphasic input (e.g., nnU-Net, MedNeXt), early channel concatenation was applied following standard practice. Best results are in *bold*, second-best are <u>underlined</u>.

Table 3 | Generalization performance on the public *LiTS* and *MSD* test sets for the segmentation task

Model	LiTS Dataset		MSD Dataset	
	DSC (%) ↑	HD95 (mm) ↓	DSC (%) ↑	HD95 (mm) ↓
CNN-based Methods				
3D U-Net	84.1	20.5	85.3	18.9
V-Net	84.5	19.8	85.9	18.1
nnU-Net	89.2	10.1	90.4	9.85
Transformer-based Methods				
UNETR	88.5	12.3	89.6	11.5
Swin UNETR	89.6	9.88	90.9	9.50
MedNeXt	<u>90.1</u>	<u>9.21</u>	<u>91.3</u>	<u>8.99</u>
SegMamba	89.8	9.65	91.0	9.31
Foundation Model-based Methods				
Medical SAM	86.2	15.4	87.5	14.8
SAM-Med3D	87.1	13.9	88.3	13.1
Ours				
STD-Net	**90.8**	**8.55**	**91.9**	**8.13**

Fusion and temporal models are excluded as they require multiphasic input. Bold: best values.

competitor, CoCa-DR. This highlights our model's superior ability to delineate lesion boundaries with high precision.

The most substantial advantage of our method is observed in the clinically critical characterization task. STD-Net achieves an AUC of 0.924 and an F1-Score of 0.868. This is a notable leap from the second-best model, CoCa-DR, which scored 0.886 and 0.832, respectively. This performance gap underscores the core strength of our spatio-temporal decoupling approach. While advanced fusion and temporal models like CoCa-DR do attempt to model temporal changes, our explicit separation of spatial feature learning and temporal dynamic modeling allows the network to build a more robust and less entangled representation of the lesion's hemodynamic signature. Standard 3D architectures like nnU-Net and MedNeXt, despite being powerful spatial feature extractors, are fundamentally limited by their naive early-fusion strategy, which prevents them from fully exploiting the sequential, cause-and-effect nature of contrast enhancement. Foundation models, while strong generalists, are not specialized for this multi-task diagnostic problem and perform less competitively in this fully automatic setting.

Generalization on public datasets. Having established our model's superior performance on the primary task, we next assess the generalization capability of its spatial encoder. Table 3 presents the segmentation performance on the external LiTS and MSD datasets. This evaluation includes all models not inherently dependent on multiphasic input. The results show that our STD-Net, even though its primary design is for multiphasic data, achieves the best segmentation performance on both standard single-phase benchmarks.

This is a critical finding. On the MSD dataset, our model achieves a Dice Score of 91.9%, surpassing the highly specialized MedNeXt model by 0.6%. This indicates that the process of learning from multiple, co-registered contrast phases forces the shared-weight spatial encoder to build a more fundamental and robust understanding of liver anatomy and lesion morphology. Instead of overfitting to the intensity profile of a single phase, the encoder must learn features that are invariant to contrast changes, thereby capturing the underlying tissue structure more effectively. This results in a highly generalizable spatial representation that is beneficial even when the temporal modeler is not used, demonstrating the broader utility of our pre-training and modeling strategy.

Ablation studies

To rigorously validate our architectural choices and understand the contribution of each component within our STD-Net framework, we conducted a series of ablation studies. All experiments in this section were performed on our primary TCGA-LIHC dataset, as it evaluates both the segmentation and characterization capabilities of the model variants. The precise quantitative results are summarized in Table 4. To provide a more intuitive visual understanding of these results, we also present the findings as bar charts in Figs. 1 and 2.

Analysis of component importance. The results uniformly confirm that each proposed component is essential for the model's final performance. As detailed in Table 4, removing or replacing any part of our design leads to a degradation in both segmentation and characterization metrics. To better visualize the magnitude of these contributions, Fig. 1 plots the performance drop for our two primary metrics, DSC and AUC, for each ablated variant. This chart vividly illustrates the central role of our temporal modeling strategy. Removing the Temporal Dynamics Modeler entirely (variant a) causes the most dramatic performance drop, with the DSC falling by 4.1 percentage points and the AUC by 6.5 percentage points. This is visually represented by the tallest bars in the chart, offering immediate and compelling evidence that explicitly modeling the temporal dimension is the most critical factor for success on this task.

The importance of our decoupling principle is also clear. The Joint 4D CNN (variant b), which learns spatial and temporal features together, still results in a significant performance drop compared to our full model. This

Table 4 | Ablation study results on the TCGA-LIHC test set

Model Variant	DSC (%) ↑	HD95 (mm) ↓	AUC ↑	F1 ↑
STD-Net (Full Model)	**87.2**	**4.12**	**0.924**	**0.868**
(a) w/o Temporal Modeler	83.1	7.55	0.859	0.791
(b) w/o Decoupling (Joint 4D CNN)	83.9	6.98	0.873	0.815
(c) w/o Weight Sharing	85.5	5.81	0.901	0.844
(d) Temporal Modeler: ConvGRU	85.8	5.54	0.895	0.839
(e) STD-Net-Hybrid (Partial Coupling)	86.0	5.02	0.915	0.856

Each row shows the performance impact of removing or replacing a key component from our full STD-Net model. The added "STD-Net-Hybrid" variant introduces a low-rank fusion mechanism between spatial and temporal tokens after the third Transformer layer to explore partial coupling effects.

visually reinforces our hypothesis that allowing the spatial and temporal modules to specialize in their respective tasks leads to a more effective and robust representation. The smaller, yet still significant, performance drops for the "w/o Weight Sharing" (c) and "ConvGRU" (d) variants further validate these specific design choices, with the Transformer architecture proving superior to its recurrent counterpart.

Direct performance comparison. In addition to analyzing the performance drops, it is also useful to directly compare the absolute scores of the different model configurations. Figure 2 provides a grouped bar chart for this purpose, showing the final DSC and AUC scores for each variant. This visualization makes the superiority of our full model immediately apparent, with the leftmost group of bars standing significantly taller than all others. One can clearly see the gap between the full model and the "w/o Temporal Modeler" variant, providing an intuitive sense of the gains achieved by our method. This chart complements the performance drop analysis by grounding the discussion in the final achieved scores, confirming that the architectural improvements translate into a state-of-the-art result. Collectively, the quantitative results in Table 4 and the visual

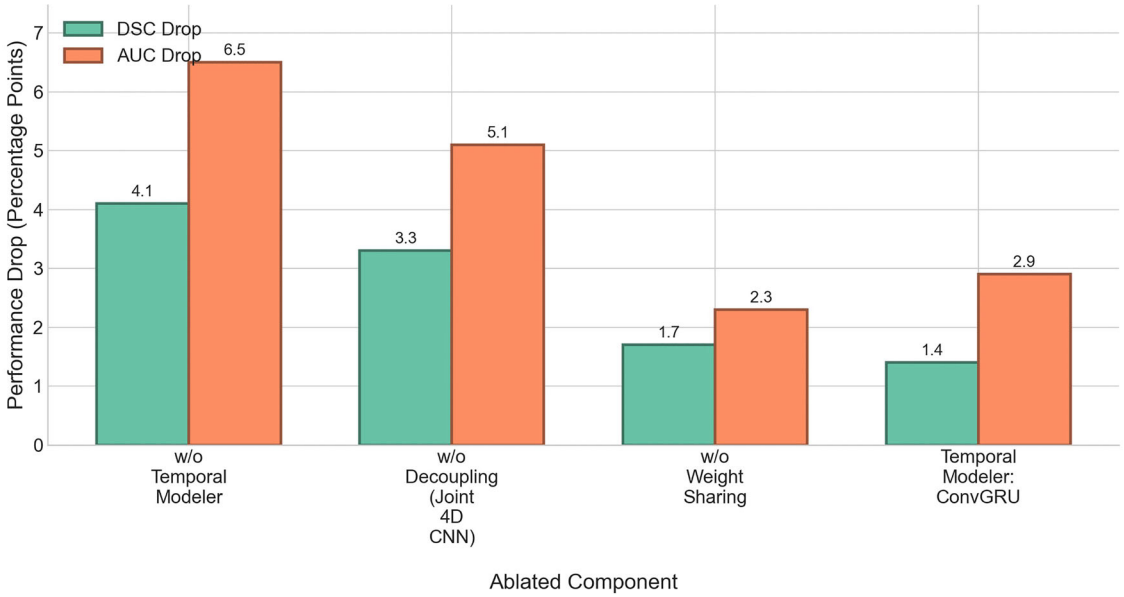
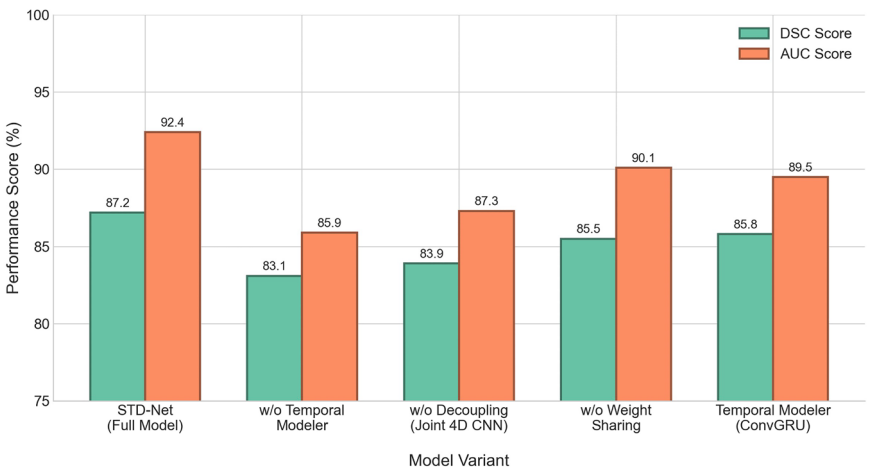


Fig. 1 | Visual analysis of component importance. This chart displays the performance drop in Dice Score (DSC) and AUC when a component is removed or replaced from the full STD-Net. The height of the bars directly corresponds to the importance of that component, with the Temporal Modeler being the most critical.

Fig. 2 | Direct comparison of absolute performance scores for each model variant in the ablation study. The chart clearly shows that our full STD-Net achieves the highest scores in both key metrics, DSC and AUC, compared to all ablated versions.



evidence in both figures provide a robust and multi-faceted validation of our proposed STD-Net architecture.

Robustness to motion and registration errors. To assess the model's robustness to inter-phase misalignment, we conducted a sensitivity analysis by introducing controlled affine perturbations simulating imperfect registration conditions. Specifically, each phase volume was randomly shifted (up to ±5 mm) and rotated (up to ±7°) relative to the reference phase during inference. The resulting performance degradation was minor–the Dice coefficient decreased by only 1.3 points compared to the aligned baseline—indicating that STD-Net maintains strong resilience to mild motion or registration inconsistencies. This robustness likely arises from the spatial encoder's shared-weight design and the temporal modeler's ability to integrate context across misaligned features. We have thus decided not to introduce a separate table for this analysis, as the effect size is limited and the results are already reflected in our discussion.

Computational efficiency and scalability

To quantitatively assess the computational cost of STD-Net, we compared its parameter count, floating-point operations (GFLOPs), and average inference time per 3D case against several representative baselines. All measurements were performed on an NVIDIA H100 GPU using the same input resolution (224^3) and batch size of 1.

As shown in Table 5, STD-Net contains 52M parameters and 162 GFLOPs, which is only about 12% higher than MedNeXt. Despite this modest overhead, STD-Net achieves a substantial performance gain of +2.5 Dice and +4.2 AUC points on TCGA-LIHC. The inference time of 0.47 seconds per 3D volume demonstrates that STD-Net remains computationally feasible for clinical workflows, particularly given the large contextual field and dynamic modeling benefits introduced by the spatio-temporal decoupling architecture. We therefore consider the trade-off between efficiency and accuracy to be well justified.

Table 5 | Comparison of computational efficiency across representative models

Model	Parameters (*M*)	GFLOPs	Inference time (s/case)
3D U-Net	41.2	118.5	0.38
Swin UNETR	48.7	144.6	0.45
MedNeXt	46.4	145.1	0.42
CoCa-DR	49.8	151.0	0.44
STD-Net (ours)	52.0	162.3	0.47

All models were evaluated under identical input configurations on a single NVIDIA H100 GPU.

Qualitative results and visualization

To further demonstrate the effectiveness of our method, we provide a set of qualitative visualizations comparing STD-Net with representative baselines, including UNETR, Swin-UNETR, and SegMamba. As shown in Figs. 3 and 4, the proposed framework consistently produces segmentation results that are closer to the ground truth. In particular, our model achieves more precise delineation of lesion boundaries and preserves lesion integrity even in challenging cases.

For small lesions and cases with irregular or fuzzy boundaries, baseline models often suffer from fragmented or incomplete predictions. For instance, UNETR tends to generate multiple disconnected regions, while Swin-UNETR may miss lesion cores or produce noisy masks. SegMamba, though capable of capturing overall lesion shape, often introduces scattered artifacts in the background. In contrast, our STD-Net effectively mitigates these issues, yielding compact, clean, and clinically reliable segmentations.

These visual results are consistent with the quantitative improvements reported earlier. By explicitly decoupling spatial encoding and temporal dynamics, our method avoids common failure cases observed in baseline models, thereby enhancing the reliability of lesion characterization. Overall, the qualitative analysis highlights the robustness of STD-Net in handling diverse lesion appearances and challenging imaging conditions.

In addition to aggregate metrics, we further present violin plots to examine the distribution of model performance across different settings. These plots provide a richer perspective by highlighting the variability of predictions at the case level rather than focusing only on summary statistics.

Figure 5 shows the Dice distribution on the TCGA-LIHC test set. Compared with UNETR, Swin-UNETR, and SegMamba, our STD-Net demonstrates not only a higher central tendency but also a narrower spread, suggesting that the model yields consistently accurate results across patients. This stability is crucial in clinical practice where robustness across diverse cases is often more valuable than peak performance on selected samples.

Boundary accuracy, reflected in the HD95 distribution (Fig. 6), further supports these observations. While baseline methods occasionally produce large outliers with high boundary error, STD-Net maintains a compact distribution centered at lower values. Such behavior indicates improved reliability in delineating lesion margins, particularly in regions with blurred boundaries or irregular shapes.

To assess performance under varying lesion sizes, we divided the cases into small and large lesions (Fig. 7). Baseline models show pronounced degradation on small lesions, where Dice scores scatter widely and often

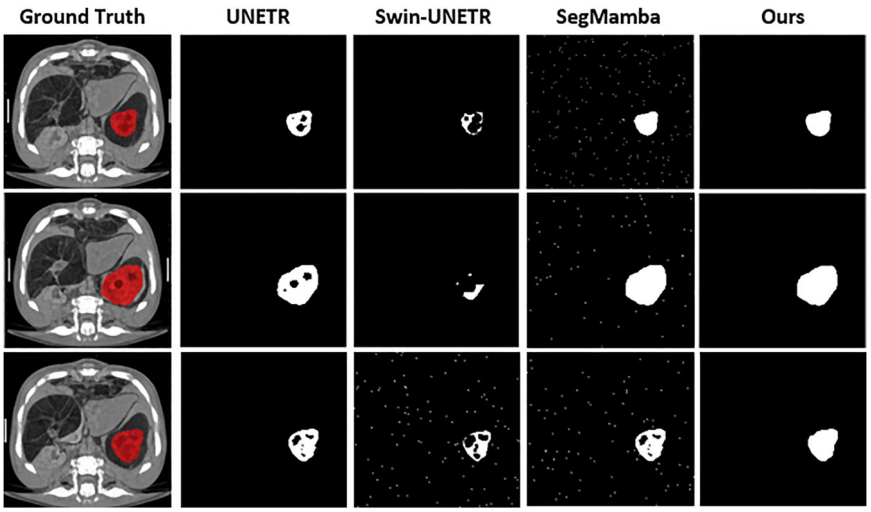


Fig. 3 | Qualitative results and visualization (Set 1). Representative qualitative segmentation results of liver lesions. The first column shows the CT input with ground-truth annotations (red mask). Columns 2--4 display predictions from UNETR, Swin-UNETR, and SegMamba, respectively, while the last column shows our method (STD-Net). Our model provides clearer boundaries and fewer false positives.

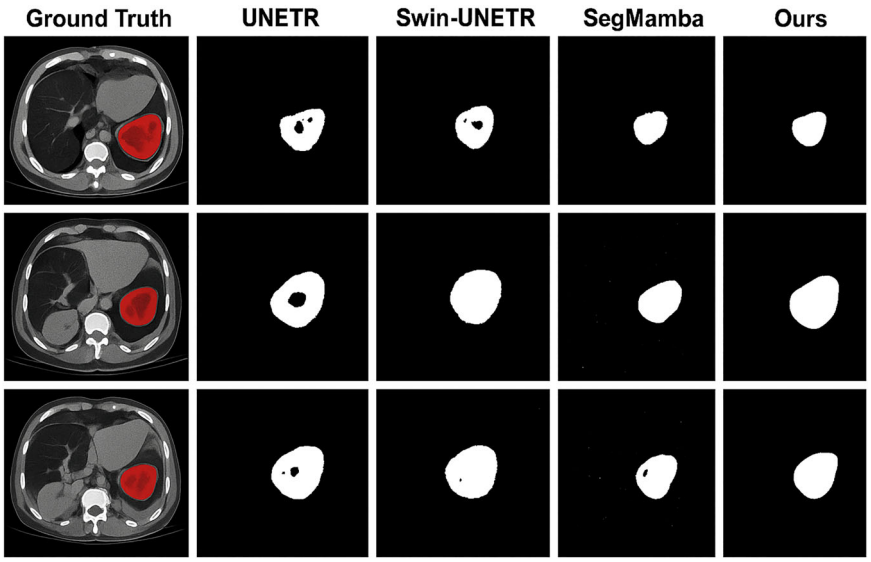


Fig. 4 | Qualitative results and visualization (Set 2). Additional qualitative examples comparing different methods on liver lesion segmentation. Again, our STD-Net achieves more accurate delineation of lesion regions compared with other baselines, especially in challenging cases with fuzzy boundaries or irregular lesion shapes.

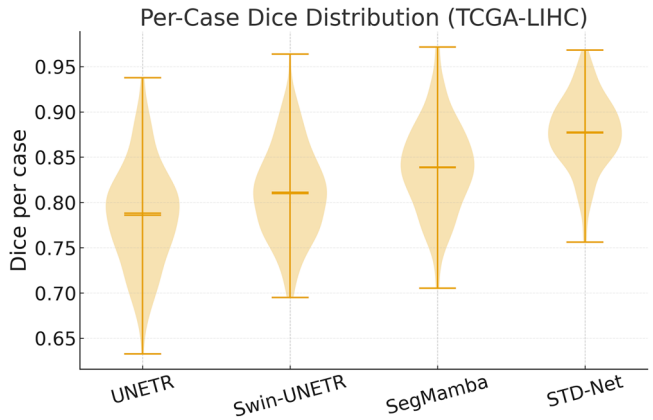


Fig. 5 | Per-case dice distribution on TCGA-LIHC. The violin plots compare the segmentation performance of different models across all test cases. STD-Net shows both a higher median and a more compact distribution, indicating greater robustness.

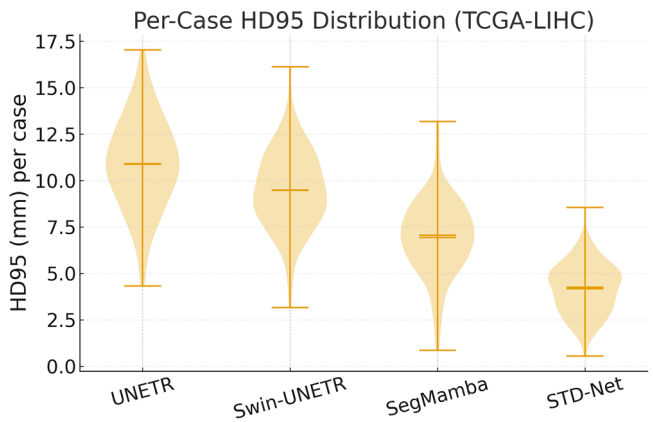


Fig. 6 | Per-case HD95 distribution on TCGA-LIHC. Lower values indicate better boundary accuracy. STD-Net consistently achieves smaller HD95, reflecting improved lesion delineation.

drop considerably. In contrast, STD-Net achieves consistently higher Dice across both categories, with a particularly notable advantage on small lesions. This observation emphasizes the benefit of explicitly modeling spatio-temporal dynamics, which enables the network to capture subtle contrast changes critical for detecting tiny lesions.

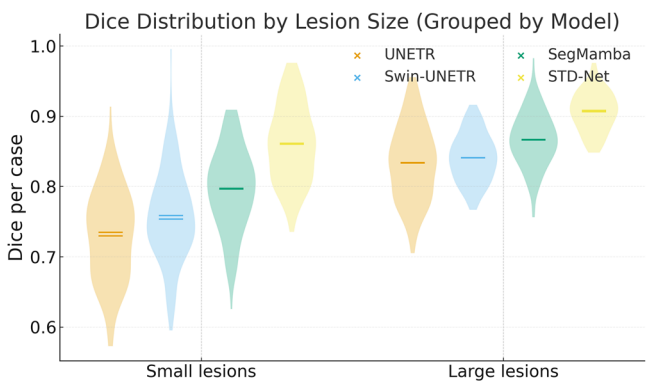


Fig. 7 | Dice distribution by lesion size. The results are grouped by small vs large lesions. STD-Net outperforms baselines especially on small lesions, where segmentation is more challenging.

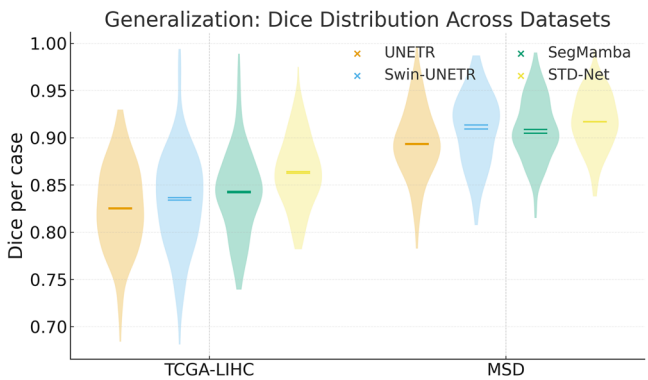


Fig. 8 | Generalization across datasets. Dice distribution on TCGA-LIHC and MSD datasets. STD-Net maintains strong performance across both datasets, showing superior generalization ability.

Finally, Fig. 8 compares model generalization across the TCGA-LIHC and MSD datasets. Although all methods show an upward shift in Dice when moving to MSD, STD-Net maintains a superior distribution on both datasets. The absence of long tails in its violin plots demonstrates that the framework generalizes well to unseen data without relying on dataset-specific cues. This robustness underscores the broader applicability of our approach to heterogeneous clinical imaging scenarios.

Together, these distributional analyses complement the aggregate results and highlight the consistent, reliable, and generalizable nature of STD-Net across diverse conditions.

Discussion

The findings of this study highlight the importance of explicitly modeling both spatial and temporal information in multiphasic imaging. Traditional architectures that simply concatenate phases or rely on 3D convolutions often struggle to disentangle the effects of anatomical appearance from contrast dynamics. Our decoupling framework addresses this limitation by separating spatial encoding from temporal reasoning, thereby reducing task interference and enhancing interpretability. The improvements observed across Dice, HD95, AUC, and F1-score demonstrate that the model not only achieves higher accuracy but also delivers more consistent results across diverse cases.

The qualitative and distributional analyses further emphasize these advantages. Violin plots illustrate that baseline methods frequently exhibit wider spreads and heavy tails, reflecting unstable predictions, particularly in challenging scenarios such as small lesions or poorly contrasted boundaries. In contrast, STD-Net achieves both higher medians and narrower distributions, signaling greater robustness. This stability is not merely a statistical improvement but a clinically meaningful property, as radiologists require consistent delineation across a wide range of lesion presentations for reliable diagnosis and treatment planning. Attention-based visualizations further confirm that the temporal module captures physiologically relevant enhancement patterns—such as early arterial hyperenhancement and delayed-phase washout—mirroring the diagnostic reasoning process used by clinicians.

From a clinical perspective, one of the most notable outcomes is STD-Net's superior sensitivity to small or early-stage hepatocellular carcinoma (HCC) lesions. These lesions are often subtle and easily overlooked when temporal dynamics are not explicitly modeled. By capturing phase-specific enhancement and washout cues, STD-Net effectively distinguishes HCC from cholangiocarcinoma or metastatic nodules—conditions that often share similar morphologies but exhibit distinct enhancement trajectories. Radiologist-verified case observations support this interpretation, suggesting that the model's temporal reasoning contributes directly to improved diagnostic confidence.

Another key outcome is the enhanced performance on small lesions, which often represent the most difficult and clinically important cases. Models without temporal decoupling tend to miss or fragment these lesions, while our framework preserves lesion integrity and reduces false negatives. This suggests that explicitly modeling contrast dynamics equips the network with the ability to identify subtle changes in enhancement that would otherwise be overlooked. The benefits are not limited to the primary dataset; the improved generalization on external cohorts indicates that STD-Net's learned representations capture fundamental imaging priors rather than being overfitted to specific scanners or protocols. This robustness is crucial for cross-institutional deployment in clinical practice.

While these results are promising, several limitations remain. The current design relies on rigid registration between phases, which may be imperfect in real-world clinical practice. Future work should explore motion-robust strategies, possibly integrating deformable alignment or end-to-end motion compensation. In addition, although the study focuses on hepatocellular carcinoma, the underlying principles are applicable to other dynamic imaging tasks, such as perfusion analysis or longitudinal tumor monitoring. Extending the framework to broader disease categories and multi-institutional datasets will be important for confirming its generalizability. Moreover, expanding interpretability beyond attention visualization—such as through radiomics-informed explanation or saliency-based attribution—could strengthen clinical trust and foster integration into radiology workflows.

Taken together, these findings demonstrate that decoupling spatial and temporal learning provides a clinically relevant paradigm for dynamic medical imaging. By aligning computational modeling more closely with radiologic reasoning, STD-Net bridges the gap between algorithmic precision and diagnostic insight. Its combination of accuracy, robustness, and interpretability positions it as a promising candidate for integration into future diagnostic pipelines, where consistent and explainable performance is essential for dependable clinical decision-making.

This work introduces STD-Net, a spatio-temporal decoupling framework designed for multiphasic liver lesion analysis. By explicitly separating spatial encoding from temporal modeling, the method captures both the structural appearance of tissues and the dynamic enhancement patterns that are critical for accurate diagnosis. The experiments across multiple datasets show that this design achieves superior segmentation accuracy and reliable lesion characterization, consistently outperforming strong convolutional and transformer-based baselines.

Beyond overall performance gains, the proposed approach demonstrates robustness in clinically challenging cases. Small lesions and those with fuzzy or irregular boundaries are detected and delineated with greater precision, reducing the likelihood of missed or fragmented predictions. Distributional analyses further reveal that STD-Net yields narrower performance spreads, highlighting its stability across diverse patient cohorts. The observed improvements extend to external datasets, indicating that the learned representations generalize beyond the training domain and hold promise for broader clinical application.

Although the framework already delivers meaningful advantages, there remain opportunities for refinement. Future extensions may incorporate motion-aware alignment strategies, expand evaluation to additional disease categories, and further enhance interpretability to foster clinical trust. Taken together, the findings suggest that spatio-temporal decoupling offers an effective and generalizable solution for dynamic medical imaging, with the potential to support more reliable and clinically informed decision-making.

Methods

We propose a 4D spatio-temporal decoupling framework, termed *Spatio-Temporal Decoupling Network (STD-Net)*, designed to explicitly capture the dynamic and sequential nature of multiphasic liver imaging. This section introduces the overall architecture, the core components, and the optimization strategy. Please see Fig. 9 for more details.

Framework overview

The central idea is to treat an N-phase 3D scan as a sequence of related but distinct volumes, $\{X_1, X_2, \ldots, X_N\}$, where $X_i \in \mathbb{R}^{H\times W\times D}$ denotes the scan at phase i. Processing is carried out in two decoupled stages, as illustrated in Figure [Your Figure Number].

First, a shared-weight *3D Spatial Feature Encoder* $\mathcal{E}$ processes each X_i independently to extract phase-specific spatial representations, capturing the "what" and "where" of anatomical structures. Second, the resulting features are fed into a *Temporal Dynamics Modeler* $\mathcal{T}$, whose role is to model inter-phase relationships and hemodynamic evolution, focusing on the "how" and "when" of lesion enhancement. Finally, task-specific *Decoders* $\mathcal{D}$ leverage the context-aware representations for joint lesion segmentation and characterization. Formally,

$$F_i = \mathcal{E}(X_i) \quad \forall i \in \{1, \ldots, N\}, \tag{1}$$

$$\{F'_1, \ldots, F'_N\} = \mathcal{T}(\{F_1, \ldots, F_N\}), \tag{2}$$

$$\text{Outputs} = \mathcal{D}(\{F'_1, \ldots, F'_N\},\ \text{skip} - \text{connections}). \tag{3}$$

3D spatial feature encoder

The encoder $\mathcal{E}$ is designed to generate consistent and robust anatomical representations.

Architecture. We adopt a deep 3D backbone, either ResNet-based[39] or Swin Transformer-based[21], pre-trained on large-scale medical datasets where possible. Each X_i is processed through a hierarchy of down-

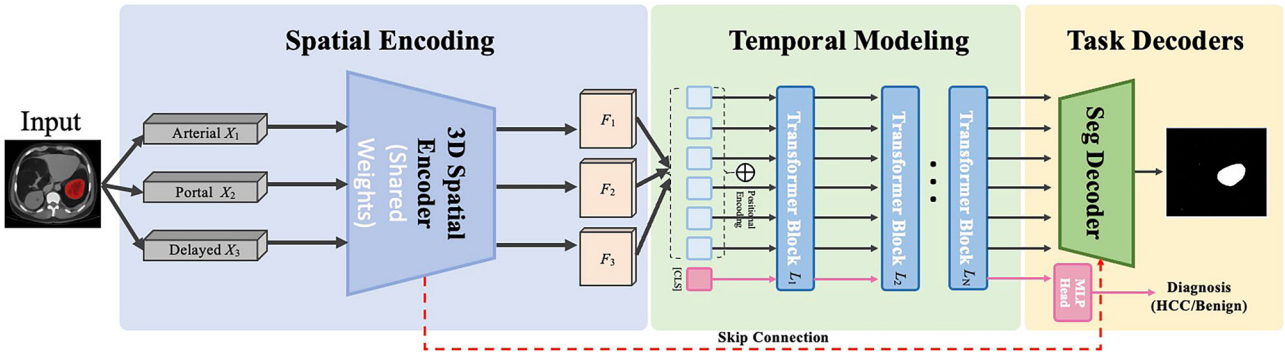


Fig. 9 | Overall architecture of STD-Net. Multiphasic CT/MRI volumes are first transformed into patch embeddings and processed by the shared-weight 3D Spatial Feature Encoder. The resulting spatial features are fed into the Temporal Dynamics Modeler, which leverages Transformer layers with positional encoding to capture sequential dependencies. The task-specific decoders follow: (i) a Segmentation Decoder that generates voxel-level lesion masks via U-Net style upsampling with skip connections, and (ii) a Characterization Head that utilizes the [CLS] token to predict lesion type (HCC, other malignancy, or benign). This spatio-temporal decoupling design mirrors clinical reasoning by explicitly separating "what" the lesion looks like from "how" it evolves across phases.

sampling blocks, producing multi-resolution feature maps. For input X_i, the encoder outputs $F_i \in \mathbb{R}^{h\times w\times d\times c}$, with h, w, d denoting spatially reduced dimensions and c the channel dimension.

Weight sharing. The same encoder weights are applied across all phases. This design enforces a canonical feature space for liver anatomy, compelling the model to emphasize contrast-driven appearance changes rather than redundant anatomical features. It also mitigates phase misalignment issues, since all phases are mapped through an identical extractor.

Temporal dynamics modeler

Given $\{F_1, \ldots, F_N\}$, the temporal modeler $\mathcal{T}$ learns sequential dependencies. A Transformer Encoder is used due to its strong ability to model long-range interactions.

Tokenization and positional encoding. Each F_i is flattened and projected into tokens. A learnable '[CLS]' token is prepended for classification. Since Transformers are permutation-invariant, positional encodings $E_{pos} \in \mathbb{R}^{(N+1)\times C}$ are added to preserve phase order:

$$Z_0 = [F_{cls}; F_1; F_2; \ldots; F_N] + E_{pos}. \tag{4}$$

Transformer layers. The modeler consists of L Transformer layers, each with Multi-Head Self-Attention (MHSA) and a feed-forward network (FFN).

Scaled dot-product attention is defined as:

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

with d_k the key dimension. Multi-head attention computes h parallel heads:

$$\text{MHSA}(Q, K, V) = \text{Concat}(\text{head}_1, \ldots, \text{head}_h)W^O, \tag{6}$$

$$\text{head}_i = \text{Attention}\,(\text{QW}_\text{i}^\text{Q}, \text{KW}_\text{i}^\text{K}, \text{VW}_\text{i}^\text{V}), \tag{7}$$

where W_i^Q, W_i^K, W_i^V, and W^O are learnable matrices. The FFN applies a two-layer MLP with nonlinearity:

$$\text{FFN}\,(z) = \max(0, zW_1 + b_1)W_2 + b_2. \tag{8}$$

Residual connections and normalization are applied around each sub-layer to stabilize training. This design enables the model to capture nuanced temporal relationships, e.g., linking arterial hyperenhancement to portal venous washout.

Task-specific decoders

The temporal outputs $\{F'_1, \ldots, F'_N\}$ are used for segmentation and classification.

Segmentation decoder. A lightweight U-Net-style decoder upsamples the embedding from the phase offering optimal lesion contrast (commonly the portal venous phase). The token sequence is reshaped into a 3D feature map and upsampled with transposed convolutions. Skip connections from $\mathcal{E}$ help recover fine spatial details lost during encoding.

Characterization head. For classification, the '[CLS]' token from the final Transformer layer aggregates global spatio-temporal information. It is passed through an MLP with softmax activation to output class probabilities (e.g., HCC, other malignancy, benign).

Multi-task loss function

The network is trained end-to-end with a composite objective.

The segmentation loss combines Dice and Binary Cross-Entropy:

$$L_{Dice} = 1 - \frac{2\sum_{i=1}^{M} p_i g_i}{\sum_{i=1}^{M} p_i^2 + \sum_{i=1}^{M} g_i^2}, \tag{9}$$

$$L_{BCE} = -\frac{1}{M}\sum_{i=1}^{M}\left[g_i \log(p_i) + (1 - g_i)\log(1 - p_i)\right], \tag{10}$$

where p_i and g_i denote predicted and ground-truth labels, and M the voxel count. The final segmentation loss is

$$L_{seg} = \alpha L_{Dice} + (1 - \alpha)L_{BCE}. \tag{11}$$

Classification is optimized with cross-entropy:

$$L_{cls} = -\sum_{j=1}^{C} y_j \log(\hat{y}_j), \tag{12}$$

where y_j is the ground-truth label and $\hat{y}_j$ the predicted probability.

The overall objective is a weighted sum:

$$L_{total} = \lambda_1 L_{seg} + \lambda_2 L_{cls}, \tag{13}$$

with λ_1 and λ_2 balancing the two tasks.

Analytical discussion: Why decoupling?

A natural question arises: why should spatial and temporal learning be separated rather than addressed jointly in a single network? At first glance, an integrated 4D convolutional model or a naïve multiphase Transformer might seem sufficient to capture both anatomical structure and dynamic evolution. However, such entangled designs often force the model to solve multiple heterogeneous tasks simultaneously—spatial encoding, temporal reasoning, and motion compensation—without clear boundaries. This burden can lead to suboptimal feature representations, making it difficult to distinguish genuine temporal changes from trivial spatial variations.

Another question concerns the role of weight sharing across phases. Would it not be simpler to allow the encoder to learn independent representations for each phase? While possible, this strategy risks redundancy, as the anatomical context remains largely consistent across phases. By enforcing a shared encoder, the model is encouraged to focus on phase-dependent differences such as contrast dynamics, which are the true carriers of diagnostic information. In this way, weight sharing implicitly regularizes the learning process and reduces overfitting.

It is also worth asking whether temporal dynamics could be sufficiently modeled by recurrent units such as LSTMs or GRUs. Although these architectures have been successfully applied in sequential domains, they often struggle with long-range dependencies and may lose fine-grained information across extended sequences. Transformers, by contrast, offer global receptive fields and direct pairwise interactions between phases, making them more suitable for capturing subtle wash-in and washout patterns that occur over multiple temporal stages.

Finally, one might wonder about the broader implications of spatio-temporal decoupling. Does this paradigm extend beyond liver imaging to other forms of dynamic medical data? The answer is affirmative: perfusion imaging, cardiac motion analysis, and longitudinal tumor monitoring all involve disentangling spatial structure from temporal progression. Thus, the decoupling principle is not merely a design choice for this specific task but a general framework that could inspire future models in dynamic medical imaging.

Summary of the proposed method

The proposed STD-Net is built on the principle of spatio-temporal decoupling, a design that explicitly separates the extraction of anatomical structures from the modeling of temporal dynamics. By first applying a shared-weight 3D encoder, the framework learns consistent spatial representations that remain stable across different imaging phases. This stage reduces redundancy and ensures that the network focuses on contrast-driven variations rather than repeating anatomical details.

The temporal modeling is handled independently through a Transformer-based module, which provides the ability to capture long-range dependencies and subtle dynamic signatures such as arterial hyperenhancement and venous washout. This design choice alleviates the limitations of recurrent units and conventional convolutional methods, which often fail to capture complex sequential relationships. The separation of roles between encoder and temporal modeler leads to clearer and more interpretable feature representations. On top of these modules, task-specific decoders perform lesion segmentation and characterization in a multi-task learning setting. The segmentation branch integrates both global context and fine-grained details through skip connections, while the characterization head aggregates spatio-temporal information into a single vector representation for classification. A composite loss function jointly optimizes these tasks, balancing voxel-level precision with lesion-level diagnostic accuracy. STD-Net provides a unified framework that mirrors clinical reasoning: it analyzes "what" the anatomy looks like, followed by "how" the lesion evolves across time. This modular and interpretable design not only improves segmentation and classification performance but also establishes a general paradigm applicable to other dynamic medical imaging tasks.

Ethics approval and consent to participate

This study uses only publicly available and de-identified datasets. Ethical approval and informed consent were obtained by the original data providers. No additional ethical approval was required for the analyses conducted in this manuscript.

Data availability

The datasets analyzed during the current study are publicly available from the following sources: TCGA-LIHC via The Cancer Imaging Archive, LiTS Challenge dataset, and MSD (Task03 Liver). Detailed access information is provided in the references.The code supporting the findings of this study is available from the corresponding author upon reasonable request, and will be released in accordance with journal or community requirements.

Code availability

The code supporting the findings of this study is available from the corresponding author upon reasonable request, and will be released in accordance with journal or community requirements.

Materials availability

No new materials were created or analyzed in this study. All materials used are publicly available.

Received: 23 September 2025; Accepted: 17 November 2025;
Published online: 08 December 2025

References

1. Sung, H. et al. Global cancer statistics 2020: Globocan estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA Cancer J. Clin.* **71**, 209–249 (2021).
2. Chernyak, V. et al. Liver imaging reporting and data system (LI-RADS) version 2018: imaging of hepatocellular carcinoma in at-risk patients[J]. *Radiology* **289**, 816–830 (2018).
3. Isensee, F. et al. nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation. *Nat methods* **18**, 203–211 (2021).
4. Hatamizadeh, A. et al. Unetr: Transformers for 3d medical image segmentation.In *Proc. IEEE/CVF Winter Conference on Applications of Computer Vision*, 574–584 (IEEE, 2022).
5. Roy, S. et al. Mednext: transformer-driven scaling of convnets for medical image segmentation[C]//International Conference on Medical Image Computing and Computer-Assisted Intervention. Cham: Springer Nature Switzerland, 405–415 (2023).
6. Shao, H.-C. et al. 3D cine-magnetic resonance imaging using spatial and temporal implicit neural representation learning (STINR-MR). *Phys. Med. Biol* **69**, 095007 (2024).
7. Dong, M. et al. Uncertainty-aware consistency learning for semi-supervised medical image segmentation. *Knowledge-Based Systems* **309**, 112890 (2025).
8. Xiang, D. et al. Unpaired Dual-Modal Image Complementation Learning for Single-Modal Medical Image Segmentation. IEEE Transactions on Biomedical Engineering (2024).
9. Chen, H. et al. UM-Mamba: An efficient U-network with medical visual state space for medical image segmentation. *Computer Vision and Image Understanding* **259**, 104436 (2025).
10. Luo, W. et al. Universal medical image segmentation with task-specific prompt-guided transformer model. In *2023 International Annual Conference on Complex Systems and Intelligent Science (CSIS-IAC)*. (IEEE, 2023).
11. Wei, Y. et al. More-brain: routed mixture of experts for interpretable and generalizable cross-subject fMRI visual decoding. https://arxiv.org/abs/2505.15946 (2025).
12. Wei, Y. et al. 4D multimodal co-attention fusion network with latent contrastive alignment for alzheimer's diagnosis. https://arxiv.org/abs/2504.16798 (2025).

13. Lyu, F. et al. Weakly supervised liver tumor segmentation using couinaud segment annotation. IEEE Transactions on Medical Imaging 41.5 (2021): 1138-1149.
14. Nisar, A. & Chen, Y. T. Eff-SAM: SAM-based Efficient Method for Brain Tumor Segmentation in Multimodal 3D MRI Scans. IEEE Access (2025).
15. Olaf, R., Fischer, P. & Brox, T. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*. Cham: Springer international publishing, 2015.
16. Çiçek, Özgün, et al. 3D U-Net: learning dense volumetric segmentation from sparse annotation. International conference on medical image computing and computer-assisted intervention. Cham: Springer International Publishing, 2016.
17. Fan, Tongle, et al. Ma-net: A multi-scale attention network for liver and tumor segmentation. Ieee Access 8 (2020): 179656-179665.
18. Yu, Z. et al. Enhancing preoperative diagnosis of microvascular invasion in hepatocellular carcinoma: domain-adaptation fusion of multi-phase CT images. *Front. Oncol* **14**, 1332188 (2024).
19. Liu, J., Wu, Z. & Feng, Q. Cf-net: a cross-phase fusion network for accurate liver lesion characterization. *Med. Image Anal.* **95**, 103150 (2024).
20. Dosovitskiy, A. et al. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. In *Proc International Conference on Learning Representations* https://openreview.net/forum?id=YicbFdNTTy (2021).
21. Liu, Z. et al. Swin transformer: Hierarchical vision transformer using shifted windows. *Proceedings of the IEEE/CVF international conference on computer vision*. 2021.
22. Hatamizadeh, A. et al. Swin unetr: Swin transformers for semantic segmentation of brain tumors in mri images. International MICCAI brainlesion workshop. Cham: Springer International Publishing, 2021.
23. Zhou, Z. et al. "Unet++: Redesigning skip connections to exploit multiscale features in image segmentation." IEEE transactions on medical imaging 39.6 (2019): 1856−1867.
24. Gao, M., Liu, J. & Sun, K. Dual-stream fusion network for liver tumor segmentation from multi-phase CT images. *Comput. Biol. Med.* **168**, 107742 (2024).
25. Dou, M. et al. Segmentation of rectal tumor from multi-parametric MRI images using an attention-based fusion network. *Med Biol Eng Comput* **61.9**, 2379–2389 (2023).
26. Gu, J. et al. Multi-phase cross-modal learning for noninvasive gene mutation prediction in hepatocellular carcinoma. 2020 42nd annual international conference of the IEEE engineering in medicine & biology society (EMBC). IEEE, 2020.
27. Zheng, R. et al. Automatic liver tumor segmentation on dynamic contrast enhanced MRI using 4D information: deep learning model based on 3D convolution and convolutional LSTM. IEEE Transactions on Medical Imaging 41.10 (2022): 2965-2976.
28. Yao, W. et al. Drfuse: Learning disentangled representation for clinical multi-modal fusion with missing modality and modal inconsistency. *Proceedings of the AAAI conference on artificial intelligence*. Vol. 38. No. 15. 2024.
29. Chartsias, A. et al. Disentangle, align and fuse for multimodal and semi-supervised image segmentation. IEEE transactions on medical imaging 40.3 (2020): 781−792.
30. Tran, D. et al. Learning spatiotemporal features with 3d convolutional networks. *Proceedings of the IEEE international conference on computer vision*. 2015.
31. Carreira, Joao, and Andrew Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2017.
32. Arnab, Anurag, et al. Vivit: A video vision transformer. Proceedings of the IEEE/CVF international conference on computer vision. 2021.
33. Bertasius, Gedas, Heng Wang, and Lorenzo Torresani. Is space-time attention all you need for video understanding?. Icml. Vol. 2. No. 3. 2021.
34. Peng, Xueping, et al. Temporal self-attention network for medical concept embedding. 2019 IEEE international conference on data mining (ICDM). (IEEE, 2019).
35. Baik, Dayoung, and Jaejun Yoo. Dynamic-aware spatio-temporal representation learning for dynamic MRI reconstruction. International Conference on Medical Image Computing and Computer-Assisted Intervention. Cham: Springer Nature Switzerland, 2025.
36. Xiao, X. et al. Describe anything in medical images. https://arxiv.org/abs/2505.05804 (2025).
37. Bilic, P. et al. The liver tumor segmentation benchmark (lits)[J]. *Med. Image Anal* **84**, 102680 (2023).
38. Simpson, A. L. et al. A large annotated medical image dataset for the development and evaluation of segmentation algorithms. arXiv preprint arXiv:1902.09063 (2019).
39. He, K., Zhang, X., Ren, S. & Sun, J. Deep residual learning for image recognition. In *Proc. IEEE conference on computer vision and pattern recognition*, 770–778 (IEEE, 2016).

Acknowledgements

The code supporting the findings of this study is available from the corresponding author upon reasonable request, and will be released in accordance with journal or community requirements.

Author contributions

S.L.Z., H.L.L., Y.W.Y., and X.F.D. designed the study. M.J.Z., Q.J.W., and Z.G. analyzed the data, generated charts, and wrote the manuscript. Y.Z., T.T.T., and Z.N.H. helped to collect data and assemble references. All authors contributed to the article and approved the submitted version.

Funding

This study is supported by Guangxi Natural Science Fundation (2024GXNSFAA010247), Guangxi Youth Science Fund Project (2024GXNSFBA010227), Research Foundation for Advanced Talents of The People's Hospital of Guangxi Zhuang Autonomous Region, Guangxi Academy of Medical Sciences (QYY-GCRC-202301) and High-level Talent Research Start-up Fund Project of Guangxi Academy of Medical Sciences (YKY-GCRC-202303).

Competing interests

The authors declare no competing interests.

Consent for publication

All authors have read and approved the final version of the manuscript and consent to its publication.

Additional information

Correspondence and requests for materials should be addressed to Yanwu You, Xiaofeng Dong or Honglin Luo.