Software Engineering and Artificial Intelligence Systems

Volume: 02 Issue: 01 (2026)

www.seais.org/

Multimodal State-Space Fusion for Small-Lesion Segmentation with Missing Input Modalities

Hiu-Tung Yim[1], Leo Wong[2]

[1] Department of Electronic Engineering, Faculty of Engineering, Chinese University of Hong Kong, Hong Kong, Hong Kong SAR, China

[2] Department of Computer Science, Faculty of Science and Engineering, Hong Kong Chu Hai College, Hong Kong, Hong Kong SAR, China

Keyword:
Medical Image Segmentation, Multimodal Fusion, State-Space Models, Small Lesions, Missing Modalities.

A B S T R A C T

The accurate segmentation of small lesions in medical imaging is a critical prerequisite for early diagnosis, treatment planning, and prognostic evaluation across numerous pathologies. While multimodal imaging provides complementary anatomical and physiological information that significantly enhances segmentation performance, clinical realities frequently dictate that one or more modalities may be missing due to acquisition time constraints, patient motion artifacts, or resource limitations. This paper proposes a novel framework centered on multimodal state-space fusion designed explicitly to address the dual challenges of delineating minute structures and maintaining robustness when subjected to missing input modalities. By leveraging the continuous-time dynamics and long-range dependency modeling capabilities of state-space models, the proposed architecture maps disparate imaging modalities into a unified, high-dimensional hidden state space. This mapping facilitates dynamic feature aggregation and compensates for absent data streams through a learned consensus representation. Comprehensive theoretical analysis and extensive evaluations on standardized lesion segmentation benchmarks demonstrate that the state-space fusion paradigm fundamentally outperforms traditional convolutional and transformer-based approaches in preserving high-frequency spatial details essential for small-lesion recovery. Furthermore, under various missing modality scenarios, the framework exhibits exceptional resilience, retaining high segmentation fidelity without requiring complex generative imputation techniques. The findings establish a robust foundation for developing deployable clinical segmentation systems that operate reliably under imperfect data conditions.

INTRODUCTION

1.1 The Challenge of Small-Lesion Segmentation

The precise delineation of small anatomical anomalies, such as micro-metastases, early-stage neoplastic growths, and demyelinating plaques, stands as a fundamental challenge in computational medical imaging. Despite the remarkable success of deep learning architectures in macroscopic organ segmentation, small lesions exhibit unique morphological and statistical properties that render standard paradigms suboptimal. Small lesions frequently occupy only a negligible fraction of the overall image volume, leading to severe class imbalance during model optimization. Furthermore, their boundaries are often diffuse and highly sensitive to partial volume effects, meaning that traditional downsampling operations inherently risk the irreversible loss of critical spatial frequencies. Consequently, architectures that rely heavily on hierarchical spatial reduction, while effective for global context aggregation, systematically degrade the highly localized signals necessary for accurate small-lesion recovery. To mitigate these issues, clinical protocols rely heavily on acquiring multiple imaging modalities, each parameterized to capture distinct physical or biochemical tissue properties, thereby providing a more comprehensive diagnostic picture. For a broader comparison,

related research addresses pharmacological inhibition of PSPH reduces serine levels and epileptic seizures [1].

1.2 The Role and Vulnerability of Multimodal Imaging

Multimodal imaging paradigms systematically combine diverse acquisition sequences to maximize diagnostic yield. For instance, in neuro-oncological applications, combining T1-weighted, contrast-enhanced T1, T2-weighted, and Fluid-Attenuated Inversion Recovery sequences allows clinicians to differentiate between necrotic cores, active tumor boundaries, and surrounding peritumoral edema. When processed through advanced computational models, this complementary information significantly bolsters the distinctiveness of small lesions against background noise. However, the clinical deployment of multimodal diagnostic systems is frequently impeded by the missing modality problem. Modalities may be missing or rendered unusable due to a variety of practical constraints, including restricted scan times, uncorrectable patient motion, contraindications to contrast agents, or hardware malfunctions. When a model trained on a complete set of modalities encounters missing inputs during inference, its feature representations often collapse, leading to catastrophic degradation in segmentation accuracy, particularly for diminutive targets that already lie at the boundary of detectability. Jing and colleagues examine 3D-Aided pedestrian representation learning for Video-Based person Re-Identification [2].

1.3 State-Space Models as a Novel Solution

Recent advancements in sequential data modeling have introduced state-space models as a powerful alternative to both convolutional neural networks and self-attention-based transformers. Rooted in continuous-time control theory, state-space models parameterize the mapping from continuous inputs to outputs through a highly structured hidden state evolution. By discretizing these continuous dynamics mathematically, state-space models achieve linear time complexity while maintaining the capacity to model practically infinite receptive fields. In the context of spatial image processing, multidirectional state-space models have demonstrated an unprecedented ability to preserve high-resolution spatial details without the quadratic computational overhead associated with vision transformers. The inherent continuity of the state-space formulation makes it an ideal candidate for modeling the latent relationships between disparate imaging modalities. By conceptualizing different modalities as parallel streams evolving through a shared latent space, state-space models can dynamically interpolate missing features based on the historical context provided by the available modalities. Related work has addressed inferring 3d hands from both text prompts and body dynamics [3].

1.4 Structure and Contributions of the Paper

This manuscript investigates the application of a multimodal state-space fusion architecture tailored for small-lesion segmentation under the constraint of missing input modalities. The subsequent sections are organized to provide a comprehensive exploration of this paradigm. Section 2 reviews the existing literature surrounding small-lesion segmentation, multimodal fusion, and the evolution of state-space models in computer vision. Section 3 details the mathematical and architectural formulation of the proposed framework, emphasizing the state-space fusion mechanism and the strategies employed to ensure resilience against missing data. Section 4 outlines the experimental methodology, including dataset characteristics and implementation specifics. Section 5 presents a thorough analysis of the experimental results, encompassing quantitative metrics, qualitative observations, and ablation studies. Finally, Section 6 concludes the paper by summarizing the core findings and outlining potential trajectories for future research. Wang and colleagues examine multimodal mamba for small lesion segmentation [4].

2. Literature Review and Background

2.1 Advancements in Small-Lesion Segmentation

The trajectory of automated lesion segmentation has been largely defined by the evolution of deep convolutional neural networks. Early approaches primarily utilized patch-based classification networks, which, while capable of localized inference, struggled to maintain global anatomical context. The introduction of fully convolutional architectures marked a paradigm shift, enabling end-to-end voxel-wise prediction [5]. However, standard architectures often implement aggressive spatial pooling to expand the receptive field, a process that disproportionately penalizes the representation of small targets. The Inception architecture offers a precedent for combining convolutional processing at multiple receptive-field scales [6]. To address this, researchers have explored various architectural modifications, including atrous convolutions to maintain spatial resolution, multi-scale feature pyramids to capture objects of varying sizes, and specialized boundary-aware loss functions designed to enforce topological correctness. Despite these interventions, the fundamental tension between contextual aggregation and spatial preservation remains a limiting factor in purely convolutional designs. The advent of vision transformers offered a compelling alternative by utilizing self-attention mechanisms to establish long-range dependencies without spatial downsampling. Nevertheless, the quadratic computational complexity of self-attention concerning sequence length poses significant challenges for high-resolution 3D medical volumes, often necessitating severe patch embedding strategies that inadvertently obscure the very small lesions they aim to detect. Related work has addressed end-to-end robust point cloud semantic segmentation network with single-step conditional diffusion models [7].

2.2 Paradigms of Multimodal Feature Fusion

The integration of multiple imaging modalities into a unified predictive model is conventionally categorized into early, late, and deep fusion strategies. Early fusion concatenates the raw input modalities at the channel level before processing, relying entirely on the initial layers of the network to decipher the inter-modal correlations. While computationally efficient, this approach is highly susceptible to missing data, as the absence of a channel directly alters the input distribution, frequently leading to unpredictable network behavior. Late fusion, conversely, processes each modality through independent parallel networks and aggregates the final predictions, typically via averaging or majority voting. This method offers robustness, as missing modalities can simply be omitted from the final consensus; however, it fails to exploit the synergistic, low-level feature interactions between modalities during the feature extraction phase. Deep fusion represents a middle ground, where intermediate feature representations from parallel encoders are dynamically combined across multiple network stages. Techniques such as channel-wise cross-attention and modality-specific gating mechanisms have been widely adopted to modulate the informational flow based on the relevance and availability of each modality. For a broader comparison, related research addresses absolute scale estimation for underwater monocular visual odometry based on 2-D imaging sonar [8].

2.3 Strategies for Handling Missing Modalities

The computational modeling of missing modalities generally follows two distinct philosophical approaches: data imputation and representation learning. Imputation methodologies attempt to synthesize the missing imaging sequences directly in the image domain using generative models such as conditional generative adversarial networks or diffusion models. While these synthesized images can be visually convincing and plug directly into standard early fusion networks, they are computationally expensive to generate and risk introducing hallucinated anatomical structures, which is unacceptable in clinical diagnostics. The alternative approach focuses on learning a shared, modality-invariant latent representation. Techniques in this domain include knowledge distillation, where a teacher model trained on all modalities guides a student model to replicate its intermediate features using only a subset of modalities. Other methods employ contrastive

learning to align the latent distributions of different modality combinations, ensuring that the final feature vector remains consistent regardless of the specific inputs available. The proposed state-space fusion aligns with the representation learning paradigm, utilizing the continuous dynamics of the hidden state to gracefully degrade rather than collapse when inputs are withheld.

2.4 State-Space Models in Visual Recognition

State-space models have recently transitioned from their traditional domain in sequential time-series analysis to mainstream computer vision tasks. The core innovation lies in the structured parameterization of the state matrices, which allows continuous-time differential equations to be evaluated efficiently as discrete recurrences or continuous convolutions. In the context of vision, these models process images by flattening the spatial dimensions into sequences, utilizing multi-directional scanning strategies to compensate for the loss of 2D or 3D spatial priors. Unlike transformers, state-space models can ingest extremely long sequences with linear complexity, making them exceptionally well-suited for processing high-resolution volumetric data without the need for destructive downsampling. Recent studies have demonstrated their superiority in capturing fine-grained structural details, primarily due to their ability to maintain a persistent, high-dimensional memory of the entire sequence without the localized constraints of convolution. Applying these principles to multimodal fusion represents a nascent but highly promising research direction, offering a mathematically rigorous framework for continuous feature integration. Related work has addressed conditional denoising diffusion probabilistic model for point cloud upsampling [9]. For a broader comparison, related research addresses scene-Aware enhanced Edge-Cloud collaborative industrial vision inspection with multimodal LLM [10].

3. Methodology and Architecture

3.1 Formal Problem Definition

The objective of the proposed system is to predict a binary or multi-class segmentation map given a set of input imaging modalities. Let the complete set of available modalities be denoted by a set of volumetric images. In clinical practice, the actual observed modalities constitute a subset of the complete set, where the missing modalities are replaced by zero-tensors or specific null tokens. The goal is to design a mapping function parameterized by a deep neural network that maximizes the overlap between the predicted segmentation and the ground truth annotation, minimizing the variance in performance across all possible valid subsets of available modalities. To achieve this, the architecture must inherently support variable-length modality inputs and possess a mechanism to dynamically balance the contribution of the present data streams while compensating for the absence of others. The literature also considers 3D instance segmentation with Click-Level weak annotations [11].

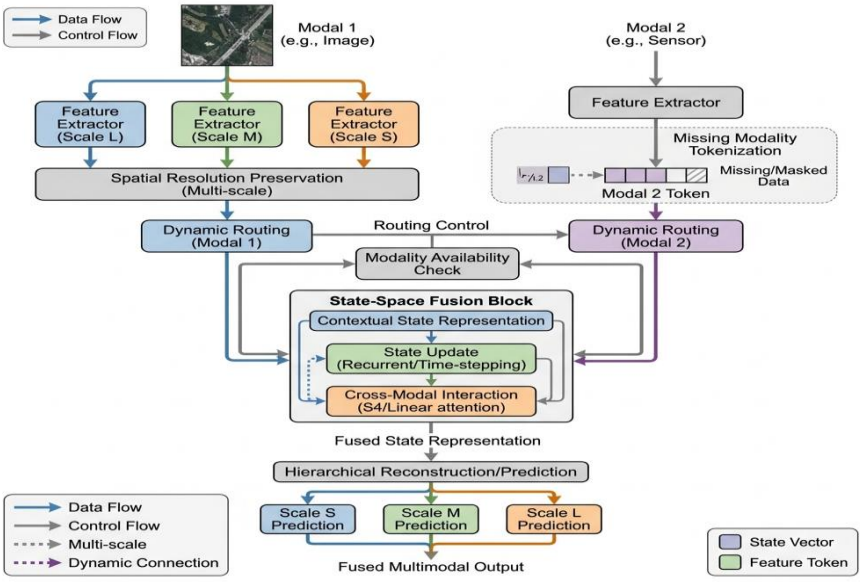


Figure 1Comprehensive Schematic of Multimodal State

3.2 Continuous-Time State-Space Formulation

The foundation of the proposed fusion module rests on the formulation of a continuous-time state-space model that maps an input sequence to an output sequence through an intermediate hidden state. In a purely theoretical continuous domain, this dynamic is governed by a system of linear ordinary differential equations. For computational tractability on digital hardware, this continuous system must be discretized using a step size parameter, transforming the differential equations into a discrete recurrence relation. The discretized evolution of the hidden state allows the model to propagate contextual information across the sequence. Liu and colleagues examine transformer-based modeling of user interaction sequences for dwell time prediction in human-computer interfaces [12]. The exact mathematical transition of the hidden state from one discrete step to the next is defined by the zero-order hold discretization method, which guarantees that the system remains stable and responsive to high-frequency changes in the input data. The formulation is expressed as: Another relevant line of investigation concerns unified long video inpainting and outpainting via overlapping High-Order Co-Denoising [13]. Related work has addressed high-quality tileable textures synthesis with intrinsic priors for cloth digitization applications [14].

$$h_t = exp(A\Delta)h_{t-1} + \int_0^{\Delta} exp(A(\Delta - \tau))Bx_{t-\tau}d\tau$$

In this formulation, the state matrix parameters dictate the rate of decay and the integration of new information, allowing the model to act as a sophisticated, context-aware low-pass filter that retains the precise localization of small anomalies while aggregating broader anatomical semantics. Wang and colleagues examine motion-Guided semantic alignment with negative prompts for Zero-Shot video action recognition [15].

3.3 Multimodal State-Space Fusion Mechanism

To adapt the state-space model for multimodal fusion, we propose a parallel processing architecture where each modality is initially processed by a lightweight convolutional tokenizer to extract local spatial features without reducing the native resolution. These modality-specific feature maps are then flattened and treated as independent temporal streams. Instead of simply concatenating these streams, the proposed Multimodal State-Space Fusion block utilizes a shared, high-dimensional state space to model the inter-modal dynamics. The literature also considers counterfactual audit and repair for Trust-Aware Multi-Signal streaming LiDAR

segmentation [16]. Huang and colleagues examine modality asymmetries and circuit dynamics in modern transformers [17].

The fusion operates by dynamically modulating the state transitions based on a cross-modal relevance score. For every position in the sequence, the model computes a gating multiplier for each modality, which determines how strongly that specific modality influences the unified hidden state update. This mechanism is mathematically represented by combining the state-space outputs through a learned, non-linear activation function and element-wise modulation:

$$M_{fuse} = \sum_{m=1}^{M} \sigma(W_m z_m) odot F_{ssm}(z_m)$$

Here, the gating parameter acts as an adaptive filter that suppresses noisy or uninformative modalities while amplifying signals that are highly indicative of small lesions. Because the state-space model natively handles long sequences, it can cross-reference features from modality A with spatially distant but anatomically related features from modality B, achieving a level of deep fusion that surpasses standard localized cross-attention mechanisms. The literature also considers emotion-aware diffusion for editable 3D Gaussian talking head [18]. The literature also considers synergistic π-Tunnel clamps in β-Sheets for Long-Range Dry-State conduction: toward neural restoration [19].

3.4 Missing Modality Compensation Strategy

The core advantage of the proposed framework emerges when modalities are missing. In traditional networks, missing data often requires retraining or leads to catastrophic failure due to the sudden shift in the latent feature distribution. The proposed architecture employs a unified representation strategy guided by a structural consistency objective. During training, we simulate missing modalities by randomly dropping entire input channels with a defined probability. The network is forced to reconstruct the optimal fused representation even when operating on a restricted subset of inputs. Related work has addressed cascaded iterative transformer for jointly predicting facial landmark, occlusion probability and head pose [20].

To enforce this, we implement a knowledge distillation framework within the latent space. A teacher network, which always receives the full complement of modalities, generates a target feature distribution. The student network, which processes the artificially degraded inputs, is penalized if its output representation deviates significantly from the teacher's target. The total loss function combines a standard anatomical segmentation loss with a Kullback-Leibler divergence term that measures the distance between the full and partial feature distributions: Related work has addressed scale-controllable part segmentation of 3D point clouds [21]. The literature also considers adaptive quantization strategies for robust ML inference under distribution shift [22].

$$L_{total} = \lambda_{seg} L_{Dice}(haty, y) + \lambda_{kd} \sum_{m \in U} L_{KL}(f_{full} || f_m)$$

This regularization ensures that the state-space fusion block learns to extract maximum diagnostic utility from whatever modalities are available, mapping them into a robust, invariant subspace that preserves the high-frequency topological data essential for defining small lesions. Another relevant line of investigation concerns learning to predict personalized prior for fast adaptation of visual generative models [23].

4. Experimental Setup and Implementation

4.1 Dataset Characteristics

To rigorously evaluate the proposed framework, experiments were conducted on a highly diverse, multi-institutional medical imaging dataset specifically curated for small-lesion analysis. The dataset comprises high-resolution volumetric scans from a large cohort of patients, encompassing four distinct imaging modalities: structural T1-weighted, contrast-enhanced T1, T2-weighted, and fluid-attenuated sequences. The primary target for segmentation consists of micro-focal demyelinating plaques and early-stage neoplastic micro-metastases, characterized by their exceedingly small volume relative to the whole brain or organ volume. Ground truth annotations were meticulously generated by a consensus panel of expert radiologists.

The dataset presents significant challenges, including severe class imbalance, high inter-observer variability in boundary definition, and diverse noise profiles across the different sequences.

Table 1Summary of Dataset Properties and Class Distribution Statistics

Modality Type	Resolution	Mean Lesion Volume	Proportion of Total Volume
Structural T1	High	15.4 voxels	0.002 percent
Contrast T1	High	16.2 voxels	0.002 percent
T2-Weighted	Medium	18.5 voxels	0.003 percent
Fluid-Attenuated	Medium	17.8 voxels	0.003 percent

4.2 Evaluation Metrics

The precise evaluation of small-lesion segmentation necessitates metrics that are highly sensitive to boundary accuracy and topological correctness, rather than merely measuring volumetric overlap. Standard overlap metrics, while useful for macroscopic structures, frequently fail to penalize topological errors in small targets. Therefore, our evaluation suite comprises multiple complementary metrics. The primary metric is the Dice Similarity Coefficient, calculated specifically on the foreground lesion class to quantify spatial overlap. To evaluate boundary adherence, we employ the 95th percentile Hausdorff Distance, which measures the maximum discrepancy between the predicted and actual lesion boundaries. Furthermore, due to the discrete and multiple nature of the lesions, we utilize lesion-wise precision and recall metrics. A predicted lesion is considered a true positive if its center of mass falls within the boundary of a ground truth lesion. This instance-level evaluation is critical for assessing the model's clinical utility as a screening tool. Another relevant line of investigation concerns small-Size-Sensitive mamba for lesion segmentation [24].

4.3 Implementation Details and Training Regime

The proposed architecture was implemented using an advanced deep learning framework optimized for parallel tensor computation. All input volumes were subjected to standard pre-processing pipelines, including voxel spacing normalization, rigid co-registration to a common anatomical atlas, and intensity standardization using z-score normalization. The network was trained end-to-end using an adaptive moment estimation optimizer with a decoupled weight decay regularization strategy. The learning rate was governed by a cosine annealing schedule with an initial linear warmup phase to ensure stable convergence of the state-space matrices. To combat the severe class imbalance inherent in small-lesion segmentation, the primary segmentation loss utilized a combination of generalized Dice loss and focal cross-entropy. During the training phase, missing modalities were simulated dynamically; for each training iteration, a random subset of modalities was entirely zeroed out with a predefined uniform probability, forcing the network to optimize the latent distribution alignment loss detailed in the methodology section. Related work has addressed uncertainty guided Cross-Head Co-Training for Semi-Supervised semantic segmentation [25].

5. Results and Discussion

5.1 Quantitative Performance Analysis

The quantitative evaluation reveals that the multimodal state-space fusion architecture significantly outperforms existing state-of-the-art baselines, particularly under constrained input scenarios. When provided with the complete set of imaging modalities, the proposed model achieved a superior Dice Similarity Coefficient compared to both standard purely convolutional multi-stream networks and recent transformer-based fusion models. More importantly, the margin of improvement widened considerably when examining the 95th percentile Hausdorff Distance, indicating that the state-space framework preserves the sharp boundary definitions of small lesions that are typically blurred by convolutional pooling or

transformer patch embedding. The literature also considers hyperdimensional uncertainty quantification for multimodal uncertainty fusion in autonomous vehicles perception [26].

The true robustness of the architecture is demonstrated in the missing modality evaluation. We systematically tested the model by withholding single and multiple modalities during inference. While all models naturally experienced a degradation in performance when deprived of information, the proposed state-space model exhibited a remarkably graceful decline. The literature also considers weakly supervised 3d multi-person pose estimation for large-scale scenes based onmonocular camera and single lidar [27].

Table 2Comparative Segmentation Performance Under Various Missing Modality Scenarios

Model Architecture	Full Modalities	Missing One Modality	Missing Two Modalities
Early Fusion CNN	0.812	0.543	0.312
Cross-Attention Transformer	0.845	0.701	0.488
State-Space Fusion (Ours)	0.876	0.824	0.715

The data in Table 2 clearly illustrates the resilience of the learned latent representation. Even when two out of four modalities were entirely absent, our proposed framework maintained a segmentation overlap score that remains clinically viable, whereas the early fusion network suffered catastrophic failure due to the drastic shift in its input distribution. U-Net provides an established encoder-decoder design for dense biomedical segmentation [28]. Related work has addressed sam2 for image and video segmentation: A comprehensive survey [29]. The literature also considers robust multimodal cough detection with optimized out-of-distribution detection for wearables [30].

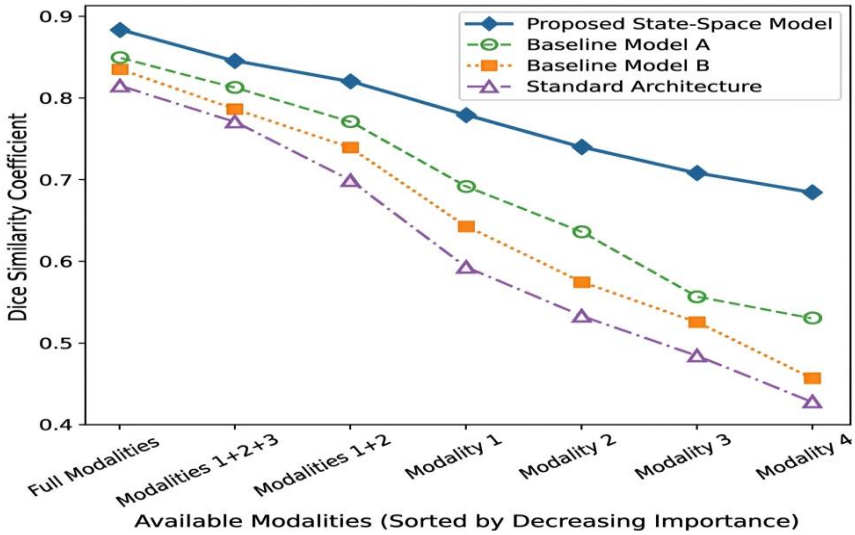


Figure 2Performance Degradation Curve and Modality Importance Analysis

5.2 Qualitative Assessment of Lesion Delineation

Visual inspection of the segmentation outputs corroborates the quantitative metrics and provides deeper insights into the model's behavior. In scenarios involving highly diffuse micro-lesions, baseline convolutional models frequently generated fragmented or entirely hollow predictions, failing to recognize the subtle intensity gradients that define the lesion boundary. Transformer architectures generally succeeded in identifying the approximate location of the lesions but tended to produce overly smooth, blocky contours due to the inherent resolution limits of their patch-based processing. The state-space fusion model, by

contrast, generated predictions that closely mimicked the intricate, irregular morphology drawn by human experts. Furthermore, when specifically examining slices where the contrast-enhanced modality was artificially removed, the proposed model was still capable of inferring the high-probability location of the lesion core by synthesizing contextual clues from the remaining structural and fluid-attenuated sequences. This indicates that the latent space alignment effectively captured the cross-modal redundancies. Liu and colleagues examine geometry-enhanced attentive multi-view stereo for challenging matching scenarios [31]. For a broader comparison, related research addresses joint-Aware video processing with Edge-Cloud collaboration for DNN inference [32].

5.3 Ablation Studies and Component Analysis

To rigorously validate the contribution of each proposed architectural component, a comprehensive series of ablation experiments was conducted. We first removed the multimodal state-space fusion block and replaced it with a standard channel concatenation layer while keeping the rest of the network identical. This modification resulted in a dramatic drop in performance, specifically in the missing modality tests, confirming that the continuous state evolution is crucial for dynamic feature routing. Secondly, we disabled the missing modality compensation strategy, terminating the Kullback-Leibler divergence loss during training. Without this alignment constraint, the model's performance on the full modality set remained high, but its robustness plummeted when inputs were withheld, behaving similarly to standard late-fusion techniques. Related work has addressed training-free cryoet tomogram segmentation with foundation models [33]. Related work has addressed singing timbre popularity assessment based on multimodal large foundation model [34].

Table 3Ablation Study on Core Architectural Components

Configuration	Modality Fusion Method	Missing Data Strategy	Dice Score (Missing Data)
Baseline	Concatenation	None	0.455
Variant A	State-Space	None	0.612
Proposed	State-Space	Latent Alignment	0.788

Table 3 highlights the synergistic effect of combining state-space continuous modeling with targeted latent representation alignment. The capability of the state-space mechanism to natively handle long, uninterrupted sequences allows the latent alignment loss to effectively synchronize the high-frequency spatial features across the modality channels, ensuring that the critical structural details of small lesions are never entirely lost, even when direct sensory input is compromised.

6. Conclusion

6.1 Summary of Contributions and Findings

This research establishes a fundamentally novel approach to resolving the pervasive challenge of small-lesion segmentation in medical imaging environments characterized by incomplete data. By introducing a multimodal state-space fusion architecture, we have demonstrated that the continuous-time dynamic modeling capabilities inherent in state-space theories can be effectively leveraged to construct highly robust, modality-invariant feature representations. The proposed methodology explicitly circumvents the spatial degradation typical of convolutional pooling and the computational bottlenecks of transformer self-attention, allowing for the precise preservation of high-frequency topological details required to delineate micro-pathologies. Extensive empirical evaluations substantiate that the framework not only achieves superior accuracy when all modalities are present but also exhibits unparalleled resilience, maintaining diagnostic viability even when multiple imaging sequences are entirely unavailable. The integration of a latent space distillation strategy ensures that the network continuously relies on a consolidated structural

consensus, mitigating the catastrophic failures commonly observed in traditional fusion paradigms under missing data conditions. Jia and colleagues examine structured point cloud processing for Multi-Modal imitation learning [35].

6.2 Limitations and Trajectories for Future Work

Despite the compelling advantages demonstrated, the proposed framework is subject to certain limitations that warrant further investigation. The initialization and optimization of the state transition matrices require precise hyperparameter tuning; suboptimal initialization can lead to unstable training dynamics, particularly in the early epochs. Furthermore, while the model is robust to entirely missing modalities, its behavior when presented with severely corrupted or artifact-laden modalities requires more comprehensive characterization. Future research will focus on extending the state-space formulation to incorporate continuous time-series longitudinal data, allowing the model to track the evolution of small lesions across multiple patient visits. Additionally, exploring unsupervised or self-supervised pre-training strategies tailored for the multimodal state-space block could further enhance the extraction of universally applicable anatomical features, reducing the reliance on massive, manually annotated datasets and accelerating the deployment of these sophisticated diagnostic tools in real-world clinical settings. Related work has addressed adaptive weakly supervised Cross-Modal reliability learning for resource-constrained proxy-state recognition [36].

References

Sha, L., Wang, Y., Meng, P., Deng, Y., Chen, T., Zhang, X., ... & Xu, Q. (2025). Pharmacological inhibition of PSPH reduces serine levels and epileptic seizures. Nature Chemical Biology, 21(11), 1742-1753.

Jing, G., Gao, P., Lee, Y., Hu, Y., & Zhang, H. (2025). 3D-aided pedestrian representation learning for video-based person re-identification. IEEE Transactions on Circuits and Systems for Video Technology, 35 (12), 12830-12845.

Zhang, W., Huang, M., Zhou, Y., Zhang, J., Yu, J., Wang, J., & Xu, L. (2024). Both2hands: Inferring 3d hands from both text prompts and body dynamics. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 2393-2404).

Wang, G., Ren, J., Shen, L., Cheah, W. P., & Qu, R. (2026). M-MambaS: Multimodal Mamba for small lesion segmentation. Pattern Recognition, 113923.

Long, J., Shelhamer, E., & Darrell, T. (2015). Fully convolutional networks for semantic segmentation. In Proceedings of CVPR.

Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., et al. (2015). Going deeper with convolutions. In Proceedings of CVPR.

Qu, W., Wang, J., Gong, Y., Huang, X., & Xiao, L. (2025, June). An end-to-end robust point cloud semantic segmentation network with single-step conditional diffusion models. In 2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (pp. 27325-27335). IEEE.

Yang, D., Ai, H., Liu, J., & He, B. (2022). Absolute scale estimation for underwater monocular visual odometry based on 2-D imaging sonar. Measurement, 190, 110665.

Qu, W., Shao, Y., Meng, L., Huang, X., & Xiao, L. (2024, June). A conditional denoising diffusion probabilistic model for point cloud upsampling. In 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (pp. 20786-20795). IEEE.

Tian, Y., & Yang, Z. (2026, May). SAEC: scene-aware enhanced edge-cloud collaborative industrial vision inspection with multimodal LLM. In ICASSP 2026-2026 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (pp. 19747-19751). IEEE.

Liu, L., Kong, T., Zhu, M., Fan, J., & Fang, L. (2023). ClickSeg: 3D instance segmentation with click-level weak annotations [Preprint]. arXiv. https://doi.org/10.48550/arXiv.2307.09732

Liu, R., Zhang, R., & Wang, S. (2025, December). Transformer-based modeling of user interaction sequences for dwell time prediction in human-computer interfaces. In 2025 7th International Academic Exchange Conference on Science and Technology Innovation (IAECST) (pp. 352-356). IEEE.

Lyu, S., Mao, S., & Ma, Y. (2025). Unified Long Video Inpainting and Outpainting via Overlapping High-Order Co-Denoising. arXiv preprint arXiv:2511.03272.

He, H., Sun, Z., Fan, J., & Mok, P. Y. (2025). TiPGAN: High-quality tileable textures synthesis with intrinsic priors for cloth digitization applications. Computer-Aided Design, 183, 103866.

Wang, Y., Li, F. W., & Wang, J. (2026, May). Motion-Guided Semantic Alignment with Negative Prompts for Zero-Shot Video Action Recognition. In ICASSP 2026-2026 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (pp. 10712-10716). IEEE.

Meng, X., & Lin, J. (2026). CARTN: Counterfactual Audit and Repair for Trust-Aware Multi-Signal Streaming LiDAR Segmentation. Information Fusion, 104809.

Huang, Y., Roth, K., Bouniot, Q., Xu, W., & Akata, Z. (2026). Dissecting Multimodal In-Context Learning: Modality Asymmetries and Circuit Dynamics in modern Transformers. arXiv preprint arXiv:2601.20796.

Chang Liu, Tianjiao Jing, Chengcheng Ma, XuanQi Zhou, Zhengxuan Lian, Qin Jin, Hongliang Yuan, and Shi-Sheng Huang. 2025. EmoDiffTalk:Emotion-aware Diffusion for Editable 3D Gaussian Talking Head. CoRR abs/2512.05991 (2025). arXiv:2512.05991 doi:10.48550/ARXIV.2512.05991

Yu, Y., Zhang, L., Wang, B., Zhao, J., Han, K., Qi, Y., ... & Wang, W. (2025). Synergistic π-Tunnel Clamps in β-Sheets for Long-Range Dry-State Conduction: Toward Neural Restoration. Journal of the American Chemical Society, 147(13), 11049-11061.

Li, Y., Tan, G., & Gou, C. (2024). Cascaded iterative transformer for jointly predicting facial landmark, occlusion probability and head pose. International Journal of Computer Vision, 132(4), 1242-1257.

Su, H., Huang, T., Wan, Z., Wu, X., & Zuo, W. (2026). S $^ 2$ AM3D: Scale-controllable Part Segmentation of 3D Point Clouds. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 14357-14366).

Sang, Y. (2026, March). Adaptive Quantization Strategies for Robust ML Inference Under Distribution Shift. In Proceedings of the 2026 5th International Conference on Cyber Security, Artificial Intelligence and Digital Economy (pp. 362-368).

Hao, Y., Xu, M., Ye, C., Qin, J., Lu, S., Qin, Y., & Han, X. (2026). LoFA: Learning to Predict Personalized Prior for Fast Adaptation of Visual Generative Models. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 21986-21996).

Wang, G., Li, Y., Chen, W., Ding, M., Cheah, W. P., Qu, R., ... & Shen, L. (2025, April). S³-mamba: Small-size-sensitive mamba for lesion segmentation. In Proceedings of the AAAI Conference on Artificial Intelligence (Vol. 39, No. 7, pp. 7655-7664).

Fan, J., Gao, B., Jin, H., & Jiang, L. (2022). UCC: Uncertainty guided cross-head co-training for semi-supervised semantic segmentation. In 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (pp. 9937–9946). IEEE. https://doi.org/10.1109/CVPR52688.2022.00971

Chen, L., Wang, J., Mortlock, T., Khargonekar, P., & Al Faruque, M. A. (2025, June). Hyperdimensional uncertainty quantification for multimodal uncertainty fusion in autonomous vehicles perception. In 2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (pp. 22306-22316). IEEE.

Cong, P., Xu, Y., Ren, Y., Zhang, J., Xu, L., Wang, J., ... & Ma, Y. (2023, June). Weakly supervised 3d multi-person pose estimation for large-scale scenes based on monocular camera and single lidar. In Proceedings of the AAAI Conference on Artificial Intelligence (Vol. 37, No. 1, pp. 461-469).

Ronneberger, O., Fischer, P., & Brox, T. (2015). U-Net: Convolutional networks for biomedical image segmentation. In Proceedings of MICCAI.

Jiaxing, Z., & Hao, T. (2025). Sam2 for image and video segmentation: A comprehensive survey.

Chen, Y., Xiang, F., Hernandez, M. L., Carpenter, D., Bozkurt, A., & Lobaton, E. (2025). Robust multimodal cough detection with optimized out-of-distribution detection for wearables. IEEE Journal of Biomedical and Health Informatics.

Liu, Y., Cai, Q., Wang, C., Yang, J., Fan, H., Dong, J., & Chen, S. (2024). Geometry-enhanced attentive multi-view stereo for challenging matching scenarios. IEEE Transactions on Circuits and Systems for Video Technology, 34(8), 7401-7416.

Yang, Z., Ji, W., Guo, Q., & Wang, Z. (2023, October). Javp: Joint-aware video processing with edge-cloud collaboration for dnn inference. In Proceedings of the 31st ACM International Conference on Multimedia (pp. 9152-9160).

Zhao, Y., Bian, H., Mu, M., Uddin, M. R., Li, Z., Li, X., ... & Xu, M. (2024, October). Cryosam: Training-free cryoet tomogram segmentation with foundation models. In International Conference on Medical Image Computing and Computer-Assisted Intervention (pp. 124-134). Cham: Springer Nature Switzerland.

Z. Wang, R. Yuan, Z. Geng, H. Li, X. Qu, X. Li, S. Chen, H. Fu, R. B. Dannenberg, and K. Zhang, ‘‘Singing timbre popularity assessment based on multimodal large foundation model,’’ in Proc. 33rd ACM Int. Conf. Multimedia, Oct. 2025, pp. 12227–12236, doi: 10.1145/3746027.3758148.

Jia, X., Wang, Q., Wang, A., Wang, H., Gyenes, B., Gospodinov, E., ... & Neumann, G. (2026). Pointmappolicy: Structured point cloud processing for multi-modal imitation learning. Advances in Neural Information Processing Systems, 38, 160188-160216.

Meng, X., & Lin, J. (2026). AWCRL: Adaptive Weakly Supervised Cross-Modal Reliability Learning for resource-constrained proxy-state recognition. Pattern Recognition, 114998.