Lightweight Temporal Models for Bolus Segmentation in VFSS

Yuanhan Chen

Abstract—Accurate bolus segmentation in Videofluoroscopic Swallow Studies (VFSS) is essential for diagnosing dysphagia, yet manual annotation is slow and inconsistent. Temporal information is valuable because the bolus is typically the only moving object, but most lightweight segmentation models process single frames. We explore two approaches for lightweight temporal models: a two-stage pipeline that reconstructs clean frames to highlight bolus motion, and a temporal model that directly processes frame sequences. Experiments on 6,337 frame pairs from 87 VFSS sequences show that a two-stage SmallUNet pipeline achieves IoU 0.519 and Dice 0.628 with AUC 0.995, while a Temporal Context Module UNet reaches IoU 0.515 and Dice 0.644 with AUC 0.996. Both deliver strong accuracy with modest parameter counts but do not surpass a single-frame MobileNetV2 UNet baseline (IoU 0.542, Dice 0.669), demonstrating that encoder capacity remains critical. Our results show temporal modeling provides the most benefit when model capacity is constrained.

Index Terms—VFSS, bolus segmentation, temporal modeling, UNet

1 Introduction

VIDEOFLUOROSCOPIC Swallow Studies (VFSS) are the gold standard for diagnosing dysphagia, a swallowing disorder affecting over 16% of the general population [1]. During VFSS examinations, clinicians track the oral contrast bolus through fluoroscopy frames to identify penetration-aspiration events and measure post-swallow residue. Manual frame-by-frame annotation is time-consuming, subjective, and prone to inter-rater variability [1], [2]. Automated bolus segmentation could reduce human error and provide objective analysis for early diagnosis and treatment planning.

Recent deep learning approaches for VFSS bolus segmentation show promising results [1], [2], [3], [4]. Most lightweight models process single frames using encoder-decoder architectures such as UNet [5], achieving good accuracy but ignoring temporal information. Heavier video models like Video-TransUNet [4] and Video-SwinUNet [6] incorporate temporal attention and improve performance, but their large parameter counts and memory requirements limit deployment on typical clinical hardware.

We explore whether lightweight models can effectively use temporal information for VFSS bolus segmentation. Our motivation comes from a simple observation: when we subtract consecutive frames and threshold the absolute difference, the result roughly matches the bolus location. Adding the previous frame's bolus mask to this difference yields a coarse estimate of the current bolus (Figure 1). This suggests that motion cues are strong indicators of bolus presence, even without complex models. Further examination reveals that the bolus and background exhibit different motion blur patterns during swallowing. The bolus moves rapidly through the pharynx while most anatomical structures remain relatively static. If a model can learn the background distribution, the residual should highlight bolus-specific motion.

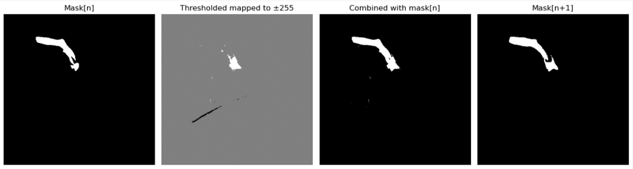

Fig. 1. Motivation for temporal modeling. From left to right: ground truth mask at frame n, thresholded frame difference $|F_{n+1} - F_n|$, combined result (thresholded difference + mask$_n$), and ground truth mask at frame $n + 1$. The simple combination closely approximates the next frame's mask, demonstrating that motion cues are strong even without learning.

Based on this insight, we propose two main approaches (Figure 2). First, a two-stage reconstruction-fusion pipeline predicts a clean version of the current frame from the adjacent frame, then fuses this reconstruction with the original frames to segment the bolus. This design avoids optical flow computation while still capturing motion information through reconstruction error. The clean frame stage uses masked loss functions that down-weight bolus pixels, encouraging the model to learn background patterns. The fusion stage then combines the clean prediction with input frames to localize the bolus. Second, we test a temporal model with Context Modules that directly process short frame sequences and output segmentation masks. This architecture aggregates temporal features at multiple scales using lightweight operations.

We also explore architectural variants within these frameworks. For the two-stage approach, we test different clean-stage backbones including small UNets, convolutional autoencoders, and ViT-style models inspired by recent deblurring work [7]. The ViT models use patch-based processing and show that transformer architectures can work for frame reconstruction without pre-training. For fusion heads,

- *Y. Chen is with the Department of Electrical and Computer Engineering, University of Toronto.*
  E-mail: yuanhan.chen@mail.utoronto.ca
  Code: https://github.com/oscarchen178/bolus_seg

we compare attention mechanisms, UNet++ [8] with dense skip connections, and windowed attention blocks.

This work makes three contributions. First, we benchmark lightweight temporal models on a consistent VFSS dataset split, showing where temporal cues help and where they reach diminishing returns. Second, we demonstrate that clean-frame reconstruction provides an effective way to expose motion without optical flow or heavy transformers. Third, we provide detailed metrics, architectural descriptions, and release all checkpoints for reproducibility.

2 Related Work

Medical Image Segmentation. UNet [5] introduced skip connections between encoder and decoder layers, becoming the de facto architecture for medical segmentation due to its ability to capture multi-scale features while maintaining a modest parameter count. UNet++ [8] extended this design with nested skip connections and deep supervision, improving gradient flow during training. More recently, transformer-based models like TransUNet [9] replaced convolutional encoders with vision transformers to capture long-range dependencies, though this increases model size and computational cost.

VFSS Bolus Segmentation. Li et al. [1] systematically compared UNet and UNet++ architectures with various encoders including MobileNetV2 [10], ResNet, and InceptionResNetV2 on VFSS images. Their best model achieved DSC 0.811 and IoU 0.683 but processed only single frames. Caliskan et al. [2] applied Mask R-CNN [11] for bolus detection, achieving mean average precision of 0.49, but bounding boxes provide less precise masks than pixel-level segmentation. Park et al. [3] proposed PECI-Net, which uses preprocessing ensembles and cascaded inference to improve robustness, demonstrating that multi-stage pipelines can be effective for bolus segmentation.

Temporal Video Segmentation. Video-TransUNet [4] and Video-SwinUNet [6] incorporate temporal information through Temporal Context Modules that blend features across frame sequences. Zeng et al. showed that temporal attention significantly improves bolus segmentation, achieving Dice 0.880 on VFSS sequences. However, these models require substantial computational resources due to multi-head attention mechanisms and transformer backbones. Our work explores whether simpler temporal aggregation can provide similar benefits with lower computational cost.

Frame Reconstruction for Motion Analysis. Denoising and deblurring models often predict clean versions of corrupted inputs. Recent work on diffusion models [7] shows that predicting clean data directly can be more effective than predicting noise, especially when data lies on low-dimensional manifolds. We adapt this insight to VFSS, where the bolus represents a moving anomaly against a relatively static background. By training models to reconstruct clean frames that ignore bolus pixels, we create a natural motion cue through reconstruction error.

3 Method

3.1 Problem Formulation

Given a sequence of grayscale VFSS frames $\{F_1, F_2, \ldots, F_T\}$ where each $F_t \in \mathbb{R}^{H\times W}$, we aim to predict a binary bolus mask $M_t \in \{0,1\}^{H\times W}$ for each frame. We explore two main strategies: using frame pairs (F_t, F_{t+1}) with a two-stage reconstruction approach, or processing short clips (F_{t-1}, F_t, F_{t+1}) with a temporal aggregation module. Figure 2 illustrates these two approaches.

3.2 Single-Frame Baseline

As a reference, we re-implement the UNet with MobileNetV2 encoder architecture from Li et al. [1]. This model processes individual frames F_t to predict masks M_t and has approximately 6.63M parameters. We train from scratch without ImageNet pre-training for 40 epochs to provide a single-frame comparison baseline on our dataset split. The model uses BCE and Dice losses:

$$\mathcal{L}_{\text{seg}} = \text{BCE}(M_t, \hat{M}_t) + \lambda_{\text{dice}}\mathcal{L}_{\text{dice}}(M_t, \hat{M}_t) \quad (1)$$

where $\lambda_{\text{dice}} = 0.5$ and the Dice loss is:

$$\mathcal{L}_{\text{dice}} = 1 - \frac{2\sum_i M_t^i \hat{M}_t^i}{\sum_i (M_t^i)^2 + \sum_i (\hat{M}_t^i)^2} \quad (2)$$

3.3 Two-Stage Reconstruction and Fusion

This approach consists of two networks trained sequentially. The clean-stage network predicts a reconstruction $\hat{F}_t$ of frame F_t from the adjacent frame F_{t+1}:

$$\hat{F}_t = \text{CleanNet}(F_{t+1}) \quad (3)$$

The reconstruction is trained with a masked L_1 loss that down-weights bolus pixels:

$$\mathcal{L}_{\text{clean}} = \sum_i (1 - M_t^i)|F_t^i - \hat{F}_t^i| \quad (4)$$

This encourages the model to learn the background distribution while ignoring bolus regions. The resulting reconstruction error $|F_t - \hat{F}_t|$ naturally highlights moving bolus pixels.

The fusion-stage network takes the original frame and the clean prediction as input:

$$\hat{M}_t = \text{FusionNet}([F_t, \hat{F}_t]) \quad (5)$$

where $[\cdot]$ denotes channel concatenation. The fusion network is trained with the segmentation loss in Eq. (1). After fusion training, both networks can be jointly fine-tuned end-to-end.

Our best-performing two-stage implementation uses SmallUNet architectures for both stages. Each SmallUNet has four encoder-decoder levels with feature dimensions [16, 32, 64, 128] and double-convolution blocks, totaling approximately 1.29M parameters per network. However, we explore various architectural choices for both clean and fusion stages, as detailed in Section 3.6.

3.4 Temporal Context Module

Instead of two stages, the temporal model directly processes a three-frame sequence to predict the middle frame's mask:

$$\hat{M}_t = \text{TemporalNet}([F_{t-1}, F_t, F_{t+1}]) \quad (6)$$

The network uses Temporal Context Modules (TCM), inspired by Video-TransUNet [4], at multiple encoder scales.

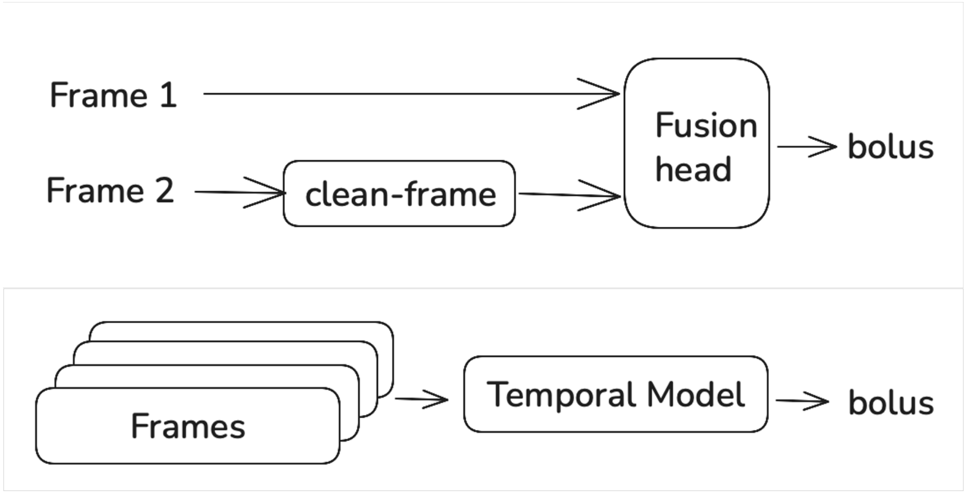

Fig. 2. Overview of our two main approaches. Top: Two-stage reconstruction-fusion pipeline where the clean stage learns to predict frame F_t from F_{t+1} using masked losses, then the fusion stage combines F_t and the clean prediction to segment the bolus. Bottom: Temporal Context Module (TCM) approach that directly processes a sequence of frames with lightweight temporal aggregation at multiple scales.

However, unlike the multi-head attention mechanism in Video-TransUNet, our lightweight TCM computes simple weighted combinations of features across frames:

$$H_t = \alpha_0 E_t + \alpha_1 (E_{t-1} + E_{t+1}) \tag{7}$$

where E represents encoder features and α are learned scalar weights. This simple aggregation avoids expensive dot-product attention while still sharing temporal information. Depthwise separable convolutions throughout the architecture keep the total parameter count near 0.6M, compared to the much larger Video-TransUNet.

3.5 Architectural Variants

Within the two-stage framework, we systematically explore different architectural choices for both clean and fusion stages. For the clean stage, we test: (1) a convolutional autoencoder with 0.42M parameters, (2) a TinyUNet with three levels (features [16, 32, 64] or [32, 64, 128]) ranging from 0.48M to 1.93M parameters, (3) a ViT-style encoder-decoder with patch size 16, embedding dimension 64, depth 6, and 4 heads (0.46M parameters), and (4) a smaller ViT variant with patch size 32, depth 4, and 0.37M parameters. We train clean models with either standard L1 loss, masked L1 that ignores bolus pixels, or L1 combined with an SSIM-like term.

For the fusion stage, we compare: (1) channel-spatial attention blocks (AttentionBolusNet, 0.029M parameters), (2) UNet++ with nested skip connections and deep supervision (2.07M parameters), (3) windowed self-attention with 8×8 windows (0.23M parameters), and (4) cross-frame attention where queries come from F_t and keys-values from $\hat{F}_t$ (0.35M parameters). These combinations allow us to assess the impact of model capacity, inductive bias, and architectural complexity on both reconstruction quality and downstream segmentation performance.

3.6 Training Details

All segmentation models use batch size 4 and train for 20-40 epochs depending on complexity. We use Adam or AdamW optimizers with learning rates of 1×10^{-3} for fusion heads and 5×10^{-4} for clean reconstructions. Some variants add a small SSIM-like term to the reconstruction loss. When jointly fine-tuning two-stage models, we reduce the clean-stage learning rate to 1×10^{-4} while keeping the fusion-stage rate at 1×10^{-3}.

3.7 Deterministic Baseline

We retain a exploratory non-learning pipeline as a reference point. The pipeline applies CLAHE, homomorphic high-pass filtering, and total variation denoising for preprocessing. Motion features include black top-hat morphology, temporal difference with exponential moving average, Farnebäck optical flow magnitude, multi-scale Laplacian edges, and a corridor mask from per-pixel standard deviation. Features are combined through a logistic function with hand-tuned weights, followed by morphological post-processing. Evaluated on the first 24 annotated frames from the notebook run, it reaches IoU 0.273, Dice 0.425, precision 0.319, and recall 0.726. The low scores and limited coverage show how much accuracy the learned models gain over deterministic priors, so we keep this baseline only for qualitative context.

4 Experimental Setup

Dataset. We use 6,424 VFSS frames from 270 swallow studies spanning 87 unique patient sequences. Each frame is a 512×512 grayscale image with a corresponding binary bolus mask annotated by trained raters. Filenames encode a 6-character sequence identifier and frame index to enable sequence-based splitting. For frame-pair experiments, we extract 6,337 consecutive pairs. We split sequences (not individual frames) into train, validation, and test sets to prevent data leakage, yielding 4,315 training pairs, 795 validation pairs, and 1,227 test pairs from 60, 13, and 14 sequences respectively. For single-frame experiments, we use 4,821 training frames, 856 validation frames, and 747 test frames. Data augmentation consists only of random horizontal and vertical flips during training to preserve anatomical structure.

Implementation. All models are implemented in PyTorch and trained on either Apple Silicon with MPS acceleration or an NVIDIA RTX 4090 GPU. Images are normalized to [0,1] range. We use batch size 4 across all experiments. Two-stage models train the clean stage first for 40 epochs, then train the fusion stage for 40 epochs, followed by optional joint fine-tuning for 10 epochs. The temporal model trains for 20 epochs. Optimization uses Adam with $\beta_1 = 0.9, \beta_2 = 0.95$ and no weight decay. We apply learning rate 1e-3 for segmentation heads and 5e-4 for reconstruction networks. During joint fine-tuning, the clean stage uses learning rate 1e-4 while the fusion stage maintains 1e-3.

Evaluation Metrics. We report three main metrics on the test set. Intersection over Union (IoU) measures the ratio of overlap to union between predicted and ground truth masks. Dice coefficient weights true positives more heavily than IoU. Area Under the ROC Curve (AUC) evaluates discrimination ability across all thresholds. All reported results use a fixed threshold of 0.5 on model outputs. For reconstruction models, we report masked L_1 error (computed only on non-bolus pixels) and Peak Signal-to-Noise Ratio (PSNR) in decibels.

5 Results

5.1 Main Results

Table 1 presents test set performance for all models. Our re-implementation of the single-frame MobileNetV2 UNet architecture from Li et al. [1], trained from scratch for 40 epochs without ImageNet pre-training, achieves IoU 0.542 and Dice 0.669 on our dataset split. This demonstrates that a stronger encoder can compensate for missing temporal information. Our lightweight temporal models do not surpass this baseline but achieve competitive results with substantially fewer parameters. The two-stage SmallUNet pipeline reaches IoU 0.519 and Dice 0.628 with AUC 0.995, while the Temporal Context Module UNet achieves IoU 0.515 and Dice 0.644 with AUC 0.996. The temporal model's higher Dice despite slightly lower IoU suggests it produces fewer false positives, as reflected in its excellent AUC of 0.996.

The ViT-based two-stage model achieves IoU 0.486 and Dice 0.609, showing that transformer architectures can work for this task but do not outperform convolutional models at this scale and parameter count. The cross-frame attention variant reaches IoU 0.448 and Dice 0.578, trailing other fusion strategies. For reference, the exploratory deterministic pipeline (evaluated on a 24-frame subset) achieves IoU 0.273, Dice 0.425, precision 0.319, and recall 0.726, highlighting the gap between learning-free and learned approaches and the much smaller evaluation set used for the deterministic run.

5.2 Reconstruction Quality

Table 2 shows clean-stage reconstruction metrics. Better reconstruction correlates with improved downstream segmentation, though the relationship is not linear. The best reconstruction comes from a small UNet with masked L1 loss, achieving masked L_1 of 0.0055 and PSNR 41.18 dB. The ViT clean model with masked training reaches masked L_1 of 0.0077 and PSNR 33.37 dB, significantly better than the unmasked ViT variant (masked L_1 0.0112, PSNR 31.36 dB). This demonstrates that masking the loss to ignore bolus pixels is crucial for learning background distributions.

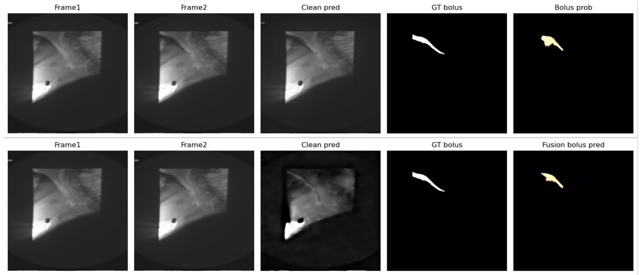

Fig. 3. Qualitative comparison of clean-stage reconstructions. Top: UNet-based clean model produces sharp, clear reconstruction with well-preserved anatomical details. Bottom: Attention-based clean model produces notably blurred reconstruction, losing fine structural information. The sharper UNet reconstruction provides better motion cues for the downstream fusion stage.

TABLE 1
Test segmentation results (threshold 0.5). Deterministic pipeline uses 24 annotated frames.

Model	IoU	Dice	AUC
MobileNetV2 UNet (single frame)	0.542	0.669	–
Two-stage clean + UNet fusion	0.519	0.628	0.995
Temporal Context UNet (seq len 3)	0.515	0.644	0.996
ViT clean + windowed fusion	0.486	0.609	0.988
UNet clean + UNet++ fusion	0.471	0.602	0.990
Masked AE clean + attention fusion	0.468	0.595	0.985
Cross-frame attention fusion	0.448	0.578	0.989
Deterministic priors (24 frames, no learning)	0.273	0.425	–

Comparing fusion strategies, UNet++ with nested skip connections achieves IoU 0.471 and Dice 0.602, slightly outperforming attention-based fusion (IoU 0.468, Dice 0.595). This suggests that dense connections between encoder and decoder features are more effective than lightweight attention at this model size. However, the gap is small, and attention-based fusion requires far fewer parameters (0.029M vs. 2.07M).

5.3 Qualitative Analysis

Figure 3 compares clean-stage reconstructions from two architectural choices: a UNet-based clean model versus an attention-based clean model. The UNet reconstruction (top row) produces sharp, clear frames with well-preserved anatomical structure, achieving masked L_1 of 0.0055 and PSNR 41.18 dB. In contrast, the attention-based reconstruction (bottom row) appears significantly blurred, with loss of fine structural details, achieving masked L_1 of 0.0059 and PSNR 38.15 dB. This blurring in the attention model's output reduces the quality of motion cues available to the fusion stage, explaining its lower downstream segmentation performance (Dice 0.595 vs. 0.628 for the UNet-based pipeline). The sharper UNet reconstruction better captures the background distribution, making bolus regions more distinctive through reconstruction error.

TABLE 2
Clean reconstruction metrics (test).

Clean model	Masked L_1	PSNR (dB)
Tiny UNet (masked)	0.0055	41.18
UNet (L1 + SSIM-like)	0.0056	40.23
Conv autoencoder (masked)	0.0059	38.15
ViT clean (masked)	0.0077	33.37
ViT clean (unmasked)	0.0112	31.36

Reconstruction quality is crucial for the two-stage approach. The masked Tiny UNet reconstruction enables the best two-stage performance, showing that accurate background modeling directly improves bolus segmentation. The ViT model's lower reconstruction quality (PSNR 33.37 dB vs. 41.18 dB for the best UNet) translates to weaker downstream segmentation, though it still substantially outperforms the weaker ablations.

Qualitatively, the two-stage models produce sharp bolus boundaries where reconstruction error is high, effectively using motion cues without explicit optical flow computation. The temporal UNet generates smoother and more temporally consistent masks across frame sequences, suggesting successful temporal feature aggregation. The deterministic baseline, evaluated on the 24-frame subset, captures gross bolus location through morphological operations and flow magnitude but misses fine structural details, explaining the large performance gap to learned approaches even before accounting for the smaller evaluation set. The failed frame-difference model produced mostly uniform predictions, unable to distinguish bolus regions from background noise in difference images.

The frame-difference experiment used a dataset of 4,761 training, 843 validation, and 733 test difference images computed from consecutive frame pairs. Despite using appropriate architectures and loss functions, the model failed to learn meaningful segmentation patterns. This failure underscores that naive differencing without access to original frame information is insufficient, even with strong supervision from ground truth masks.

6 Discussion

6.1 Effectiveness of Temporal Models

Temporal information consistently improves performance within lightweight models. The temporal UNet achieves Dice 0.644, and the two-stage pipeline reaches Dice 0.628 despite using only simple SmallUNet architectures with 1.29M parameters each. Both approaches demonstrate that temporal cues can be exploited without heavy transformer attention mechanisms or optical flow computation.

However, encoder capacity remains crucial. Our re-implementation of the MobileNetV2 UNet architecture from Li et al. [1], trained from scratch for 40 epochs, achieves the highest Dice (0.669) with 6.63M parameters, outperforming our best temporal model (Dice 0.644). This shows that a stronger encoder can compensate for missing temporal information. The advantage of temporal models becomes clearer when comparing models of similar capacity: the temporal UNet (0.6M parameters) and two-stage pipeline (2.58M total) provide the best accuracy among sub-3M parameter options. The key insight is that temporal modeling provides the most benefit when model capacity is limited.

6.2 Two-Stage Design Insights

The clean-frame reconstruction approach works because it exploits the different motion characteristics of bolus and background. By training with masked losses that ignore bolus pixels, the clean stage learns to predict static or slowly-moving anatomical structures. The reconstruction error $|F_t - \hat{F}_t|$ naturally highlights the rapidly-moving bolus without requiring explicit motion modeling. This design is conceptually similar to background subtraction but learns the background model from data rather than using simple temporal averaging.

The choice of clean-stage architecture significantly affects reconstruction quality and downstream performance. UNet-based clean stages produce sharp, detailed reconstructions with well-preserved anatomical structure (Figure 3, top). In contrast, attention-based clean models produce noticeably blurred outputs that lose fine structural details (Figure 3, bottom). This blurring degrades the quality of motion cues available to the fusion stage, directly impacting final segmentation accuracy. The convolutional inductive bias in UNets appears better suited for capturing local anatomical patterns than the global receptive fields in attention mechanisms at this model scale.

For fusion strategies, UNet++ slightly outperforms attention-based fusion despite requiring 70× more parameters. This suggests diminishing returns from architectural complexity at this dataset size. The attention-based fusion provides a good efficiency-accuracy tradeoff for deployment scenarios with limited computational resources.

6.3 Clinical Implications

All learned models achieve AUC above 0.98, indicating they separate positive and negative pixels well when sweeping thresholds, but their Dice/IoU remain in the 0.51–0.67 range. This level of overlap is below what would be needed for clinical deployment and should be interpreted as research-stage performance. The temporal models produce more temporally consistent predictions that reduce flickering, yet overall mask quality is not sufficient for real-time clinical use without further improvements and validation.

The lightweight architectures (temporal UNet at 0.6M parameters, two-stage pipeline at 2.58M total) are suitable for deployment on standard clinical workstations without specialized hardware. Training times of 20-40 epochs on consumer hardware demonstrate accessibility for researchers without access to large GPU clusters.

However, the single-center dataset and limited fluid viscosity range constrain generalizability. The models were trained on frames from one institution using specific barium preparations. Performance on different centers, equipment, or contrast agents remains unknown and requires validation studies.

6.4 Limitations

Our lightweight temporal models do not surpass our re-implementation of the MobileNetV2 UNet architecture from

Li et al. [1], which achieves Dice 0.669 compared to our best temporal model at Dice 0.644. This demonstrates that encoder capacity remains a critical factor, and simply adding temporal information to small models is insufficient to match larger single-frame encoders. The tradeoff is model size: our temporal approaches use 0.6M to 2.58M parameters compared to 6.63M for the baseline.

Runtime and memory consumption were not systematically measured. While model sizes suggest real-time feasibility, actual inference speed depends on implementation details and hardware. Future work should benchmark frame rates to confirm real-time applicability during live VFSS examinations.

The cross-frame attention variant achieved lower performance (IoU 0.448, Dice 0.578) compared to other fusion strategies, suggesting that this particular attention mechanism may not be well-suited for this task or requires different training procedures to reach optimal performance.

6.5 Future Directions

The clean-frame reconstruction idea could extend to multi-frame prediction, where the model predicts frame F_t from both F_{t-1} and F_{t+1}. This bidirectional reconstruction might provide stronger motion cues. Combining the temporal UNet architecture with a moderately larger encoder like EfficientNet could potentially approach or exceed the single-frame baseline while maintaining temporal consistency.

Alternative temporal aggregation mechanisms warrant exploration. While our simple weighted combination works well, learned attention across frames or 3D convolutions might capture more complex motion patterns. The optimal sequence length (currently 3 frames) also deserves investigation, as longer sequences could provide additional context at the cost of increased computation.

Multi-center validation is essential for clinical translation. Testing on frames from different institutions, equipment models, and patient populations would reveal whether the learned representations generalize or require domain adaptation techniques.

7 Conclusion

We explore lightweight temporal models for VFSS bolus segmentation without requiring heavy transformer architectures or explicit optical flow computation. Our two-stage reconstruction-fusion pipeline achieves Dice 0.628 using simple SmallUNets (2.58M total parameters), while a Temporal Context Module UNet reaches Dice 0.644 with only 0.6M parameters. Both approaches provide strong temporal gains at small model sizes but remain below the single-frame MobileNetV2 baseline, confirming that temporal cues help most when capacity is constrained.

The key insight is that learning background distributions through masked reconstruction naturally exposes motion cues. By ignoring bolus pixels during training, the clean stage learns to predict static anatomical structures, and reconstruction error highlights the moving bolus. This design avoids the computational cost of optical flow while still capturing motion information.

However, encoder capacity remains critical. Our re-implementation of the MobileNetV2 UNet architecture from Li et al. [1], trained from scratch for 40 epochs, achieves Dice 0.669 with 6.63M parameters, outperforming our best temporal model. This demonstrates that stronger encoders can compensate for missing temporal information. Our results suggest that temporal modeling provides the most benefit when model capacity is constrained: for lightweight deployments, temporal architectures offer better accuracy than similarly-sized single-frame models, but larger single-frame encoders remain the top performers when computational resources allow.

All training scripts are publicly released for reproducibility. Future work should validate these approaches on multi-center data and explore combinations of temporal architectures with moderately larger encoders to potentially match or exceed the single-frame baseline while maintaining temporal consistency.

Acknowledgments

The author thanks collaborators and annotators who supported data preparation and experimentation.

References

[1] W. Li, S. Mao, A. S. Mahoney, S. Petkovic, J. L. Coyle, and E. Sejdić, “Deep learning models for bolus segmentation in videofluoroscopic swallow studies,” *Journal of Real-Time Image Processing*, vol. 21, p. 18, 2024.

[2] H. Caliskan, A. S. Mahoney, J. L. Coyle, and E. Sejdić, “Automated bolus detection in videofluoroscopic images of swallowing using Mask-RCNN,” in *2020 42nd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*, 2020, pp. 2173–2177.

[3] D. Park, Y. Kim, H. Kang, J. Lee, J. Choi, T. Kim, S. Lee, S. Son, M. Kim, and I. Kim, “PECI-Net: Bolus segmentation from video fluoroscopic swallowing study images using preprocessing ensemble and cascaded inference,” *Computers in Biology and Medicine*, vol. 172, p. 108241, 2024.

[4] C. Zeng, X. Yang, M. Mirmehdi, A. M. Gambaruto, and T. Burghardt, “Video-transunet: Temporally blended vision transformer for CT VFSS instance segmentation,” in *Proc. SPIE International Conference on Machine Vision*, vol. 12701, 2023, p. 127010D.

[5] O. Ronneberger, P. Fischer, and T. Brox, “U-net: Convolutional networks for biomedical image segmentation,” in *Medical Image Computing and Computer-Assisted Intervention*, 2015, pp. 234–241.

[6] C. Zeng, X. Yang, D. Smithard, M. Mirmehdi, A. M. Gambaruto, and T. Burghardt, “Video-SwinUNet: Spatio-temporal deep learning framework for VFSS instance segmentation,” in *2023 IEEE International Conference on Image Processing (ICIP)*, 2023, pp. 2470–2474.

[7] T. Li and K. He, “Back to basics: Let denoising generative models denoise,” *arXiv preprint arXiv:2511.13720*, 2025.

[8] Z. Zhou, M. M. Rahman Siddiquee, N. Tajbakhsh, and J. Liang, “UNet++: A nested U-Net architecture for medical image segmentation,” *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support*, pp. 3–11, 2018.

[9] J. Chen, Y. Lu, Q. Yu, X. Luo, E. Adeli, Y. Wang, L. Lu, A. L. Yuille, and Y. Zhou, “Transunet: Transformers make strong encoders for medical image segmentation,” in *Medical Image Computing and Computer-Assisted Intervention*, 2021.

[10] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen, “MobileNetV2: Inverted residuals and linear bottlenecks,” in *IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 4510–4520.

[11] K. He, G. Gkioxari, P. Dollár, and R. Girshick, “Mask R-CNN,” in *IEEE International Conference on Computer Vision*, 2017, pp. 2961–2969.

Yuanhan Chen Biography text omitted for brevity.