Temporal Representation Learning for Ultrasound Analysis using Masked Modeling

Yves Stebler [1] **Thomas M. Sutter** [*1] **Ece Ozkan** [*1 2] **Julia E. Vogt** [1]

Abstract

Ultrasound (US) imaging is a critical tool in medical diagnostics, offering real-time visualization of physiological processes. One of its major advantages is its ability to capture temporal dynamics, which is essential for assessing motion patterns in applications such as cardiac monitoring, fetal development, and vascular imaging. Despite its importance, current deep learning models often overlook the temporal continuity of ultrasound sequences, analyzing frames independently and missing key temporal dependencies. To address this gap, we propose a method for learning effective temporal representations from ultrasound videos, with a focus on echocardiography-based ejection fraction (EF) estimation. EF prediction serves as an ideal case study to demonstrate the necessity of temporal learning, as it requires capturing the rhythmic contraction and relaxation of the heart. Our approach leverages temporally consistent masking and contrastive learning to enforce temporal coherence across video frames, enhancing the model's ability to represent motion patterns. Evaluated on the EchoNet-Dynamic dataset, our method achieves a substantial improvement in EF prediction accuracy, highlighting the importance of temporally-aware representation learning for real-time ultrasound analysis.

1. Introduction

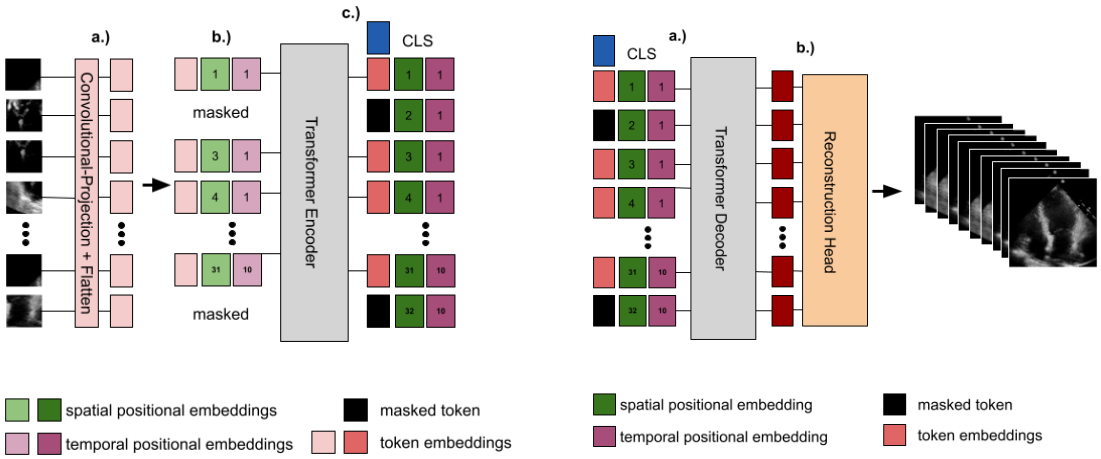
Ultrasound (US) imaging is one of the most widely used diagnostic tools in medicine due to its non-invasive nature, real-time feedback, and cost-effectiveness. It enables clinicians to visualize internal structures dynamically, making it indispensable for a wide range of medical applications, including obstetrics, cardiology, and emergency care (Jensen, 2007; Edler & Lindström, 2004). One of the primary advantages of ultrasound is its ability to capture temporal information—sequential images over time that reflect physiological processes in real-time. This temporal aspect is crucial for assessing organ motion, blood flow, and dynamic physiological events. These applications rely heavily on understanding motion and changes over time, which are not easily captured by static image-based models (Thomas & Popović, 2006). To fully harness the temporal richness of ultrasound, it is critical to learn effective temporal representations that encode motion patterns and sequential dependencies.

Recent advancements in self-supervised learning have introduced Masked Autoencoders (MAEs), which have demonstrated strong capabilities in learning spatial representations by reconstructing masked input data (He et al., 2022). MAEs achieve this by randomly masking portions of an image and training the model to predict the missing parts, effectively learning rich feature representations in an unsupervised manner. However, their application has largely been restricted to frame-level analysis, where each frame is treated as an independent sample. This approach overlooks the sequential and continuous nature of ultrasound imaging, where physiological changes evolve smoothly over time. To extend MAEs for video understanding, VideoMAE was recently introduced, applying a similar masked reconstruction concept but optimized for video sequences, allowing for temporal feature extraction (Tong et al., 2022). While VideoMAE improves temporal learning over naive frame-based methods, its current implementations do not fully exploit the unique temporal dynamics of medical video sequences like ultrasound. As a result, these models are limited in their ability to capture temporal dependencies critical for real-time assessment and clinical decision-making (Bertasius et al., 2021). See Appendix A for more discussion on related work.

A prime example of the importance of temporal learning in ultrasound is echocardiography, where the goal is to measure cardiac function by analyzing sequences of heartbeats (Ouyang et al., 2020; Zhang et al., 2018). In particular, Ejection Fraction (EF) estimation, quantifying the percent-

[*]Equal contribution [1]Department of Computer Science, ETH Zurich, Switzerland [2]Department of Biomedical Engineering, University of Basel, Switzerland. Correspondence to: Ece Ozkan <ece.oezkanelsen@unibas.ch>.

Proceedings of the ICML 2025 Workshop on Multi-modal Foundation Models and Large Language Models for Life Sciences, Vancouver, Canada. 2025. Copyright 2025 by the author(s).



(a) Encoder of the Temporal Model.

(b) Decoder of the Temporal Model.

Figure 1. Overview of the Temporal Model. (i) The encoder extracts patches from the input frames, flattens them, and applies a learned spatial and temporal positional embedding to the unmasked patches, while removing masked patches from the sequence. (ii) The decoder reconstructs the original video by filling in masked tokens, reapplying positional embeddings, and passing through a transformer-based reconstruction process.

age of blood ejected from the ventricles during each heartbeat—serves as a critical marker for heart health. Accurate EF prediction demands an understanding of the heart's motion across multiple frames, highlighting the need for effective temporal representation learning.

In this work, we specifically evaluate EF estimation as a case study to showcase the importance of temporal modeling in ultrasound imaging. Our method extends the MAE framework by incorporating temporally-aware mechanisms that enable the model to capture cardiac motion across sequences of frames. By learning temporally-aware representations, our approach significantly improves EF prediction accuracy, demonstrating the potential of temporal representation learning for real-time ultrasound analysis.

2. Methods

Our approach extends the original Masked Autoencoder (MAE) architecture (He et al., 2022) to effectively capture temporal dynamics and spatial features from video-based ultrasound sequences. The key novelty of our approach lies in how we handle temporal information during encoding, allowing the model to learn not just spatial features, but also temporal dynamics critical for medical video analysis.

2.1. Preprocessing and Video-Based Token Extraction

To efficiently encode temporal data, we preprocess the ultrasound video by stacking frames sequentially and dividing each frame into non-overlapping patches. Specifically, each video frame X_t is processed independently, ensuring non-overlapping spatial regions. This results in a set of spatial patches P_t for frame $t \in [1, T]$, where T is the number of frames. These patches are then flattened and linearly embedded into the encoder's latent space, preserving the spatial and temporal structure across frames. The entire video sequence is represented as a flattened tensor of shape $X_{\text{flattened}} \in \mathbb{R}^{(T \cdot N) \times D}$, where N is the total number of patches per frame, and D is the embedding size. This structured representation allows the transformer to learn dependencies both within individual frames and across the temporal axis.

To explicitly encode temporal relationships, we introduce *temporal-positional embeddings* that encode both the frame order and patch positions as $E_t = E_{\text{pos}}(t) + E_{\text{time}}(t)$, where $E_{\text{pos}}(t)$ represents the spatial position within the frame and $E_{\text{time}}(t)$ encodes the temporal sequence across frames.

2.2. Frame-wise Random Masking

Unlike traditional tube-like masking strategies (Tong et al., 2022; Kim et al., 2024), we adopt a *frame-wise random masking strategy*. For each frame t, we randomly select a subset of patches to be masked

$$M_t = \text{Mask}_t(P_t), \quad \forall t \in [1, T]. \tag{1}$$

The randomness in masking is applied independently across frames t, ensuring that different spatial regions are masked over time. This design encourages the model to reconstruct not only the missing patches but also the motion dynamics that link frames, reinforcing spatiotemporal learning during pretraining.

2.3. Pretraining Objective

Our pretraining process optimizes two complementary objectives.

Table 1. Binary classification results comparing frame-based and temporal models. The temporal model achieves better AUROC and recall, indicating stronger temporal feature capture.

Model Type	Training Mode	Resolution	Model	F1 Score	Recall	Precision	Accuracy	AUROC
Frame-based	Base	32×32	ViT-T	0.87	0.90	0.85	0.80	0.79
	End-to-End	32×32	ViT-T	0.89	0.89	0.83	0.83	0.86
	End-to-End, Oracle	32×32	ViT-T	0.89	0.89	0.88	0.82	0.84
Temporal	Base	32×32	ViT-T	0.88	0.85	0.92	0.80	0.77
	End-to-End	32×32	ViT-T	0.89	0.87	0.90	0.82	0.83
	End-to-End, Oracle	32×32	ViT-T	0.89	0.85	0.93	0.82	0.82
	End-to-End, Contrastive, Oracle	32×32	ViT-T	0.89	0.87	0.92	0.84	0.88

Reconstruction Loss: The reconstruction loss is applied over the masked patches, compelling the model to restore the original spatial details with

$$L_{\text{rec}} = \frac{1}{|M|} \sum_{t=1}^{T} \sum_{i \in M} \|\hat{P}_t[i] - P_t[i]\|^2, \tag{2}$$

where $P_t[i]$ is the original patch at index i in frame t, $\hat{P}_t[i]$ is its reconstructed version, and M is the set of masked patches. This loss encourages high-fidelity reconstruction of local image features while learning robust spatial representations.

Temporal Contrastive Loss: To capture temporal coherence, we compute frame-level representations as the average of all patch tokens in each frame as

$$f_t = \frac{1}{N} \sum_{i=1}^{N} p_t[i] \in \mathbb{R}^D. \tag{3}$$

For any pair of frames $(t, t + \Delta t)$, we calculate their cosine similarity

$$\cos(f_t, f_{t+\Delta t}) = \frac{f_t \cdot f_{t+\Delta t}}{\|f_t\| \|f_{t+\Delta t}\|}. \tag{4}$$

We further derive the cosine distance as

$$d_{t,\Delta t} = 1 - \cos(f_t, f_{t+\Delta t}). \tag{5}$$

The contrastive loss encourages temporally close frames to have similar representations while enforcing a margin for distant frames with

$$L_{\text{contrast}} = \frac{1}{C} \sum_{t=1}^{T} \sum_{\Delta t=1}^{T-t} \begin{cases} d_{t,\Delta t}^2 & \text{if } \Delta t \leq \tau_p \\ [\tau_m - d_{t,\Delta t}]_+^2 & \text{if } \Delta t > \tau_p \end{cases} \tag{6}$$

where τ_p is the threshold for positive temporal consistency, τ_m is the margin for negative temporal separation and C is the total number of temporal comparisons, where

$$C = \sum_{t=1}^{T} (T - t). \tag{7}$$

The overall pretraining objective is a weighted sum of the reconstruction and contrastive losses

$$L_{\text{total}} = L_{\text{rec}} + \lambda L_{\text{contrast}}, \tag{8}$$

where $\lambda \in [0, 1]$ balances the influence of spatial reconstruction and temporal alignment.

2.4. Downstream Tasks

For downstream tasks, we utilize the encoded *CLS token* from the final transformer block in the encoder as the input to a lightweight regression head. This CLS token, enriched with temporal and spatial representations, serves as a summary of the video sequence, enabling high accuracy in clinical predictions.

3. Experiments and Results

3.1. Experimental Setup

Our experimental setup is designed to evaluate the effectiveness of temporal representation learning in real-time ultrasound video analysis, specifically targeting Ejection Fraction (EF) estimation. All experiments are conducted on the EchoNet-Dynamic dataset (Ouyang et al., 2020), which comprises approximately 10,000 echocardiogram videos, each annotated with EF values.

We employ a temporal backbone model that processes sequences of 10 frames per input video, with each frame downsampled to 32×32 resolution for computational reasons. The backbone follows the Vision Transformer (ViT) architecture, using the ViT-Tiny and ViT-Base variants as proposed by (Dosovitskiy et al., 2020; Wu et al., 2022). The frames are uniformly sampled over a one-second interval of the cardiac cycle, ensuring the capture of critical moments such as End Diastolic Volume (EDV) and End Systolic Volume (ESV). This sampling strategy is intended to preserve the temporal structure of heart motion for better feature extraction.

Pretraining is performed using a combination of a masked reconstruction objective and our proposed temporal contrastive loss, which encourages the model to learn both

Table 2. Binary classification results from Zhang et al. (2024) compared with our temporal contrastive loss model.

Model	Dataset	Params	Video-Input	F1	Accuracy	AUROC
VideoMAE	∼200,000	∼98M	$16 \times 224 \times 224$	0.92	0.88	0.91
ECHO-VISION-FM	∼200,000	∼98M	$16 \times 224 \times 224$	0.93	0.89	0.93
Ours (Temporal, End-to-end, Oracle)	∼10,000	∼8M	$10 \times 32 \times 32$	0.89	0.84	0.88

spatial and temporal representations effectively. The model is trained using the AdamW optimizer with a base learning rate of 1.5×10^{-4}, adjusted by the batch size, and a weight decay of 0.05. The learning rate schedule is managed by a LambdaLR scheduler, which applies a warm-up period during the first 200 epochs, followed by a cosine decay. We implement early stopping with a patience threshold of 75 epochs, terminating the training if the reconstruction loss does not improve by at least 5×10^{-5}.

For the downstream classification task, we perform binary classification to distinguish between normal (EF $> 50\%$) and reduced EF (EF $\leq 50\%$). The classification head consists of two fully connected layers of sizes 256 and 128, respectively, and uses the CLS token output from the encoder transformer. The model is evaluated using standard classification metrics, including F1 Score, Recall, Precision, Accuracy, and AUROC.

Training Configurations We evaluate following training configurations:

Base Training: In this setting, the encoder is frozen and only the classification head is trained. This setup serves as a lower-bound baseline to isolate the quality of the learned representations without further fine-tuning.

End-to-End Training: In this configuration, both the encoder and the classification head are jointly optimized during the fine-tuning phase. This allows for simultaneous gradient updates across the entire architecture, enhancing feature extraction and classification alignment.

Contrastive Training: This mode uses our temporal contrastive loss during pretraining to encourage temporal consistency in learned representations, complementing the spatial reconstruction objective.

Oracle Setting: This setup assumes optimal frame selection during pretraining and inference, where frames are perfectly aligned with key cardiac phases, such as systole and diastole. This configuration serves as an upper bound on achievable performance, providing insight into the maximum potential of our temporal modeling approach.

Contrastive + Oracle combines both the contrastive pretraining and oracle-aligned input, reflecting the best-case temporal modeling performance.

3.2. Binary Classification Results

Table 1 summarizes the performance of all evaluated models and training configurations for the binary classification task. Notably, the temporal model trained with our contrastive loss under the oracle setting achieves the highest AUROC of 0.88, outperforming all frame-based counterparts, including those trained end-to-end. This result underscores the importance of modeling temporal dependencies explicitly, as the contrastive objective effectively encourages the model to capture the motion dynamics across frames.

3.3. Comparison with State-of-the-Art

We compare our model with the work of Zhang et al. (2024), as shown in Table 2. For pretraining, the authors utilized 40% of the MIMIC-IV-ECHO dataset (Gow et al., 2023), which comprises approximately 500,000 echocardiogram videos. Their classifier was fine-tuned on the EchoNet-Dynamic dataset. They also employed an input resolution of 224×224 and processed 16 frames per forward pass using a ViT-B backbone.

Although specific parameter counts are not provided in their work, the authors mention that their configuration closely follows the original ViT-B/16 VideoMAE design, on which our parameter estimation is based. Despite using significantly less training data, a smaller model size, lower input resolution, and fewer frames, our model achieves competitive performance with ECHO-Vision-FM.

4. Discussion

Our experimental results demonstrate that the proposed temporal MAE-based model effectively captures the temporal dynamics of cardiac cycles, outperforming frame-based baselines in EF estimation. By leveraging ViT-Tiny and ViT-Base backbones, our model achieves competitive performance with state-of-the-art methods while requiring significantly less data and computational resources. The introduction of the Temporal Contrastive Loss further enhances the temporal consistency of learned representations, contributing to improved classification accuracy. Although our model shows promising results, future work could explore adaptive frame selection and multi-view echocardiography to further enhance temporal feature extraction. Overall, our findings underscore the importance of temporally-aware self-supervised learning for real-time ultrasound analysis.

References

Bertasius, G., Wang, H., and Torresani, L. Is space-time attention all you need for video understanding? In *Proceedings of the International Conference on Machine Learning (ICML)*, July 2021.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

Droste, R., Cai, Y., Sharma, H., Chatelain, P., Drukker, L., Papageorghiou, A. T., and Noble, J. A. *Ultrasound Image Representation Learning by Modeling Sonographer Visual Attention*, pp. 592–604. Springer International Publishing, 2019. ISBN 9783030203511. doi: 10.1007/978-3-030-20351-1_46. URL http://dx.doi.org/10.1007/978-3-030-20351-1_46.

Edler, I. and Lindström, K. The history of echocardiography. *Ultrasound in Medicine & Biology*, 30(12): 1565–1644, December 2004. ISSN 0301-5629. doi: 10.1016/s0301-5629(99)00056-3. URL http://dx.doi.org/10.1016/S0301-5629(99)00056-3.

Gow, B., Pollard, T., Greenbaum, N., Moody, B., Johnson, A., Herbst, E., Waks, J. W., Eslami, P., Chaudhari, A., Carbonati, T., Berkowitz, S., Mark, R., and Horng, S. Mimic-iv-echo: Echocardiogram matched subset, 2023. URL https://physionet.org/content/mimic-iv-echo/0.1/.

He, K., Chen, X., Xie, S., Li, Y., Dollár, P., and Girshick, R. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 16000–16009, 2022.

Jensen, J. A. Medical ultrasound imaging. *Progress in Biophysics and Molecular Biology*, 93(1–3):153–165, January 2007. ISSN 0079-6107. doi: 10.1016/j.pbiomolbio.2006.07.025. URL http://dx.doi.org/10.1016/j.pbiomolbio.2006.07.025.

Jiao, J., Droste, R., Drukker, L., Papageorghiou, A. T., and Noble, J. A. Self-supervised representation learning for ultrasound video, 2020. URL https://arxiv.org/abs/2003.00105.

Kim, S., Jin, P., Song, S., Chen, C., Li, Y., Ren, H., Li, X., Liu, T., and Li, Q. Echofm: Foundation model for generalizable echocardiogram analysis. *arXiv preprint arXiv:2410.23413*, 2024.

Ouyang, D., He, B., Ghorbani, A., Yuan, N., Ebinger, J., Langlotz, C. P., Heidenreich, P. A., Harrington, R. A., Liang, D. H., Ashley, E. A., and Zou, J. Y. Video-based ai for beat-to-beat assessment of cardiac function. *Nature*, 580(7802):252–256, March 2020. ISSN 1476-4687. doi: 10.1038/s41586-020-2145-8. URL http://dx.doi.org/10.1038/s41586-020-2145-8.

Thomas, J. D. and Popović, Z. B. Assessment of left ventricular function by cardiac ultrasound. *Journal of the American College of Cardiology*, 48(10):2012–2025, November 2006. ISSN 0735-1097. doi: 10.1016/j.jacc.2006.06.071. URL http://dx.doi.org/10.1016/j.jacc.2006.06.071.

Tong, Z., Song, Y., Wang, J., and Wang, L. VideoMAE: Masked autoencoders are data-efficient learners for self-supervised video pre-training. In *Advances in Neural Information Processing Systems*, 2022.

Wu, K., Zhang, J., Peng, H., Liu, M., Xiao, B., Fu, J., and Yuan, L. Tinyvit: Fast pretraining distillation for small vision transformers. In *European conference on computer vision (ECCV)*, 2022.

Zeyu, F., Jianbo, J., Robail, Y., Lior, D., Papageorghiou, A. T., and Alison, N. Anatomy-aware contrastive representation learning for fetal ultrasound. In *European Conference on Computer Vision Workshop*, 2022.

Zhang, J., Gajjala, S., Agrawal, P., Tison, G. H., Hallock, L. A., Beussink-Nelson, L., Lassen, M. H., Fan, E., Aras, M. A., Jordan, C., Fleischmann, K. E., Melisko, M., Qasim, A., Shah, S. J., Bajcsy, R., and Deo, R. C. Fully automated echocardiogram interpretation in clinical practice: Feasibility and diagnostic accuracy. *Circulation*, 138(16):1623–1635, October 2018. ISSN 1524-4539. doi: 10.1161/circulationaha.118.034338. URL http://dx.doi.org/10.1161/CIRCULATIONAHA.118.034338.

Zhang, Z., Wu, Q., Ding, S., Wang, X., and Ye, J. Echo-vision-fm: A pre-training and fine-tuning framework for echocardiogram video vision foundation model. October 2024. doi: 10.1101/2024.10.09.24315195. URL http://dx.doi.org/10.1101/2024.10.09.24315195.

A. Related Work

Research in ultrasound representation learning has explored a range of methods to improve downstream task performance. Traditional approaches predominantly focused on frame-based learning, where spatial features are extracted independently from each frame without consideration of temporal continuity. More recent works, however, have shifted towards learning representations that incorporate temporal dynamics, recognizing the importance of capturing motion and periodicity in ultrasound sequences.

Ultrasound Representation Learning Recent works have demonstrated the advantages of learning robust representations from ultrasound images for downstream tasks. Droste et al. (2019) introduced a self-supervised learning method that pretrains a CNN using a visual-tracking dataset to predict saliency maps corresponding to sonographer focus points. Pretraining with visual-tracking information improved F1 performance on a standard plane-detection task, outperforming direct fine-tuning, highlighting the importance of ultrasound-specific feature learning. Similarly, Zeyu et al. (2022) proposed Anatomy-Aware Contrastive Learning for self-supervised ultrasound representation learning. By grouping positive pairs based on anatomical similarity, their method effectively captured granular information, improving performance in standard plane classification and fetal biometry estimation compared to ImageNet-pretrained models.

Temporal Representation Learning While frame-based methods have shown promising results, recent research emphasizes the importance of modeling temporal dynamics for tasks like cardiac monitoring and fetal development analysis. Authors in (Jiao et al., 2020) introduced a self-supervised framework for ultrasound video representation learning that leverages video-frame sequence ordering and geometric transformation as pretext tasks. This approach effectively captured temporal dependencies, outperforming static frame-based baselines. Zhang et al. (2024) proposed ECHO-Vision-FM, which combined VideoMAE with a Spatio-Temporal Feature Fusion Network (STFF-Net). VideoMAE, serving as the backbone, utilized tube-masking strategies to maintain temporal consistency across frames, significantly improving performance on EF prediction tasks. Authors in (Kim et al., 2024) extended this concept with EchoFM, introducing a periodic contrastive loss that enforces temporal consistency within cardiac cycles. This strategy enhanced the learning of periodic heart motion, resulting in improved segmentation and EF prediction accuracy.