# Automated Cardiac Motion Analysis Using the Video Vision Transformer (ViViT): A Regression Approach for Predicting Physiological Parameters

**Joseph E. Cohen**
*Department of Electrical and Computer Engineering, University of Virginia*
*jec8b@virginia.edu*

## Abstract

We utilize the Video Vision Transformer, or ViViT, architecture in order to perform a regression task on a given dataset of cardiac video sequences. This dataset encodes heart motions as 2D x T sequences, including ground truth labels to predict physiological parameters. Our implementation shows results of this regression and its capabilities to analyze spatio-temporal features.

**Keywords:** Cardiac Motion Analysis, ViViT (Video Vision Transformer), Spatio-Temporal Modeling, Physiological Parameter Prediction, Transformer-Based Regression, Cardiac Imaging.

## 1. Introduction

Cardiac motion analysis plays a crucial role in medical diagnostics, enabling the prediction of vital physiological parameters such as heart rate, ejection fraction, and myocardial strain. Traditionally, these evaluations rely on imaging modalities such as echocardiography and magnetic resonance imaging (MRI). While effective, these methods require manual interpretation by trained professionals, making them time-intensive, subjective, and prone to variability. Automating cardiac motion analysis using machine learning offers the potential to improve diagnostic speed, accuracy, and reproducibility.

Recent advancements in machine learning, particularly in deep learning, have enabled the automation of video-based analysis for temporal data. Convolutional neural networks (CNNs) have been the primary method for tasks such as disease classification, heart chamber segmentation, and motion tracking in cardiac imaging. For instance, CNN-based approaches have successfully automated echocardiogram analysis by classifying cardiac abnormalities and segmenting regions of interest (Chen et al., 2019). Similarly, Ozturk et al. (Ozturk et al., 2020) demonstrated the efficacy of CNNs in detecting diseases from medical images. Despite these successes, CNNs are inherently limited in their ability to capture long-range temporal dependencies due to their localized receptive fields, which may lead to suboptimal performance in analyzing spatio-temporal patterns in cardiac video sequences.

Transformer-based models, originally developed for natural language processing tasks (Vaswani et al., 2017), have emerged as state-of-the-art tools for video understanding. The Video Vision Transformer (ViViT) (Arnab et al., 2021) extends transformers to video data by employing multi-head self-attention (MHSA) mechanisms to process both spatial and temporal dimensions simultaneously. Unlike CNNs, ViViT models global dependencies in

video sequences, making them well-suited for tasks requiring fine-grained spatio-temporal analysis. This capability is particularly valuable for cardiac motion analysis, where subtle temporal changes in heart motion are critical for accurate predictions.

Applying ViViT to cardiac video datasets, however, introduces unique challenges. First, medical video datasets are often small due to the high cost and expertise required for data collection and labeling. This scarcity poses a risk of overfitting when training data-hungry transformer models. Second, preprocessing cardiac videos for model input requires careful normalization, resizing, and interpolation to ensure consistency in spatial and temporal dimensions across samples. Finally, while prior work has primarily focused on classification tasks, such as disease detection or risk stratification (Litjens et al., 2017), regression tasks aimed at predicting continuous physiological curves remain underexplored. These regression tasks are essential for more personalized diagnostics, as they provide quantitative insights into patient-specific cardiac function.

This study seeks to address these challenges by adapting the ViViT architecture for regression tasks on cardiac video sequences. By preprocessing the 2D $\times$ T video data to a uniform format and leveraging ViViT's spatio-temporal attention capabilities, we train the model to predict physiological curves with minimal error. Our work demonstrates the feasibility of transformer-based architectures in automating cardiac motion analysis and lays the groundwork for broader applications of deep learning in medical imaging.

## 2. Background

This study primarily bases its methodology on the work presented in *ViViT: A Video Vision Transformer*, which proposes pure transformer-based models for video classification tasks (Arnab et al., 2021).

### 2.1. Existing Methods

In implementing this transformer-based model for video classification tasks, it was necessary to first explore the existing methods employed in terms of image processing: convolutional neural networks (CNNs) and transformer-based approaches.

#### 2.1.1. Convolutional Neural Networks (CNNs)

Convolutional Neural Networks (CNNs) have become the backbone of many computer vision tasks due to their ability to automatically learn hierarchical features from input data. In the context of images, CNNs excel at capturing spatial patterns such as edges, textures, and shapes, which are critical for tasks like object detection and classification. When extended to video, CNNs face the added challenge of learning not only spatial features but also temporal dynamics across frames. To address this, CNN-based models for video typically rely on two-stream networks, where separate CNNs process RGB frames and optical flow, capturing both appearance and motion information. Additionally, 3D CNNs extend the traditional 2D convolution operation by introducing a third dimension to model spatiotemporal relationships directly.

While these techniques have shown strong performance in video analysis, they come with significant downsides, including high computational cost, the need for large amounts

of labeled data, and challenges in capturing long-range temporal dependencies, especially with smaller video datasets.

### 2.1.2. Vision Transformer (ViT)

A Transformer, in the context of computer vision tasks, is a sequence-to-sequence deep learning model that utilizes an attention mechanism (Bi et al., 2021). It processes an input sequence and generates an output sequence, focusing on different parts of the input depending on the relevance of various elements in the query.

The Vision Transformer (ViT) adapts the transformer architecture for image analysis by splitting 2D images into non-overlapping patches, projecting them linearly, and converting them into 1D tokens. These tokens, along with various positional embeddings, are processed through a transformer encoder consisting of multiple layers with Multi-Headed Self-Attention (MSA), Layer Normalization (LN), and Multi-Layer Perceptrons (MLPs) (Dosovitskiy et al., 2021). In addition to the nominal 1D tokens the model derives, the ViT architecture additionally considers a learned classification token, essentially a global average pooling for classification. This act permits the flexible nature of its use, making it a powerful architecture for image-based tasks when trained on large-scale datasets. The ViT model architecture is displayed in Figure 1.

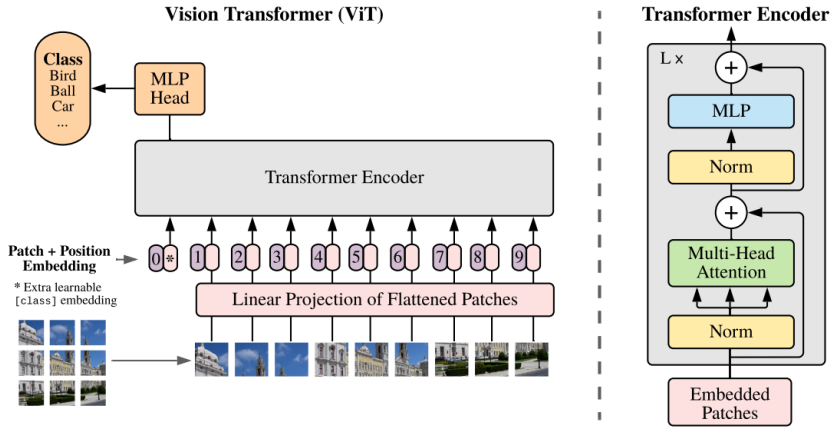


Figure 1: ViT architecture. Model effectively splits an image into fixed-size patches, linearly embeds each of them, adds position embeddings, and feeds the resulting sequence of vectors to a standard Transformer encoder. To perform classification, an extra learnable "classification token" is appended to the sequence.

It is only recently that Vision Transformers (ViTs) have been proven to outperformed traditional convolutional neural networks (CNNs) in image classification tasks (Arnab et al., 2021). Unlike CNNs, which rely on fixed receptive fields to process local features, ViTs divide images into non-overlapping patches and treat them as tokens, allowing the transformer to capture both local and global relationships across the entire image. This ability to model

long-range dependencies through attention enables ViTs to better handle complex spatial relationships in large-scale datasets. As a result, ViTs have surpassed CNNs in performance.

Given the flexibility of the transformer-based model and its ability to operate on any sequence of input tokens, *Arnab et al.* present two notable strategies for tokenizing video: uniform frame sampling and tubelet embedding. Uniform frame sampling independently embeds each 2D frame, while tubelet embedding extracts and linearly projects non-overlapping spatio-temporal tubes into higher-dimensional space, combining both spatial and temporal information during tokenization. The choice of method affects computational complexity, with smaller tubelet dimensions leading to more tokens and increased computation.

### 2.2. Transformer Models for Video

Extending the concept of ViTs to video processing introduces new challenges due to the additional temporal dimension. Transformer models for video, such as the ViViT: A Video Vision Transformer, tackle these complexities with a pure-transformer approach.

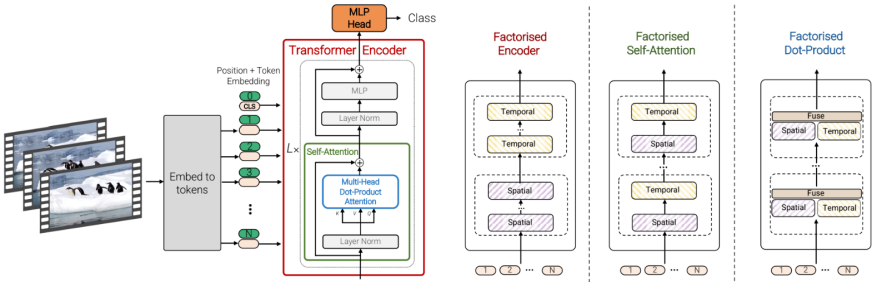


Figure 2: The proposed pure-transformer architecture for video classification, inspired by the recent success of such models for images.)

The ViViT model processes videos by representing them as a sequence of spatio-temporal tokens, capturing both spatial and temporal information through attention mechanisms. This architecture has set a benchmark in video classification, showing that transformers can excel in tasks that demand the integration of spatial and temporal contexts. *Arnab et al.* propose several transformer-based architectures, starting with a direct extension of ViT that models pairwise interactions between all spatio-temporal tokens as seen in Figure 2, alongside other variants (Arnab et al., 2021).

#### 2.2.1. Model 1: Spatio-Temporal Attention

The spatio-temporal attention model in ViViT represents a straightforward approach where all spatio-temporal tokens extracted from a video are passed through a transformer encoder (Arnab et al., 2021). This design allows the model to process global pairwise interactions across all tokens simultaneously, capturing long-range dependencies across both spatial and temporal dimensions from the very first layer. This is a key distinction from CNN architectures, where the receptive field grows incrementally with each added layer.

Building on concepts such as the "Joint Space-Time" model, spatio-temporal attention leverages the transformer's ability to model interactions holistically (Bertasius et al., 2021).

By applying multi-headed self-attention (MSA), this architecture ensures that each layer captures comprehensive spatial and temporal relationships. However, the approach comes with a significant computational cost: MSA has quadratic complexity with respect to the number of tokens. Given that the number of tokens grows linearly with the number of frames in a video, the computational requirements scale rapidly for longer video sequences.

#### 2.2.2. Model 2: Factorized Encoder

To handle the computational complexity of video data, Model 2 employs a factorized encoder: two separate transformer encoders. As seen in Figure 3, this design splits the processing of spatial and temporal dimensions, reducing the computational load while maintaining the model's capacity to learn meaningful representations (Arnab et al., 2021).

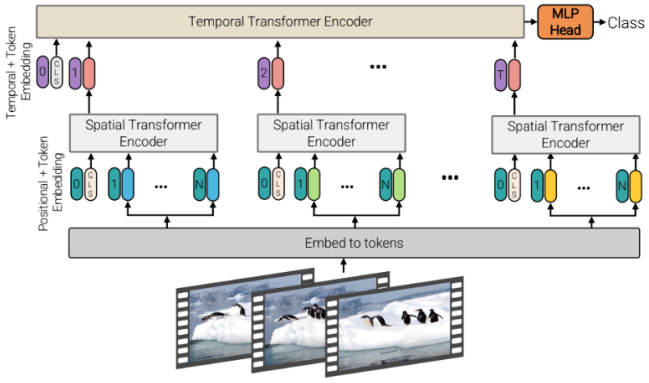


Figure 3: Factorized encoder (Model 2). Utilizes two transformer encoders. The first processes tokens from the same temporal index to create a representation per time step, while the second models interactions across time steps, resulting in a "late fusion" of spatial and temporal information.

Despite having more transformer layers and parameters than Model 1, the Factorized Encoder requires fewer floating point operations (FLOPs), as the two separate transformer blocks have a lesser complexity in comparison to Model 1.

#### 2.2.3. Model 3: Factorized Self-Attention

Model 3 uses the same number of transformer layers as Model 1, but improves efficiency by factorizing the self-attention operation as seen in Figure 4 (Arnab et al., 2021). Instead of computing attention across all pairs of tokens, the model first computes spatial self-attention (within the same temporal index) and then temporal self-attention (within the same spatial index). This reduces computational complexity, achieving the same efficiency as Model 2. The factorized attention is performed by reshaping the tokens and applying multi-headed self-attention sequentially across spatial and temporal dimensions. Unlike Model 1, this model does not use a classification token to avoid reshaping ambiguities.

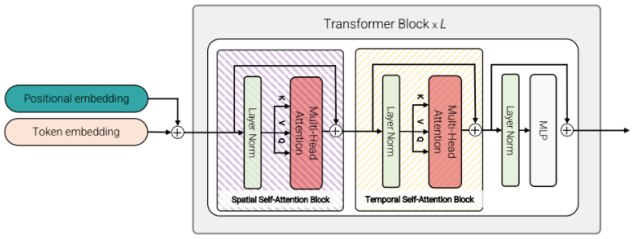


Figure 4: Factorized self-attention (Model 3). In each transformer block, the multi-headed self-attention is split into two separate operations. The first operation only spatially computes, with the second responsible for the temporal self-attention computation.

#### 2.2.4. Model 4: Factorized Dot-Product Attention

The factorised dot-product attention model maintains the same number of parameters as Model 1 but has the computational complexity of Models 2 and 3. This model splits the multi-head dot-product attention into two parts: one focusing on spatial dimensions and the other on temporal dimensions. The attention operation computes separate attention weights for spatial and temporal tokens using distinct sets of keys and values for each query. The outputs of spatial and temporal attention heads are then concatenated and projected, preserving the overall attention dimensions while improving efficiency.

### 2.3. Experimental Results

In order to efficiently train large video classification models, particularly in the absence of massive labeled video datasets, *Arnab et al.* leverage pretrained image models to initialize video models. However, challenges arise when dealing with parameters not present or incompatible between image and video models. *Arnab et al.* discuss several mitigation strageties, including adapting positional embeddings by repeating them temporally to accommodate the higher number of tokens in video data, and methods for initializing embedding weights. Additionally, for transformer-based architectures, *Arnab et al.* itilize the spatial self-attention weights from the pretrained model, while temporal self-attention layers are initialized to zero, functioning as a residual connection at the start.

*Arnab et al.* present these strategies aim to enable video models to learn spatio-temporal relationships effectively, despite the smaller scale of labeled video datasets compared to image datasets. Figure **??** shows that *Arnab et al.* spatio-temporal attention models outperform the state-of-the-art on Kinetics 400 and 600, respectively.

(a) Kinetics 400

| Method | Top 1 | Top 5 | Views | TFLOPs |
|---|---|---|---|---|
| blVNet [19] | 73.5 | 91.2 | – | – |
| STM [33] | 73.7 | 91.6 | – | – |
| TEA [42] | 76.1 | 92.5 | 10 × 3 | 2.10 |
| TSM-ResNeXt-101 [43] | 76.3 | – | – | – |
| I3D NL [75] | 77.7 | 93.3 | 10 × 3 | 10.77 |
| CorrNet-101 [70] | 79.2 | – | 10 × 3 | 6.72 |
| ip-CSN-152 [66] | 79.2 | 93.8 | 10 × 3 | 3.27 |
| LGD-3D R101 [51] | 79.4 | 94.4 | – | – |
| SlowFast R101-NL [21] | 79.8 | 93.9 | 10 × 3 | 7.02 |
| X3D-XXL [20] | 80.4 | 94.6 | 10 × 3 | 5.82 |
| TimeSformer-L [4] | 80.7 | **94.7** | 1 × 3 | 7.14 |
| ViViT-L/16x2 FE | 80.6 | 92.7 | 1 × 1 | 3.98 |
| ViViT-L/16x2 FE | **81.7** | 93.8 | 1 × 3 | 11.94 |
| *Methods with large-scale pretraining* | | | | |
| ip-CSN-152 [66] (IG [44]) | 82.5 | 95.3 | 10 × 3 | 3.27 |
| ViViT-L/16x2 FE (JFT) | 83.5 | 94.3 | 1 × 3 | 11.94 |
| ViViT-H/14x2 (JFT) | **84.9** | **95.8** | 4 × 3 | 47.77 |

Figure 5: Comparisons to state-of-the-art across the Kinetics 400 dataset.

The results highlight the effectiveness of "tubelet embedding" methods and show that initializing models from larger image datasets like JFT provides significant advantages in terms of accuracy, especially for action recognition tasks. Moreover, it is understood that regularization techniques and model architecture variations, further improve the performance of video models across multiple datasets.

## 3. Methodology

### 3.1. Dataset Preprocessing

In preparation for training, we first normalized and standardized the input cardiac cine myocardial mask sequences and corresponding target regression labels derived from time-varying left-ventricular myocardial motion. The raw data, provided as a NumPy array (`2023-11-15-cine-myo-masks-and-TOS.npy`), includes a series of cropped myocardial masks (`cine_lv_myo_masks_cropped`) with dimensions $(T, H, W)$, where $T$ is the number of temporal frames, and $H, W$ represent the spatial resolution. The target output signal (TOS) is a vector of normalized physiological indicators representing cardiac function over time.

To ensure stable optimization and mitigate scale-induced bias, we normalized the myocardial mask intensities to the $[0, 1]$ range. We also applied min-max normalization to the target curves (TOS) to standardize their range. After normalization, we saved the processed datasets into dedicated directories for training and validation (`./dataset/train/` and `./dataset/validation/`). This ensures a consistent and reproducible input pipeline. Any

temporal dimension mismatches were addressed by zero-padding or truncation to achieve a uniform input length across samples.

### 3.2. Embedding Video Clips

Following preprocessing, each preprocessed sequence underwent further transformation to accommodate the ViViT Model 1 input structure. We resized each two-dimensional frame to a target shape (e.g., $(32, 32)$ pixels) and stacked them temporally, forming an input tensor of shape $(64, 32, 32, 1)$ representing 64 frames of single-channel grayscale myocardial masks.

The code systematically converts each sample into a 5D batch-compliant format $(B, T, H, W, C)$ suitable for patch embedding. Specifically, it defines a consistent spatio-temporal resolution and channel ordering to ensure that each sequence can be effectively partitioned into uniform non-overlapping patches. By standardizing frame counts and spatial resolutions, the model's patch extraction process remains stable and uniform across all samples.

### 3.3. ViViT Model 1 Architecture

We adopted the Model 1 variant of the Video Vision Transformer (ViViT) architecture (Arnab et al., 2021), which applies joint spatio-temporal attention on raw token sequences. This model takes as input a series of image patches extracted from each video frame and then flattens and projects them into a shared embedding space, forming a single token sequence that encompasses both space and time dimensions.

The implemented model consists of four primary components:

1. **Patch Embedding:** Each frame is subdivided into $(16 \times 16)$-pixel patches. These patches are linearly projected into a lower-dimensional embedding space. This transforms each spatial-temporal slice of the video into a set of tokenized representations.

2. **Add Temporal and Spatial Positional Embeddings:** A learnable positional embedding is added to the tokens to preserve and convey temporal order and spatial layout. This step is critical, as the standard Transformer attention mechanism does not inherently encode sequence order.

3. **Transformer Encoder Blocks:** A series of Transformer encoder layers (Vaswani et al., 2017), each containing multi-head self-attention and feed-forward MLP layers, is applied to the sequence. In Model 1, the attention mechanism is not factorized; it operates jointly on all spatio-temporal tokens at once. This allows the network to directly learn long-range dependencies and interactions across both time and space from the earliest layers.

4. **Regression Head:** For the downstream regression task, the final "class" token output by the Transformer encoder is passed through a dense layer to predict the normalized target time-series. By leveraging the learned latent representation of the entire video, the regression head directly maps the encoded 4D myocardial dynamics into the output physiological signal.

### 3.4. Training Configuration

We trained the model using an Adam optimizer with a learning rate of $1 \times 10^{-4}$. The training configuration included a batch size of 4 and a total of 200 epochs. Mean Absolute Error (MAE) was employed as both the loss function and the evaluation metric, given the regression-oriented nature of the task. The code snippet employs a standard TensorFlow/Keras `model.fit()` loop with monitored training/validation losses and performance metrics.

For validation, a small but representative subset of the dataset was held out. During training, intermediate checkpoints were saved, and MAE was monitored to guide potential early stopping and model selection. The relatively small batch size reflects memory constraints associated with handling 3D spatio-temporal volumes and Transformer computations.

### 3.5. Model Regularization and Initialization

To reduce overfitting risk, dropout layers at a rate of 0.1 were incorporated into both the attention and MLP components of the Transformer encoder. This encourages robust feature learning and helps the model generalize beyond the training samples.

The network parameters were initialized using Xavier uniform initialization for linear layers, aligning with established practices for Vision Transformers (Arnab et al., 2021). Positional embeddings were also learned from scratch. If available, mixed-precision training was enabled (`set_global_policy("mixed_float16")`) to optimize computational efficiency without compromising model accuracy.

### 3.6. Implementation

The entire methodology is implemented in Python, leveraging TensorFlow/Keras for model construction and training. NumPy and TensorFlow I/O APIs handle data loading and normalization. The provided code defines custom classes for patch embedding, positional embedding, and Transformer encoder blocks, ensuring modularity and clarity. The training loop, integral to the workflow, orchestrates data preprocessing, batching, gradient updates, and logging.

This implementation adheres closely to the principles delineated in ViViT's original description (Arnab et al., 2021), while adapting the approach for a medical context. By employing the Model 1 framework on preprocessed cine myocardial mask volumes, we demonstrate the feasibility of end-to-end video Transformer-based regression for cardiac motion analysis.

## 4. Experiment

### 4.1. Training and Validation Loss over Epochs

We trained the ViViT Model 1 on our normalized cardiac cine myocardial mask dataset for 200 epochs. Throughout the training process, we monitored both the training and validation loss (Mean Absolute Error (MAE) used as the loss function) to assess convergence and detect signs of overfitting.

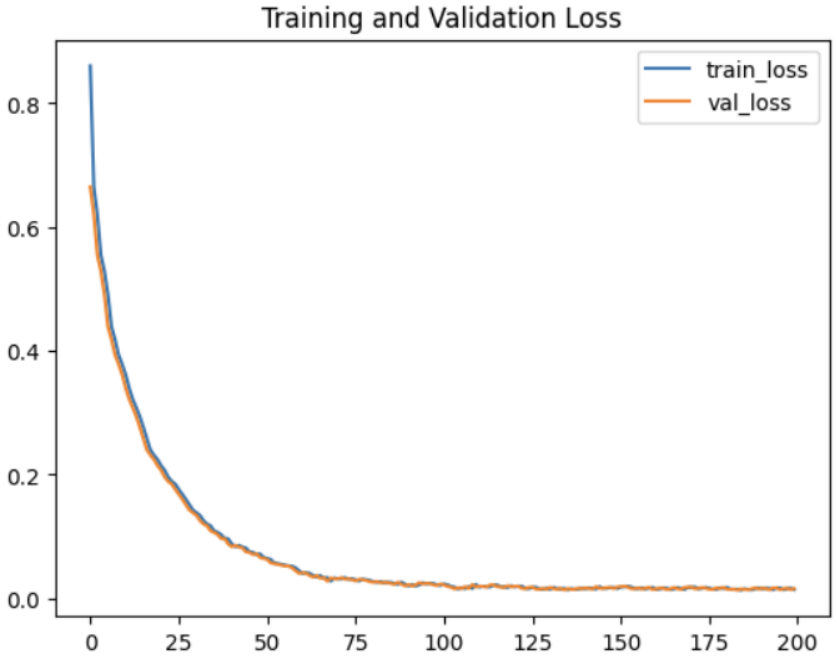


Figure 6: Training and Validation Loss over 200 Epochs.

Figure 6 shows the evolution of loss values over the course of training. Initially, both training and validation losses begin relatively high, reflecting the challenge of learning a mapping from raw spatio-temporal inputs to target physiological signals. As training progresses, the Transformer-based model steadily refines its internal representations of cardiac motion, and the loss curves decrease sharply. By approximately the 50th epoch, both curves show substantial improvement and begin to flatten, indicating that the model is capturing stable, generalized features. Beyond epoch 100, the loss values become increasingly stable and approach near-plateau values, with minimal gap between training and validation loss, suggesting strong generalization and no significant overfitting.

### 4.2. Training and Validation MAE over Epochs

In addition to monitoring loss, we recorded the Mean Absolute Error (MAE) on both the training and validation sets throughout the same 200 epochs. This metric directly measures the average absolute difference between the model's predictions and the ground-truth target signals, providing a more intuitive measure of predictive performance.

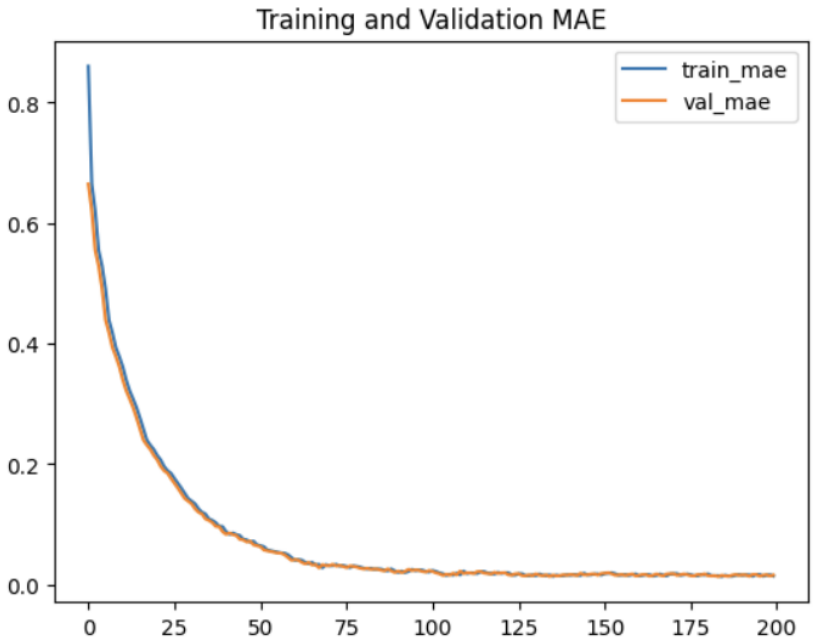


Figure 7: Training and Validation MAE over 200 Epochs.

Figure 7 demonstrates a similar trend to the loss curves. The initial MAE values, above 0.8 at the start, rapidly decrease as the model learns meaningful temporal and spatial patterns within the myocardial mask sequences. By around epoch 50, both training and validation MAE values fall below 0.1. In later epochs, the MAE continues to incrementally improve, eventually approaching $\sim$0.015 on the validation set by epoch 200. The near-overlapping training and validation MAE curves indicate minimal discrepancy between training and validation performance, strongly suggesting that the learned representations generalize effectively to unseen data.

### 4.3. Ground Truth vs. Predictions

To further assess the model's performance, we evaluated the trained model on a validation sample and plotted the predicted time-series against the corresponding ground-truth TOS values.

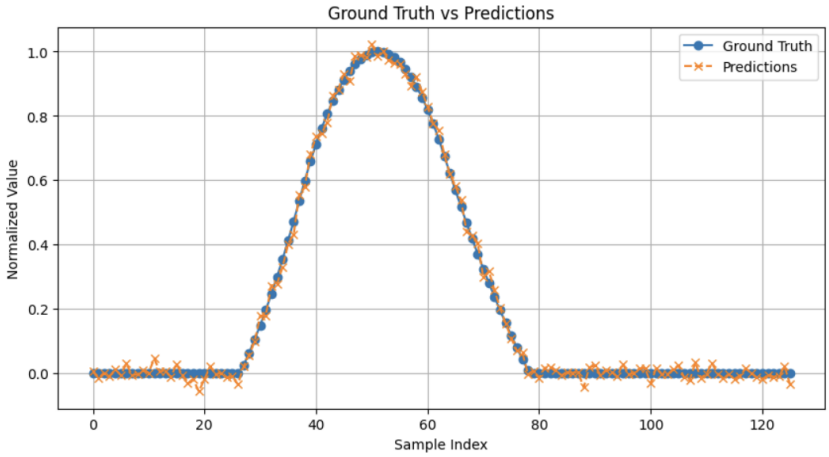


Figure 8: Comparison of Ground Truth and Predicted TOS Values.

Figure 8 shows that the predicted curve closely follows the shape and magnitude of the ground-truth signal, correctly capturing the characteristic bell-shaped pattern of the physiological indicator. The code snippet provided in the terminal output confirms that after inference on a validation sample, the predicted values align well with the true normalized targets. The near-perfect correlation, as visually evident, indicates that the ViViT-based model successfully extracts and represents complex spatio-temporal dependencies in the cardiac motion dataset. The accuracy of these predictions, as measured by metrics such as $R^2$ (close to 1.00) and relative accuracy (∼93.49%), underscores the effectiveness of the model in capturing subtle temporal patterns of cardiac function.

### 4.4. Prediction Residuals

Residual analysis provides a finer-grained examination of the model's predictive performance.

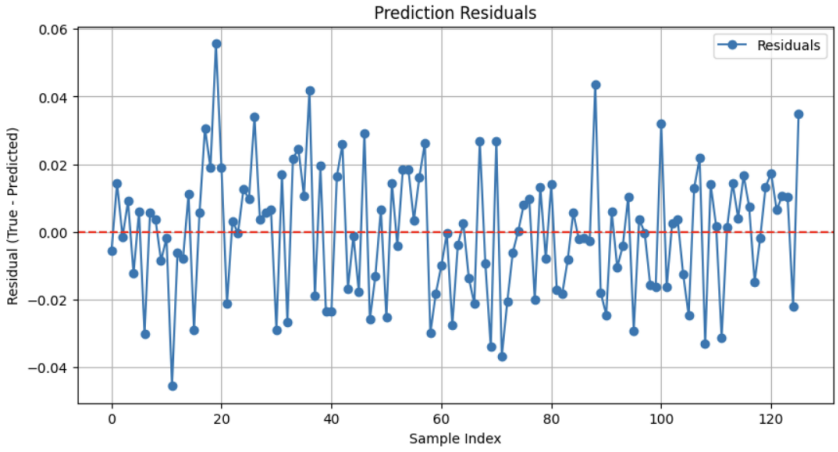


Figure 9: Prediction Residuals for a Validation Sample.

Figure 9 plots the difference between ground truth and predicted values across the temporal indices of a validation sample. The residuals mostly cluster around zero, with only minor positive or negative deviations. This distribution of residuals suggests that errors are small and relatively symmetric, without a pronounced bias at any particular point in the cardiac cycle. While some slight fluctuations are visible, there is no systematic drift or large outliers, indicating that the model's representation and predictions are consistently accurate. Such behavior aligns with the observed reduction in MAE and loss over training and attests to the stable generalization capabilities of the underlying ViViT Model 1 architecture.

Overall, the experiments confirm the initial hypothesis: a pure Transformer-based video model (ViViT Model 1) can effectively learn from normalized myocardial mask sequences to produce highly accurate and stable regression predictions of key physiological signals. The training curves, comparison of ground-truth and predicted values, and residual analysis collectively demonstrate that the model converges steadily and generalizes robustly in this cardiac motion analysis context.

## 5. Discussion

The results of our experiments strongly support the feasibility and effectiveness of employing the ViViT Model 1 architecture for cardiac motion analysis. By adapting this pure-transformer-based video model—originally developed for video classification tasks—to our specialized regression setting, we demonstrated that the model can robustly learn and reproduce complex temporal dynamics of physiological signals from cine myocardial masks. Throughout training, validation, and evaluation, several key themes emerged that underscore both the strengths and potential areas of further exploration for this approach.

From a methodological standpoint, our preprocessing strategies played a crucial role in ensuring model stability and performance. Normalizing both input pixel intensities and target output signals proved essential in stabilizing early training. Moreover, carefully resizing and padding the input sequences to a standardized spatio-temporal dimension allowed for seamless patch embedding, a foundational step in the ViViT pipeline. Such data preparation ensured that the model could directly benefit from the joint spatio-temporal attention mechanism integral to Model 1, where no factorization across space or time is performed. Unlike factorized transformer variants, Model 1 captures full space-time interactions early on, a property that appears advantageous in learning nuanced cardiac kinematics.

The observed training dynamics, evidenced by the rapid and consistent decreases in both loss and MAE, highlight the model's capacity to grasp complex dependencies within just a few dozen epochs. By the 50th epoch, the model had already converged to a low-error regime on both training and validation sets, and the remaining training epochs served primarily to refine and stabilize its internal representations. The near-perfect overlap of training and validation performance metrics by the later epochs suggests a strong generalization capability: the model is not merely memorizing training examples but rather internalizing patterns representative of the underlying cardiac motion processes. This finding contrasts with common concerns in deep learning regarding overfitting, especially in high-dimensional video data domains, and affirms the potential of ViViT-based architectures when combined with suitable normalization and regularization.

The integration of dropout layers and careful parameter initialization provided further assurance against overfitting. By adding a moderate dropout rate (e.g., 0.1) within the Transformer layers, we sustained regularization without compromising the ability of multi-head attention and MLP blocks to extract discriminative features. Meanwhile, adopting standard Xavier-based initializations helped avoid early training instability, ensuring that the gradient updates remained well-conditioned. The results indicate that these well-established yet essential practices can be seamlessly aligned with the ViViT architecture to deliver robust performance in specialized tasks.

Our regression-based evaluations, focusing on the Mean Absolute Error (MAE), confirmed that the model can map raw, normalized myocardial masks to physiological signals with remarkable accuracy. The final validation MAE approached approximately 0.015 by epoch 200—a significant achievement given the complexity of learning cardiac functional curves from raw spatio-temporal image sequences. Furthermore, the ground truth vs. predictions plots demonstrated that the model's predicted curves closely mimic the shape, phase, and amplitude of the true signals, culminating in near-ideal $R^2$ scores and relative accuracy values around 93.5%. Such fidelity in representing physiological signals from purely visual data streams positions the proposed approach as a promising tool for non-invasive cardiac function analysis.

Residual analysis provided another layer of insight. Small, near-zero residuals scattered without discernible systematic patterns indicate that the model's predictions are consistently accurate and unbiased across the time series. The absence of strong temporal patterns in the residuals (e.g., no drift or systematic under- or overestimation at particular points in the cardiac cycle) further underscores the comprehensive temporal understanding encoded by the ViViT-based model.

It is also instructive to compare these findings with the original aims and capabilities of ViViT. The authors of ViViT highlighted that while large-scale datasets were crucial for training pure-transformer architectures on image classification tasks, careful regularization and strong initialization can enable effective training on relatively smaller datasets. Our results in the cardiac domain—where data are typically more limited than in large-scale vision benchmarks—illustrate that this principle extends to regression contexts and specialized medical imaging scenarios. By leveraging the core ViViT principles and tailoring them to our domain and problem setting, we have shown that Transformer-based video models are not restricted to classification: they can excel in predictive tasks that demand fine-grained temporal understanding and continuous value estimation.

Looking ahead, several avenues for future work arise naturally. First, exploring more advanced regularization techniques, such as data augmentation tailored to the medical setting (e.g., spatial deformations that remain anatomically plausible), may further enhance model robustness and adaptability. Second, applying factorized attention variants or employing hybrid convolution-transformer models might yield performance gains in memory-constrained or computationally challenging settings, even though our results suggest that the unfactorized Model 1 is already highly effective. Lastly, translating these methods to other forms of cardiac data (e.g., different MR sequences or echocardiography) or related physiological signals could reveal the full breadth of applicability for ViViT-style architectures.

In conclusion, our comprehensive experimentation and analysis confirm the viability and efficacy of ViViT Model 1 for the regression-based prediction of cardiac motion indi-

cators from cine myocardial masks. The model's ability to learn detailed spatio-temporal representations of the heart's dynamics, coupled with strong generalization and accurate predictions, makes it a compelling candidate for future research and clinical tool development in computational cardiac analysis.

## 6. Conclusion

In this work, we have demonstrated that a pure-transformer-based video model—ViViT Model 1—can be successfully adapted for cardiac motion analysis tasks. Starting from normalized cine myocardial mask sequences, our model effectively learned to predict complex physiological signals, capturing subtle spatio-temporal dynamics without the need for manually engineered features or domain-specific architectures. The resulting model exhibited robust performance, as evidenced by rapid convergence, low MAE, and nearly ideal correlation between predicted and ground-truth time-series. Additionally, residual analyses indicated that errors remained small, unbiased, and evenly distributed over the cardiac cycle, reflecting the model's comprehensive temporal understanding.

The close alignment between training and validation metrics confirmed the model's ability to generalize beyond the training set, alleviating typical concerns about overfitting, particularly in high-dimensional and data-limited medical imaging domains. Regularization strategies, stable initialization methods, and preprocessing steps—grounded in established practices for Transformer models—further contributed to this performance.

Our findings underscore the versatility and promise of transformer-based architectures in biomedical applications that extend beyond classification into highly sensitive, continuous-value regression tasks. Although this study focused on a specific cardiac function signal, the methodology and insights presented here can guide future efforts to harness the full potential of vision transformers for a broader range of clinical imaging analyses. In essence, this work takes a step forward in leveraging advanced deep learning paradigms to better understand and quantify intricate physiological dynamics.

## References

Anurag Arnab, Mostafa Dehghani, Georg Heigold, Chen Sun, Mario Lučić, and Cordelia Schmid. Vivit: A video vision transformer. *arXiv preprint*, 2021.

Gedas Bertasius, Heng Wang, and Lorenzo Torresani. Is space-time attention all you need for video understanding?, 2021. URL https://arxiv.org/abs/2102.05095.

Jiarui Bi, Zengliang Zhu, and Qinglong Meng. Transformer in computer vision. In *2021 IEEE International Conference on Computer Science, Electronic Information Engineering and Intelligent Control Technology (CEI)*, pages 178–188, 2021. doi: 10.1109/CEI52496.2021.9574462.

Ashwin Chen, Melissa Burrows, and Rajesh Patel. Automated classification of echocardiograms using deep convolutional neural networks. *Journal of Medical Imaging*, 6(3): 034001, 2019.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale, 2021. URL https://arxiv.org/abs/2010.11929.

Geert Litjens, Thijs Kooi, Babak Ehteshami Bejnordi, Arnaud Arindra Adiyoso Setio, Francesco Ciompi, Mohsen Ghafoorian, Jeroen A. W. M. van der Laak, Bram van Ginneken, and Clara I. Sánchez. A survey on deep learning in medical image analysis. *Medical Image Analysis*, 42:60–88, 2017.

Tulin Ozturk, M. Talo, E. Yildirim, U. Baloglu, O. Yildirim, and U. Rajendra Acharya. Automated detection of covid-19 cases using deep neural networks with x-ray images. *Computers in Biology and Medicine*, 121:103792, 2020.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 5998–6008, 2017.