Regularizing and Interpreting Vision Transformers by Patch Selection on Echocardiography Data

Alfred Nilsson ALFREDN@KTH.SE
Hossein Azizpour AZIZPOUR@KTH.SE
KTH Royal Institute of Technology, Sweden

Abstract

This work introduces a novel approach to model regularization and explanation in Vision Transformers (ViTs), particularly beneficial for small-scale but high-dimensional data regimes, such as in healthcare. We introduce stochastic embedded feature selection in the context of echocardiography video analysis, specifically focusing on the EchoNet-Dynamic dataset for the prediction of Left Ventricular Ejection Fraction (LVEF). Our proposed method, termed Gumbel Video Vision-Transformers (G-ViTs), augments Video Vision-Transformers (V-ViTs), a performant transformer architecture for videos with Concrete Autoencoders (CAEs), a common dataset-level feature selection technique, to enhance V-ViT's generalization and interpretability. The key contribution lies in the incorporation of stochastic token selection individually for each video frame during training. Such token selection regularizes the training of V-ViT, improves its interpretability, and is achieved by differentiable sampling of categoricals using the Gumbel-Softmax distribution. Our experiments on EchoNet-Dynamic demonstrate a consistent and notable regularization effect. The G-ViT model outperforms both a random selection baseline and standard V-ViT. The G-ViT is also compared against recent works on EchoNet-Dynamic where it exhibits state-of-the-art performance among end-to-end learned methods. Finally, we explore model explainability by visualizing selected patches, providing insights into how the G-ViT utilizes regions known to be crucial for LVEF prediction for humans. This proposed approach, therefore, extends beyond regularization, offering enhanced interpretability for ViTs.

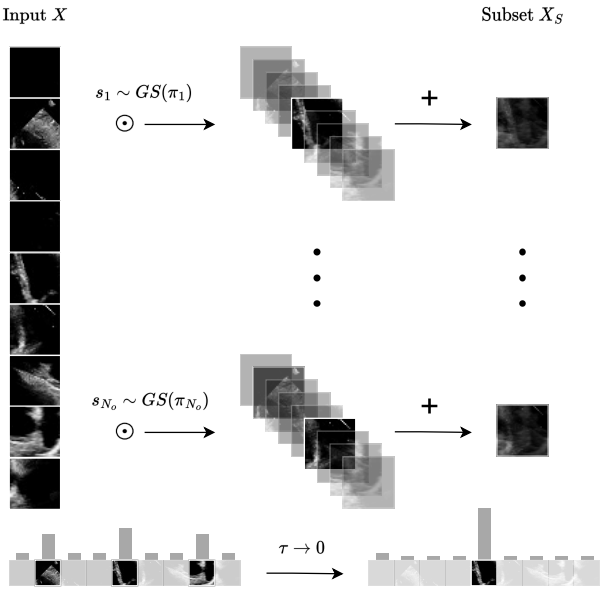


Figure 1: **G-ViT performs per-frame patch selection for regularization and interpretability.** It involves sampling multiple Gumbel-Softmax distributions. For each sample s_i, we multiply its entries with corresponding patches, resulting in N_O linear combinations of patches. As depicted in the lower part of the figure, these selections transition to a discrete state as the temperature τ undergoes annealing towards zero.

Data and Code Availability This work utilizes the public EchoNet-Dynamic dataset which comprises 10,030 labeled echocardiogram videos. Each video captures several cardiac cycles, annotated with critical measurements such as the Left Ventricular Ejection Fraction (LVEF). The dataset, freely available for academic and research purposes, is detailed by Ouyang et al. (2020).

Our code, which encompasses all data preprocessing routines, model implementations, and evaluation pro-

tocols used in our research, is made available as a supplemental zip file. We have taken steps to ensure reproducibility, with precise instructions for how to reproduce the results of this paper.

Institutional Review Board (IRB) Our research does not require IRB approval.

1. Introduction

Human assessment of cardiac function is fundamentally constrained by the limitation that humans can only observe a limited number of cardiac cycles, which results in significant inter-observer variability. In their 2020 paper Ouyang et al. (2020), acknowledged the significance of utilizing machine learning methods to facilitate the assessment of cardiac function. They introduced EchoNet, a factorized 2D + 1D convolutional approach, and to facilitate progress, they made available a dataset named EchoNet-Dynamic consisting of ultrasound videos of hearts and multiple verified human annotations of the Left Ventricular Ejection Fraction (LVEF), a measure of the heart's ability to pump. We develop a novel state-of-the-art architecture for this task based on vision transformers.

Vision Transformers (ViTs) (Dosovitskiy et al., 2021), a recent computer vision architecture, have demonstrated performance that is at least comparable, if not superior, to convolutional networks (Dosovitskiy et al., 2021; Touvron et al., 2020; Liu et al., 2021). Importantly, derivative works of the ViT have applied a strategy of randomly masked training, a technique commonly used with language models, to images (He et al., 2021) for self-supervised learning. Interestingly, ViTs have also demonstrated the ability to generate accurate predictions even when only partial images are observed, particularly for high-resolution medical images (Liu et al., 2023). In this work, inspired by Concrete Autoencoders (CAEs), we aim to automatically *learn to select frame patches during training of video ViTs* to improve their generalization and interpretability.

CAEs employ end-to-end differentiable networks to select discrete features (Abid et al., 2019). The advent of differentiable feature selection techniques opens up the possibility of adapting them for other purposes than selecting dataset-level features. Inspired by regularization techniques such as LASSO (Tibshirani, 1996), that restrict the number of input features for regularization purposes, this work explores employing embedded feature selection as a regularization technique with the ultimate purpose of improving generalization and interpretability.

The impressive track-record of ViTs and their inherent token-based architecture make them suitable for learnable patch-based feature selection. Therefore, their merger with embedded feature selection techniques poses a promising direction which is precisely what we pursue in this work; particularly for the automated prediction of the Left Ventricular Ejection Fraction using echocardiography which has high-dimensional input and relatively low training data.

The contributions of this work are as follows:

- We introduce factorized Vision Transformer architecture for video analysis (V-ViT), establishing a baseline for our experiments. The factorization into spatial and temporal components facilitates the integration of patch token selection on video frames.
- We enhancement model generalization through learned token selection, resulting in the G-ViT model: our proposed integration of this feature selection technique with V-ViT.
- We achieve state-of-the-art performance in end-to-end learned LVEF prediction on the EchoNet-Dynamic dataset with G-ViT.
- We discuss the interpretability of G-ViT, demonstrating how G-ViT's selective focus on informative patches improves model transparency and sheds light on the underlying regularization mechanism.

2. Method

Our proposed method, G-ViT, consists of two key ingredients: (i) the embedded feature selection mechanism and (ii) the patch-based V-ViT architecture. Here, we first describe the underlying feature selection technique that we adopted in section 2.1 and then present our V-ViT architecture and the novel incorporation of patch selection in section 2.2.

2.1. Embedded Feature Selection

Embedded feature selection methods are methods that jointly learn to select features and train a model (Chandrashekar and Sahin, 2014). While classical embedded methods such as LASSO (Tibshirani,

1996) have been extensively studied in linear regression models, far less attention has been devoted to embedded feature selectors for supervised learning with deep learning models. We base our work on a recent approach, called CAE (Abid et al., 2019), which uses Gumbel Softmax distributions.

2.1.1. Gumbel-Softmax

The Gumbel-Softmax (GS) (Jang et al., 2016) and Concrete (Maddison et al., 2016) distributions emerged as solutions to the challenge of embedding stochastic feature selection in models, due to the lack of a differentiable sampling method from categorical distributions. Extending the ideas of the Gumbel-Max trick, a known reparameterization trick for the sampling from categoricals using noise drawn from the Gumbel distribution, they put forth a way to reparameterize the sampling step from their relaxation of the categorical distribution, see Definition 1.

Definition 1 *A sample $\boldsymbol{s}$ can be drawn from a Gumbel-Softmax distribution with categorical probabilities π_j, by perturbing each $\log \pi_j$ with Gumbel noise, where j represents each category. Thus, first samples are drawn $g_j \sim Gumbel(0,1)$ for all j, and then a tempered Softmax transformation is applied to the samples (Jang et al., 2016; Maddison et al., 2016)*

$$s_j = \frac{exp((log(\pi_j) + g_j)/\tau)}{\sum_{j=1}^{k} exp((log(\pi_j) + g_j)/\tau)}, for\ all\ j \quad (1)$$

where τ is a temperature parameter.

This differentiable reparameterization is a highly important result because the distributional parameters can now be learned in an end-to-end fashion.

2.1.2. Concrete Autoencoders

CAEs (Abid et al., 2019) is an embedded feature selection technique that exploits the relaxed sampling procedure in Definition 1. CAE utilizes k GS distributions to select k features from input vectors $\mathbf{x}$ of dimension d, where $k << d$, by drawing a sample $\mathbf{s}_i$ from each distribution with learned parameter vectors $\boldsymbol{\pi}_i \in \mathbb{R}^d$. Taking the dot product of $\mathbf{s}_i$ with a training sample $\mathbf{x}$ yields a "soft" selection of a feature. A soft selection refers to the fact that $\mathbf{s}_i \cdot \mathbf{x}$ will be a convex combination of each input feature, with combination weights determined by $\mathbf{s}_i$. Furthermore, they anneal the temperature τ gradually towards 0 during training, which makes the samples approach one-hot and, therefore, at the end of training, the selections approach discrete samples. By forming a matrix whose rows contain $\{\mathbf{s}_i\}_{i=1}^{k}$ and denoting it by $S \in \mathbb{R}^{k \times d}$, we can express the complete subset selection according to CAE as

$$\mathbf{x}_S = S\mathbf{x}, \text{where } \mathbf{x}_S \in \mathbb{R}^k \quad (2)$$

The parameters $\boldsymbol{\pi}_i$ are jointly optimized through a reconstruction objective.

2.1.3. Enhancing CAEs

Previous work has identified that CAEs may select duplicate features, a sign of model degeneration leading to suboptimal local minima, and this is correlated with increased reconstruction error (Anonymous, 2024). To alleviate this, two mechanisms are proposed that we also use in this work: (i) GJS regularization, (ii) Indirect Parametrization.

GJS Regularization. The Generalized Jensen-Shannon Divergence (D_{GJS}) has previously been employed to measure the diversity among the mixture components (Kviman et al., 2022a,b) and can be utilized as a loss function (Englesson and Azizpour, 2021). As the goal is to learn distinct distributions that converge to unique features, maximizing the D_{GJS} can help prevent degeneration to repeat selections (Anonymous, 2024).

Definition 2 *The Generalized Jensen-Shannon Divergence (D_{GJS}) for M categorical distributions with probabilities $\{\boldsymbol{\pi}_i\}_{i=1}^{M}$, and weights $\mathbf{w}$*

$$D_{GJS}(\{\boldsymbol{\pi}_i\}_{i=1}^{M}) = \sum_{i=1}^{M} w_i D_{KL}(\boldsymbol{\pi}_i \,||\, \sum_{l=1}^{M} w_l \boldsymbol{\pi}_l) \quad (3)$$

The regularization strength is controlled by a parameter $\lambda_{GJS} > 0$:

$$\mathcal{L} = \mathcal{L}_{MSE} - \lambda_{GJS} D_{GJS}(\boldsymbol{\pi}, ..., \boldsymbol{\pi}_{N_o}) \quad (4)$$

Indirect Parameterization

Alternative approaches have been proposed for the parameterization of the Gumbel-Softmax distributions in CAE. Anonymous (2024) shows that the categorical probabilities π_i can be reparameterized with a matrix of learnable parameters $\boldsymbol{\psi} \in \Psi = \mathbb{R}^{k \times p}$, along with a learned linear transformation $(\boldsymbol{W}, \boldsymbol{b})$ leading to an *indirect parameterization* of the categorical probabilities of the GS distributions:

$$\boldsymbol{\pi}_i = \boldsymbol{W}\boldsymbol{\psi}_i + \boldsymbol{b}, \quad i = 1, 2, \ldots, k, \quad (5)$$

where $\boldsymbol{W} \in \mathbb{R}^{d \times p}$ and $\boldsymbol{b} \in \mathbb{R}^{d}$. This can be interpreted as a feature embedding with embedding dimensionality p. The setting of p is arbitrary; thus, this parameterization allows one to use an arbitrary number of learnable parameters for the distribution. This embedding has been shown to facilitate smooth training and improve convergence speed (Anonymous, 2024).

Regardless of parameterization, we normalize each probability vector $\boldsymbol{\pi}_i$ through a square-sum activation that improves the numerical stability by avoiding *exp* and ensures normalized probabilities:

$$\pi_{i,j} = \frac{\pi_{i,j}^2}{\sum_{j=1}^{k} \pi_{i,j}^2} \tag{6}$$

2.2. Video Vision Transformer for Echocardiography Analysis

This section introduces the baseline architecture for LVEF prediction, which is based on the ViT architecture. The model consists of primary components, a ViT vision model, and a transformer sequence model part. This factorization of spatial and temporal modeling was chosen because it facilitates token selection using the mechanism proposed in section 2.2.4. The full model consisting of both the spatial and temporal components is termed Video ViT (V-ViT).

2.2.1. ViT Frame Encoder

The core of the model consists of a frame encoder of the standard ViT architecture (Dosovitskiy et al., 2021). Performance on smaller datasets, such as biomedical data, has been lacking with ViTs due to the lack of spatial inductive biases of Convolutional Neural Network (CNN)s. This has been addressed by employing off-the-shelf pre-trained ViTs, and fine-tuning them on tasks with limited data. Particularly, this has been demonstrated to achieve performance equal to or better than CNNs for medical datasets, even though the domain of the pre-training data is completely unrelated (Matsoukas et al., 2021). In this work, we initialize the ViT with publicly available[1] pre-trained weights obtained using the unsupervised pre-training regimen of the DinoV2 method (Oquab et al., 2024) on ImageNet1k (Deng et al., 2009). We use a patch size of 14x14 pixels. Since the video frames of EchoNet-Dynamic are 112x112 pixels, this results in 64 total patches per frame. Here, standard 2D sinusoidal positional encoding is added to the tokenized image frames to provide positional information (Dosovitskiy et al., 2021).

Final Frame Representation. By construction of the attention mechanism, transformers output a sequence of the same dimensions as the input sequence. But here, a compressed representation of each image frame is desired. We attach an additional learnable class token (denoted CLS token) which encodes a global hidden frame representation. The CLS token holds a latent representation Z_t of the full image frame at a time-step t.

The frame encoder is applied to each frame in a video independently and thus converts each image sequence $\{I_t\}_{t=1}^{T}$ into a sequence of CLS tokens $\{Z_t\}_{t=1}^{T}$.

2.2.2. Temporal Component

The sequence model complements the spatial representations provided by the ViT frame encoder by addressing the temporal aspects of echocardiography videos. Unlike transformers which incorporate both encoder and decoder components, this sequence model relies on a vanilla Transformer encoder block (Vaswani et al., 2017).

The sequence model receives the sequence of encoded frames from the frame encoder. Similarly to the frame encoder, the sequence model utilizes an auxiliary CLS token. This CLS token, however, stores the representation of the full video. Lastly, it is transformed by a linear layer, followed by a Sigmoid activation function to perform the LVEF prediction for the video.

Time is encoded into the sequence encoder by adding standard 1D sinusoidal position encodings to the input tokens (Vaswani et al., 2017).

2.2.3. Video ViT

The complete factorized V-ViT architecture, which includes both the ViT frame encoder (E_F) and the sequence model (E_S), is highlighted with dashed lines in Figure 2. Denoting the sequence of image frames by $I_{1:T}$ the MSE regression objective can be expressed as follows.

$$\begin{aligned} \mathcal{L}_{MSE} &= \|y - \sigma(E_S(Z_{1:T}))\|_2 \\ &= \|y - \sigma(E_S(E_F(I_{1:T})))\|_2 \end{aligned} \tag{7}$$

That is, the architecture is trained end-to-end to minimize the MSE between the prediction $\hat{y}$ and

1. https://timm.fast.ai/

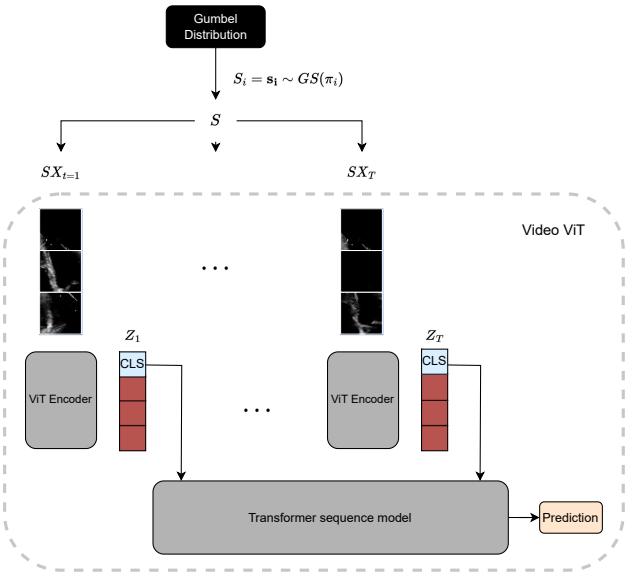


Figure 2: The proposed Video Vision Transformer (V-ViT). The dashed line highlights the basic architecture that is used regardless of feature selection. The Gumbel distribution module represents the case where learned feature selection is used.

ground-truth label y. Here, σ denotes the Sigmoid activation function.

$$\sigma(\hat{y}) = \frac{1}{1 + e^{-\hat{y}}}$$

This choice of activation is made to limit predictions within [0, 1] which is the bounds of the LVEF target variable y.

2.2.4. Learned Token Selection

CAE selects individual features through $\mathbf{x}_S = S\mathbf{x}$. Since ViTs operate on patches, it is natural to extend the feature selection to patches, instead of individual pixels. Performing selection on patches also drastically reduces the search space of possible selections, since selections are made in the order of 10 patches instead of the order of 10^4 pixels, likely leading to a more tractable optimization problem. We extend CAE to perform selection on image patches. Figure 1 highlights this extension.

Let X denote the matrix of patches of an individual video frame where $X \in \mathbb{R}^{N_p \times d_e}$, N_p is the total number of patches, and d_e is the patch embedding dimension. Samples $\mathbf{s_i}$ are drawn as defined in Eq. 1, but each distribution now represents a distribution over patches. Denoting N_o as the number of observed (selected) patches, S now has the form $S \in \mathbb{R}^{N_o \times N_p}$ and we can represent the differentiable *patch* selection as

$$X_S = SX, \text{ where } X_S \in \mathbb{R}^{N_o \times d_e} \tag{8}$$

The architecture is illustrated in Figure 2 and represents the incorporation of patch-level feature selection through the *Gumbel Distribution* block.

As an equivalent but more interpretable hyperparameter to N_o, we introduce the *Mask Ratio* (MR). This ratio is defined by

$$N_o = floor(N_p \cdot (1 - \text{MR}))$$

Our porposed Gumbel Video Vision Transformer (G-ViT) applies such stochastic image patch selection within the V-ViT framework for regluarization.

2.2.5. Token Selection as Regularization

Regularization for LVEF. In echocardiography, the left ventricle is crucial for determining LVEF, signaling the importance of certain ultrasound video frame regions. It highlights a potential problem with the data: namely, some regions are likely more

noisy, and some regions are likely more informative. Overfitting to noisy, uninformative regions should be avoided.

A General Rationale. Drawing an analogy with LASSO, where L1 norm encourages sparsity to improve model generalization by omitting nuisance variables, we propose a similar approach for discarding less informative patches. Unlike LASSO, our method employs embedded feature selection on ViT input tokens *during training only*, via Gumbel-Softmax distributions, allowing a relaxed, annealed approach towards discrete token selection. This approach aims to enhance generalization by focusing on informative features while remaining sensitive to potential nuances in test data by utilizing the full input during inference. Unlike deterministic selection methods, this stochastic feature selection ensures every patch has a chance to be observed during training, likely increasing the robustness of the model.

3. Related Works

3.1. Embedded Feature Selection

Stochastic Gates (Yamada et al., 2020) is a supervised feature selection approach based on a probabilistic relaxation of the features l_0 norm. More specifically, they use a continuous relaxation of the Bernoulli distribution and can therefore be optimized directly with gradient descent. (Zhang et al., 2021) puts forth Unsupervised Feature Selection via Transformed Auto-Encoders where an indicator matrix is used for feature selection. The indicator matrix is constrained by non-negativity and orthogonality via deep auto-encoders. The LassoNet (Lemhadri et al., 2021) architecture is a generalization of the LASSO method to neural networks which applies a L_1 norm to the last layer and utilizes residual connections.

3.2. Masked Autoencoders

While masked training has been extensively employed in Natural Language Processing (NLP) related pre-training tasks, masked autoencoders (He et al., 2021) have extended its applicability to computer vision, specifically for images. He et al. (2021) demonstrated that ViTs can achieve accurate reconstructions with only a subset of image patches as input, suggesting the potential for efficient use of image data through masked image-frame representations. A key result is that ViTs exhibit strong inference capabilities even when trained on a random fraction of the complete image. This suggests a possibility for even more efficient utilization of image data through stochastically selected patches.

3.3. Deep Learning for LVEF Prediction

EchoNet beat-by-beat (Ouyang et al., 2020), by the publishers of the EchoNet-Dynamic dataset, designed a pipeline mimicking human workflow for LVEF prediction. EchoNet employs a semantic segmentation model to identify the left ventricle, supervised with human tracings. The segmentations are then used to identify individual heartbeat cycles, and the predictions are averaged over all per-beat predictions. EchoNet's pipeline includes a semantic segmentation network (DeepLabV3) and a factorized ResNet2D + 1D convolution (R2+1D) LVEF prediction model. This pipeline achieved state-of-the-art in 2020, and more importantly, demonstrated an error rate similar to human experts (Ouyang et al., 2020).

The EchoNet-Dynamic authors also provide several end-to-end learned models trained with the objective of directly inferring LVEF from the videos. These models are therefore directly comparable to the G-ViT. They include three architectures: R2+1D, ResNet3D, and MC3 mixed convolution.

Recent works utilizing transformers have demonstrated improvements in end-to-end learned LVEF prediction on Echonet-Dynamic. They draw advantage of the most recent developments in computer vision with transformers and observe a higher predictive power than previously achieved with end-to-end learnable methods on EchoNet-Dynamic. One such work, named Ultrasound Vision Transformers (UVT) utilizes a combination of a CNN frame encoder with a transformer sequence model (Reynaud et al., 2021). Most recent is the application of the Shifted-Window ViT (Swin), a modification of ViT to achieve linear complexity, to EchoNet-Dynamic in the method named UltraSwin (Fazry et al., 2022). The UltraSwin provides results for two separate size variants of their architecture: UltraSwin-Small and UltraSwin-Base. Each of these end-to-end learnable methods is highly relevant for comparison with G-ViT.

4. Experiments

In this section, we evaluate our proposed G-ViT method on the EchoNet-Dynamic dataset through an extensive ablation on feature selection versus random selection and full-input training, where the V-ViT

serves as a baseline. We additionally compare against recent related works.

4.1. Hyperparameters

All models employed early stopping with a 20-epoch tolerance, up to a maximum of 100 epochs, utilizing the AdamW optimizer, and a weight decay of 0.05. A batch size of 60 videos was selected to maximize available GPU memory. The learning rate was annealed using a cosine schedule from lr_{max} (5e-4 and 1e-4 for V-ViT and G-ViT respectively) to lr_{min} =1e-7, after 10 warmup epochs from lr_{min} to lr_{max}.

The temperature of the GS distributions τ was initially annealed in an exponential schedule following Abid et al. (2019). τ_{min} was set to 0.1. τ_{max} was tuned in the range of {0.1, 1, 5, 10} and found optimal at 0.1 w.r.t. validation loss, resulting in a fixed temperature, producing highly one-hot selections.

We use the indirect parameterization of Equation 5 with $p = 1000$, and D_{GJS} regularization with $\lambda_{GJS} = 0.05$, see Equation 4. Refer to Appendix B for more detail and tuning procedures.

4.2. On the Significance of Learned Patch Selection During Training

4.2.1. Experimental Design

To demonstrate the effectiveness of patch selection during training, we design an ablation study by comparing two identical architectures[2], where one observes the full token sequence of each video frame (the V-ViT), and where one is trained with embedded feature selection on tokens (the G-ViT) each video frame.

The aim is to assess whether limiting input tokens through learned feature selection can induce a regularization effect, potentially lowering test error. We hypothesized that the regularization effect, if present, will manifest at certain mask ratios (MRs), and possibly lead to underfitting for high mask ratios. This experiment explores a spectrum of MRs in

$$\text{MR} \in \{0, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9\}$$

to identify the presence and optimal point of this effect, with all other hyperparameters held constant as per subsection 4.1.

Furthermore, the study extends to compare the G-ViT across various mask ratios against V-ViT subjected to both the full input and randomly masked input. This serves as an additional ablation study that is designed to evaluate the significance of selecting features over simply randomly masking the input.

During inference, no feature selection is used for either model. Instead, each model observes the full, discrete token sequence of each video frame. The aim is to regularize the models during training to learn robust inference rules during training, but not to perform predictions on subsets of the input during inference, in contrast to techniques such as LASSO (Tibshirani, 1996; Lemhadri et al., 2021).

For each setting of mask ratio and each model, the result was repeated for five different random seeds and thus different random initializations[3] of the models. This is to account for model variance with respect to both the random initialization of the architecture as well as the stochastic nature of the feature selection method. Results are recorded by observing predictions on the unseen official test partition of the dataset. Following prior work on EchoNet-Dynamic (Ouyang et al., 2020; Reynaud et al., 2021; Fazry et al., 2022), we include the following metrics: the Root Mean Square Error (RMSE), Mean Absolute Error (MAE) and the Coefficient of Determination (R^2).

4.2.2. Results

We observe that the G-ViT outperforms the V-ViT in RMSE, MAE, as well as R^2. Particularly, the optimal validation loss was found for an MR of 0.3. Table 1 provides a side-by-side comparison of the baseline V-ViT model training using the full input, to the G-ViT model trained with an MR of 0.3. There, the proposed technique is verified to have a regularizing effect, improving all metrics on the held-out test set, and was additionally found to reduce model variance compared to training on the full input. The latter is reflected in the significant difference in the measured standard deviations between the models.

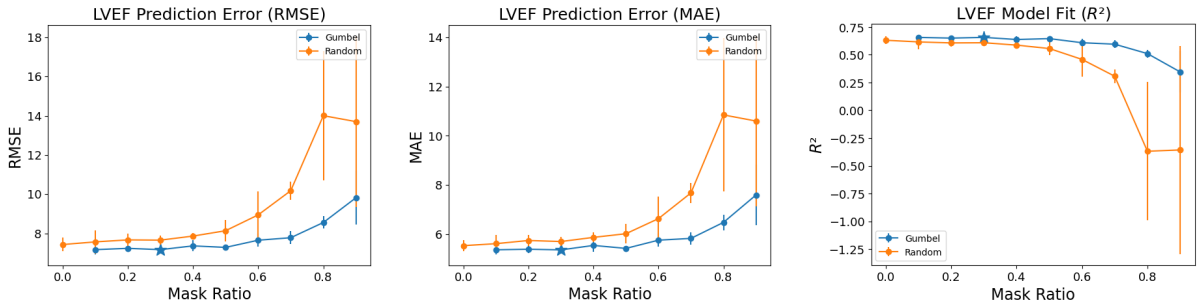
Figure 3 reports test set results for a range of different MRs. Each figure depicts the mean results alongside the standard deviations represented by error bars. The overall best-performing setting is highlighted by the star ($\star$) symbol. Generalization wors-

2. The video models are identical, but the G-ViT also has learnable parameters for token selection

3. The ViT frame encoder is pre-trained and is not reinitialized, but the transformer sequence encoder, as well as the distributional parameters, initialized differently with different random seeds.

Table 1: The strongest regularizing effect of the G-ViT was found at MR=0.3. It demonstrates a significant improvement over the V-ViT baseline on the unseen test dataset in all metrics. Additionally, model variance is reduced with G-ViT.

Model	MAE	RMSE	R^2	Params
V-ViT (MR=0)	5.53 ± 2.13	7.43 ± 3.53	0.63 ± 0.0350	30.3M
G-ViT (MR=0.3)	5.36 ± 0.0671	7.17 ± 0.12	0.656 ± 0.0115	30.4M



(*a*) RMSE (*b*) MAE (*c*) R^2

Figure 3: Test set metrics with mean and standard deviation across 5 random seeds, comparing feature selection (Gumbel) with the random masking baseline (Random). Training with G-ViT demonstrates an improvement over the random selection or full input V-ViT that is consistent across all metrics (RMSE, MAE, R). Note how the model with random selections loses predictive power at a much faster rate than with learned feature selection as the mask ratio increases, while the model with learned selections retains a non-trivial R^2 score even at an MR of 0.9. Unlike random selections, the variance of G-ViT remains low even at high mask ratios. The overall best-performing setting is marked by a star.

ens for higher as well as lower mask ratios. This indicates that there is a trade-off between overfitting and underfitting controlled by the mask ratio.

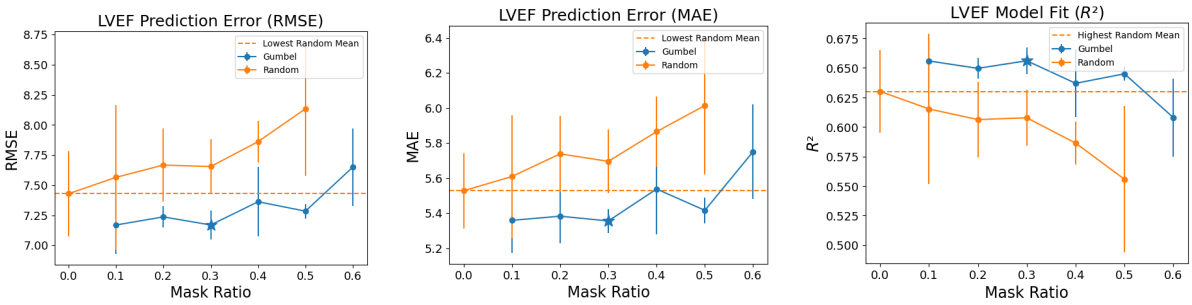
For the baseline V-ViT, we found that random masking during training did not help generalization. As seen in Figure 3, there is a monotonous decrease in performance as the MR increases from 0, unlike G-ViT.

Figure 4 shows more clearly that the regularizing effect is present not only for one specific setting of the MR but up to 0.5, compared to the baseline V-ViT observing the full input. This essentially means that test error can be reduced while only needing to process half of the input data in the frame encoder.

4.3. Predictive Power Compared to Related Work

To contextualize the predictive power of the G-ViT model, we compare it to existing works, by the performance in predicting LVEF from the echocardiography videos in the EchoNet-Dynamic's official test dataset, see Table 2.

EchoNet pipeline, upper bound. The EchoNet beat-by-beat prediction pipeline (see subsection 3.3), is noted as a "soft upper bound" for the G-ViT due to its reliance on additional steps and annotations. It necessitates training a segmentation model for the left ventricle and using signal processing to detect individual beat cycles, leading to a modeling strategy that averages predictions across these cycles. This approach introduces specific inductive biases such as invariance to the order of video snippets, and the assumption that each cycle contributes equally to the



(*a*) RMSE (*b*) MAE (*c*) R^2

Figure 4: A closer look limited to the range of mask ratios in $[0, 0.7]$ reveals that the improvement with learned feature selection (Gumbel) during training over the base V-ViT architecture trained using the full input (Random, MR = 0), persists at MRs as high as 0.5. The dashed line represents the best test set mean result achieved with random masking, which is the case with no masking (mask ratio = 0).

Table 2: Comparison to related works. ↑/↓ means higher/lower values are better. Our mean result of the G-ViT is reported alongside the corresponding best random seed in the parenthesis. UltraSwin-S/B refer to the Small and Base versions of the UltraSwin architecture respectively. * The EchoNet beat-by-beat pipeline consists in part of a DeepLabV3 segmentation model of 43.9 M parameters and a R2+1D LVEF prediction model of 31.5M parameters.

	Model	MAE ↓	RMSE ↓	R^2 ↑	Params
EchoNet beat-by-beat	EchoNet(1)	4.22	5.56	0.79	75,4 M*
End-to-end	R2+1D	7.35	9.53	0.40	31.5M
	R3D	7.63	9.75	0.37	33.4M
	MC3	6.59	9.39	0.42	11.6M
	UVT	5.95	8.38	0.52	346.8M
	UltraSwin-S	5.72	7.63	0.58	49.7M
	UltraSwin-B	5.59	7.59	0.59	88.2M
	G-ViT	**5.36** (5.24)	**7.17** (7.03)	**0.66** (0.67)	30.4M

overall LVEF estimate. We explore the potential of a fully end-to-end learned method, without the requirement of training a segmentation model and identifying beat cycles. This pipeline is denoted by EchoNet(1).

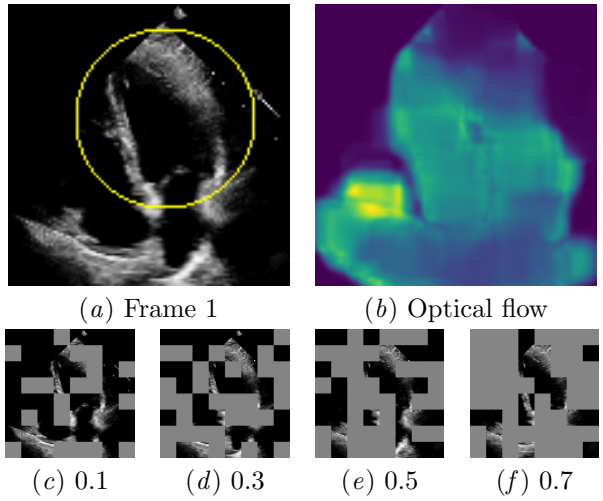
EchoNet end-to-end. Direct comparisons include EchoNet-Dynamic's end-to-end learnable models: factorized ResNet2D+1D, ResNet3D, and MC3 mixed convolution.

Transformer-based, end-to-end. Recent advancements with transformers, notably Ultrasound Vision Transformers (UVT) and Shifted-Window ViT (Swin) in UltraSwin, underscore the evolving landscape of LVEF prediction. UltraSwin's small and base variants, alongside UVT, provide important benchmarks for G-ViT.

G-ViT, improved generalization. The G-ViT demonstrates a significant leap in test set generalization over previous end-to-end methods consistently across RMSE, MAE and R^2 (Table 2).

4.4. Visualizing Selected Features

We provide a visual representation of the patches selected during training by masking certain image patches not included in the arg max of the learned



(a) Frame 1 (b) Optical flow

(c) 0.1 (d) 0.3 (e) 0.5 (f) 0.7

Figure 5: (a) An example video frame, with the left ventricle highlighted by the yellow circle.
(b) Avg. magnitude of optical flow, illustrating areas with much movement in the video. Yellow corresponds to much movement, and dark blue no movement.
(c)-(f) Visualization of the tokens with the highest probability of being selected (visible) during training, for varying mask ratio (MR). The model typically selects several tokens corresponding to the region of the left ventricle, as well as regions with high optical flow.

Gumbel-Softmax distributions. A patch is visible if its flattened location index j appears in:

$$\arg\max_j \pi_{i,j}$$

otherwise, it is masked (greyed out). Here, learned weights π are the learned GS probabilities.

We select a single video frame from EchoNet-Dynamic and apply this procedure, which can be seen in Figure 5. To convey the motion of the same video, we visualize the average magnitude of optical flow in the form of a heatmap[4].

4.4.1. Mechanism of Regularization

The first aim of this visualization is to identify if high-probability tokens align with medically significant areas, particularly the left ventricle.Hence, the visualizations are accompanied by the original image (Figure 5(a)), with the left ventricle highlighted by the yellow circle. We find that the model increasingly prioritizes the inclusion of the left ventricle as the MR increases. A second aim is to determine if the model prioritizes regions with significant temporal changes, vital for LVEF estimation, which relies on observing the heart's contraction and expansion. By examining the flow heatmap (Figure 5(b)), we observe that the model learns to prefer areas of higher motion for higher MRs. We conclude that this learned inclusion of regions of known significance strengthens our initial hypothesis that feature selection improves generalization by discarding nuisance features.

4. Optical flow was calculated using the Farneback method.

4.4.2. Model Interpretability

Figure 5 not only facilitates understanding of the regularization mechanism but also serves as a novel interpretability tool. This can be likened to data attribution techniques that attribute a model's prediction to input features. The difference here is that the selections are made unconditionally of the input, representing a fixed distribution for the entire dataset. Consequently, this method offers an indication of the regions generally needed by the model for LVEF prediction, instead of per-example. Ouyang et al. (2020) emphasize the critical role of the left ventricle in cardiomyopathy assessment, specifically its use in calculating end systolic and end diastolic volumes for LVEF determination. It is affirming to see the G-ViT's feature selection process naturally prioritizing these clinically significant regions, reminiscent of the methodologies trusted by medical practitioners, see Figure 5 (c)-(f).

5. Discussion

G-ViT, introduced in this work, is a novel and effective tool for the regularization of ViTs in the context of end-to-end learned LVEF prediction from echocardiography videos. Notably, using learned feature selection during training with a moderate mask ratio results in significant generalization improvement.
The G-ViT improves over related end-to-end trained ViTs recently applied on EchoNet-Dynamic, including UVT, UltraSwin-Small, and UltraSwin-Base, and earlier convolutional approaches and notably achieves an R^2 of 0.66. The baseline architecture V-ViT also outperforms previous architectures with an R^2 of 0.63, an unexpectedly high baseline performance.

Remarkably, our G-ViT model, using just 50% of the image patches, matches the full-input V-ViT.
Importantly, our method not only improves model performance but also offers enhanced interpretability for ViTs. We show how the selected features can be visualized and how learned selections correspond to regions of known relevance to the determination of LVEF. Our visualizations shed light on the regularization mechanism of G-ViT, revealing its focus on informative patches.
Finally, the potential scope of our method extends beyond regularization, and the outcomes of this work can lead to downstream applications beyond our primary focus.

References

Abubakar Abid, Muhammad Fatih Balin, and James Y. Zou. Concrete autoencoders for differentiable feature selection and reconstruction. *CoRR*, abs/1901.09346, 2019. URL http://arxiv.org/abs/1901.09346.

Anonymous. Indirectly parameterized concrete autoencoders. *Concurrent submission, anonymous pdf attached as supplemental material*, 2024.

Girish Chandrashekar and Ferat Sahin. A survey on feature selection methods. *Computers & Electrical Engineering*, 40(1):16–28, 2014.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009.

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. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=YicbFdNTTy.

Erik Englesson and Hossein Azizpour. Generalized jensen-shannon divergence loss for learning with noisy labels. *CoRR*, abs/2105.04522, 2021. URL https://arxiv.org/abs/2105.04522.

Lhuqita Fazry, Asep Haryono, Nuzulul Khairu Nissa, Sunarno, Naufal Muhammad Hirzi, Muhammad Febrian Rachmadi, and Wisnu Jatmiko. Hierarchical vision transformers for cardiac ejection fraction estimation. In *2022 7th International Workshop on Big Data and Information Security (IWBIS)*, pages 39–44, 2022. doi: 10.1109/IWBIS56557.2022.9924664.

Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross B. Girshick. Masked autoencoders are scalable vision learners. *CoRR*, abs/2111.06377, 2021. URL https://arxiv.org/abs/2111.06377.

Eric Jang, Shixiang Gu, and Ben Poole. Categorical reparameterization with gumbel-softmax, 2016. URL https://arxiv.org/abs/1611.01144.

Oskar Kviman, Harald Melin, Hazal Koptagel, Victor Elvira, and Jens Lagergren. Multiple importance sampling elbo and deep ensembles of variational approximations. In *International Conference on Artificial Intelligence and Statistics*, pages 10687–10702. PMLR, 2022a.

Oskar Kviman, Ricky Molén, Alexandra Hotti, Semih Kurt, Víctor Elvira, and Jens Lagergren. Learning with miselbo: The mixture cookbook. *arXiv preprint arXiv:2209.15514*, 2022b.

Ismael Lemhadri, Feng Ruan, and Rob Tibshirani. LassoNet: Neural Networks with Feature Sparsity. In *Proceedings of The 24th International Conference on Artificial Intelligence and Statistics*, pages 10–18. PMLR, March 2021. URL https://proceedings.mlr.press/v130/lemhadri21a.html. ISSN: 2640-3498.

Yue Liu, Christos Matsoukas, Fredrik Strand, Hossein Azizpour, and Kevin Smith. Patchdropout: Economizing vision transformers using patch dropout. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 3953–3962, 2023.

Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. *CoRR*, abs/2103.14030, 2021. URL https://arxiv.org/abs/2103.14030.

Chris J. Maddison, Andriy Mnih, and Yee Whye Teh. The concrete distribution: A continuous relaxation of discrete random variables. *CoRR*, abs/1611.00712, 2016. URL http://arxiv.org/abs/1611.00712.

Christos Matsoukas, Johan Fredin Haslum, Magnus Söderberg, and Kevin Smith. Is it time to replace cnns with transformers for medical images? *CoRR*, abs/2108.09038, 2021. URL https://arxiv.org/abs/2108.09038.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy V. Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel HAZIZA, Francisco Massa, Alaaeldin El-Nouby, Mido Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Herve Jegou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. DINOv2: Learning robust

visual features without supervision. *Transactions on Machine Learning Research*, 2024. ISSN 2835-8856. URL https://openreview.net/forum?id=a68SUt6zFt.

David Ouyang, Bryan He, Amirata Ghorbani, Neal Yuan, Joseph Ebinger, Curtis P. Langlotz, Paul A. Heidenreich, Robert A. Harrington, David H. Liang, Euan A. Ashley, and James Y. Zou. Video-based ai for beat-to-beat assessment of cardiac function. *Nature*, 2020.

Hadrien Reynaud, Athanasios Vlontzos, Benjamin Hou, Arian Beqiri, Paul Leeson, and Bernhard Kainz. Ultrasound video transformers for cardiac ejection fraction estimation. In Marleen de Bruijne, Philippe C. Cattin, Stéphane Cotin, Nicolas Padoy, Stefanie Speidel, Yefeng Zheng, and Caroline Essert, editors, *Medical Image Computing and Computer Assisted Intervention – MICCAI 2021*, pages 495–505, Cham, 2021. Springer International Publishing. ISBN 978-3-030-87231-1.

Robert Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, 58(1):267–288, 1996. ISSN 00359246. URL http://www.jstor.org/stable/2346178.

Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Hervé Jégou. Training data-efficient image transformers & distillation through attention. *CoRR*, abs/2012.12877, 2020. URL https://arxiv.org/abs/2012.12877.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. Attention is all you need. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. URL https://proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf.

Yutaro Yamada, Ofir Lindenbaum, Sahand Negahban, and Yuval Kluger. Feature selection using stochastic gates. In Hal Daumé III and Aarti Singh, editors, *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 10648–10659. PMLR, 13–18 Jul 2020. URL https://proceedings.mlr.press/v119/yamada20a.html.

Yunhe Zhang, Zhoumin Lu, and Shiping Wang. Unsupervised feature selection via transformed auto-encoder. *Knowledge-Based Systems*, 215:106748, 2021. ISSN 0950-7051. doi: https://doi.org/10.1016/j.knosys.2021.106748. URL https://www.sciencedirect.com/science/article/pii/S0950705121000113.

Appendix A. Code and Reproducibility

We have taken extensive care to facilitate the reproduction of our results. Our attached code comes complete with a `README.md` file, providing detailed instructions for downloading the dataset, installing the Python environment, and running the code. We provide two configuration files `configs/config_baseline_timm.yaml` and `configs/config_gumbel_timm.yaml` for reproducing the results of V-ViT and G-ViT respectively, with optimal hyperparameter settings. The code will automatically download the pretrained weights of the frame encoder when running the main program.

Appendix B. Hyperparameters

B.1. Fixed hyperparameters

The AdamW optimizer was used with running average coefficients $\beta = (0.9, 0.9999)$.

B.2. V-ViT hyperparameters

For the baseline V-ViT, hyperparameter tuning was performed at a mask ratio of 0, which means using the full input, to give the baseline model the fairest advantage to the G-ViT. The maximum learning rate lr_{max} was varied in the range of {5e-3, 1e-3, 5e-4, 1e-4} for a set of two different pre-trained weights.

Two sets of pretrained weights were tested for the V-ViT baseline: one obtained through discriminative training on ImageNet1k and the other using the unsupervised pre-training regimen of the DinoV2 method on ImageNet1k. The precise ViT architectures used for the frame encoder network

can be found in the TIMM source code[5], as well as the source code associated with this paper. They are denoted `vit_small_patch14_dinov2` and `vit_tiny_patch16_224`.

The optimal settings were found to be lr_{max}=5e-4 together with the DinoV2 weights `vit_small_patch14_dinov2`, slightly outperforming `vit_tiny_patch16_224`.

B.3. G-ViT hyperparameters

For the proposed G-ViT model, hyperparameter tuning was performed at a mask ratio of 0.5, which means the model was allowed to select a token sequence length of 50% of the full input for each frame. This was done to ensure that the learning dynamics with feature selection engaged were captured well. Even though changing the mask ratio (in later experiments) means altering the architecture and parameter count of the G-ViT model slightly, unlike with random masking, hyperparameter searches were not performed on other mask ratios. This was meant to keep the comparison to the baseline V-ViT fair and save computational resources.

The optimally performing pre-trained weights were kept from the V-ViT sweep, and that experiment was not reiterated for the G-ViT model. The maximum learning rate lr_{max} was varied in the range of {1e-3, 5e-4, 1e-4}. The maximum temperature τ_{max} of the GS distributions was varied in the range of {0.1, 1, 5, 10}, with the annealing schedule (with respect to 100 epochs) varied between the exponential schedule proposed by CAE (Abid et al., 2019) and a simply linearly decreasing schedule. The optimal setting of τ_{max} was found to be 0.1, which means the temperature was held fixed at 0.1, and the type of schedule is irrelevant. The optimal lr_{max} was found to be 1e-4 for the G-ViT.

B.3.1. CAE optimization tricks

A limited test using the optimization tricks (detailed in Section 2.1.3) of Indirect Parameterization (IP) parameterization and D_{GJS} regularization was done. This was done after the search for lr_{max} and τ_{max} to keep the search non-combinatorial and save computational power. IP vectors with dimension $p = 1000$ were tried versus the original parameterization of CAE. This setting of vector dimension was based on the observation that the improvements saw diminishing returns for very large dimensions Anonymous (2024), so a sufficiently large setting ($\geq$ number of possible selections $= N_p$) was selected somewhat arbitrarily. D_{GJS} regularization with a regularization strength coefficient of 0.05 was tried versus no D_{GJS} regularization. The optimally performing setting was found to be a combination of both tricks: namely IP vectors of dimension 1000 and D_{GJS} regularization with a strength coefficient of 0.05.

Appendix C. Efficiency benefits

As touched upon, our method holds the potential for substantial performance benefits in terms of training speeds and the amount of computation required. Figure 6 demonstrates the decrease in required floating-point operations[6] (FLOPs) in each forward pass drastically decreases with increasing MR.

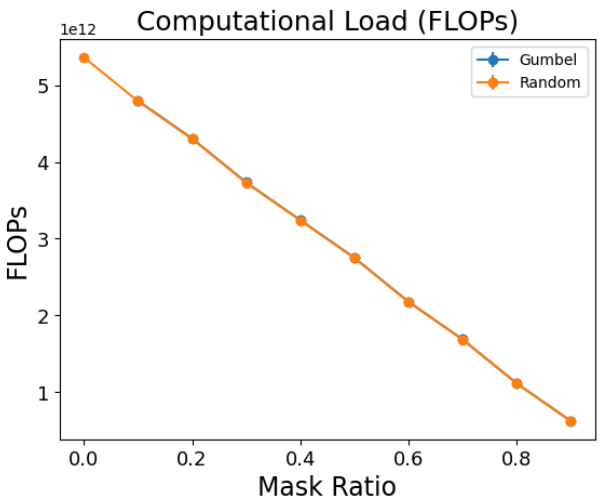


Figure 6: Reducing the input drastically reduces floating point operations (FLOPs). The overlapping graphs indicate that feature selection introduces almost zero overhead compared to random masking.

5. https://github.com/huggingface/pytorch-image-models/blob/main/timm/models/vision_transformer.py

6. Measured on a single forward pass using the DeepSpeed package (https://www.deepspeed.ai/)