# GRS CS 640: Project Proposal

***Reshab Chhabra**
Boston University
reshabc@bu.edu

***Mark D. Yang**
Boston University
mdyang@bu.edu

## 1 Topic

The project we would like to propose is *"Algorithmic Video Reduction by Estimating Chunk Significance Using Spatial-Temporal Masked Auto Encoders"*. In this study, we offer to utilize spatial-temporal (ST) masked autoencoders (MAEs) [1][2] to facilitate the development of information-preserving video reduction techniques. Although transformer-based video compression [3], among others, exist, they do not preserve ST properties, i.e. the result is no longer a video. Sampling a ST MAE will be used to estimate the "importance" of ST chunks. We will then employ an algorithmic compression approach that prioritizes the preservation of the "important" chunks. By preserving the informational content, these reduced videos can further be used for other purposes, like training ML models.

**Inspiration** The demand for video reduction techniques that retain information is vast. Currently, the inclusion of video data is frequently hindered by its excessively high computational cost [2]. Many video compression processes exist, but either lose the ST aspect [3] (i.e. it's not a video anymore) or are not temporal reduction [5]. By enabling downsized datasets to effectively train video machine learning models, a video reduction method that maintains the information, we propose, could be highly advantageous — we would have reduced worry about the computational cost when provided a reduced and information-preserving video. These ST MAEs extend the principle of self-attention to higher dimensions, thereby providing invaluable insight concerning each video chunk's relative significance.

## 2 Goal

The goal is to produce qualitatively good video reductions that give quantitatively good results as training data. If successful, this approach could facilitate the creation of lightweight video training datasets while retaining essential information. The proposed methodology holds the potential to be highly advantageous in quickening numerous video-based training tasks.

## 3 Hypotheses

**ST Chunk Significance** We hypothesize that ST MAEs are capable of capturing the connections between ST chunks and that it is feasible to approximate the relative significance of these chunks by conducting a sampling of the marginal mask losses.

**Information Retention** Our project postulates that incorporating chunk importance in video reduction can retain valuable information that would be lost in conventional video reduction methods like downsampling. This approach can potentially produce reduced videos that retain sufficient informational content and be used to train video-oriented models more quickly without compromising their performance.

## 4 Data Collection

The Kinetics-400 dataset [6] (see `https://www.deepmind.com/open-source/kinetics`) provides a wide variety of short, high-quality single person-object action clips. The paper [1] additionally uses the K400 as its validation set.

## 5 Formulation

The video reduction algorithm preprocesses the video to a standardized format to be passed into the ST MAE model [2]. Then the video is partitioned into chunks of a to-be-determined size. For example, paper [1] uses $2 \times 16 \times 16$ pixel chunks (2 is time dimension), which we may replicate.

To initialize the procedure, the sampler takes an initial randomized spatial-temporal agnostic chunk mask of the video chunks (i.e. taking a random fraction of all chunks). This mask would then be passed into the pre-trained ST MAE model [1], which would provide a reconstructed video whose loss is computed from the original video via MSE. The mask introduces a new chunk — the chunk of interest — then the newly reconstructed video is made from the new mask and the new MSE loss is recorded. The marginal loss $(new - original)$ is noted by the chunk included in the mask. We would then do this for each video chunk except chunks that are already included in the mask.

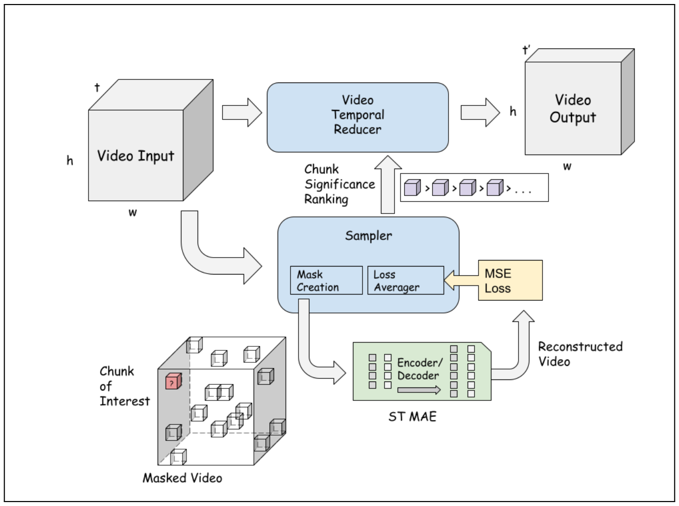


Figure 1: Proposed Model Layout

We would then iterate over this process again with new randomized masks while keeping the chunk loss association values, and repeat it until sufficient data is obtained to rank these chunk loss metrics created.

The reductions for each chunk are averaged over each sample, and the sampler algorithm will proceed to the video temporal reduction phase. The video will be reduced by an algorithmic technique. A naive implementation would be temporal max pooling, where temporal slices are related to the chunk of max significance. Another implementation could be content-aware video re-scaling [4], where the significance is used to create the shrink ability maps.

The reduced videos would then be bench-marked as training data for a variety of simple video machine-learning models. The performance of the models will be compared with models trained on videos traditionally reduced through techniques like downsampling.

# 6 Evaluation Criteria

As this project consists of two parts i.e. video reduction algorithm and comparing reduced video training input to the non-reduced video input, we plan to evaluate our progress through answering the following questions:

1. Were we able to successfully parse the output from the pre-trained model?
2. Were we able to define in a video what are the most important chunks?
3. Were we able to make a video reduction algorithm from these chunks?
4. Were we able to compare our reduced video input to non-reduced input?

When evaluating these answers, we will additionally consider restraints such as whether we had enough time and processing power. We would also look into else we can work on in the future with additional time, as we value conceptually understanding the project from a low level and its potential over short term results.

# References

[1] Christoph Feichtenhofer & Haoqi Fan et al. (2021) Masked Autoencoders As Spatiotemporal Learners. arXiv:2205.09113v2

[2] Zhan Tong, Yibing Song, Jue Wang, and Limin Wang. VideoMAE: Masked autoencoders are data-efficient learners for self-supervised video pre-training. arXiv:2203.12602, 2022.

[3] Fabian Mentzer, George Toderici, David Minnen, Sung-Jin Hwang, Sergi Caelles, Mario Lucic, Eirikur Agustsson. (2022) VCT: A Video Compression Transformer. arXiv:2206.07307

[4] Zhang, Y.-F., Hu, S.-M. and Martin, R.R. (2008), Shrinkability Maps for Content-Aware Video Resizing. Computer Graphics Forum, 27: 1797-1804. https://doi.org/10.1111/j.1467-8659.2008.01325.x

[5] He, K., Chen, X., Xie, S., Li, Y., Dollár, P., & Girshick, R. (2022). Masked autoencoders are scalable vision learners. arXiv:2111.06377

[6] Will Kay, Joao Carreira, Karen Simonyan, Brian Zhang, Chloe Hillier, Sudheendra Vijayanarasimhan, Fabio Viola, Tim Green, Trevor Back, Paul Natsev, Mustafa Suleyman, Andrew Zisserman. (2017) The Kinetics Human Action Video Dataset. arXiv:1705.06950