

# **BOSCH MODEL EXTRACTION**

Team 6

# KEY CHALLENGES

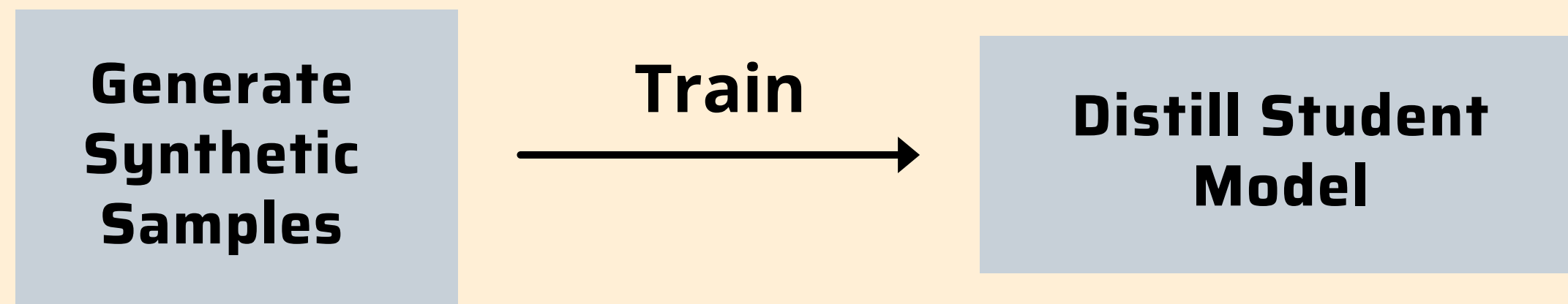
- The complexity of action recognition and the number of classes
- Lack of relevant literature on video classification model extraction
- Scale of datasets being used

## Additional Challenges in Black Box Extraction

- No access to any dataset
- Lack of pre-trained Video GANS

# **Black Box Model Extraction**

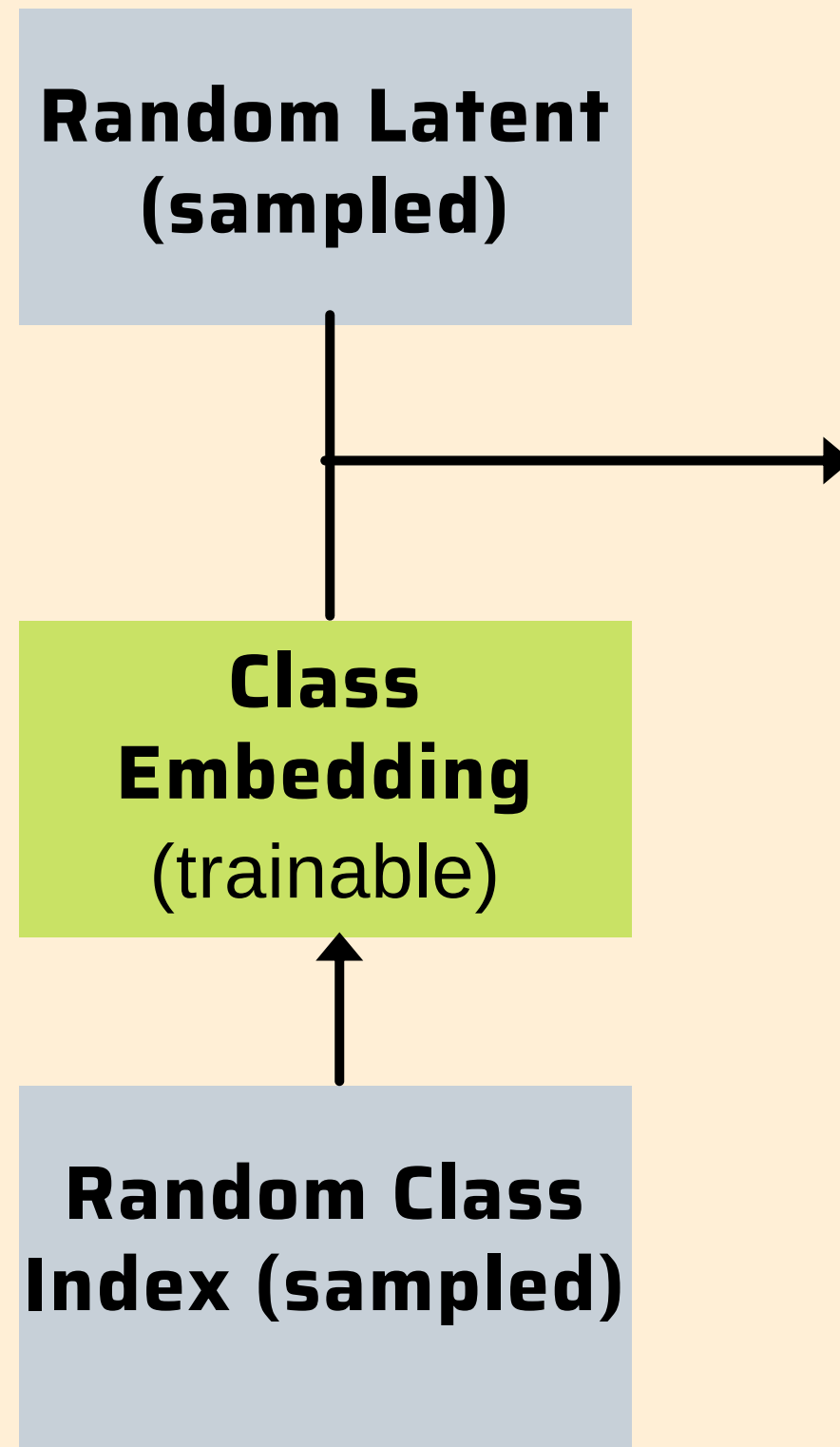
# BLACK BOX APPROACH



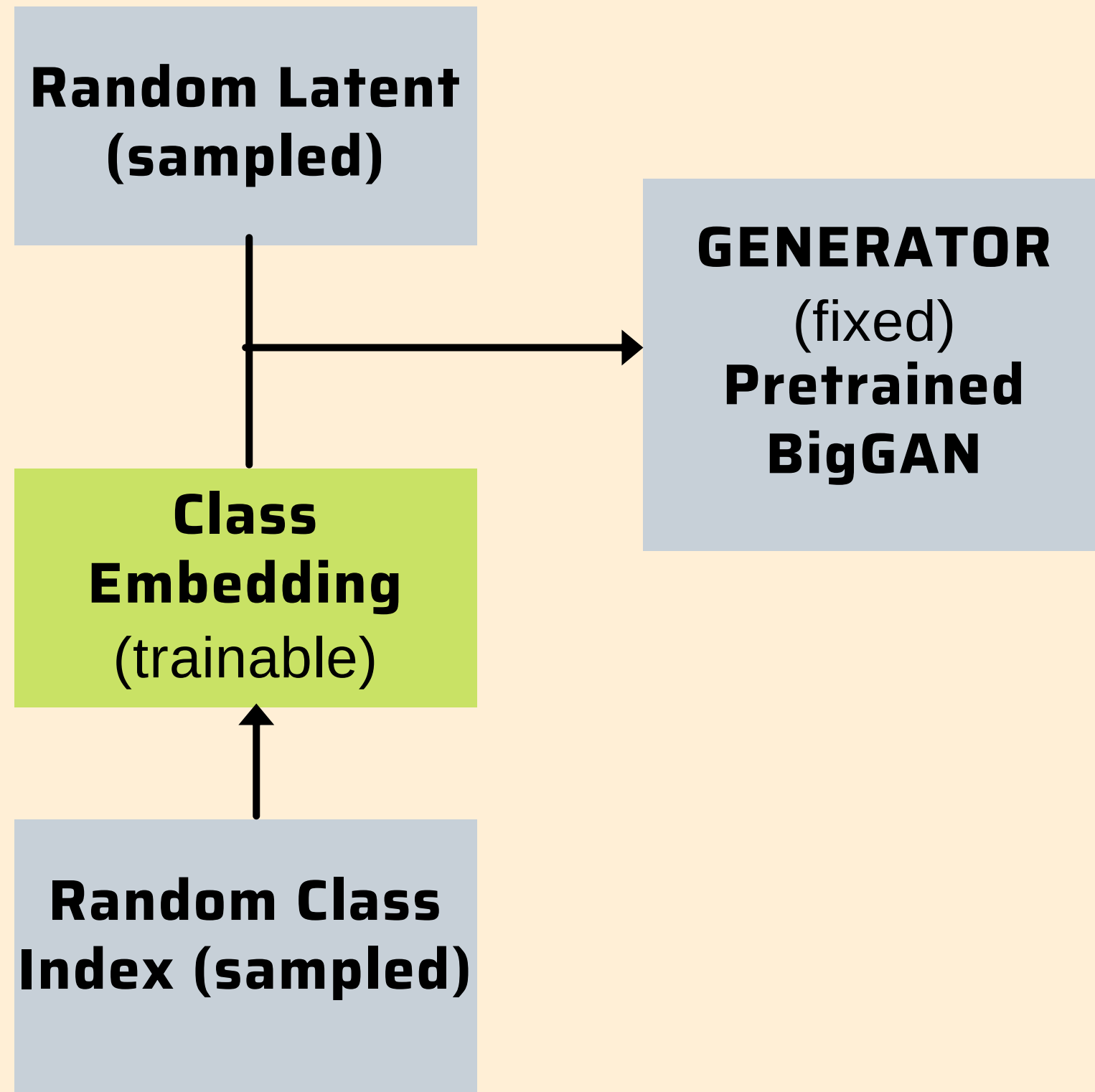
# SYNTHETIC DATA GENERATION

- Due to the absence of class-conditioned Video GANs, we re-purpose a pre-trained BigGAN (ImageNet) to generate fake samples.
- Learn embeddings for each class index predicted by the teacher.
- Freeze generator, teacher model and finetune to minimise loss between teacher predicted class index and sampled class index.
- Finetune for a few steps until predicted class confidence  $> 90\%$

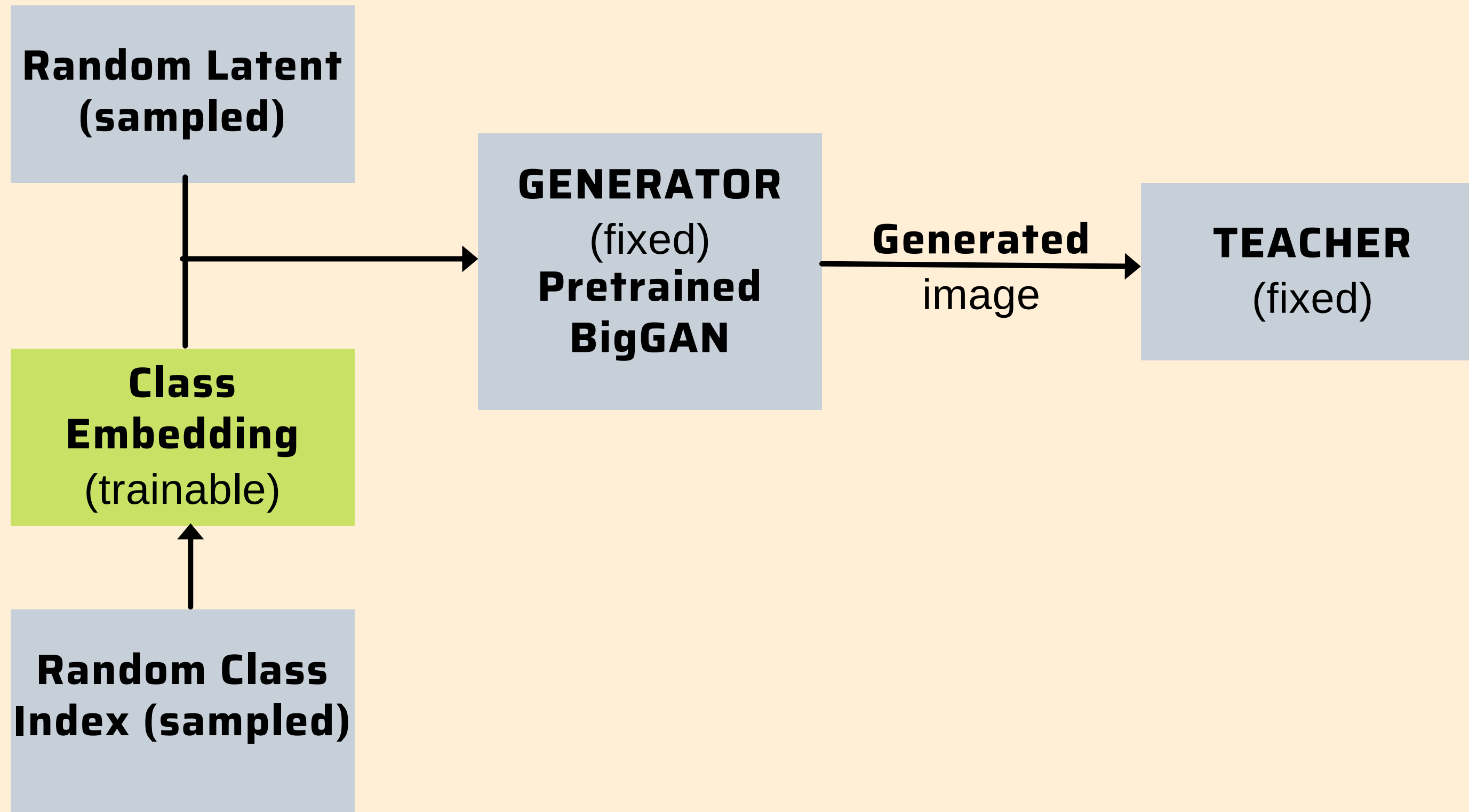
# SYNTHETIC DATA GENERATION



# SYNTHETIC DATA GENERATION

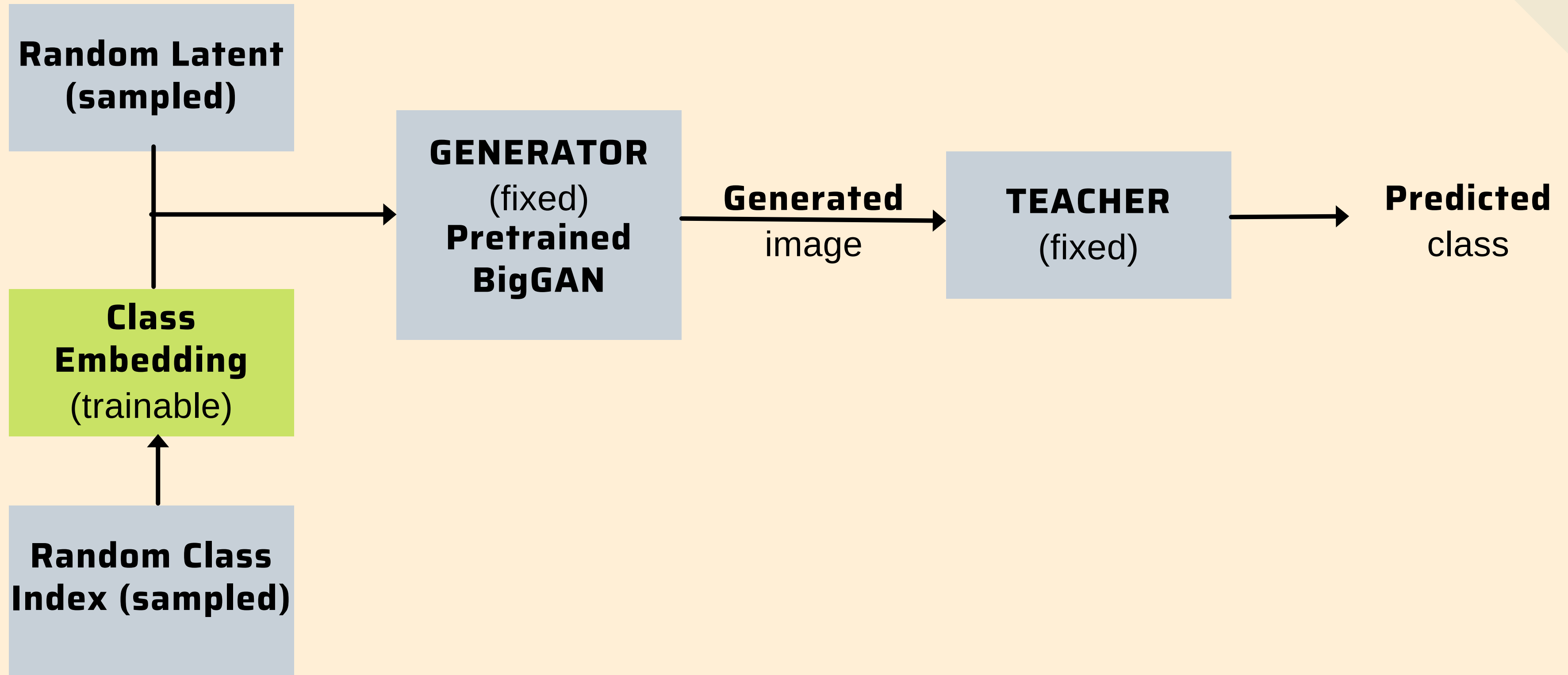


# SYNTHETIC DATA GENERATION

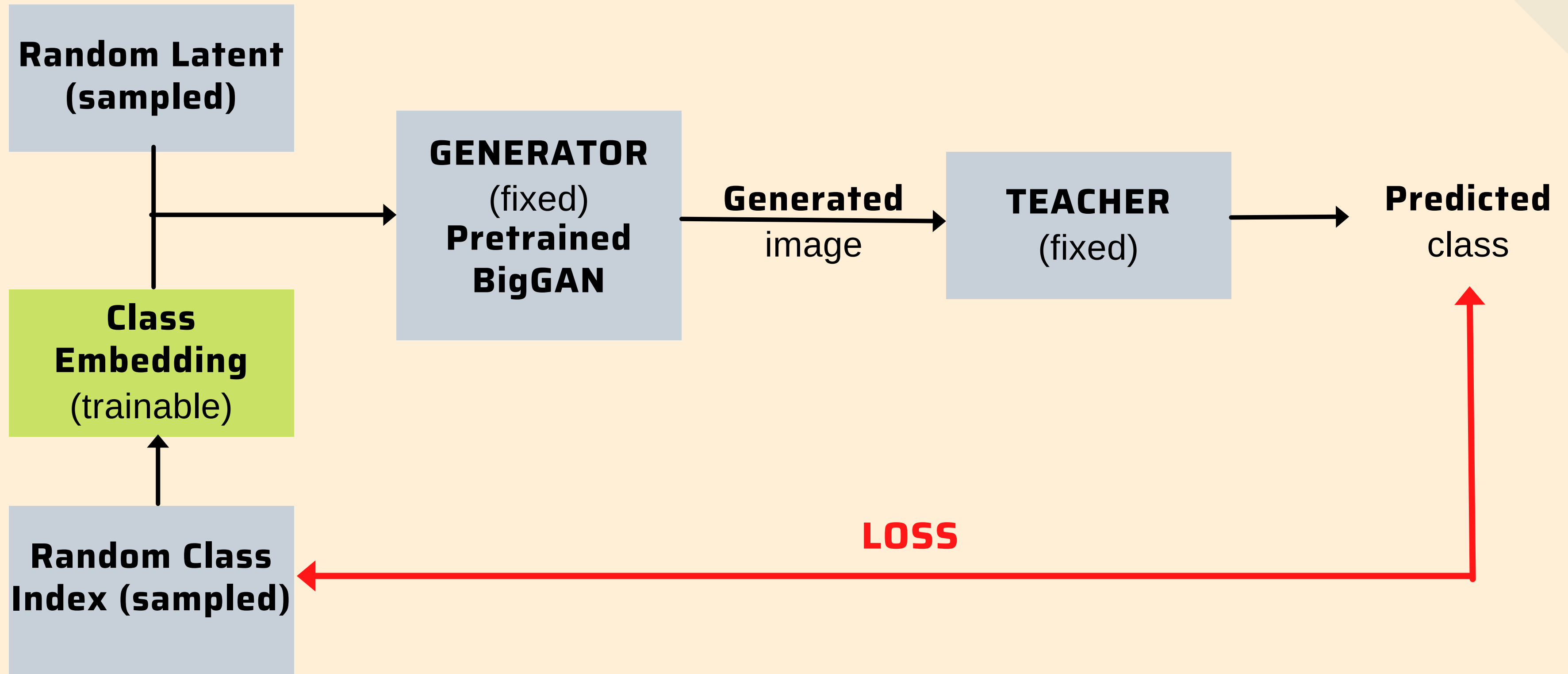




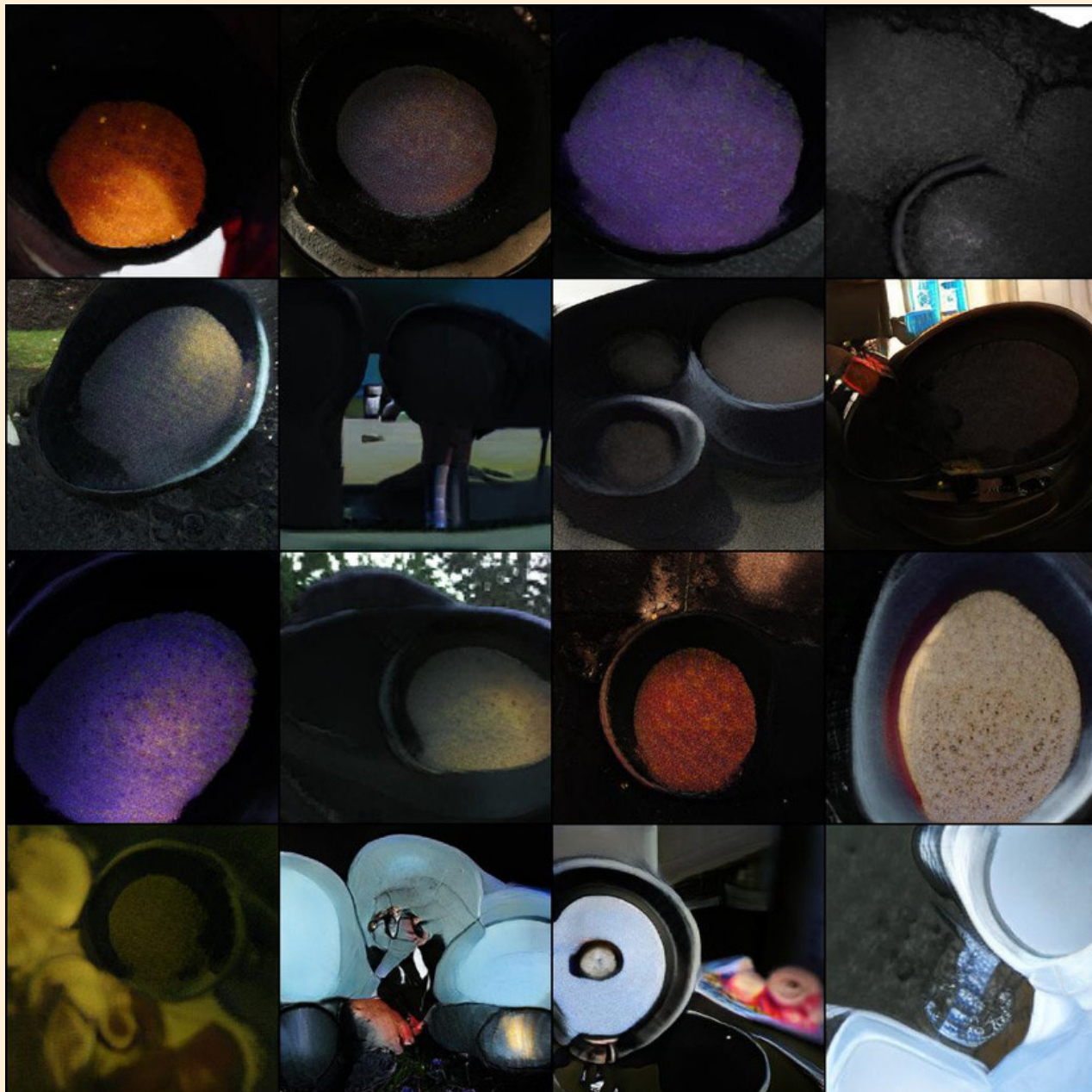
# SYNTHETIC DATA GENERATION



# SYNTHETIC DATA GENERATION



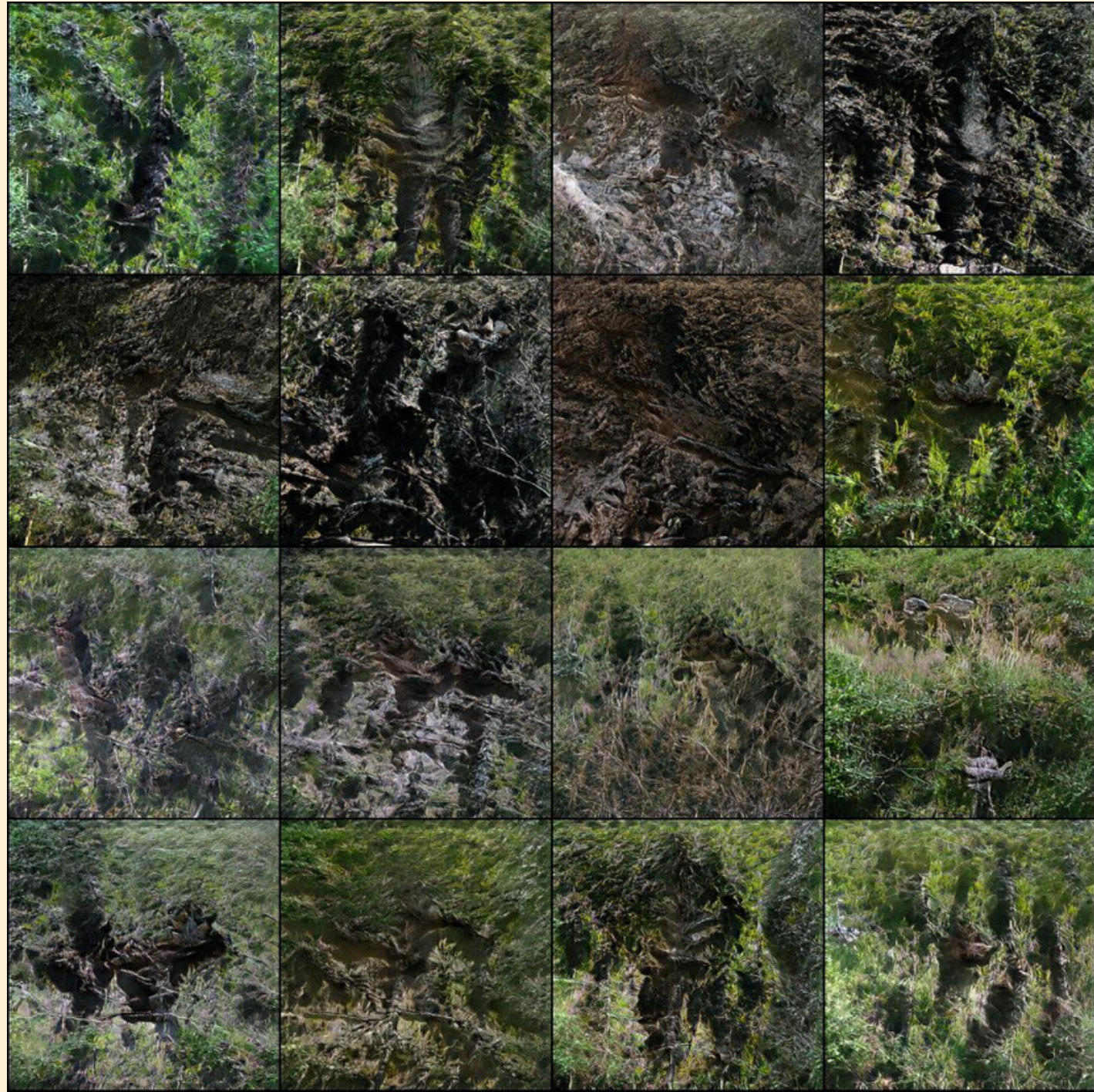
# SYNTHETIC DATA - EXAMPLE



- The image corresponds to samples generated for class index 1 or "air drumming"
- Although the images are not visually appealing, they fool the teacher classifier with high confidence.
- Thus the generated samples belong to the training distribution of the teacher network.

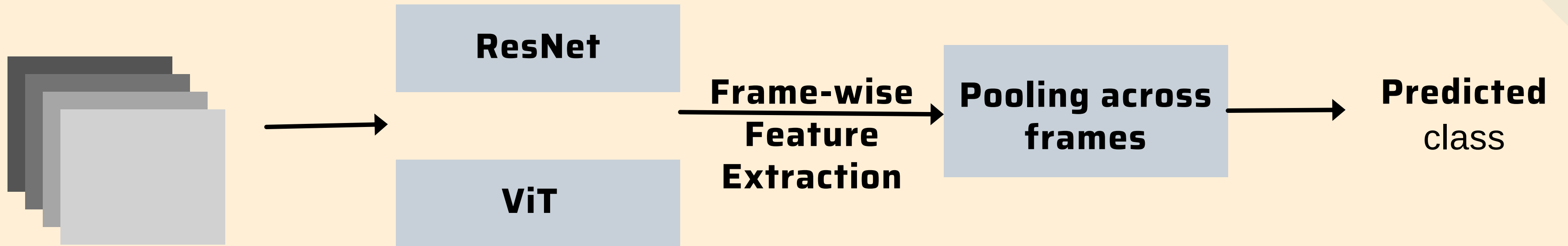


# SYNTHETIC DATA - EXAMPLE



class index 0: abseiling

# STUDENT MODEL

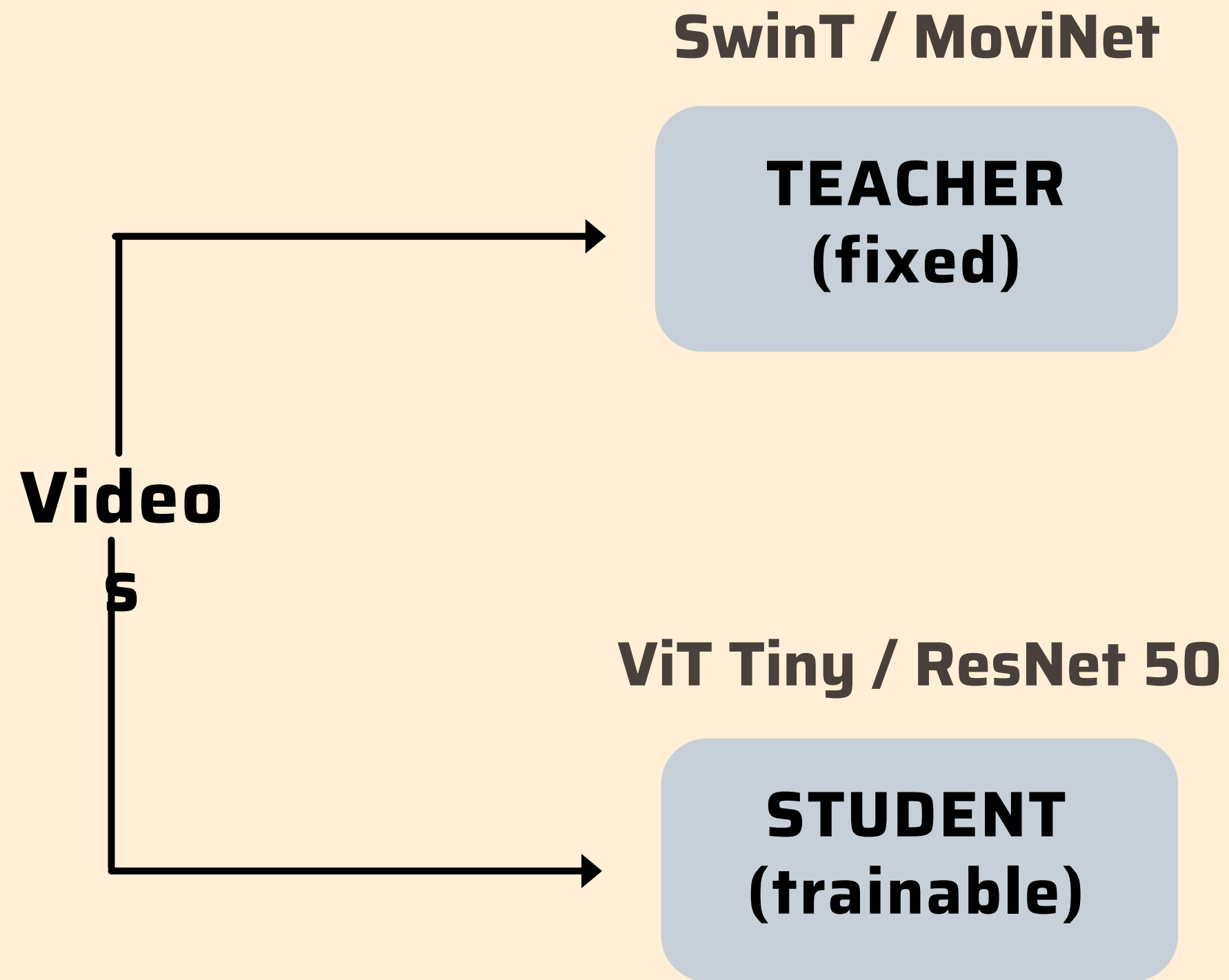


Frame  
S

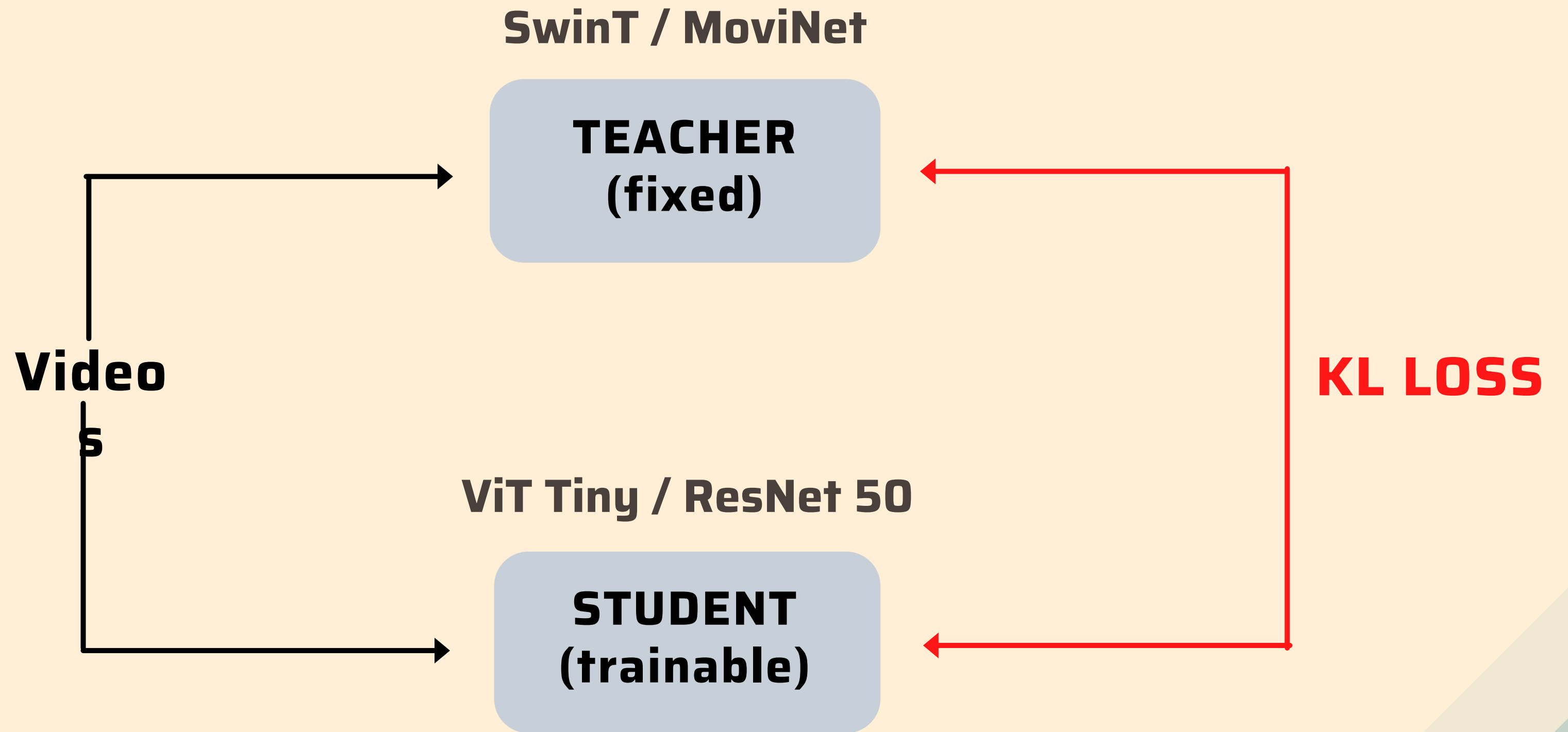
Dataset	Teacher	Student
Kinetics 600	MoviNet	ResNet-50
Kinetics 400	Swin-T	VIT-T



# KNOWLEDGE DISTILLATION



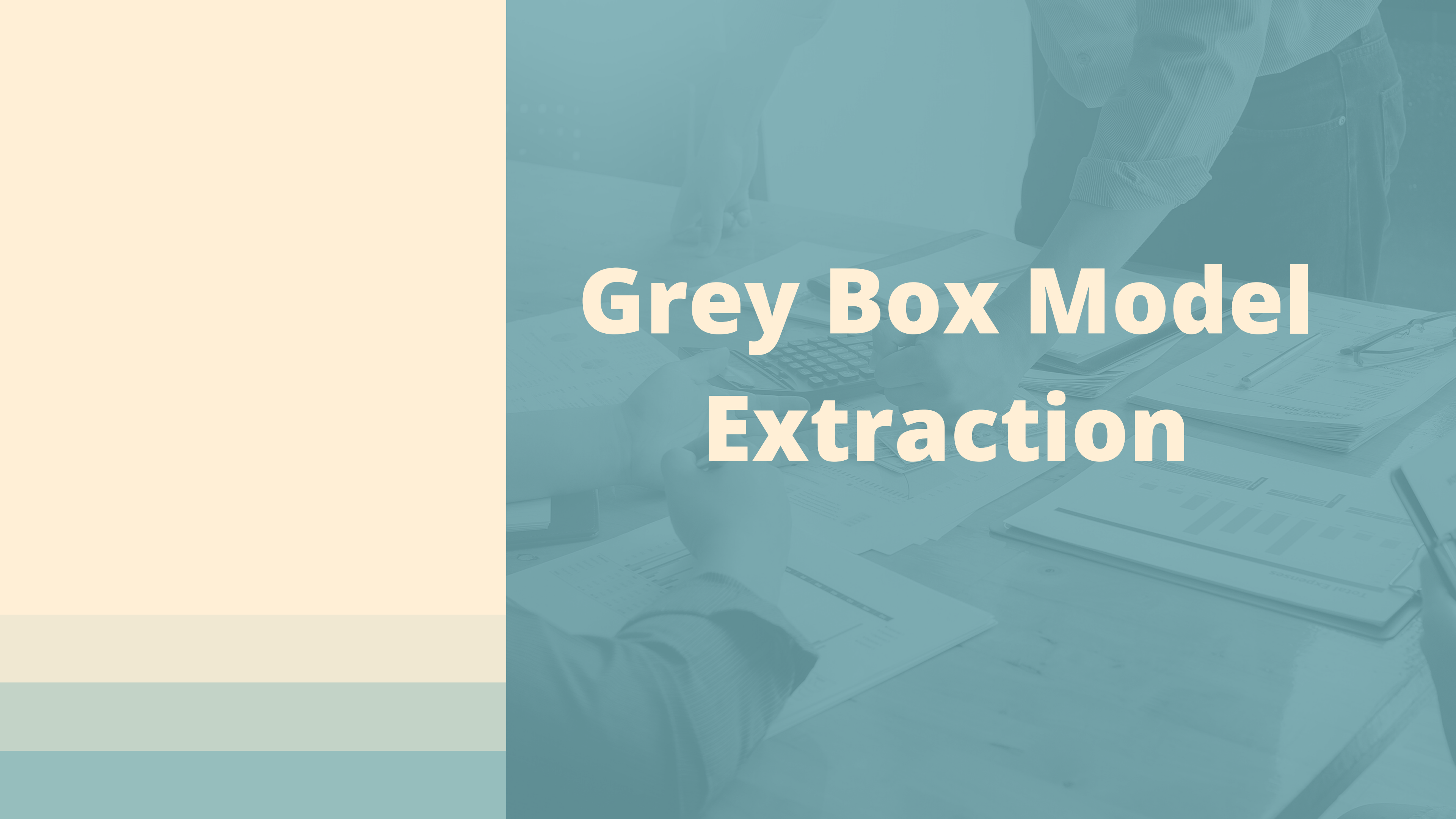
# KNOWLEDGE DISTILLATION



**Black Box  
Accuracies**

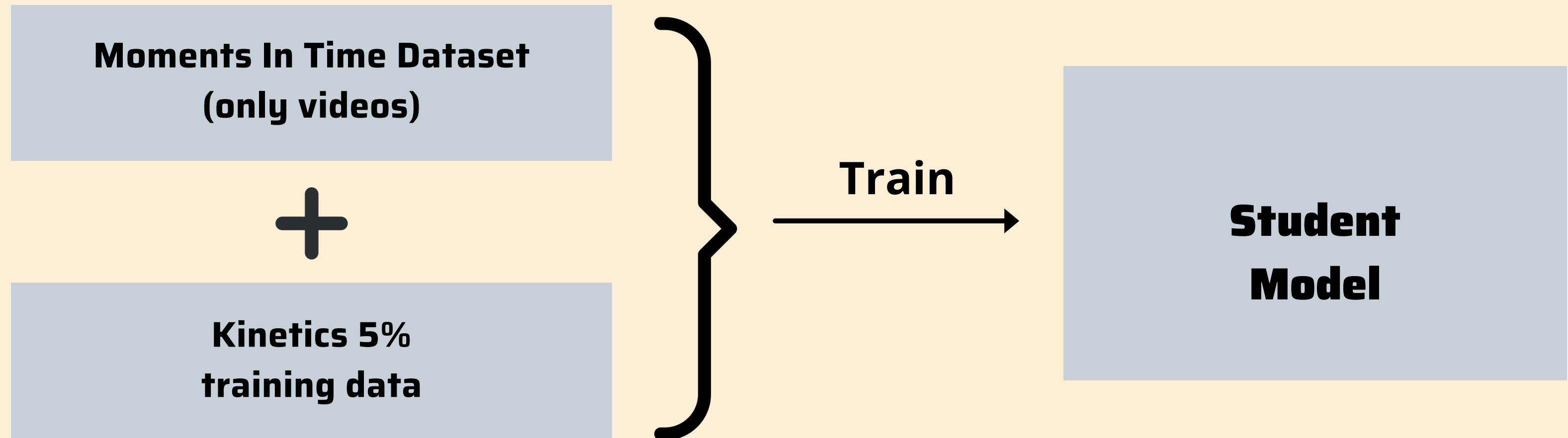
	VideoSwin <b>Transformer</b>	<b>Movinet</b>
<b>Distill</b>	<b>7.8%</b>	<b>12.43%</b>



The background of the slide features a blue-tinted photograph of several people in business attire working at a desk. They are surrounded by various financial documents, including spreadsheets and reports. One document in the foreground is clearly labeled 'Total Expenses'. A calculator is also visible on the desk. The overall scene suggests a professional financial or accounting environment.

# Grey Box Model Extraction

# GREY BOX APPROACH



# KNOWLEDGE DISTILLATION

SwinT / MoviNet

**TEACHER**  
**(fixed)**

ViT Tiny / ResNet 50

**STUDENT**  
**(trainable)**

**Auxiliary  
Images**

```
graph LR; AI[Auxiliary Images] --> T[TEACHER (fixed)]; AI --> S[STUDENT (trainable)];
```

The diagram illustrates a knowledge distillation setup. On the left, the text 'Auxiliary Images' is positioned. Two arrows originate from this text: one points horizontally to the right towards a light blue rounded rectangle labeled 'TEACHER (fixed)', and the other points horizontally to the right towards a similar rectangle labeled 'STUDENT (trainable)'. Above the teacher box is the text 'SwinT / MoviNet', and below the student box is 'ViT Tiny / ResNet 50'. To the right of the student box, there is a section titled 'Auxillary Dataset KD' followed by a list of dataset details.

**Auxillary Dataset KD**

Irrelevant dataset: Moments in Time

- Number of classes: 305
- **Minimal** overlap with Kinetics 400 / Kinetics 600 Dataset

# FINE TUNING ON 5% TRAINING DATA

## Lottery Ticket Hypothesis, Why?

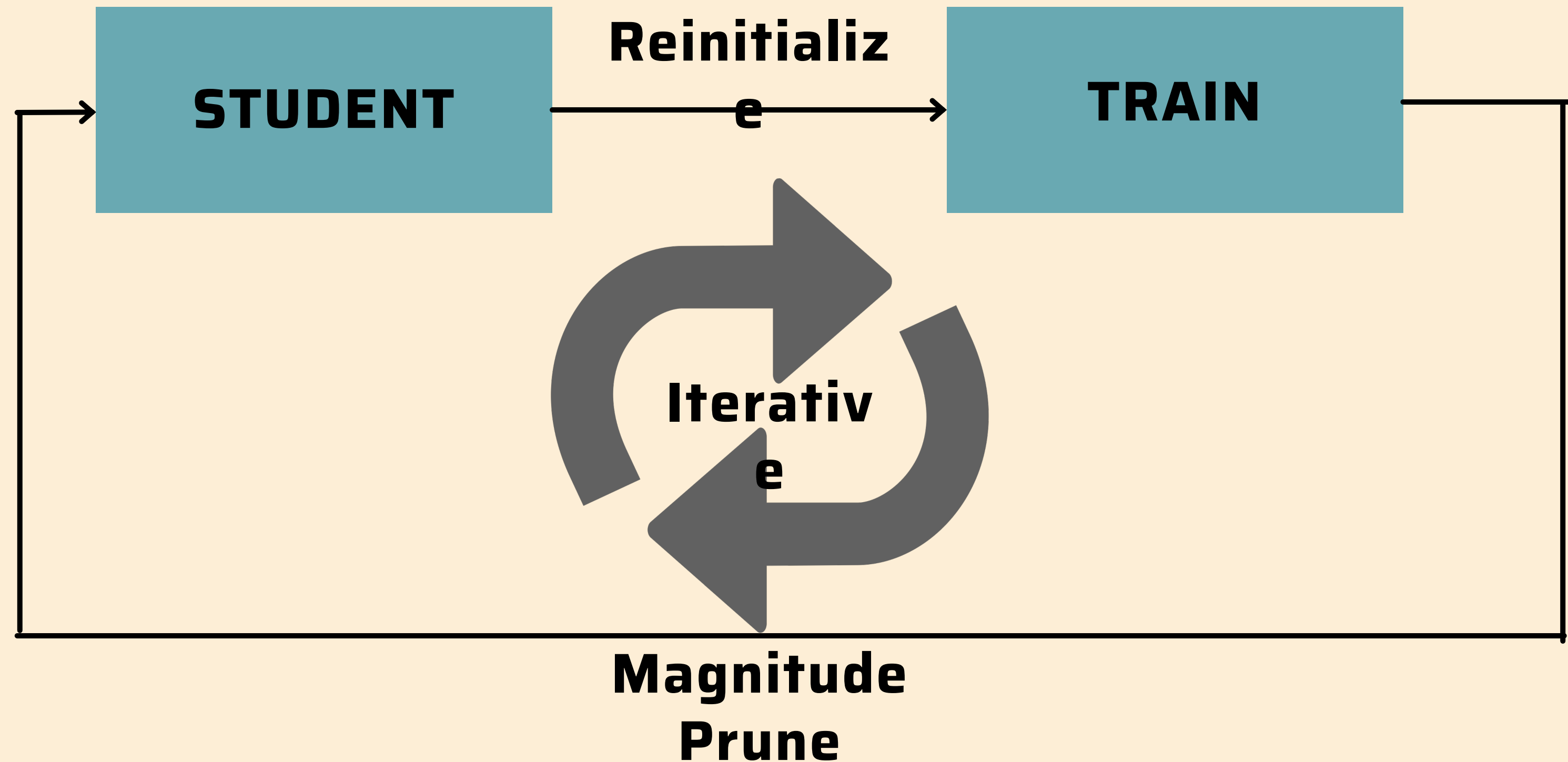
*Every network contains a sparse network which can potentially outperform the dense network*

- Sparsity is a regularization to avoid over-fitting.
- Induces inductive bias specific to fine-tuned task, hence improved performance.

## Iterative Magnitude Pruning

- Initialize network
- Train for a few epochs
- Prune least magnitude weights
- Re-initialize and perform steps 2-3 till required sparsity.

# FINE TUNING ON 5% TRAINING DATA



# Grey Box Accuracies

	VideoSwin Transformer	Movinet
Knowledge Distillation	8.3%	13.8%
Fine-Tune 5% Kinetics (Dense)	12.6%	22.3%
Fine-Tune 5% Kinetics (Winning Ticket)	17.3%	33.5%



# Thanks

Team 6

