LAMARR

Institute for Machine Learning and Artificial Intelligence

UNIVERSITÄT BONN

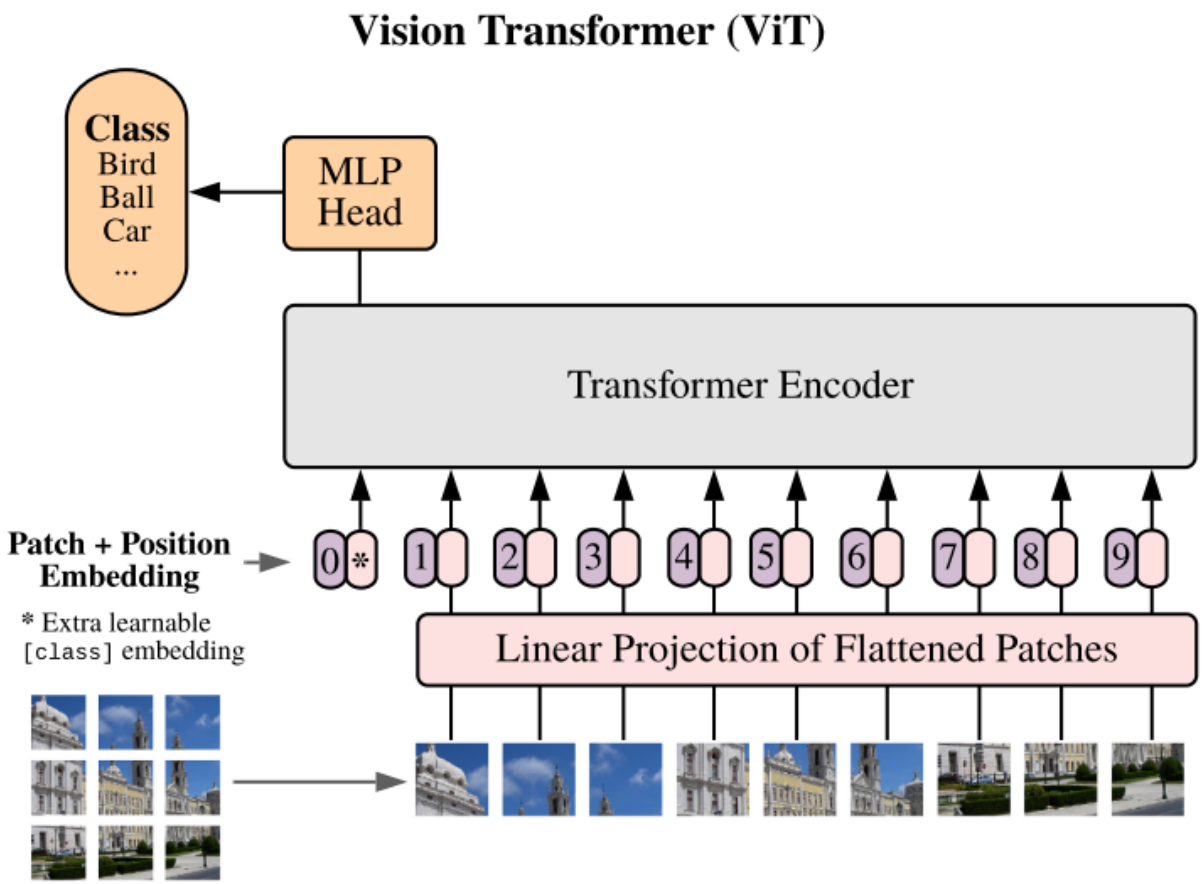
Juergen Gall

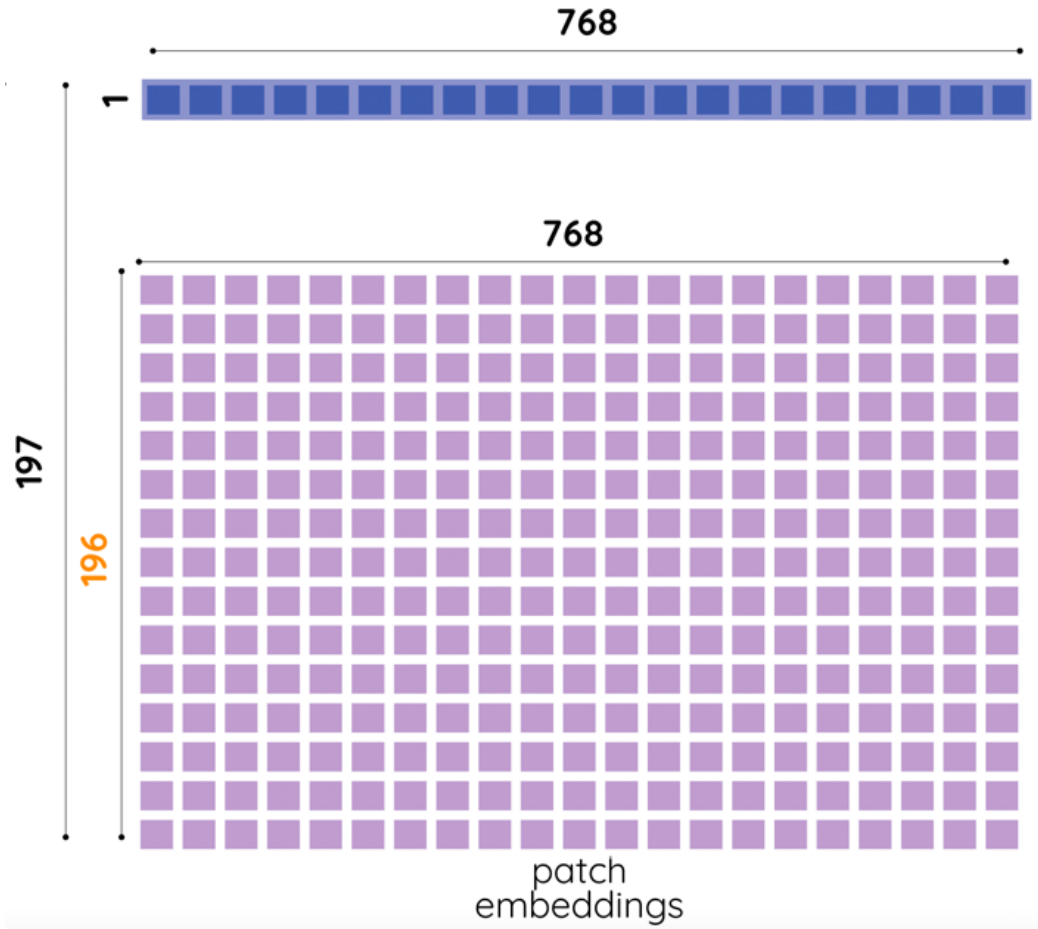
Recent Advances in Video Understanding and Forecasting

Vision Transformers for Video Understanding

[E. Bahrami et al. **How Much Temporal Long-Term Context is Needed for Action Segmentation?** ICCV 2023]

Vision Transformers for Images or Videos

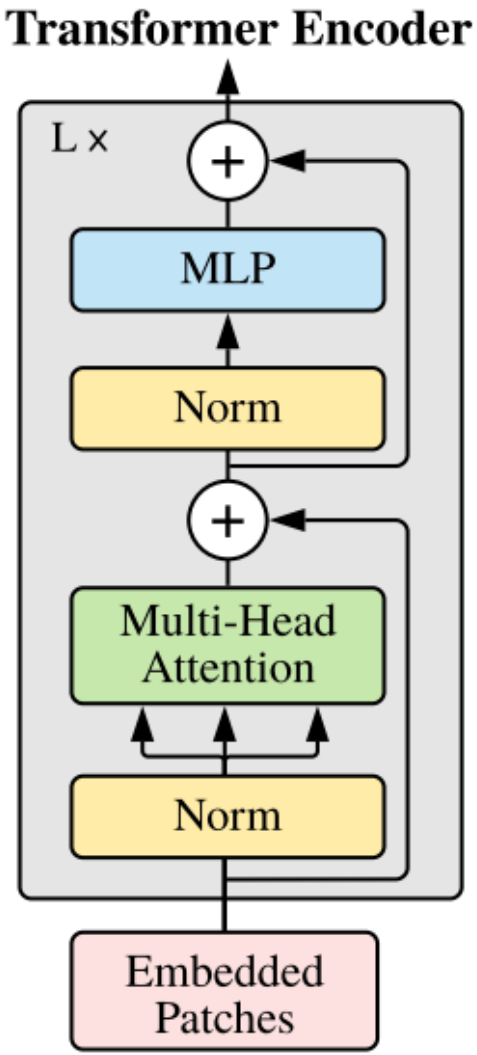




[A. Dosovitskiy et al. **An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale.** ICLR 2021]

[https://amaarora.github.io/2021/01/18/ViT.html]

Vision Transformers for Images or Videos



$$\mathcal{Q} \in \mathbb{R}^{(N+1)\times d}$$

$$\mathcal{K} \in \mathbb{R}^{(N+1)\times d}$$

$$\mathcal{V} \in \mathbb{R}^{(N+1)\times d}$$

$$\mathcal{A} = \mathrm{Softmax}\left(\mathcal{Q}\mathcal{K}^T/\sqrt{d}\right)$$

$$\mathcal{A} \in \mathbb{R}^{(N+1)\times(N+1)}$$


How can we make transformer architectures more efficient?

Efficient Image and Video Understanding

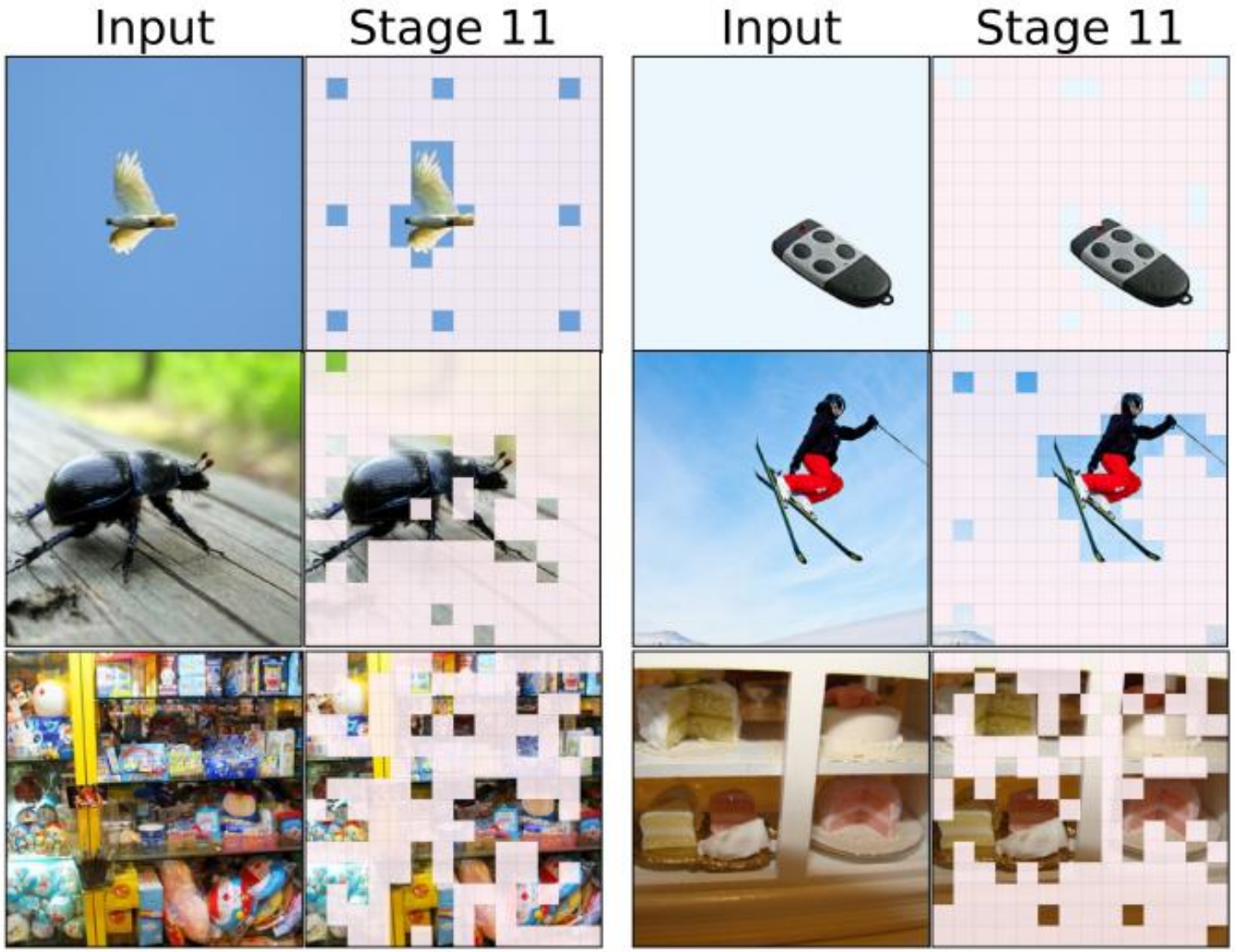
- Transformers are very expensive (high GFLOP)
- Not all GFLOPS are needed for all videos or images

[M. Fayyaz et al. **Adaptive Token Sampling For Efficient Vision Transformers.** ECCV 2022]

Reduce number of tokens for inference

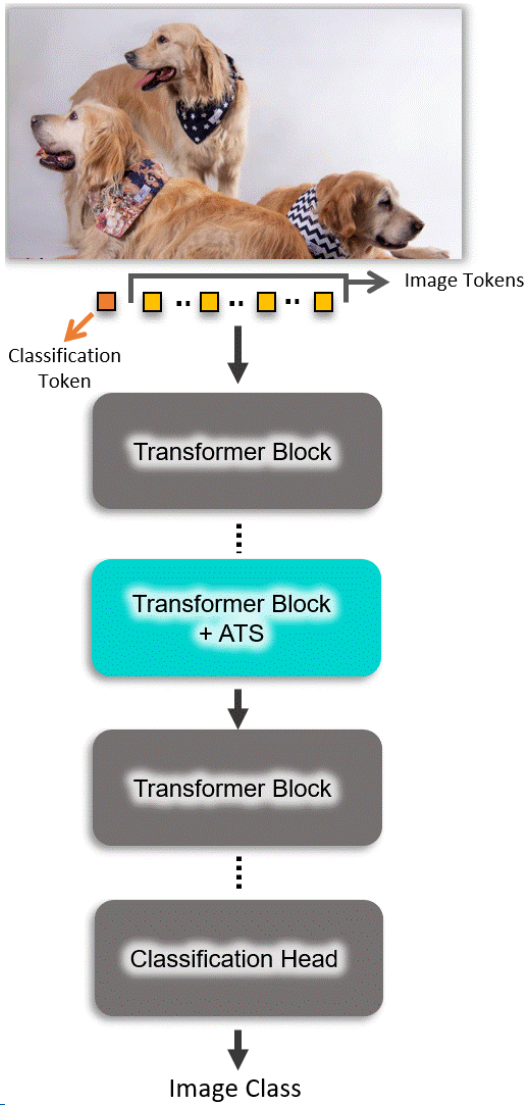
Efficient Vision Transformers

Use only as much tokens as needed for inference



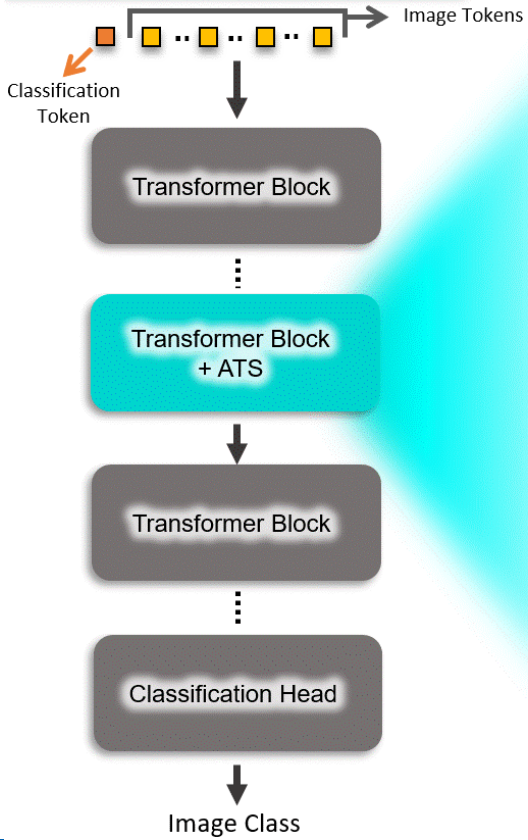
[M. Fayyaz et al. **Adaptive Token Sampling For Efficient Vision Transformers.** ECCV 2022]

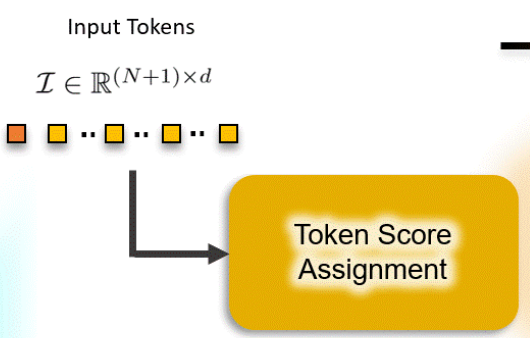
Efficient Vision Transformers

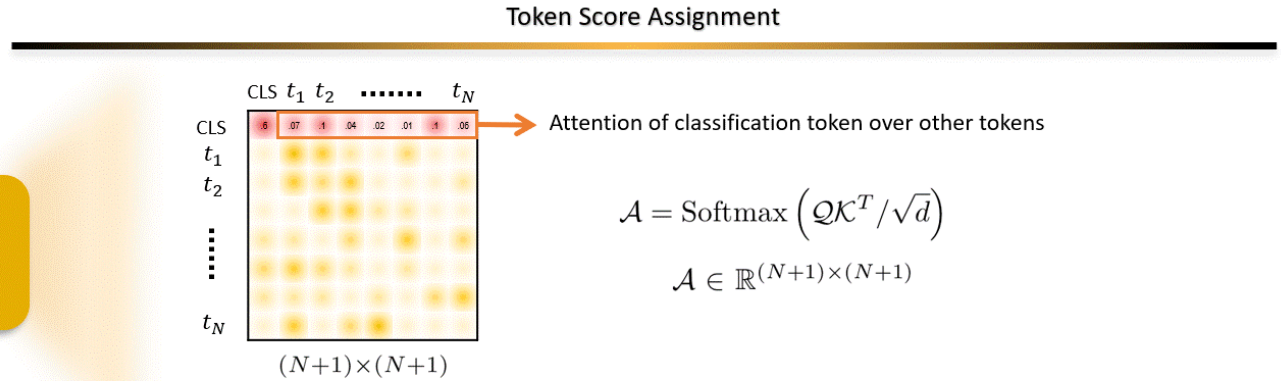


[M. Fayyaz et al. **Adaptive Token Sampling For Efficient Vision Transformers.** ECCV 2022]

Efficient Vision Transformers

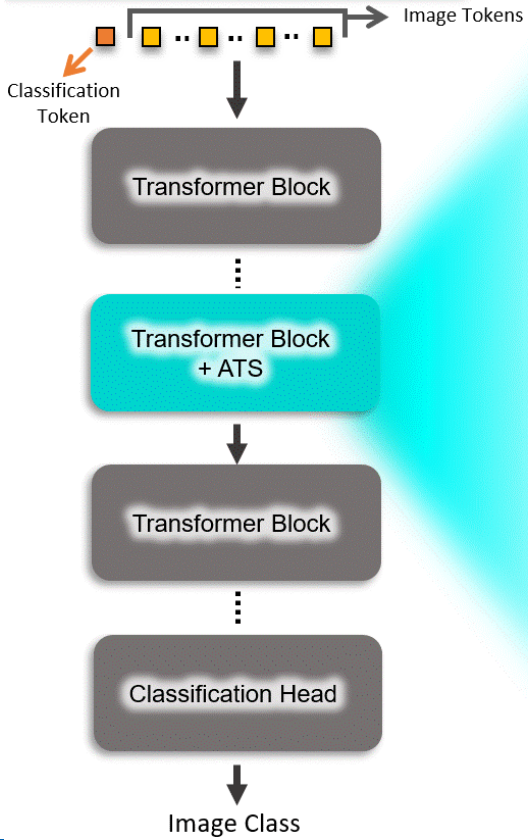


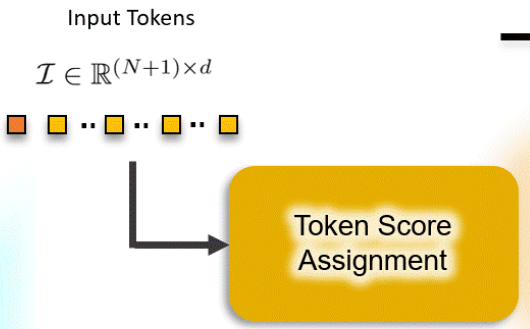


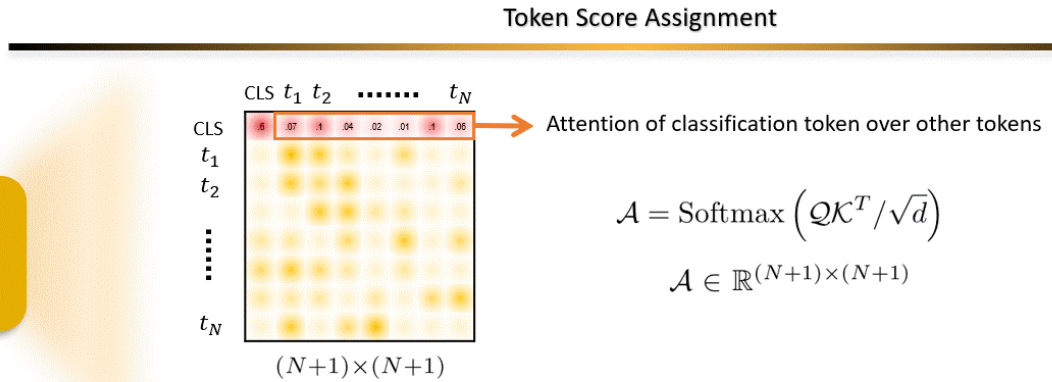


[M. Fayyaz et al. **Adaptive Token Sampling For Efficient Vision Transformers.** ECCV 2022]

Efficient Vision Transformers







Score of token, j=2, ..., N+1

$$\mathcal{S}_j = \frac{\mathcal{A}_{1,j} \times ||\mathcal{V}_j||}{\sum_{i=2} \mathcal{A}_{1,i} \times ||\mathcal{V}_i||}$$

[M. Fayyaz et al. **Adaptive Token Sampling For Efficient Vision Transformers.** ECCV 2022]

Efficient Vision Transformers

Image Tokens

Classification Token

Transformer Block

Transformer Block + ATS

Transformer Block

Classification Head

Image Class

Input Tokens

$\mathcal{I} \in \mathbb{R}^{(N+1)\times d}$

Token Score Assignment

Inverse Transform Sampling

Token Score Assignment

CLS t_1 t_2 ..... t_N

CLS
t_1
t_2
t_N

$(N+1)\times(N+1)$

Attention of classification token over other tokens

$$\mathcal{A} = \text{Softmax}\left(\mathcal{Q}\mathcal{K}^T/\sqrt{d}\right)$$

$$\mathcal{A} \in \mathbb{R}^{(N+1)\times(N+1)}$$

Inverse Transform Sampling

Scores

Input Tokens

$$\mathcal{S}_j = \frac{\mathcal{A}_{1,j} \times ||\mathcal{V}_j||}{\sum_{i=2} \mathcal{A}_{1,i} \times ||\mathcal{V}_i||}$$

[M. Fayyaz et al. **Adaptive Token Sampling For Efficient Vision Transformers.** ECCV 2022]

Efficient Vision Transformers

Image Tokens

Classification Token

Transformer Block

Transformer Block + ATS

Transformer Block

Classification Head

Image Class

Input Tokens

$\mathcal{I} \in \mathbb{R}^{(N+1)\times d}$

Token Score Assignment

Inverse Transform Sampling

Token Score Assignment

CLS t_1 t_2 ..... t_N

CLS t_1 t_2 ⋮ t_N

Attention of classification token over other tokens

$\mathcal{A} = \text{Softmax}\left(\mathcal{Q}\mathcal{K}^T/\sqrt{d}\right)$

$\mathcal{A} \in \mathbb{R}^{(N+1)\times(N+1)}$

$(N+1)\times(N+1)$

Inverse Transform Sampling

Scores

Input Tokens

Cumulative Sum

Sample K points uniformly from y axis

Inverse Transform Sampling

Input Tokens

[M. Fayyaz et al. **Adaptive Token Sampling For Efficient Vision Transformers.** ECCV 2022]

Efficient Vision Transformers

Image Tokens

Classification Token

Transformer Block

Transformer Block + ATS

Transformer Block

Classification Head

Image Class

Input Tokens

$\mathcal{I} \in \mathbb{R}^{(N+1)\times d}$

Token Score Assignment

Inverse Transform Sampling

Token Score Assignment

CLS t_1 t_2 ..... t_N

Attention of classification token over other tokens

$$\mathcal{A} = \text{Softmax}\left(\mathcal{Q}\mathcal{K}^T/\sqrt{d}\right)$$

$\mathcal{A} \in \mathbb{R}^{(N+1)\times(N+1)}$

$(N+1)\times(N+1)$

Inverse Transform Sampling

Scores

Cumulative Sum

Sample K points uniformly from y axis

Input Tokens

Inverse Transform Sampling

Sample unique K' tokens and refine $\mathcal{A}$ to get $\mathcal{A}^s$

$K' \leq K \leq N$

Input Tokens

$\mathcal{A} \rightarrow \mathcal{A}^s$

$\mathcal{A}^s$

$(K'+1)\times(N+1)$

[M. Fayyaz et al. **Adaptive Token Sampling For Efficient Vision Transformers.** ECCV 2022]

Efficient Vision Transformers

Image Tokens

Classification Token

Transformer Block

Transformer Block + ATS

Transformer Block

Classification Head

Image Class

Input Tokens

$\mathcal{I} \in \mathbb{R}^{(N+1)\times d}$

Token Score Assignment

Inverse Transform Sampling

Output Tokens

$\mathcal{O} \in \mathbb{R}^{(K'+1)\times d}$

Token Score Assignment

CLS t_1 t_2 ..... t_N

CLS
t_1
t_2
t_N

$(N+1)\times(N+1)$

Attention of classification token over other tokens

$$\mathcal{A} = \text{Softmax}\left(\mathcal{Q}\mathcal{K}^T/\sqrt{d}\right)$$

$$\mathcal{A} \in \mathbb{R}^{(N+1)\times(N+1)}$$

Inverse Transform Sampling

Scores

Input Tokens

Cumulative Sum

Sample K points uniformly from y axis

Inverse Transform Sampling

Input Tokens

Sample unique K' tokens and refine $\mathcal{A}$ to get $\mathcal{A}^s$

$K' \leq K \leq N$

$\mathcal{A} \rightarrow \mathcal{A}^s$

$\mathcal{A}^s$

$(K'+1)\times(N+1)$

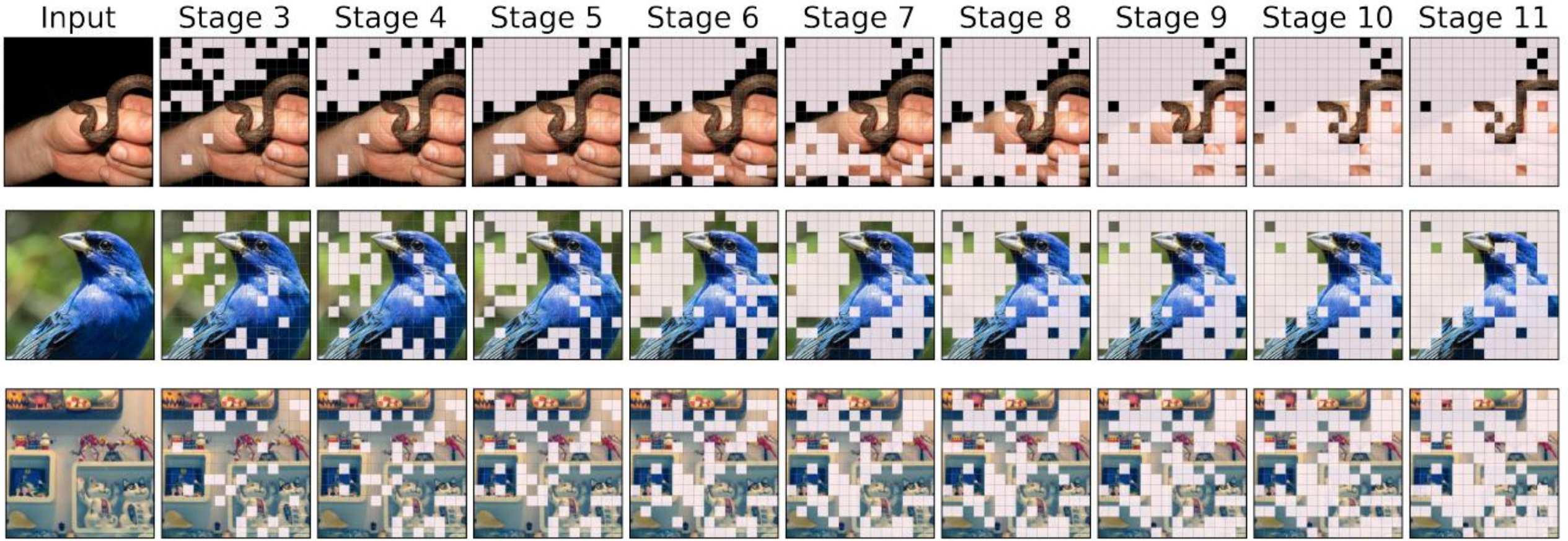
$\mathcal{V}_1$
$\mathcal{V}_2$
$\mathcal{V}_{N+1}$

$(N+1)\times d$

Output Tokens

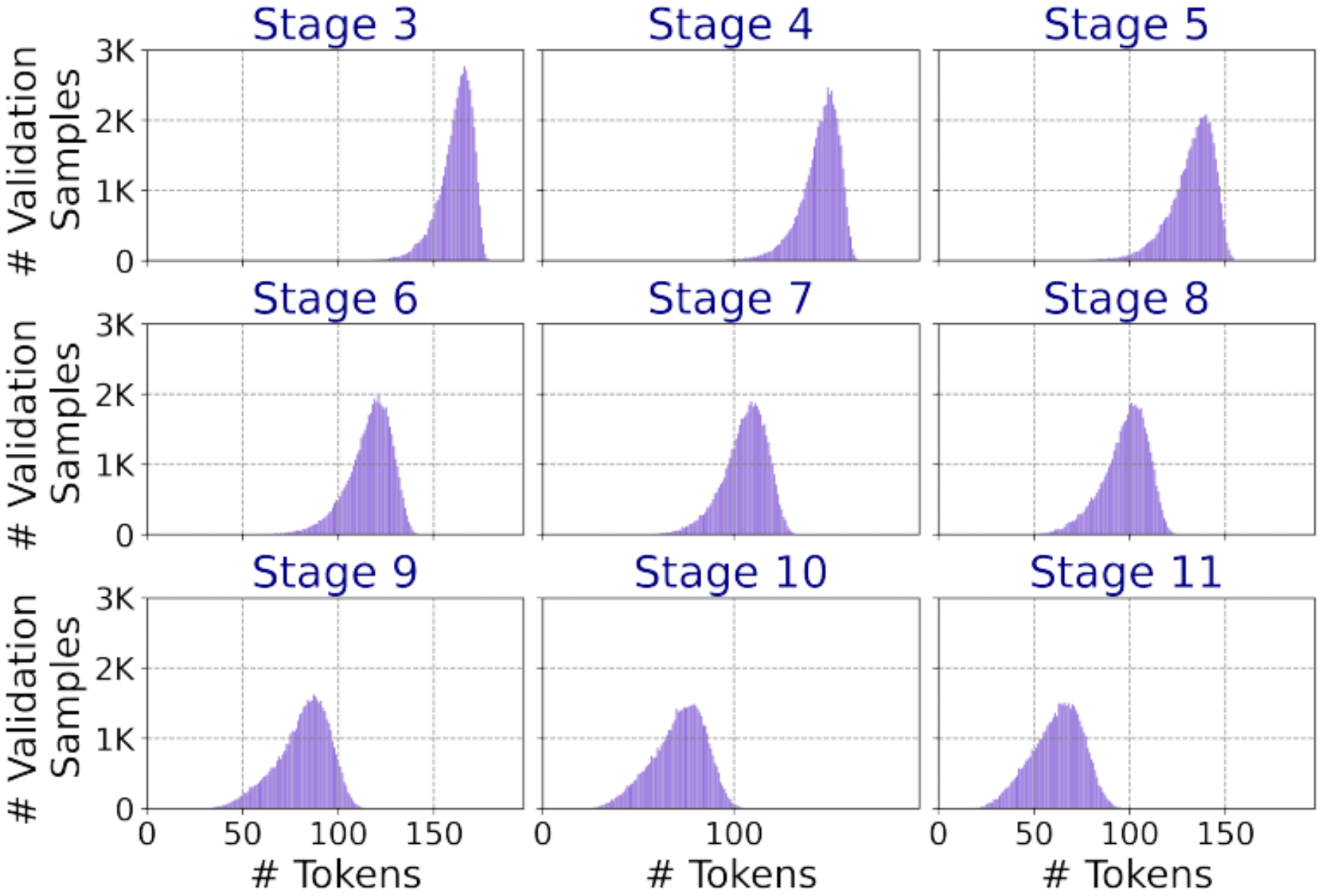
$\mathcal{O} \in \mathbb{R}^{(K'+1)\times d}$

Qualitative Results

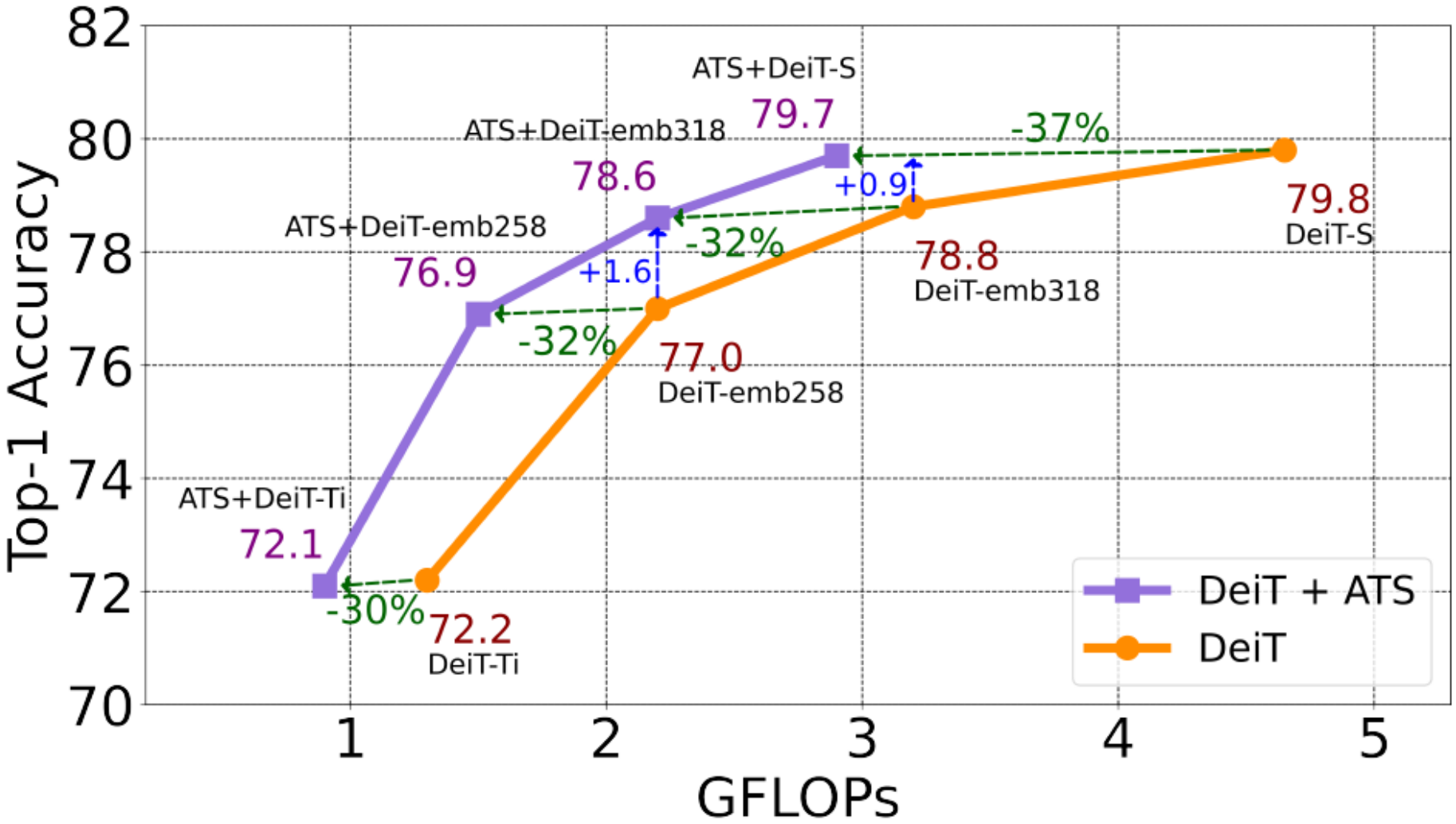


[M. Fayyaz et al. **Adaptive Token Sampling For Efficient Vision Transformers.** ECCV 2022]

Quantitative Results



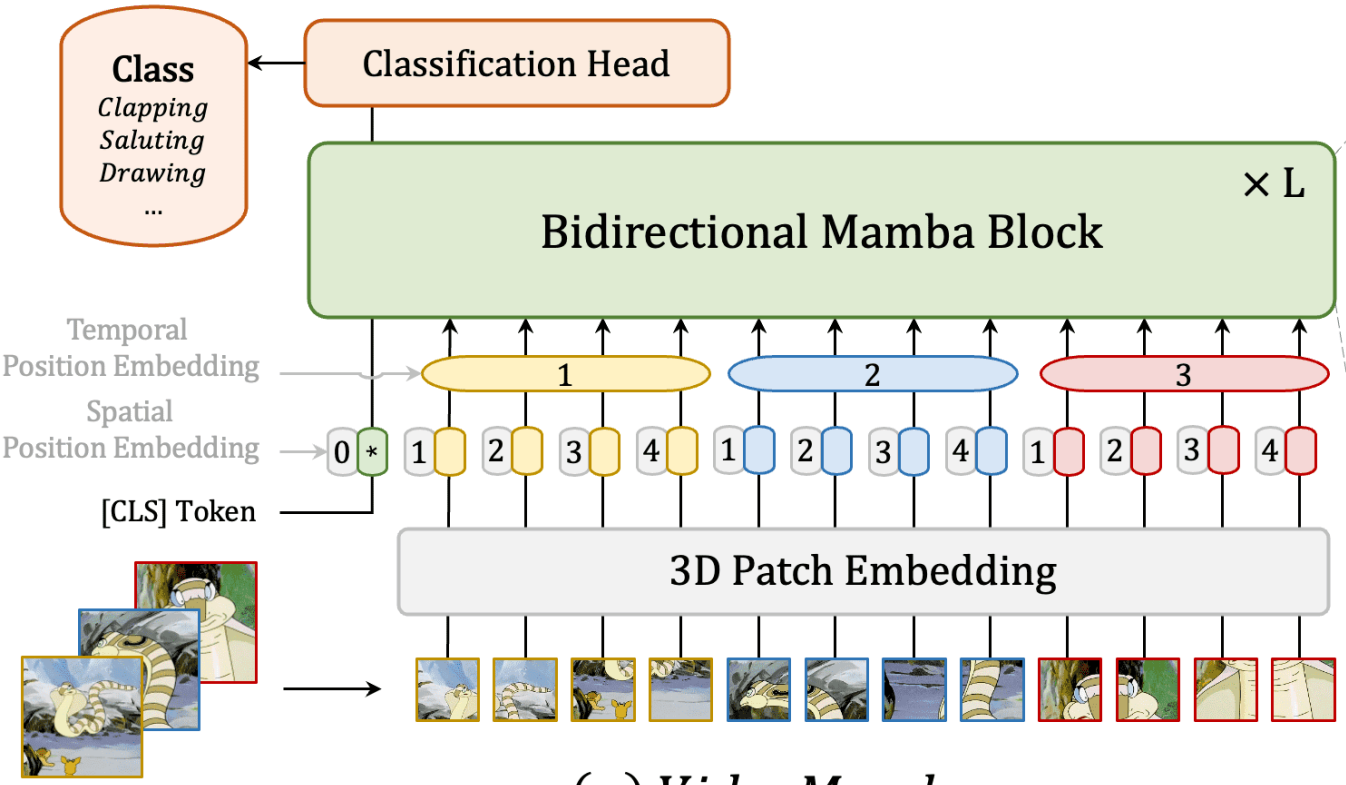
Quantitative Results



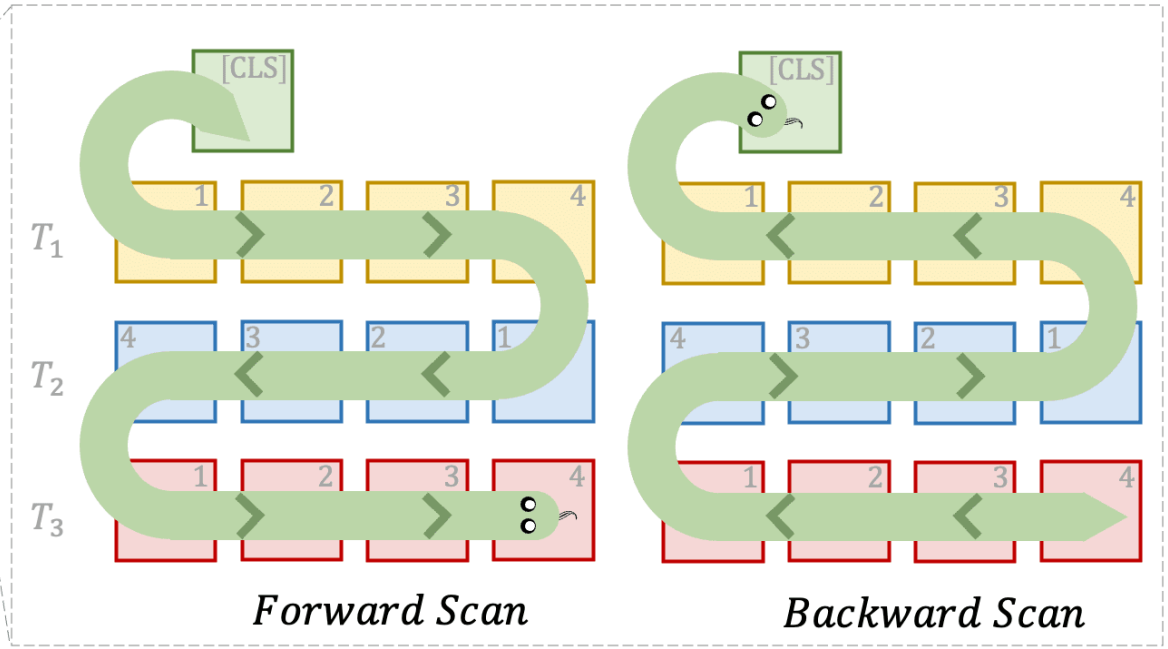
[M. Fayyaz et al. **Adaptive Token Sampling For Efficient Vision Transformers.** ECCV 2022]

Replace attention by state space model

State Space Model for Video Understanding



(a) VideoMamba

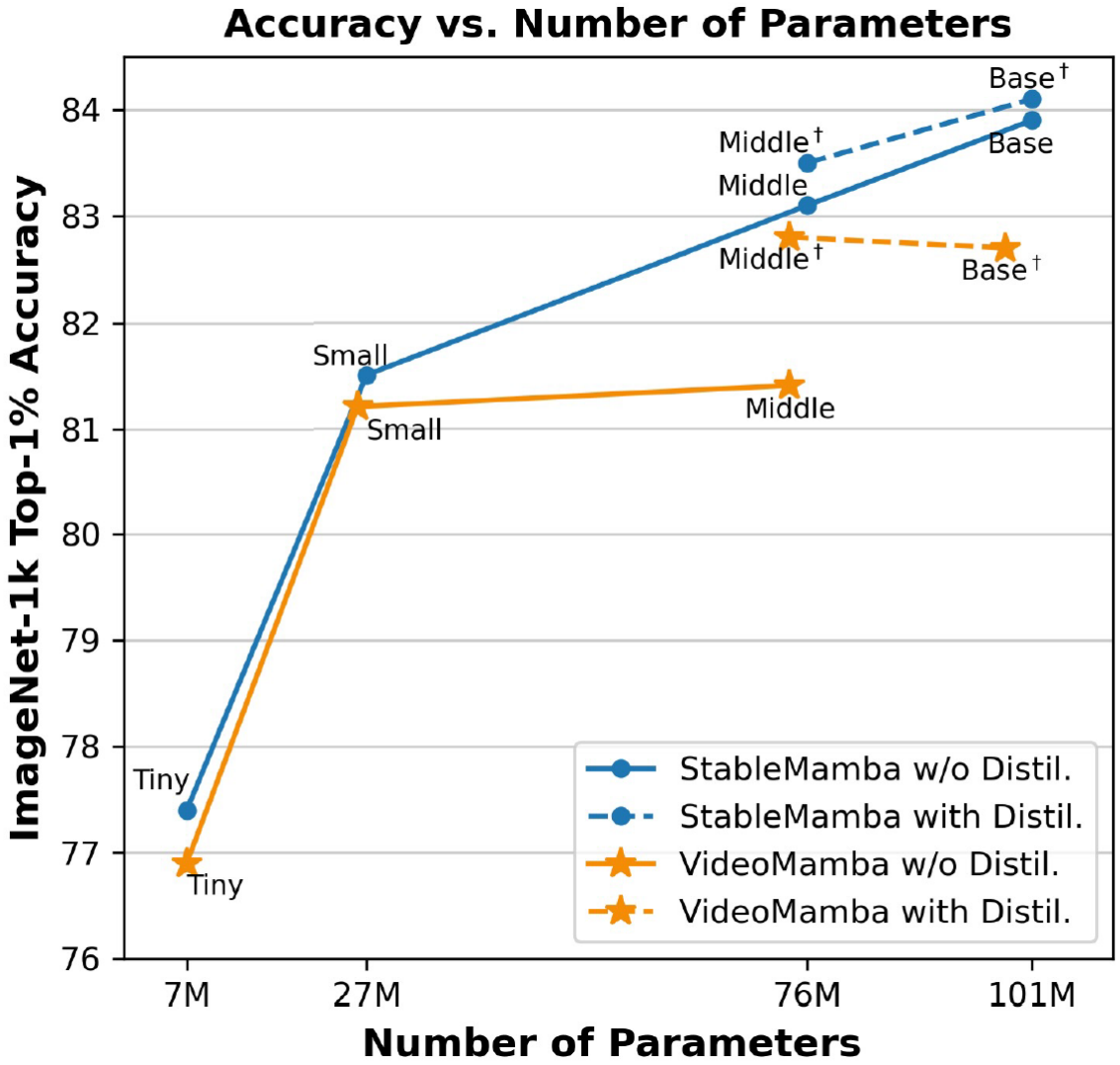


(b) Spatiotemporal Scan

[K. Li et al. **VideoMamba: State Space Model for Efficient Video Understanding.** ECCV 2024]

State Space Model for Video Understanding

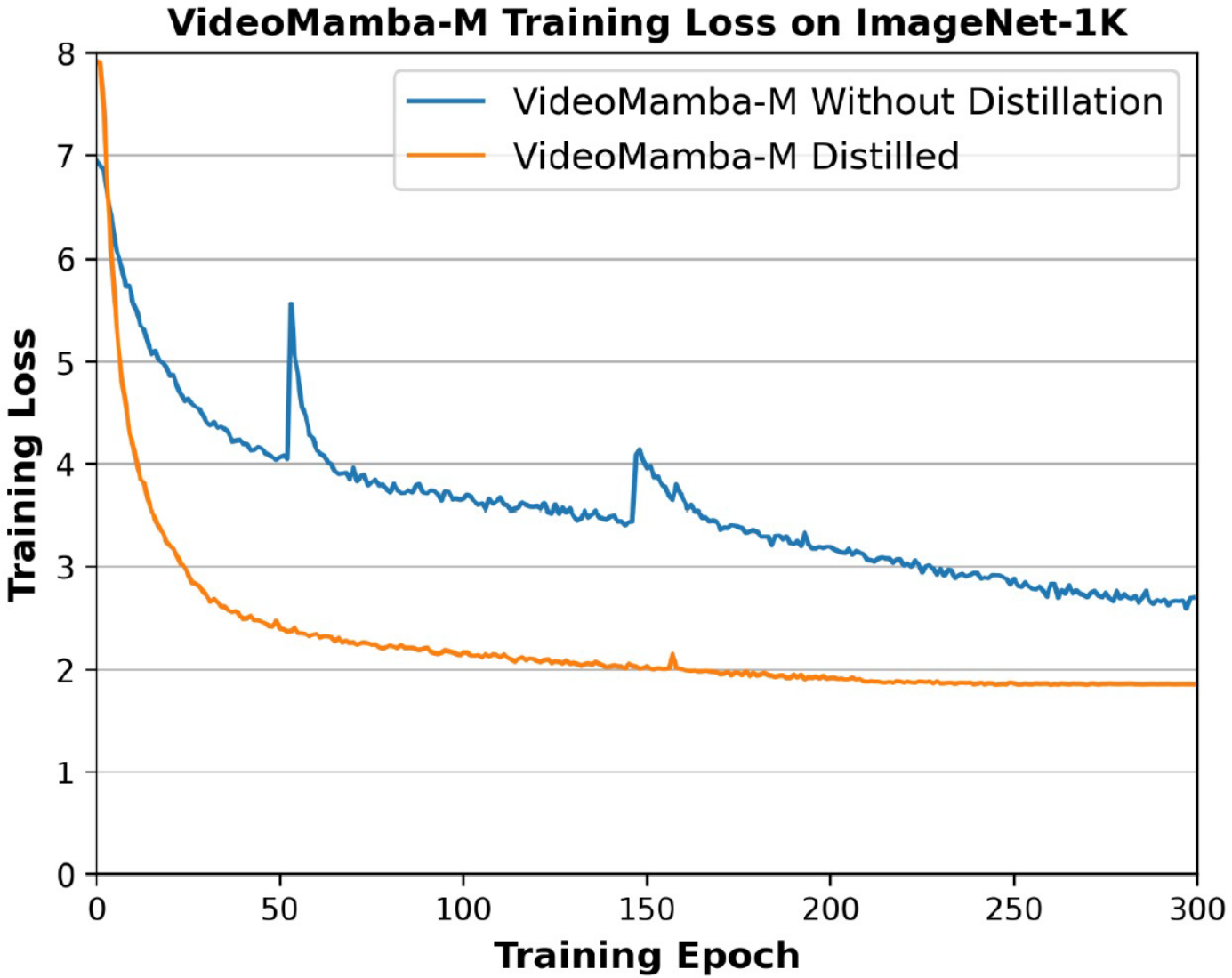
- VideoMamba has major shortcomings
  - Does not scale



[H. Suleman et al. **Distillation-free Scaling of Large SSMs for Images and Videos.** arxiv 2024]

State Space Model for Video Understanding

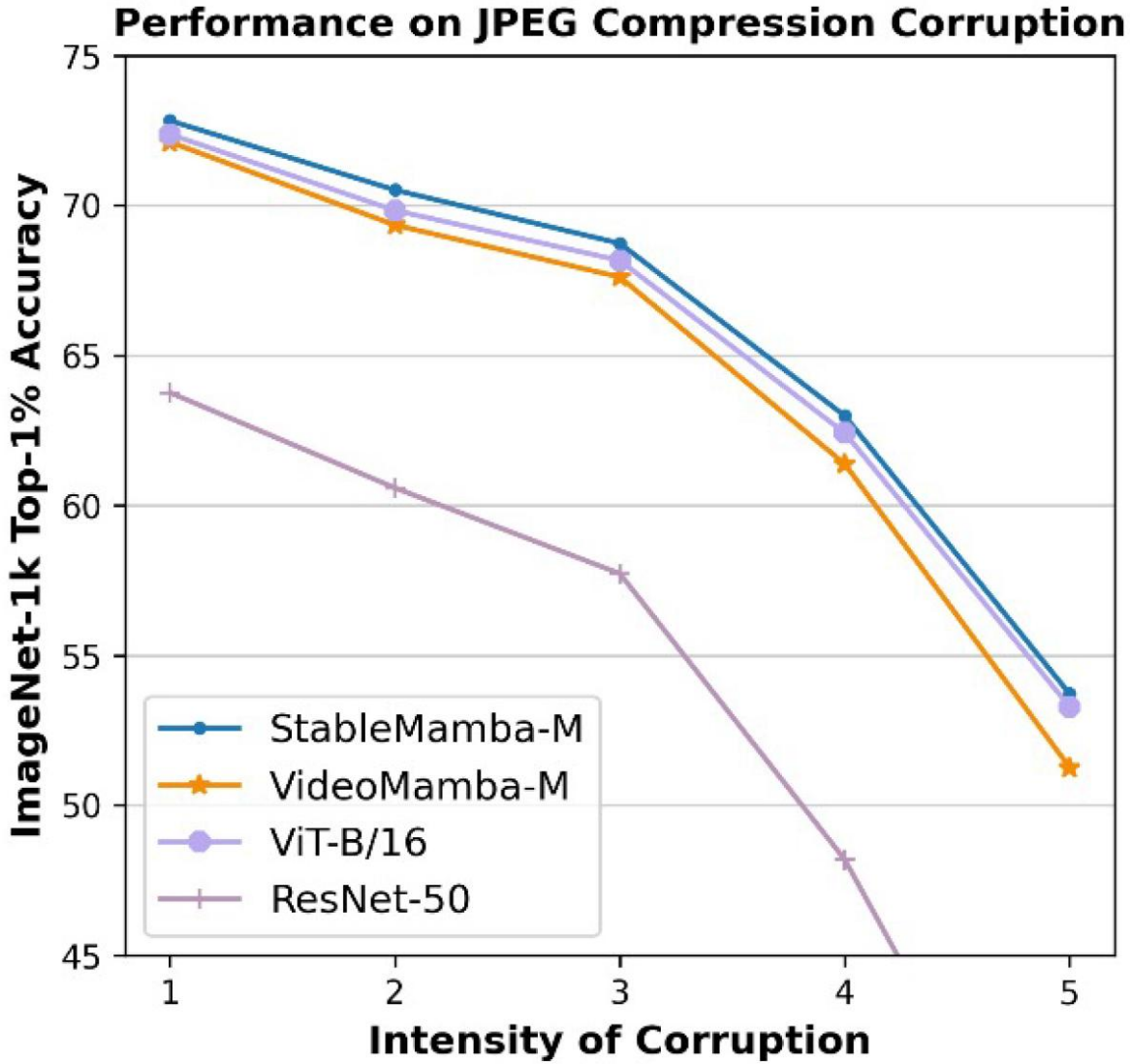
- VideoMamba has major shortcomings
  - Does not scale
  - Instable training



[H. Suleman et al. **Distillation-free Scaling of Large SSMs for Images and Videos.** arxiv 2024]

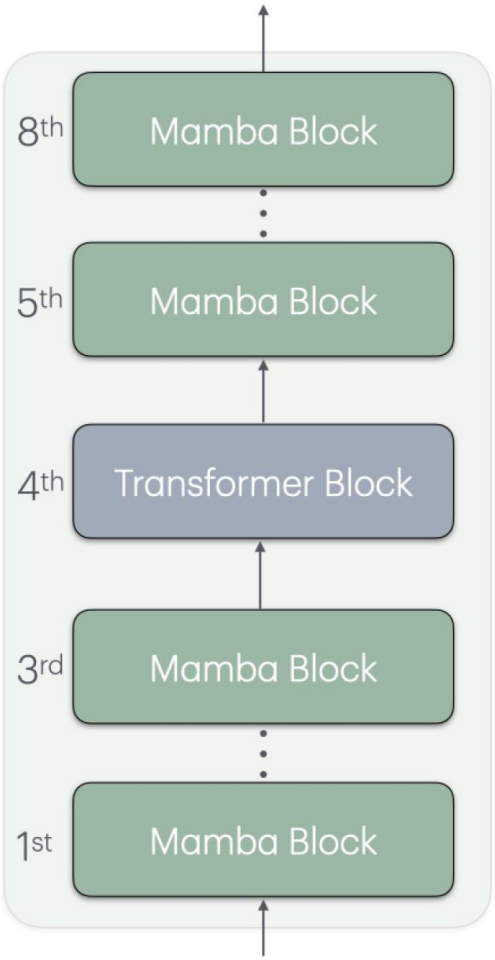
State Space Model for Video Understanding

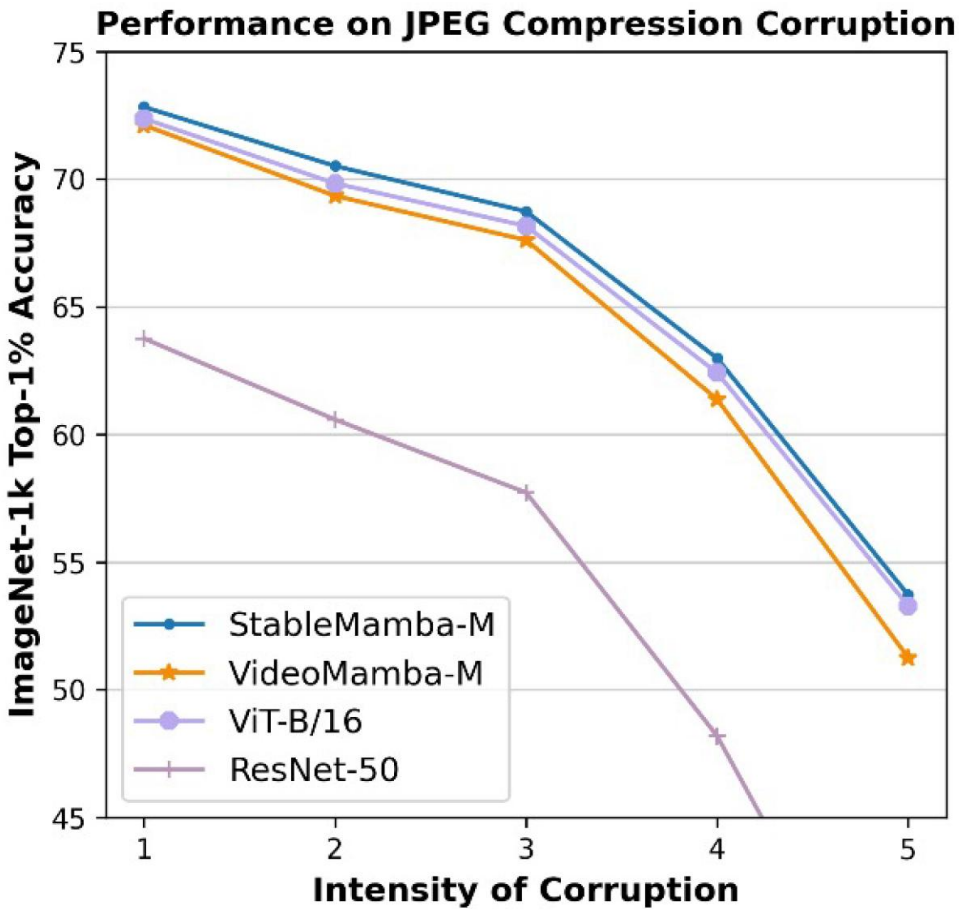
- VideoMamba has major shortcomings
  - Does not scale
  - Instable training
  - More sensitive to image corruption compared to transformer

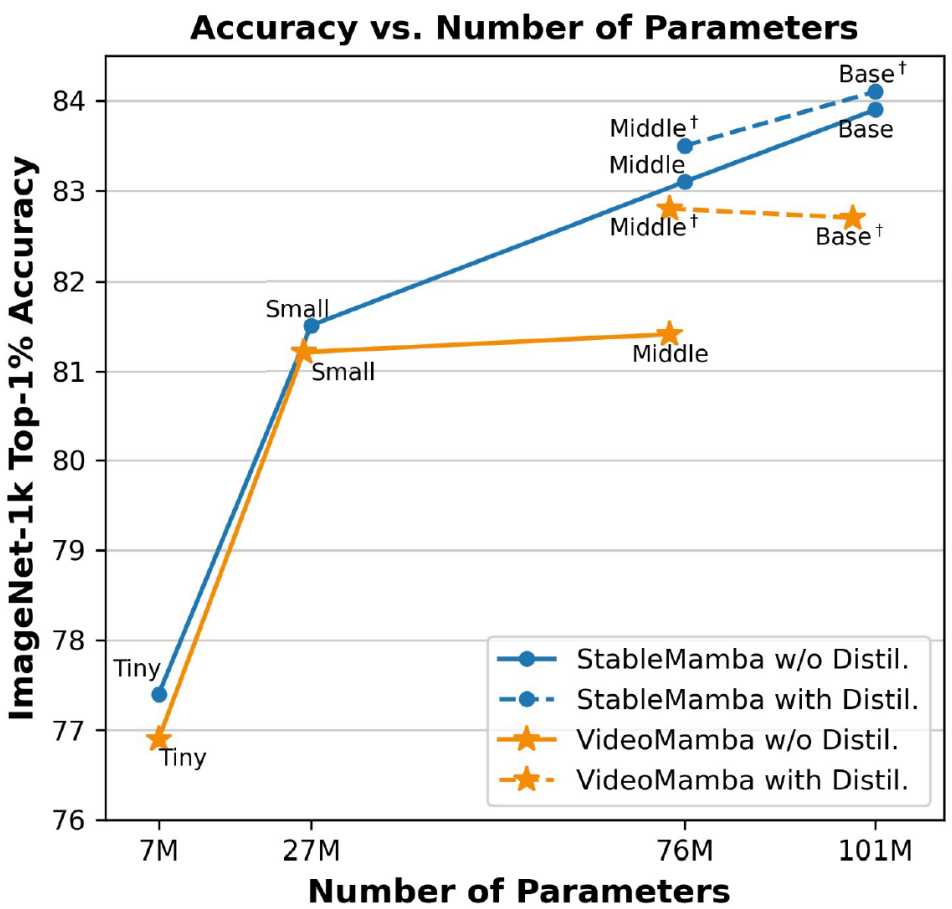


[H. Suleman et al. **Distillation-free Scaling of Large SSMs for Images and Videos.** arxiv 2024]

State Space Model for Video Understanding







[H. Suleman et al. **Distillation-free Scaling of Large SSMs for Images and Videos.** arxiv 2024]


Avoid vision encoder for
video-language models

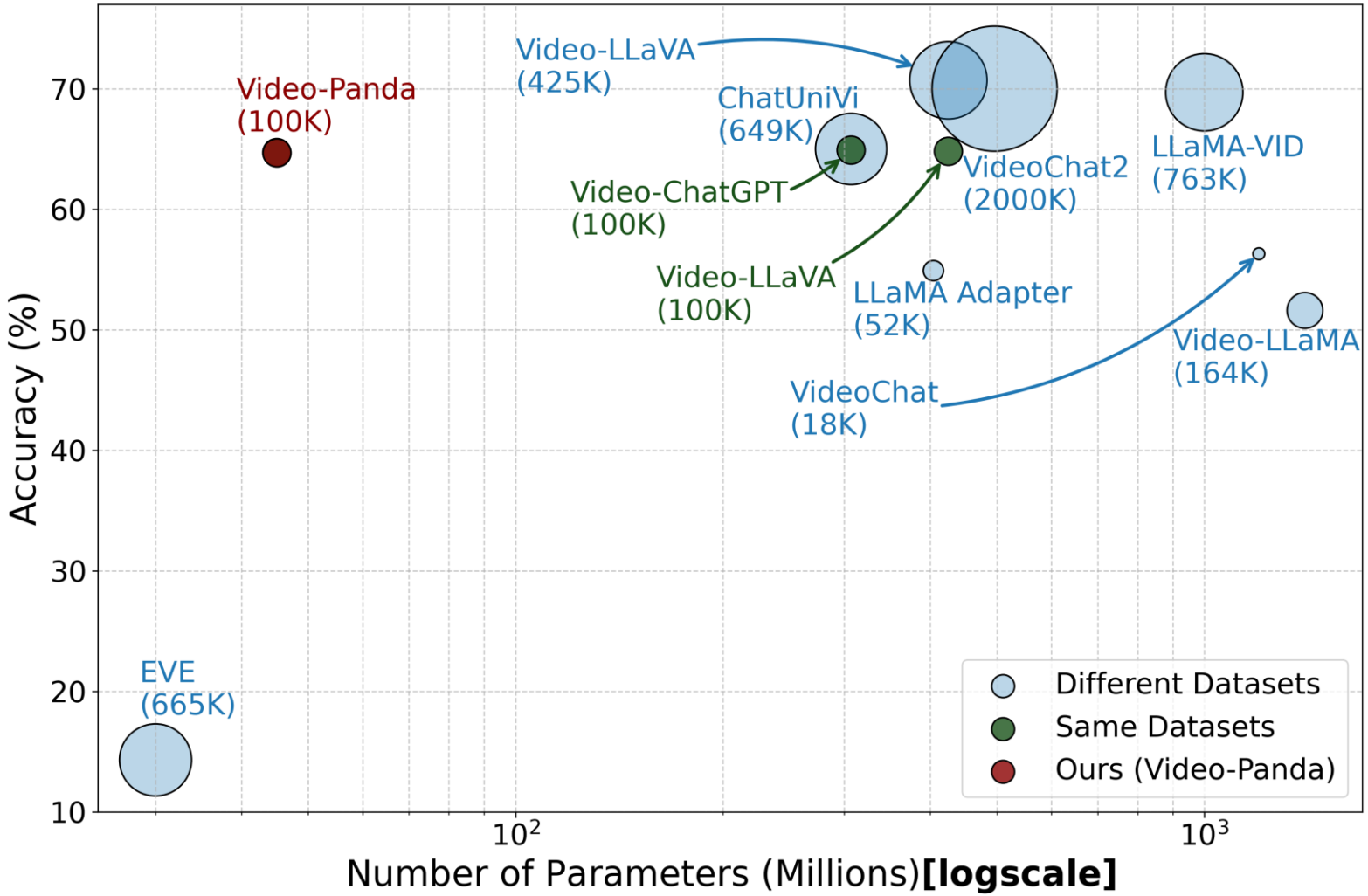
Video Question Answering

Video-Panda

Demos from MSVD-QA

[J. Yi et al. **Video-Panda: Parameter-efficient Alignment for Encoder-free Video-Language Models.** CVPR 2025]

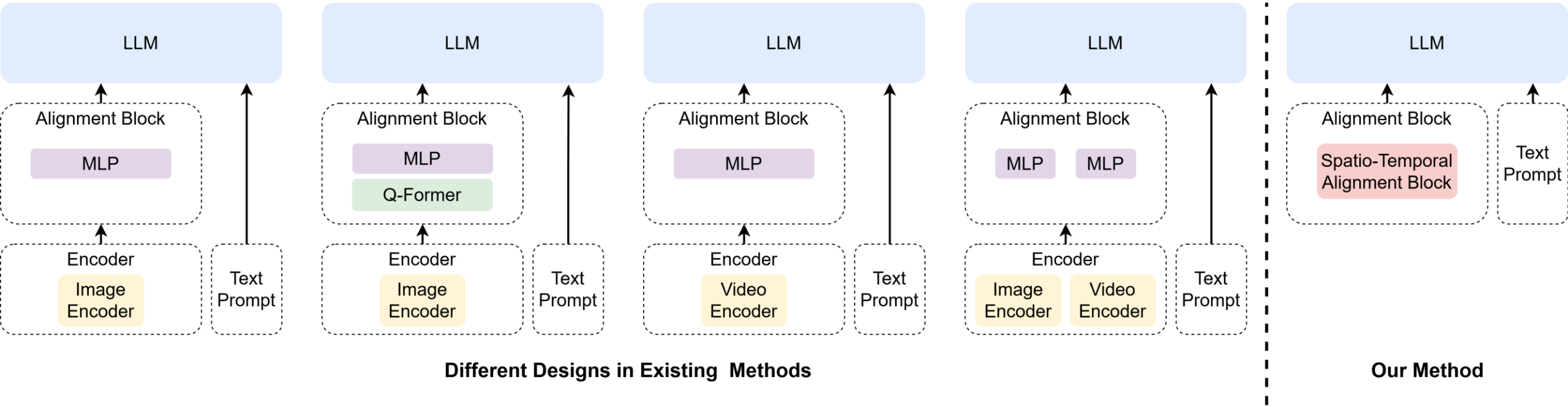
Parameter-Efficient Video-Language Models



[J. Yi et al. **Video-Panda: Parameter-efficient Alignment for Encoder-free Video-Language Models.** CVPR 2025]

Parameter-Efficient Video-Language Models

- No need for additional vision encoders



[J. Yi et al. **Video-Panda: Parameter-efficient Alignment for Encoder-free Video-Language Models.** CVPR 2025]

Parameter-Efficient Video-Language Models

- No need for additional vision encoders

Model	#Param.(M)	Inference time (ms)
VideoChatGPT [26]	307	171
Video-LLaVA [22]	425	125
Video-Panda	45	41

[J. Yi et al. **Video-Panda: Parameter-efficient Alignment for Encoder-free Video-Language Models.** CVPR 2025]

Video Question Answering

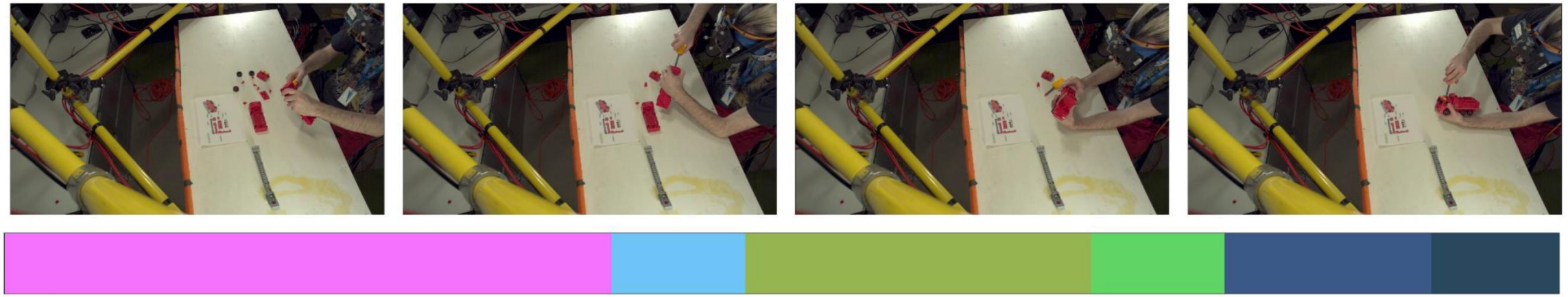
Video-Panda

Demos from TGIF-QA

[J. Yi et al. **Video-Panda: Parameter-efficient Alignment for Encoder-free Video-Language Models.** CVPR 2025]

Combine local and global attention

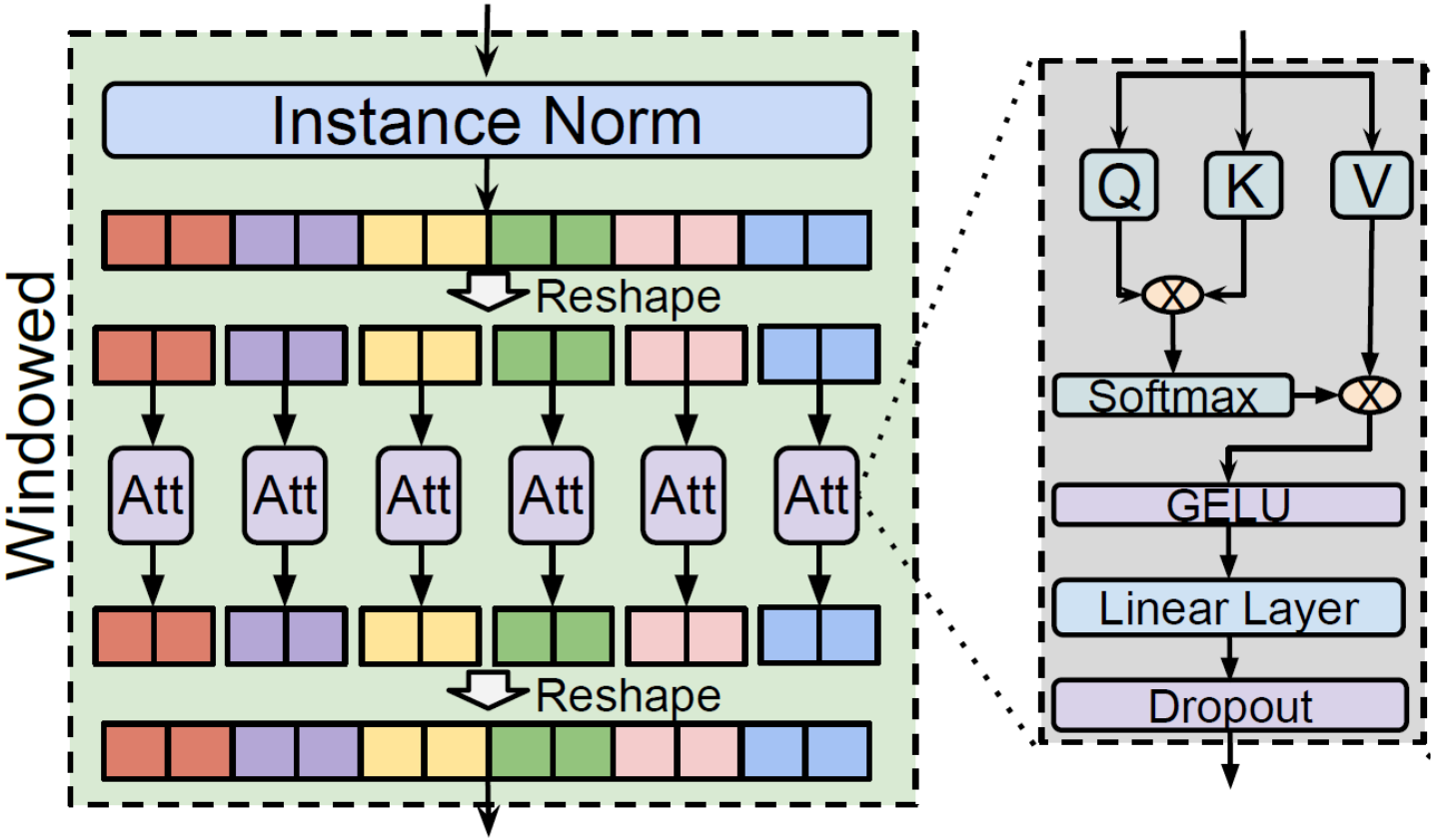
Vision Transformers for Action Segmentation



[E. Bahrami et al. **How Much Temporal Long-Term Context is Needed for Action Segmentation?** ICCV 2023]

Vision Transformers for Action Segmentation

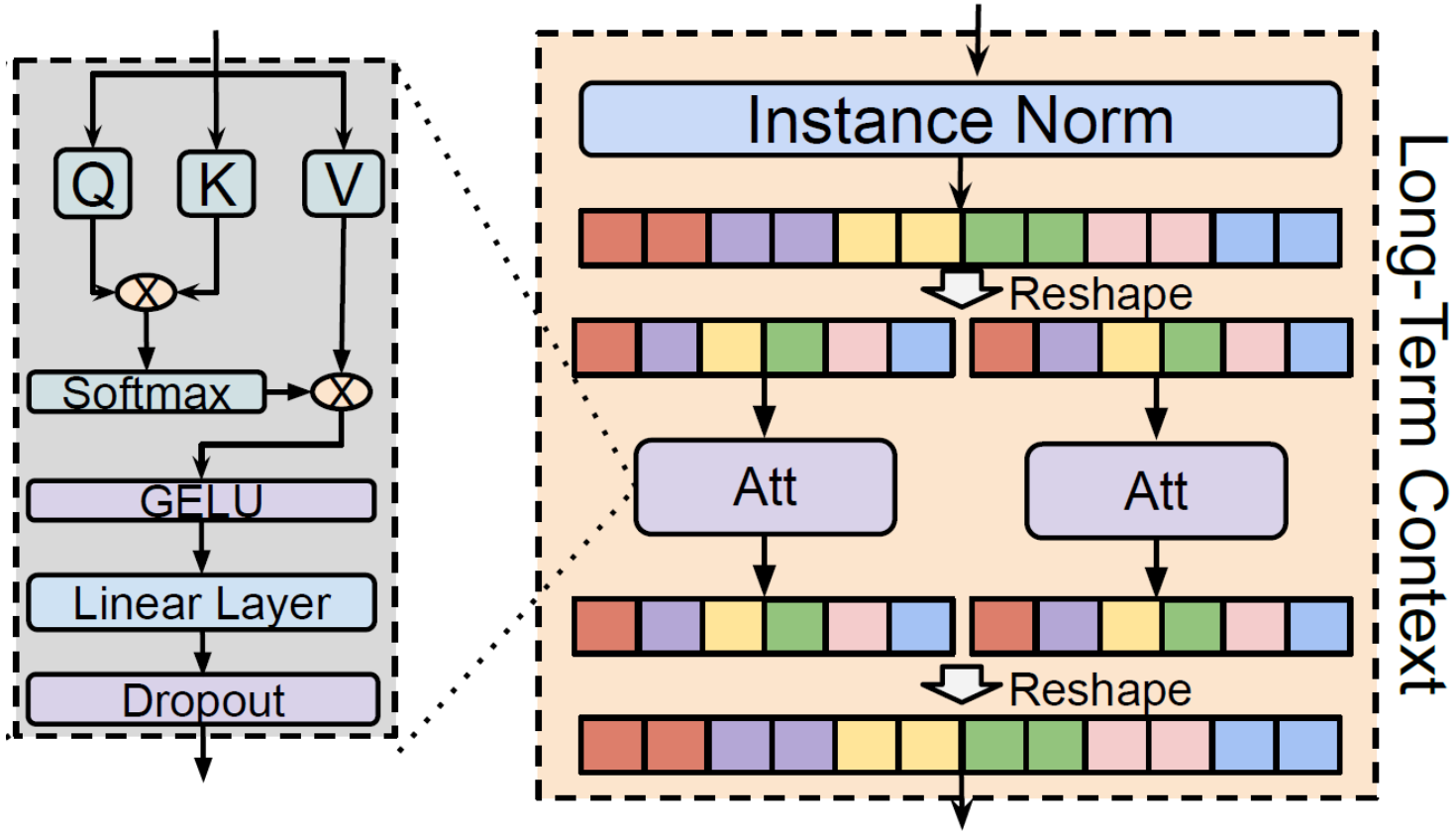
Attention over local window



[E. Bahrami et al. **How Much Temporal Long-Term Context is Needed for Action Segmentation?** ICCV 2023]

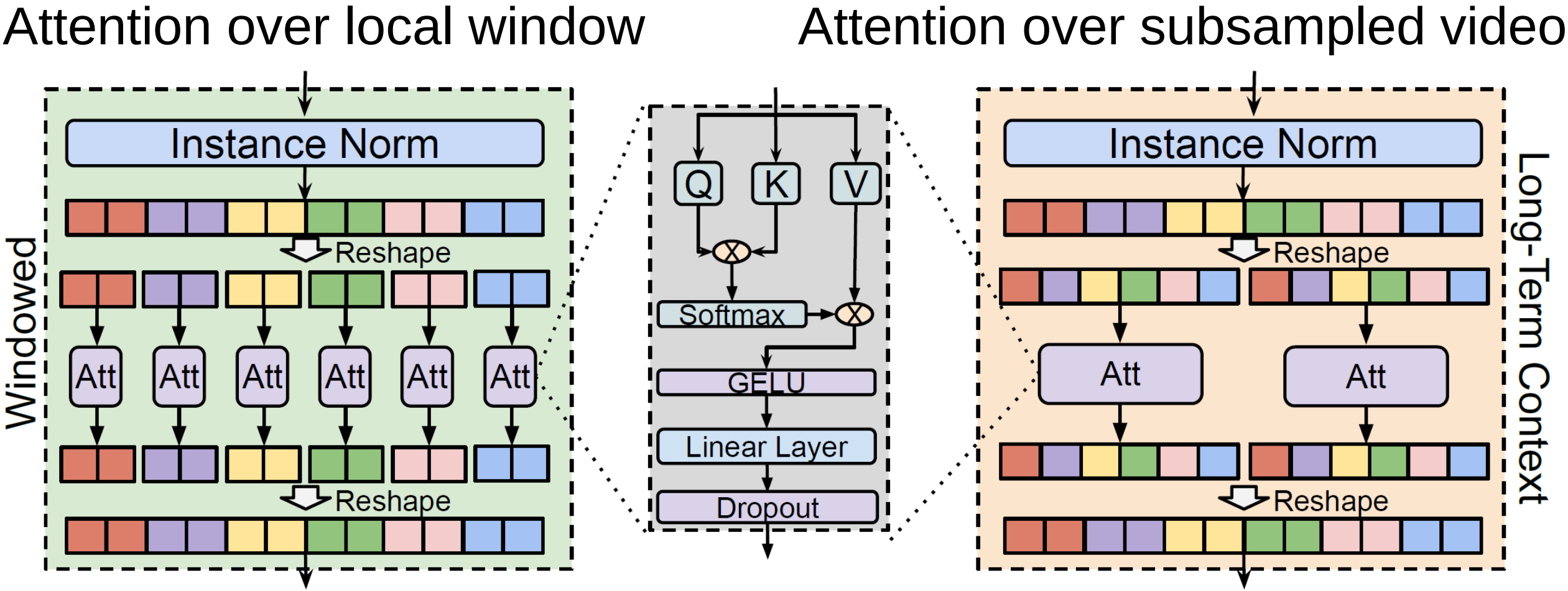
Vision Transformers for Action Segmentation

Attention over subsampled video



[E. Bahrami et al. **How Much Temporal Long-Term Context is Needed for Action Segmentation?** ICCV 2023]

Vision Transformers for Action Segmentation

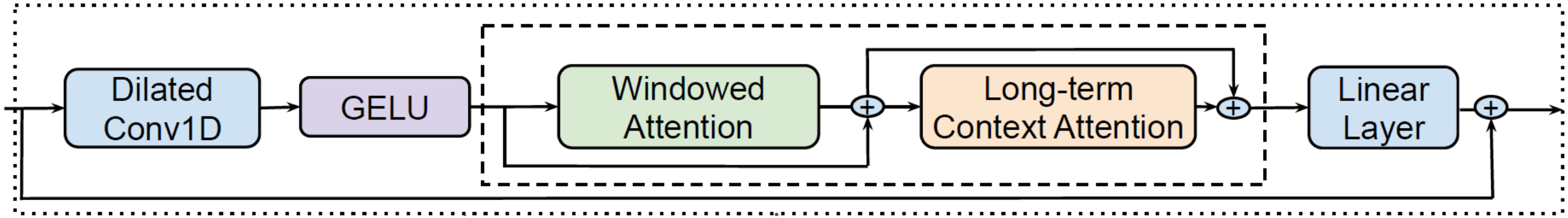


[E. Bahrami et al. **How Much Temporal Long-Term Context is Needed for Action Segmentation?** ICCV 2023]

Vision Transformers for Action Segmentation

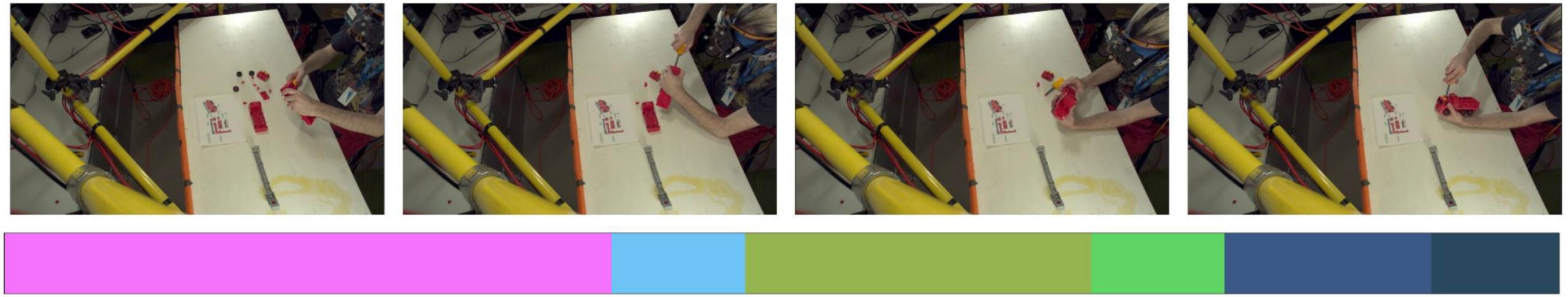
Attention over local window

Attention over subsampled video



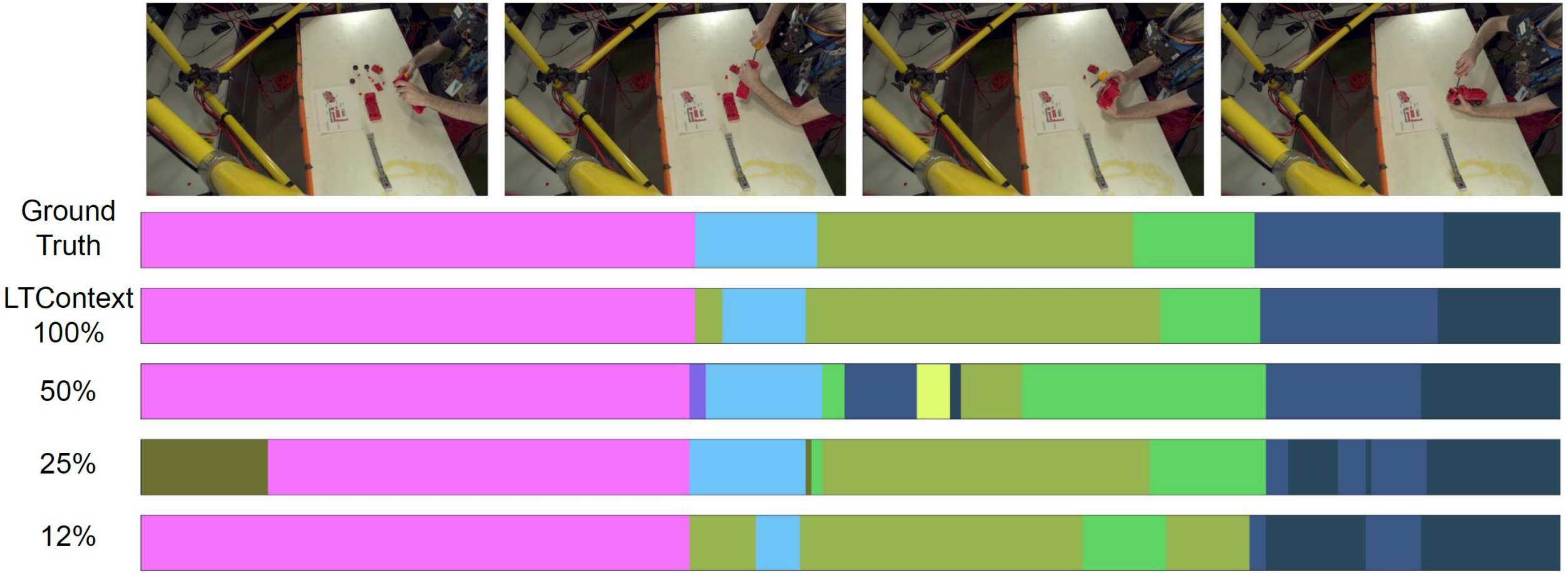
[E. Bahrami et al. **How Much Temporal Long-Term Context is Needed for Action Segmentation?** ICCV 2023]

Vision Transformers for Action Segmentation



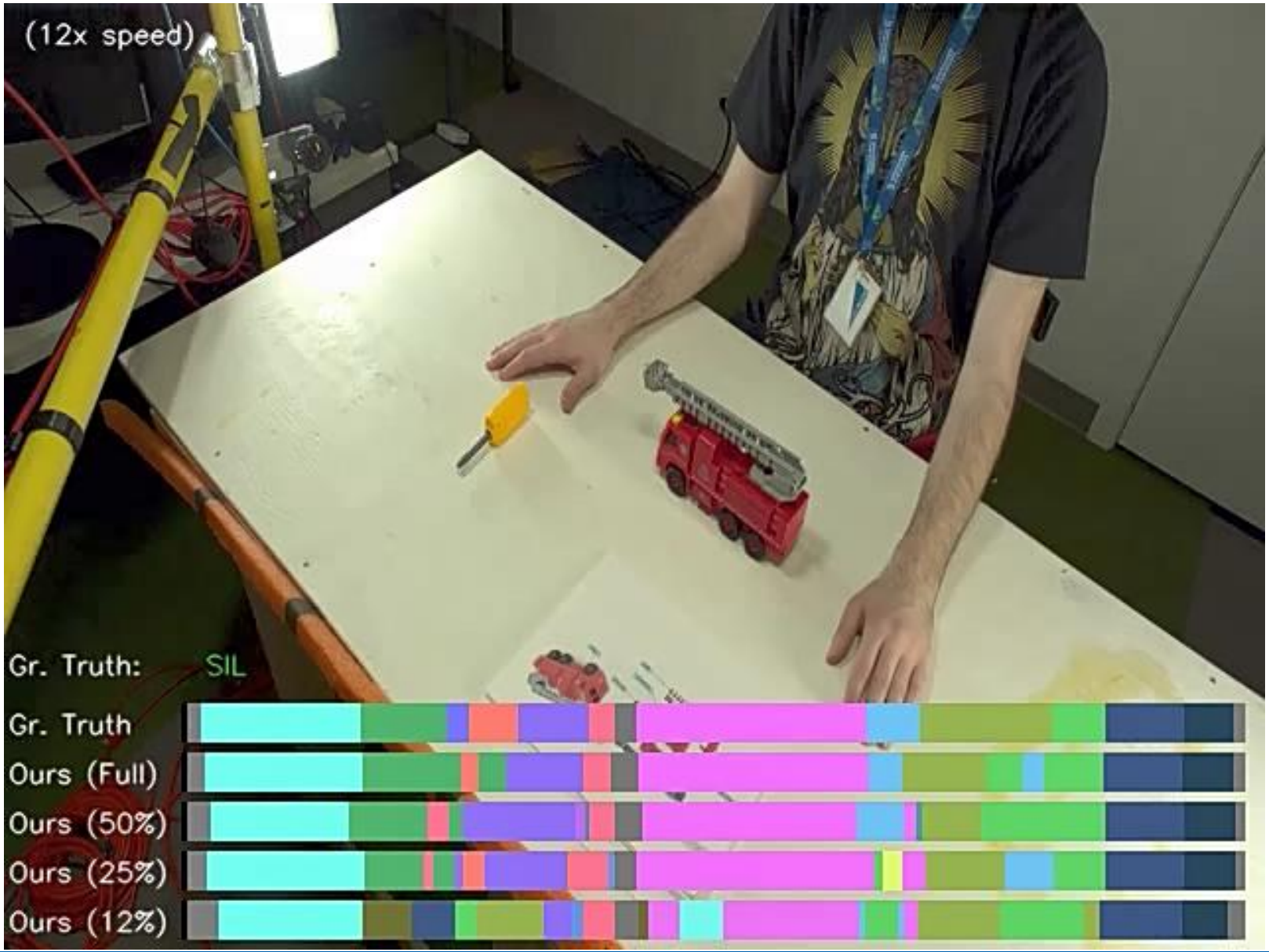
[E. Bahrami et al. **How Much Temporal Long-Term Context is Needed for Action Segmentation?** ICCV 2023]

Vision Transformers for Action Segmentation



[E. Bahrami et al. **How Much Temporal Long-Term Context is Needed for Action Segmentation?** ICCV 2023]

Vision Transformers for Action Segmentation



[E. Bahrami et al. **How Much Temporal Long-Term Context is Needed for Action Segmentation?** ICCV 2023]

Anticipating Behavior

Past → Anticipate → Future

[Y. Abu Farha et al. **When will you do what? Anticipating Temporal Occurrences of Activities.** CVPR 2018]

Gated Temporal Diffusion for Anticipation

- Forecast future actions and their durations

[O. Zatsarynna et al. **Gated Temporal Diffusion for Stochastic Long-term Dense Anticipation.** ECCV 2024]

Gated Temporal Diffusion for Anticipation

- Forecast future actions and their durations
- Model uncertainty of the future

[O. Zatsarynna et al. **Gated Temporal Diffusion for Stochastic Long-term Dense Anticipation.** ECCV 2024]

Gated Temporal Diffusion for Anticipation

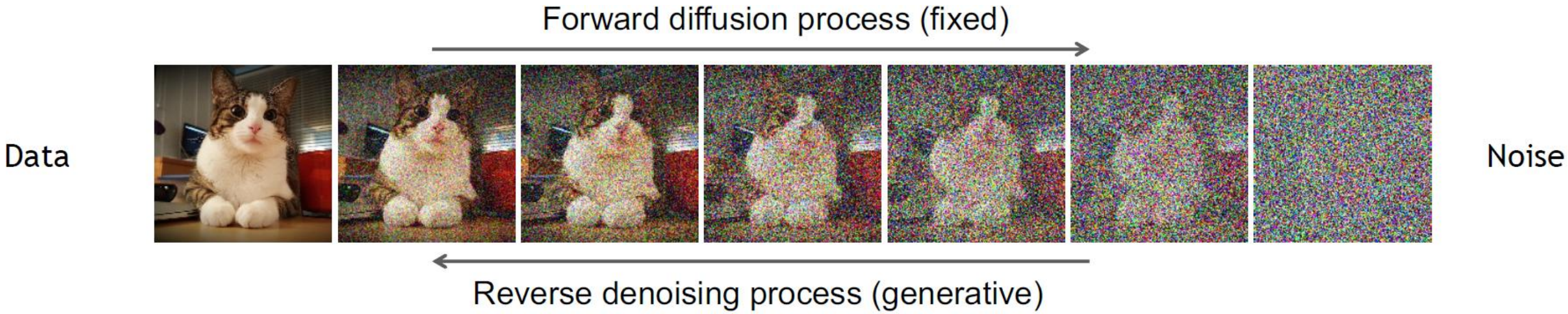
- Forecast future actions and their durations
- Model uncertainty of the future
- Model uncertainty of the observation

[O. Zatsarynna et al. **Gated Temporal Diffusion for Stochastic Long-term Dense Anticipation.** ECCV 2024]

Diffusion Model

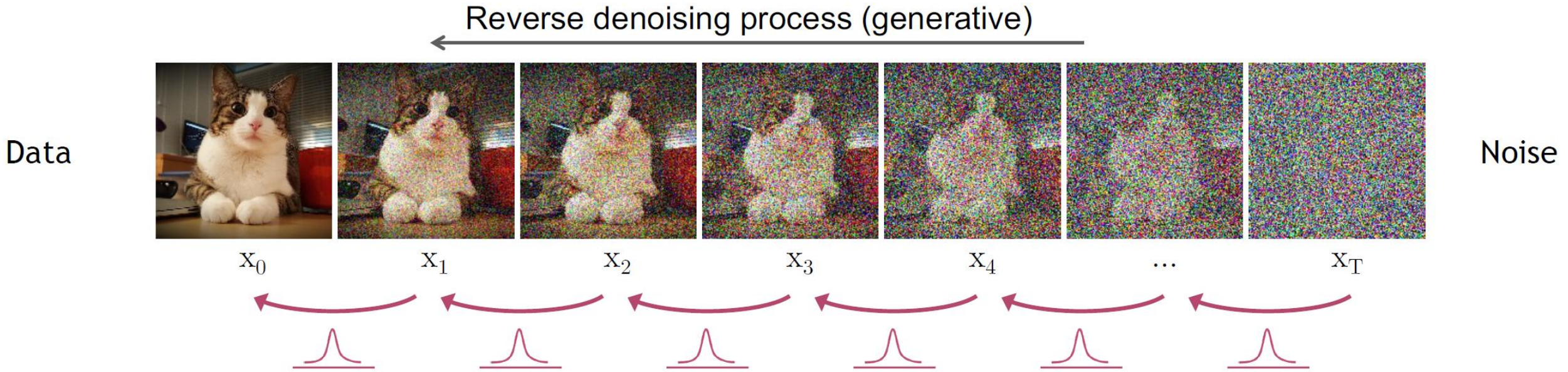
Diffusion models consist of two processes:

- Forward diffusion process that gradually adds noise to input
- Reverse denoising process that learns to generate data by denoising



Generative Process

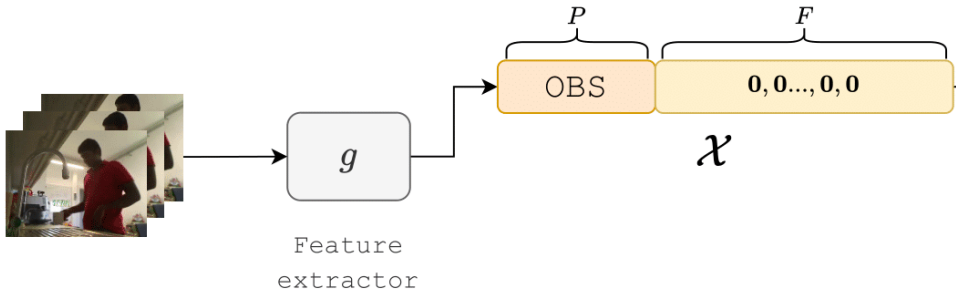
Reverse denoising process



$$\hat{\mathbf{X}}_{0,t-1} = g(q(\hat{\mathbf{X}}_{t-1}|\hat{\mathbf{X}}_{0,t}), t-1)$$

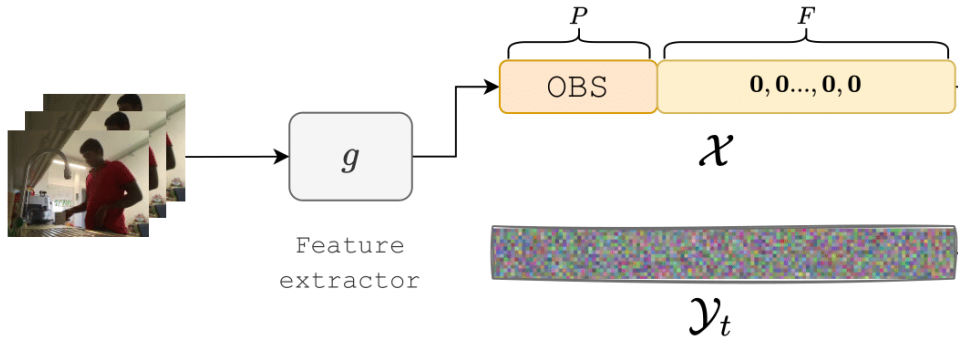
g is a network

MANTA: Diffusion Mamba for Anticipation



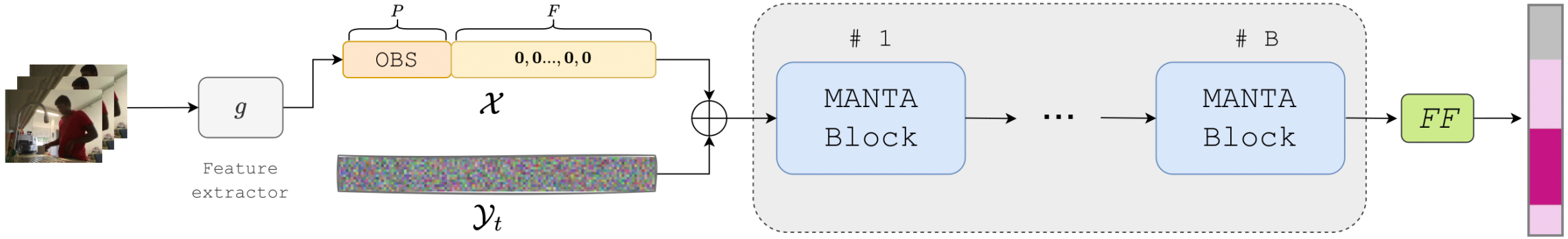
[O. Zatsarynna et al. **MANTA: Diffusion Mamba for Efficient and Effective Stochastic Long-Term Dense Anticipation.** CVPR 2025]

MANTA: Diffusion Mamba for Anticipation



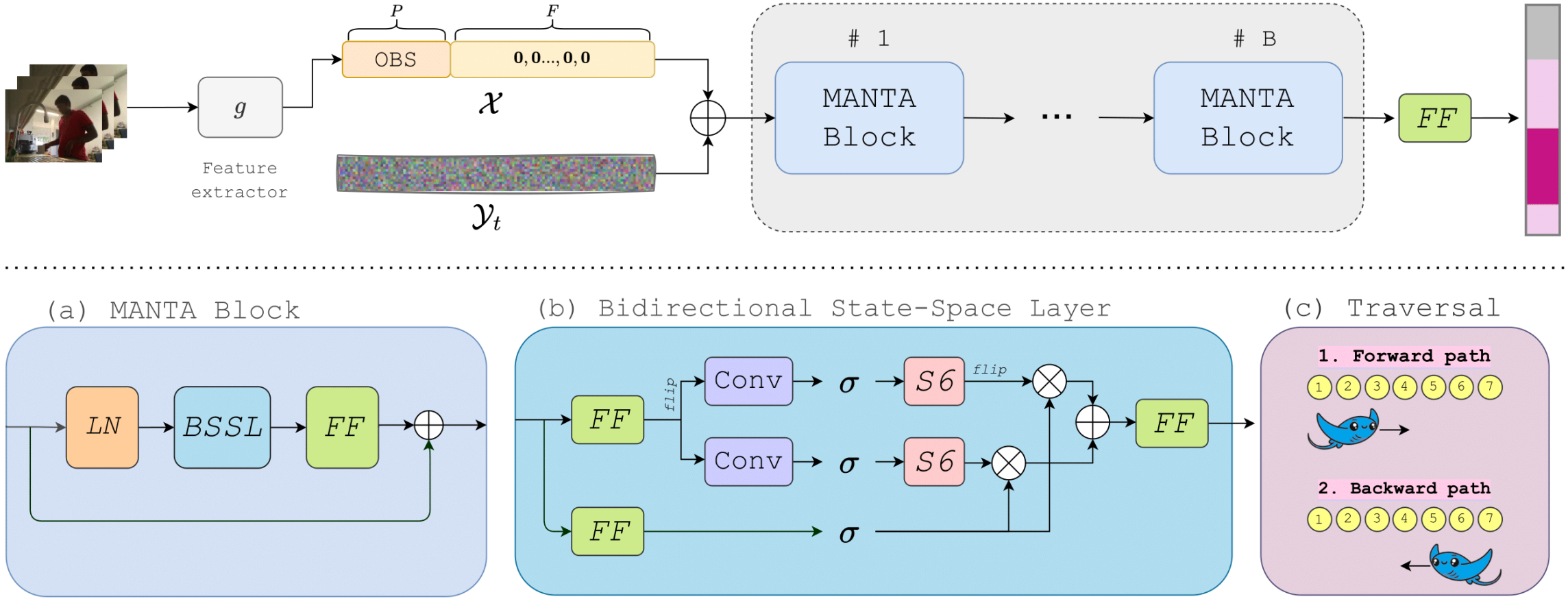
[O. Zatsarynna et al. **MANTA: Diffusion Mamba for Efficient and Effective Stochastic Long-Term Dense Anticipation.** CVPR 2025]

MANTA: Diffusion Mamba for Anticipation



[O. Zatsarynna et al. **MANTA: Diffusion Mamba for Efficient and Effective Stochastic Long-Term Dense Anticipation.** CVPR 2025]

MANTA: Diffusion Mamba for Anticipation

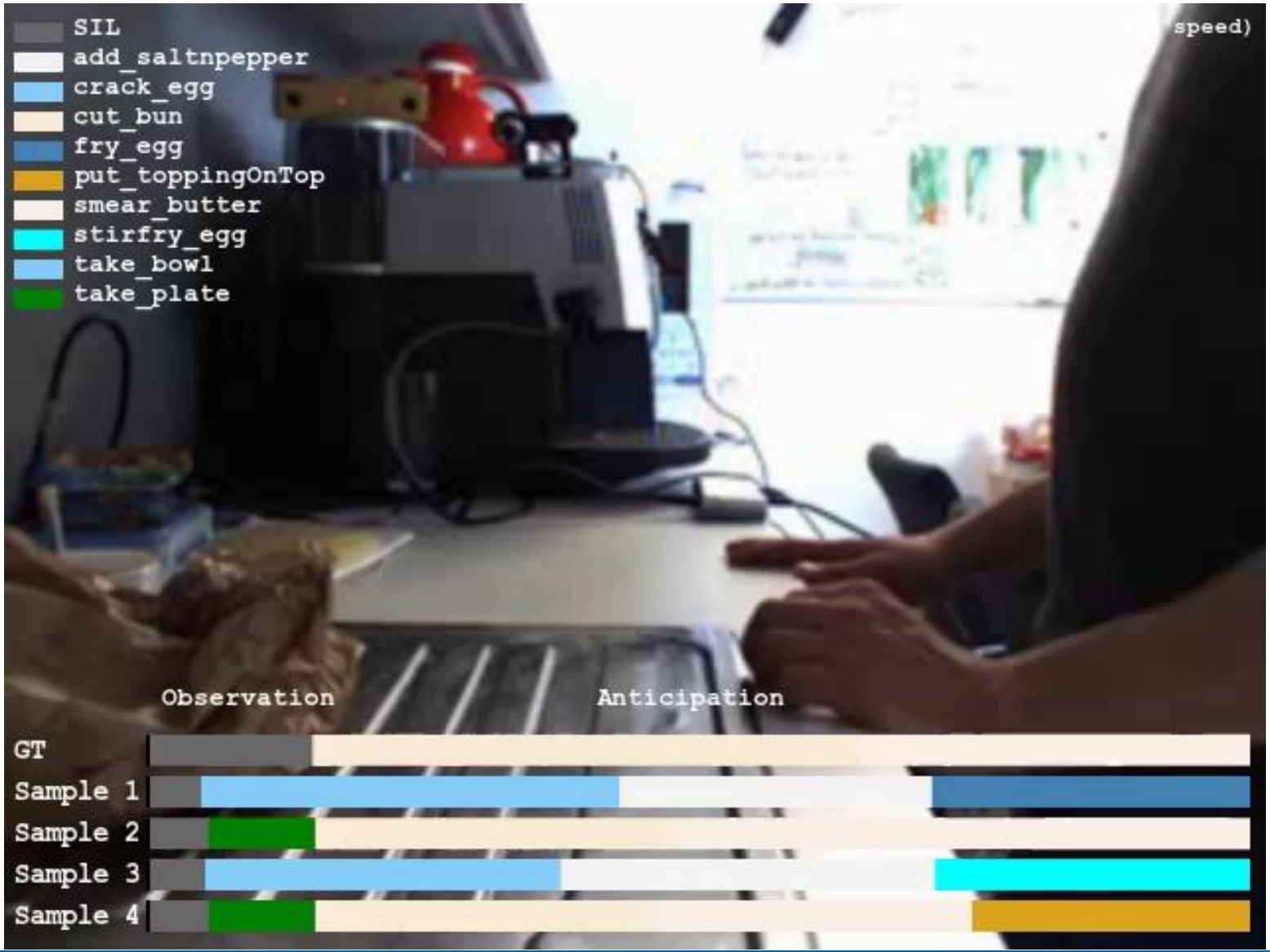


[O. Zatsarynna et al. **MANTA: Diffusion Mamba for Efficient and Effective Stochastic Long-Term Dense Anticipation.** CVPR 2025]

MANTA: Diffusion Mamba for Anticipation

[O. Zatsarynna et al. **MANTA: Diffusion Mamba for Efficient and Effective Stochastic Long-Term Dense Anticipation.** CVPR 2025]

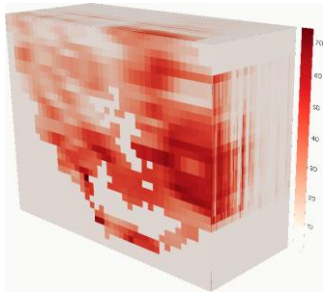
MANTA: Diffusion Mamba for Anticipation



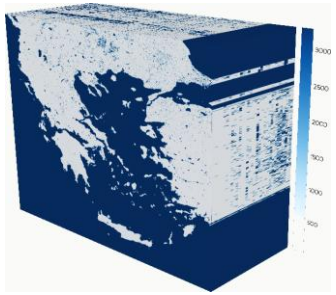
[O. Zatsarynna et al. **MANTA: Diffusion Mamba for Efficient and Effective Stochastic Long-Term Dense Anticipation.** CVPR 2025]


Forecasting
Hundreds left homeless as Greek fires rage

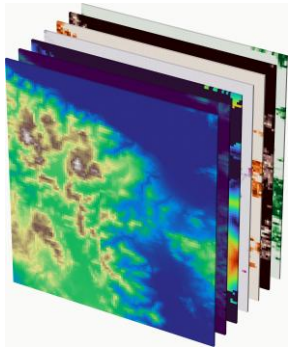
Wildfire Forecasting



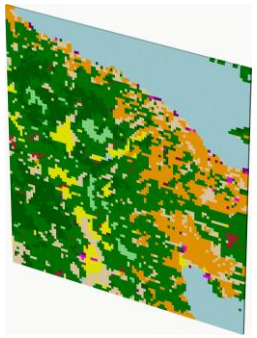
Meteorological data



Satellite products



Topographic variables

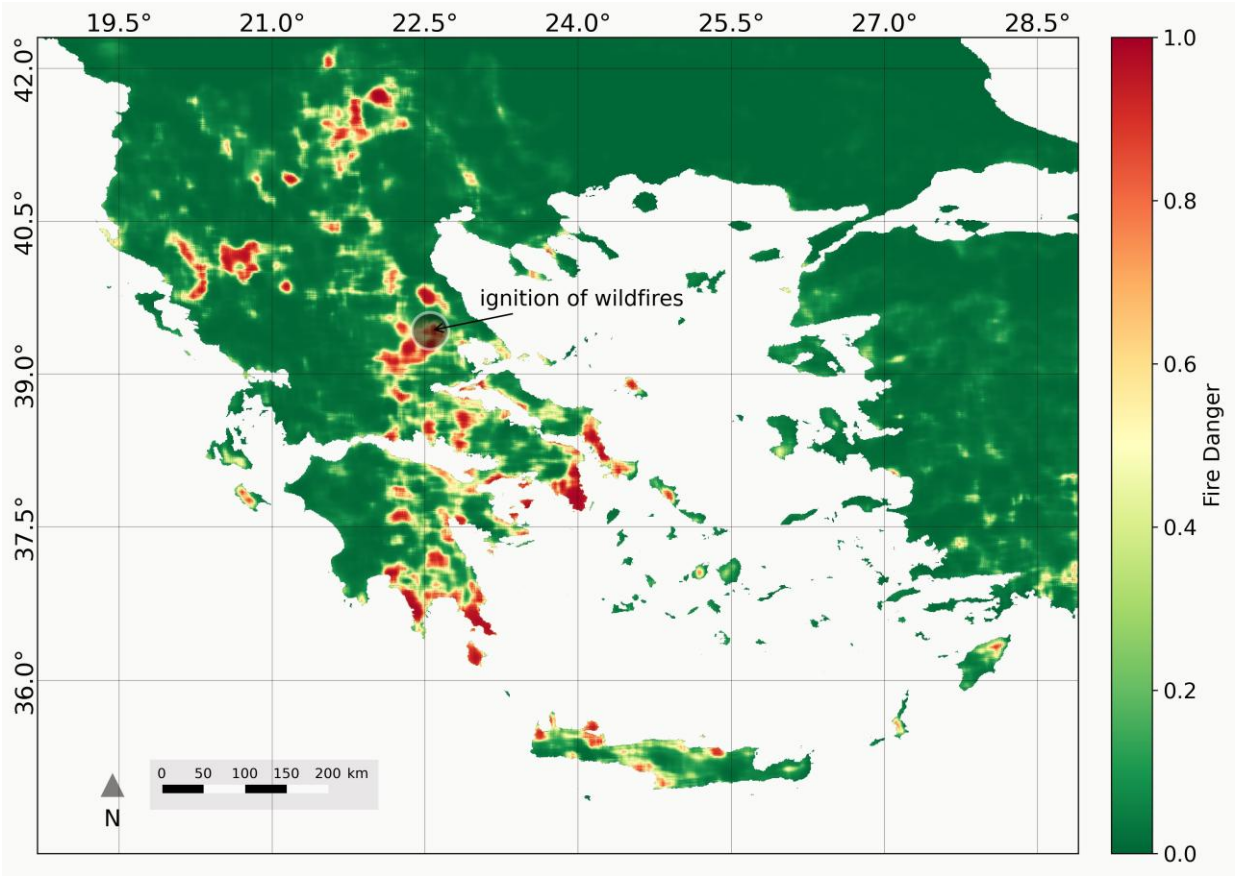


Land cover

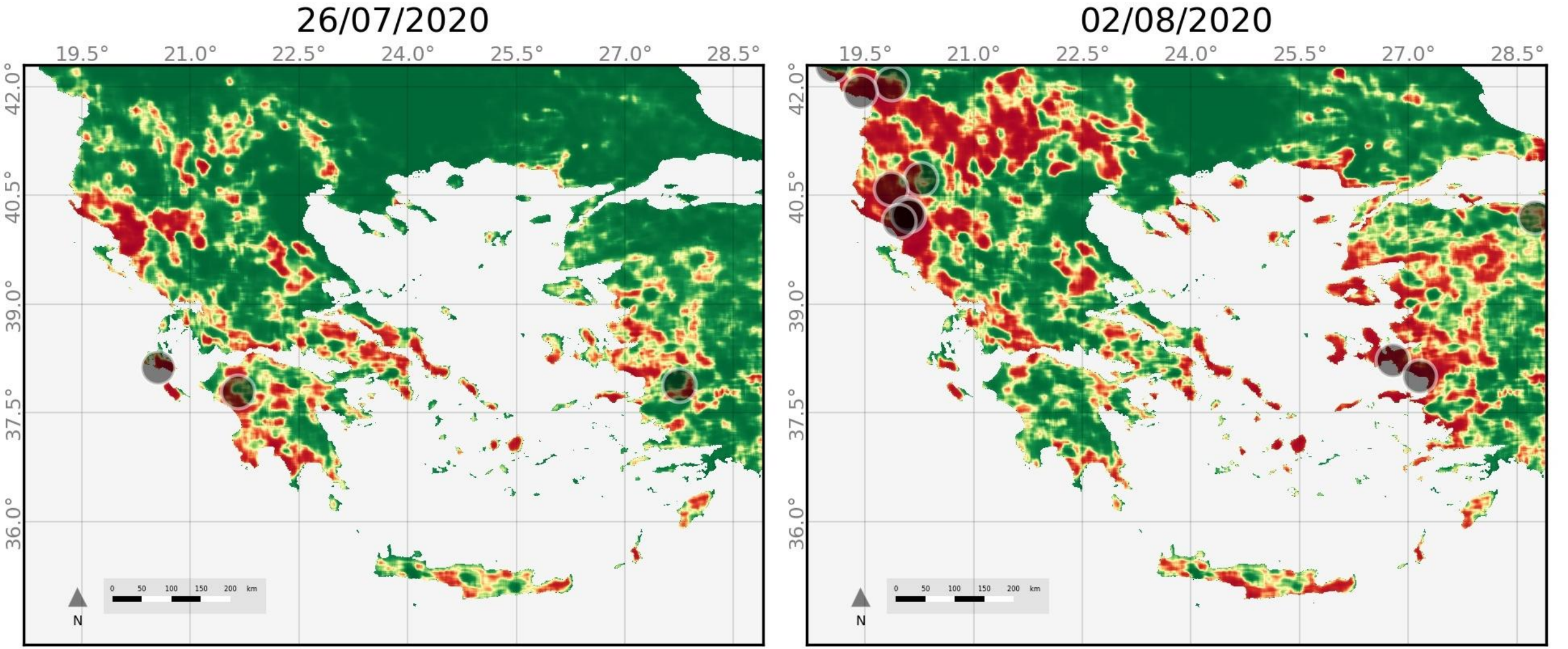
Resolution:
1km × 1km × 1day

years - 2009-2021

Wildfire susceptibility map for 02/07/2021

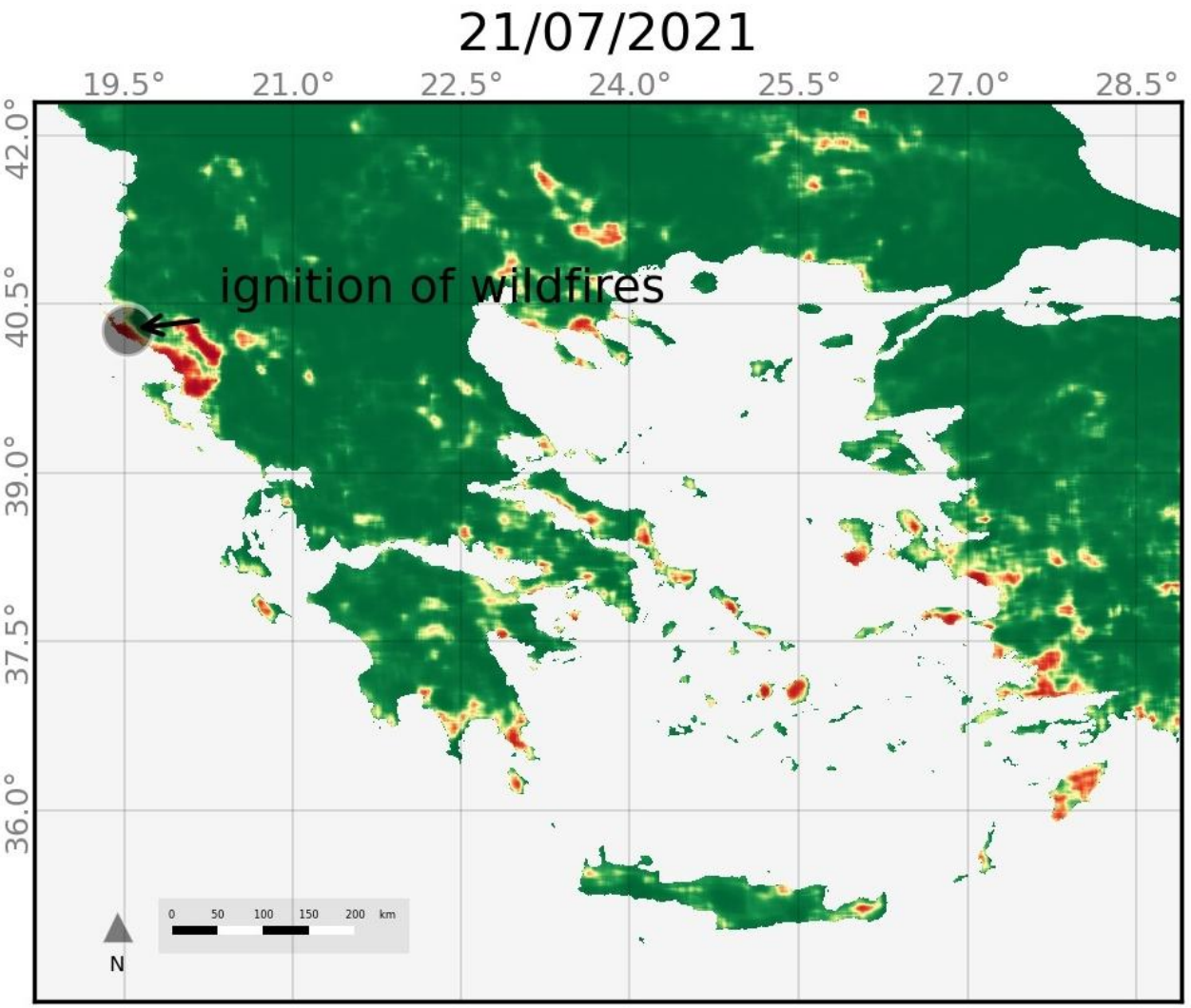


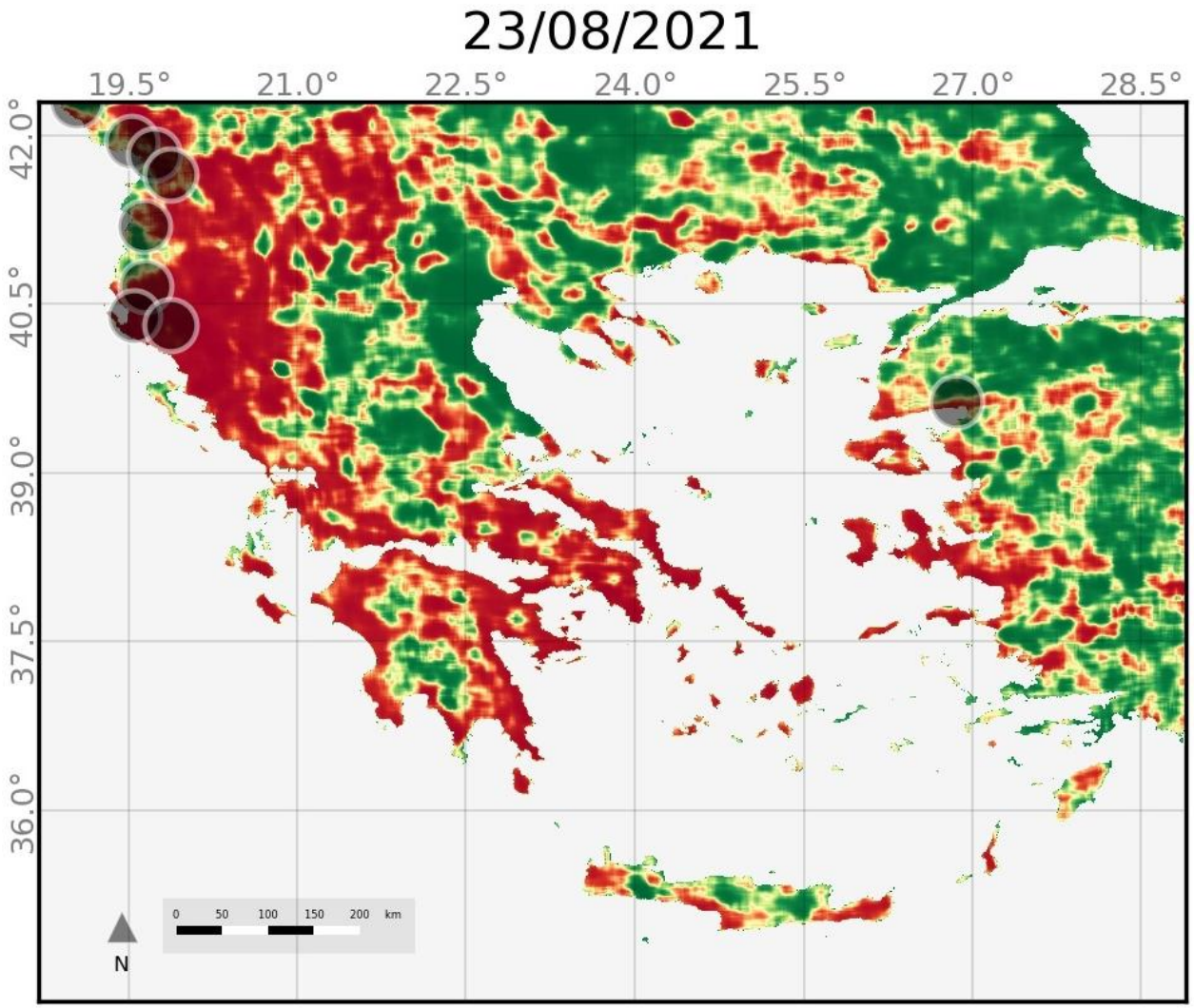
Qualitative Results



[M.H.S. Eddin et al. **Location-Aware Adaptive Normalization: A Deep Learning Approach for Wildfire Danger Forecasting.** IEEE Transactions on Geoscience and Remote Sensing 2023]

Qualitative Results

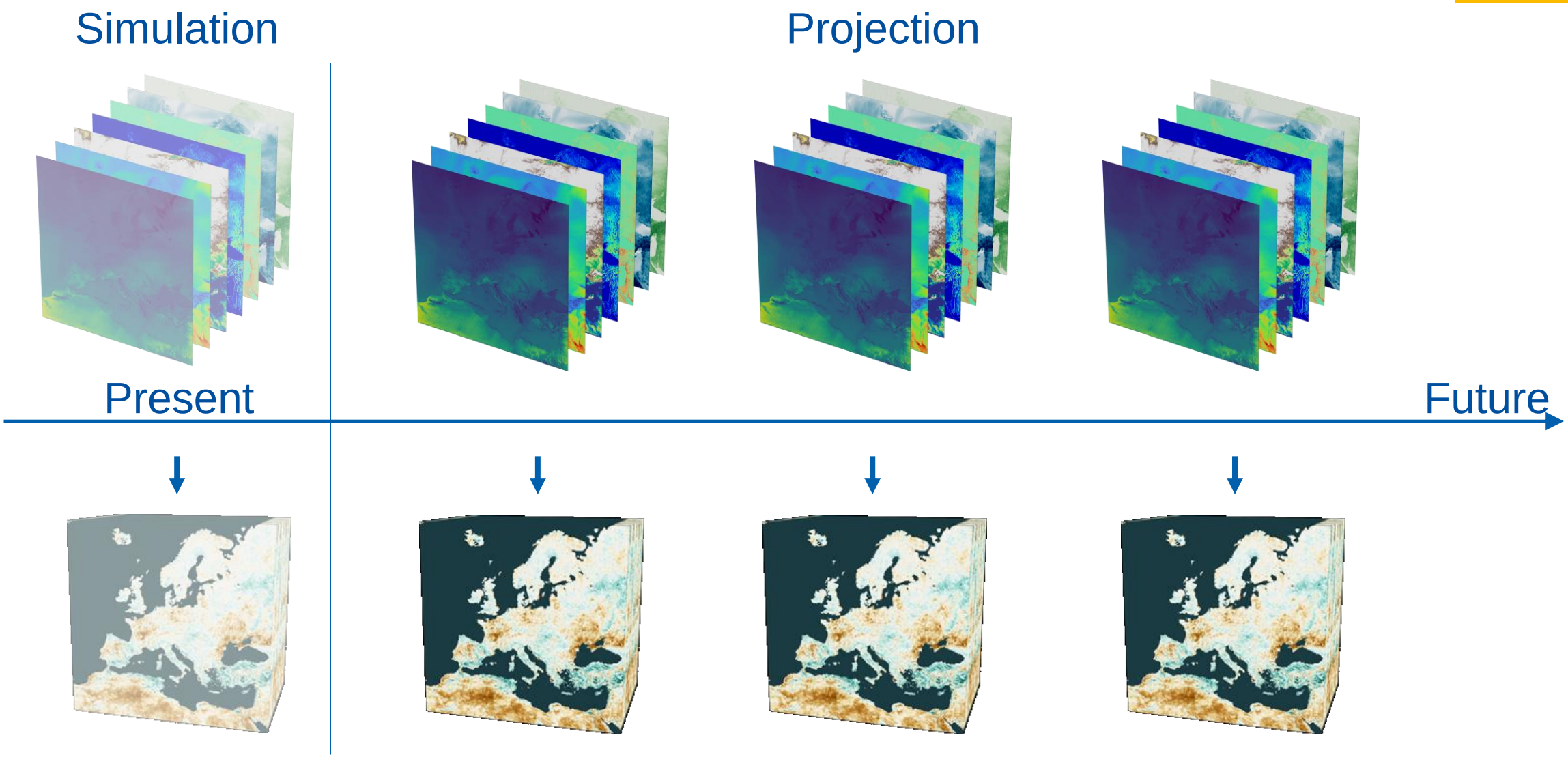




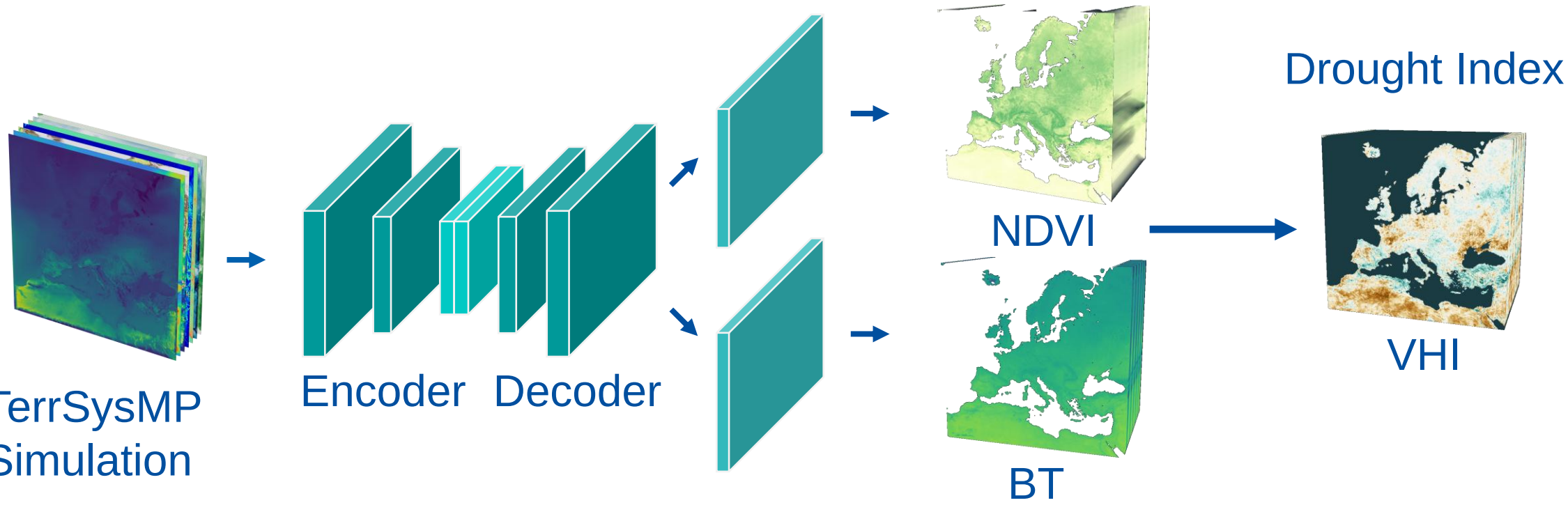
[M.H.S. Eddin et al. **Location-Aware Adaptive Normalization: A Deep Learning Approach for Wildfire Danger Forecasting.** IEEE Transactions on Geoscience and Remote Sensing 2023]

Forecasting Agricultural Drought

Forecasting Agricultural Drought

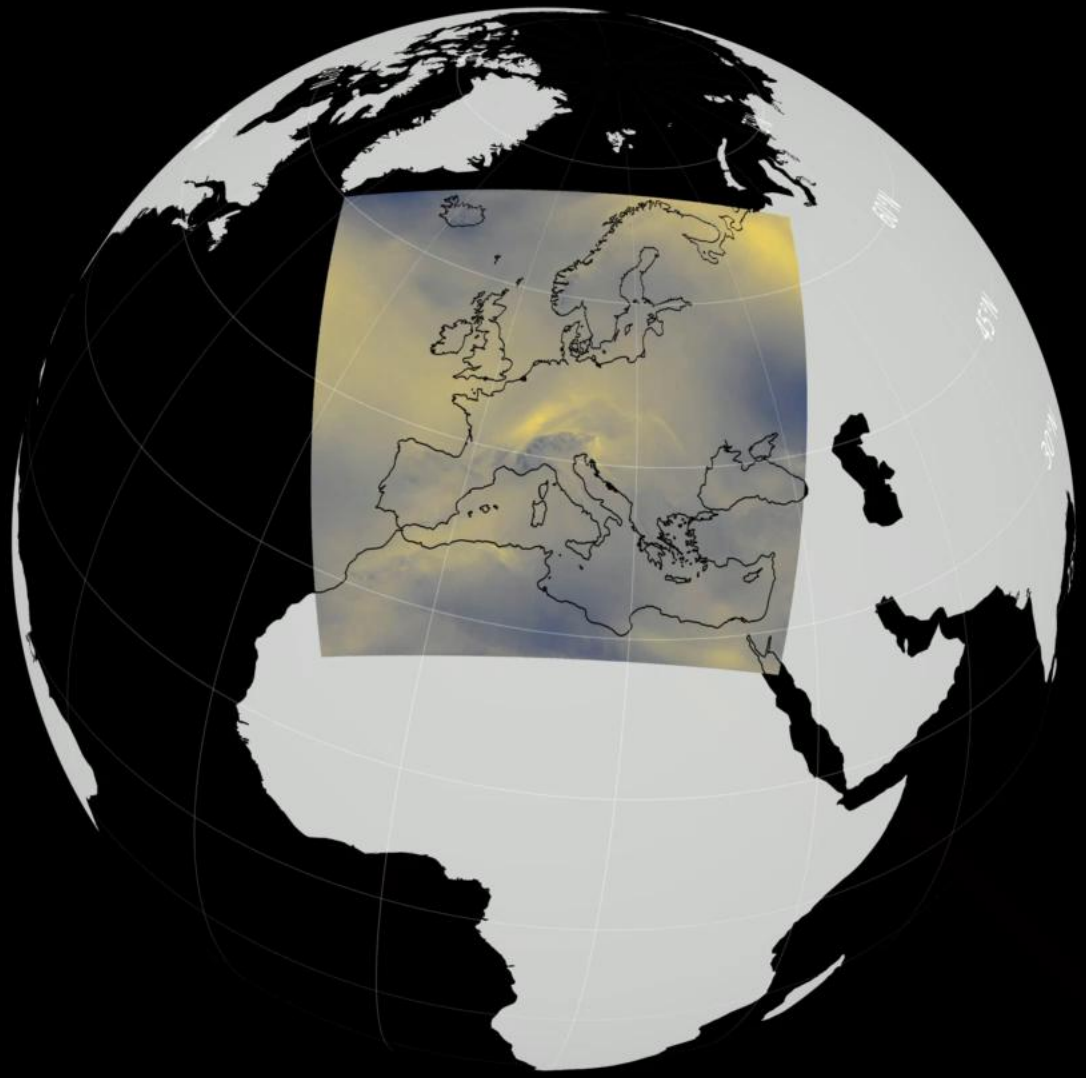


Forecasting Agricultural Drought



[M.H.S. Eddin et al. **Focal-TSMP: Deep learning for vegetation health prediction and agricultural drought assessment from a regional climate simulation.** Geoscientific Model Development 2024]

Forecasting Agricultural Droughts



Summary

- There are many ways to improve the efficiency of transformer architectures
- State space models / Mamba architectures are more efficient, but have some disadvantages compared to transformers
- HPC is crucial for all experiments:
  - Marvin (University of Bonn)
  - JUWELS (WestAI, Jülich Supercomputing Centre)
  - Leonardo (EuroHPC, CINECA, Italy)

Thank you for your attention.

SFB
1502

WEST AI
KI-Servicezentrum

LAMARR
Institute for Machine Learning
and Artificial Intelligence

Ministerium für
Kultur und Wissenschaft
des Landes Nordrhein-Westfalen

UNIVERSITÄT BONN