Is Space-Time Attention All You Need for Video Understanding?

ICML 2021

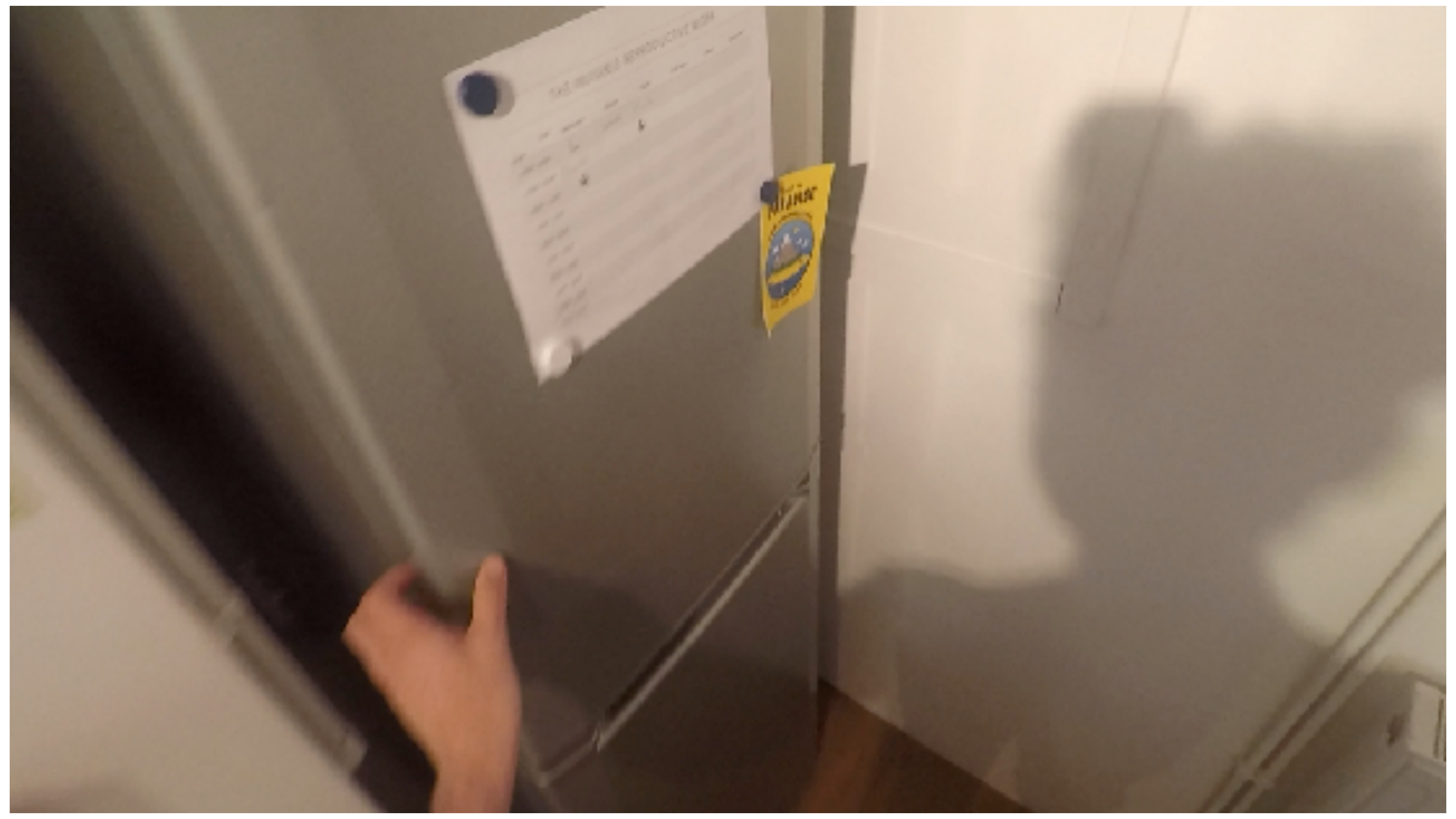
Gedas Bertasius, Heng Wang, Lorenzo Torresani

Video Classification

- Given a video, we want to classify it into one of the action categories.

Cartwheeling

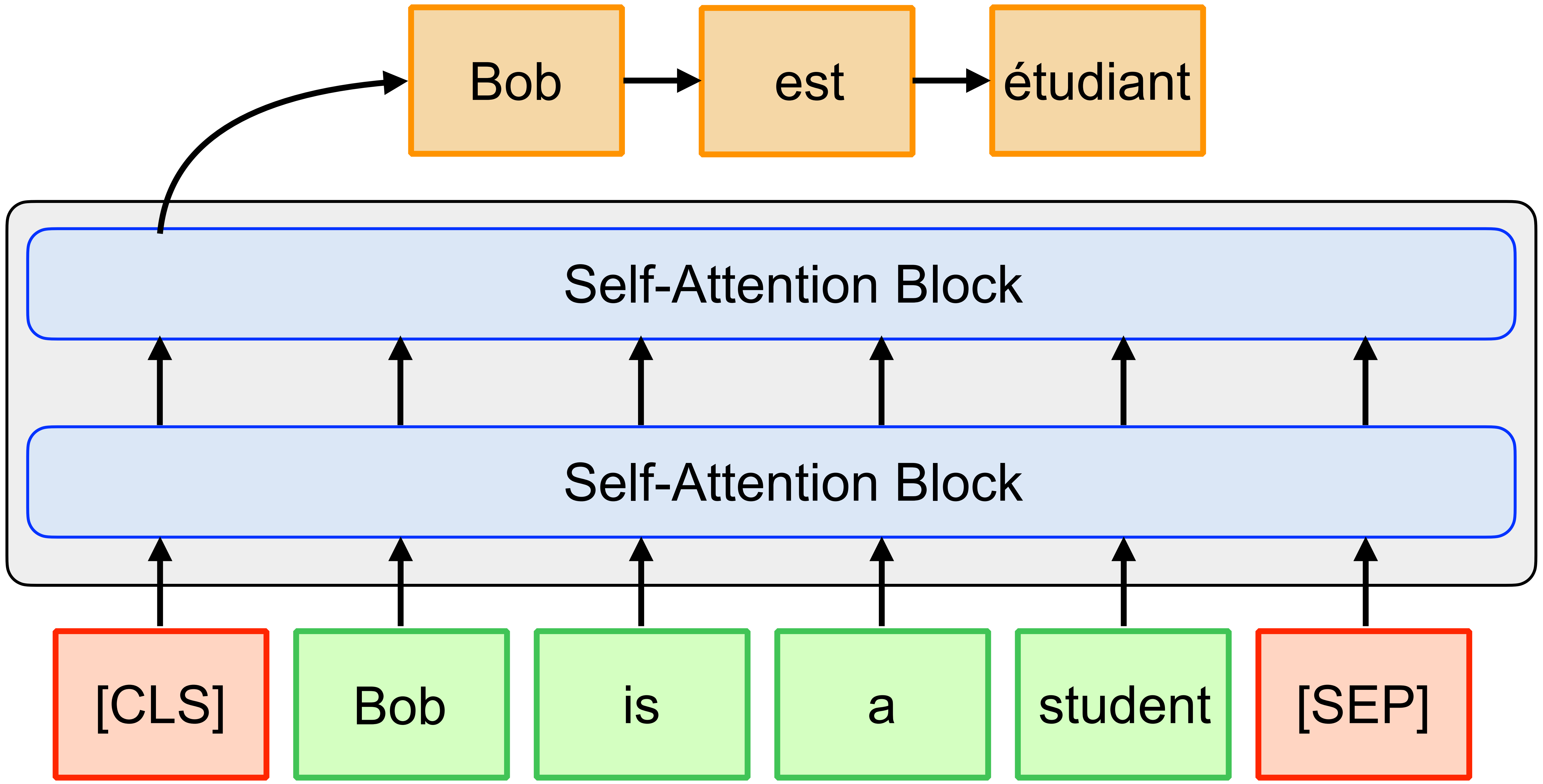
Braiding Hair



Opening a Fridge

Modern Language Models

- Self-attention enables capturing long-range dependencies among words.



"Attention is All You Need", Vaswani et al., NIPS, 2017

Modern Language Models

- Self-attention enables capturing long-range dependencies among words.

Self-Attention

"Attention is All You Need", Vaswani et al., NIPS, 2017

Modern Language Models

- Self-attention enables capturing long-range dependencies among words.

Self-Attention

"Attention is All You Need", Vaswani et al., NIPS, 2017

Modern Language Models

- Self-attention enables capturing long-range dependencies among words.

"Attention is All You Need", Vaswani et al., NIPS, 2017

Modern Language Models

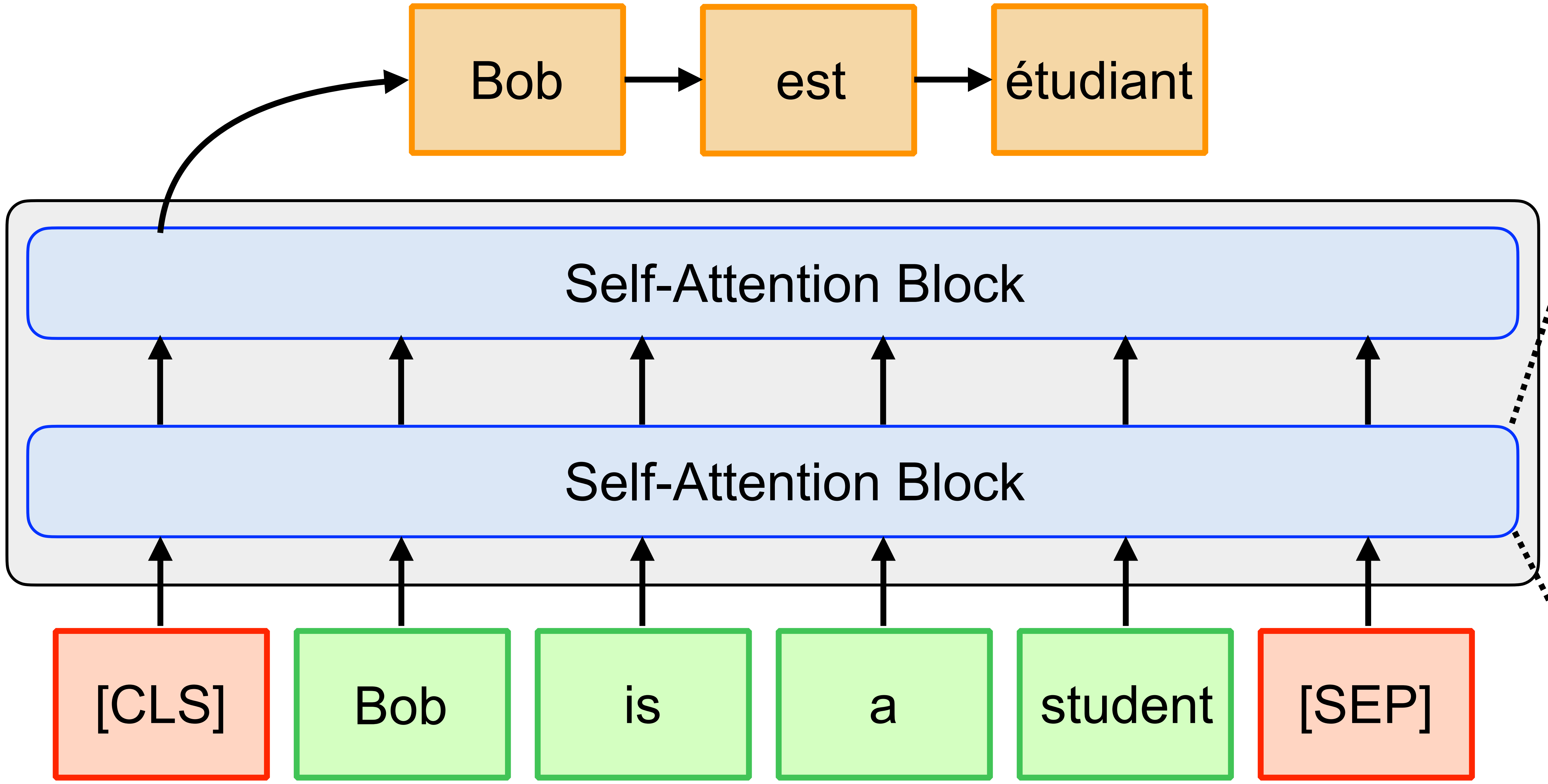
- Self-attention enables capturing long-range dependencies among words.

Self-Attention

"Attention is All You Need", Vaswani et al., NIPS, 2017

Modern Language Models

- Self-attention enables capturing long-range dependencies among words.



$$X^{(\ell)} \in \mathbb{R}^{n \times d}$$

$$W_Q \in \mathbb{R}^{d \times d_q} \qquad Q = X^{(\ell)} W_Q$$

$$W_K \in \mathbb{R}^{d \times d_k} \qquad K = X^{(\ell)} W_K$$

$$W_V \in \mathbb{R}^{d \times d_v} \qquad V = X^{(\ell)} W_V$$

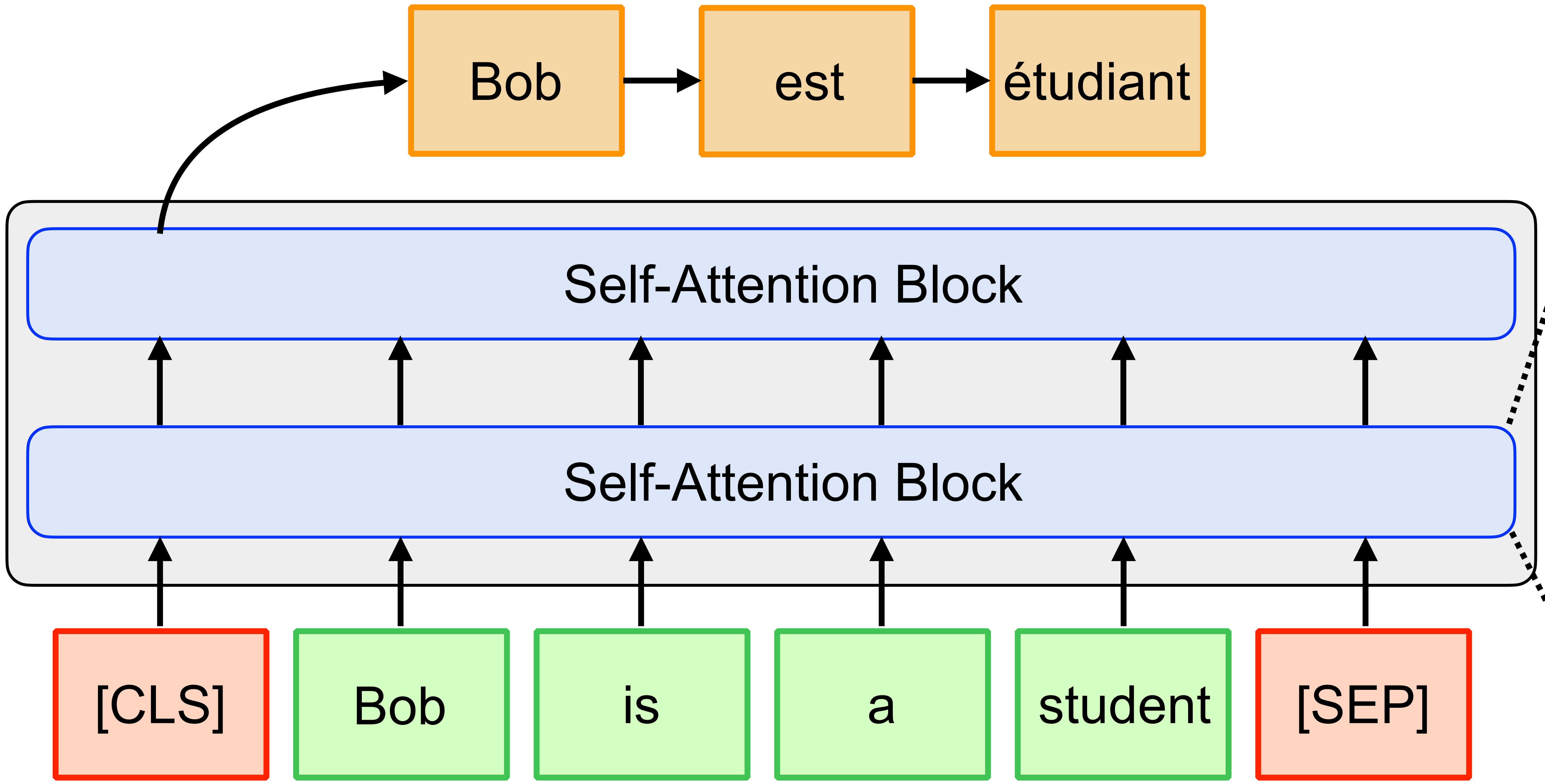
$$X^{(\ell+1)} = \mathrm{softmax}\left(\frac{QK^T}{\sqrt{d_q}}\right) V$$

Self-Attention Block

"Attention is All You Need", Vaswani et al., NIPS, 2017

Modern Language Models

- Self-attention enables capturing long-range dependencies among words.



$$\boxed{X^{(\ell)} \in \mathbb{R}^{n \times d}}$$

$$W_Q \in \mathbb{R}^{d \times d_q} \qquad Q = X^{(\ell)} W_Q$$

$$W_K \in \mathbb{R}^{d \times d_k} \qquad K = X^{(\ell)} W_K$$

$$W_V \in \mathbb{R}^{d \times d_v} \qquad V = X^{(\ell)} W_V$$

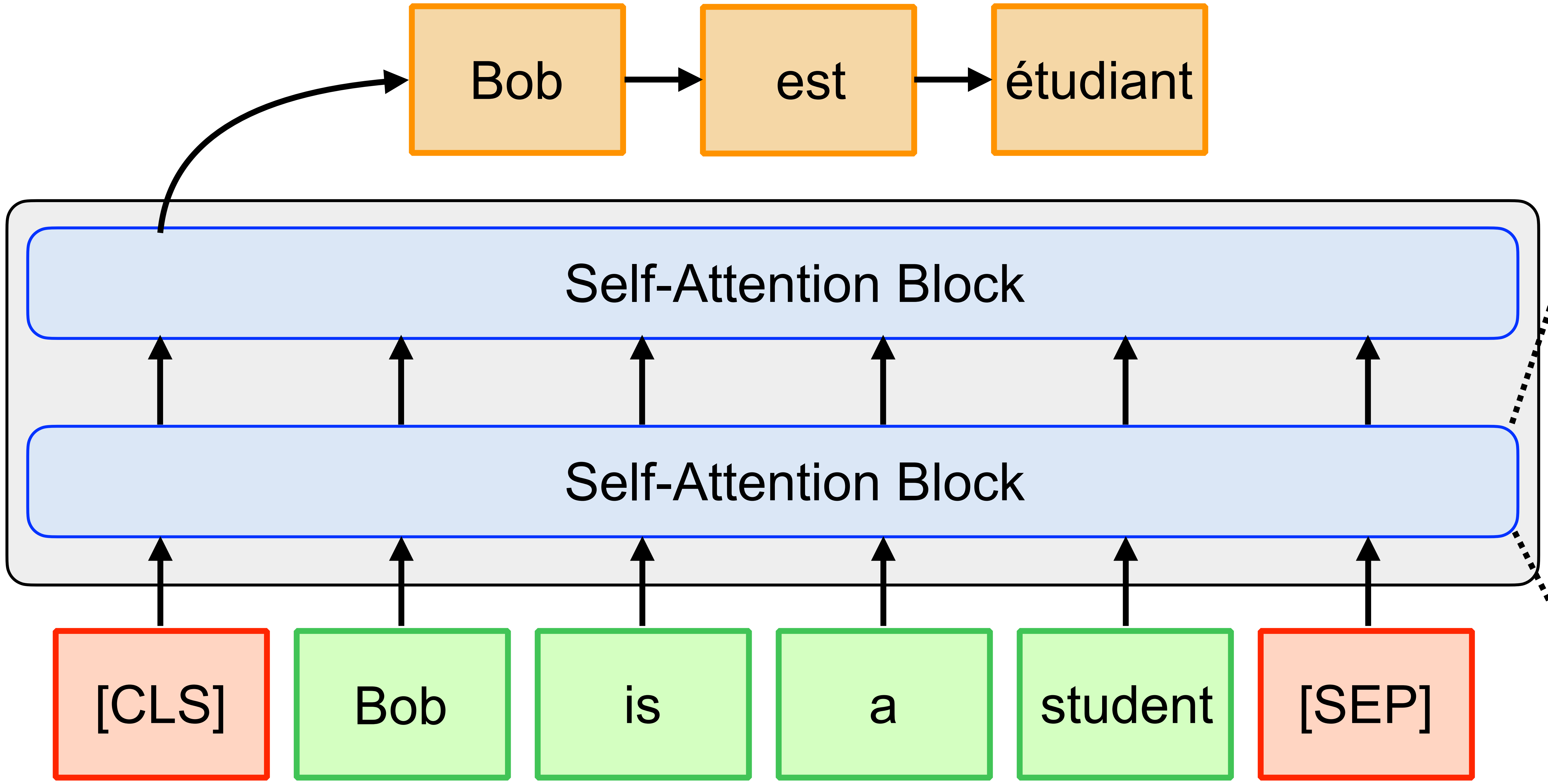
$$X^{(\ell+1)} = \mathrm{softmax}\left(\frac{QK^T}{\sqrt{d_q}}\right) V$$

Self-Attention Block

"Attention is All You Need", Vaswani et al., NIPS, 2017

Modern Language Models

- Self-attention enables capturing long-range dependencies among words.



$$X^{(\ell)} \in \mathbb{R}^{n \times d}$$

$$\boxed{\begin{aligned} W_Q &\in \mathbb{R}^{d \times d_q} \\ W_K &\in \mathbb{R}^{d \times d_k} \\ W_V &\in \mathbb{R}^{d \times d_v} \end{aligned}} \qquad \begin{aligned} Q &= X^{(\ell)} W_Q \\ K &= X^{(\ell)} W_K \\ V &= X^{(\ell)} W_V \end{aligned}$$

$$X^{(\ell+1)} = \mathrm{softmax}\left(\frac{QK^T}{\sqrt{d_q}}\right) V$$

Self-Attention Block

"Attention is All You Need", Vaswani et al., NIPS, 2017

Modern Language Models

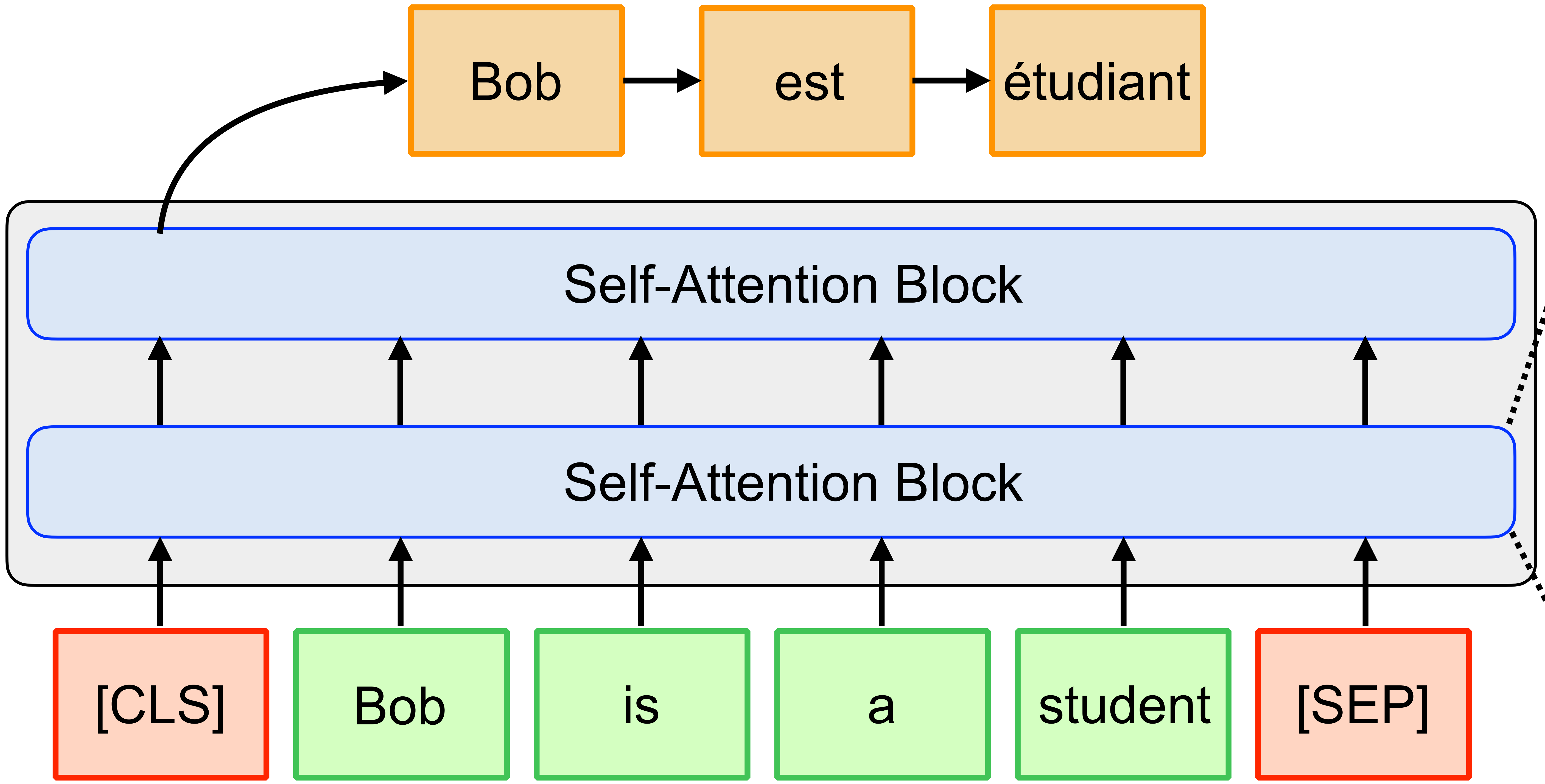
- Self-attention enables capturing long-range dependencies among words.

Self-Attention Block

"Attention is All You Need", Vaswani et al., NIPS, 2017

Modern Language Models

- Self-attention enables capturing long-range dependencies among words.



$$X^{(\ell)} \in \mathbb{R}^{n \times d}$$

$$W_Q \in \mathbb{R}^{d \times d_q} \qquad Q = X^{(\ell)} W_Q$$

$$W_K \in \mathbb{R}^{d \times d_k} \qquad K = X^{(\ell)} W_K$$

$$W_V \in \mathbb{R}^{d \times d_v} \qquad V = X^{(\ell)} W_V$$

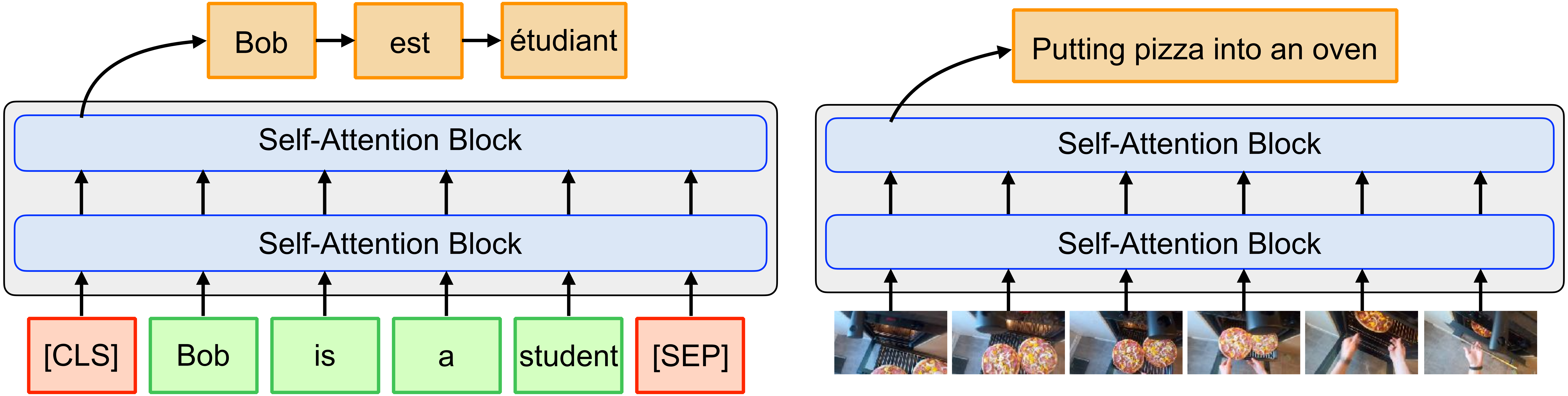
$$\boxed{X^{(\ell+1)} = \mathrm{softmax}\left(\frac{QK^T}{\sqrt{d_q}}\right) V}$$

Self-Attention Block

"Attention is All You Need", Vaswani et al., NIPS, 2017

Modern Language Models

- Self-attention enables capturing long-range dependencies among words.

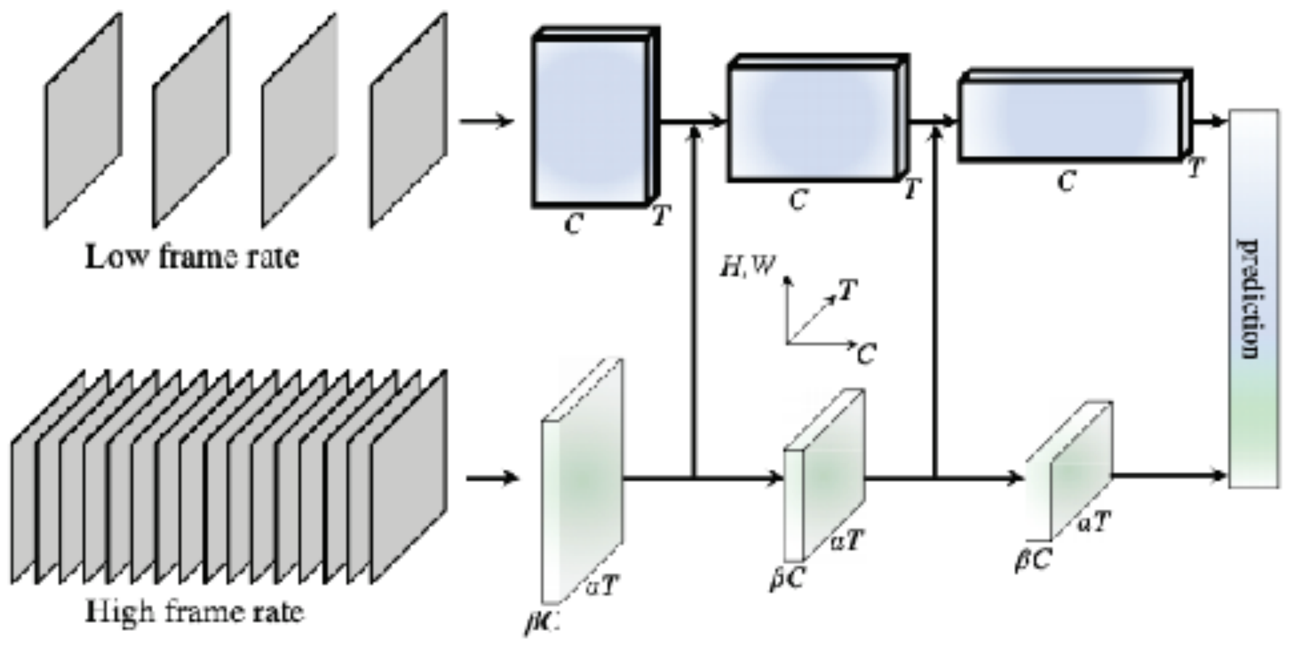


a) Language Model

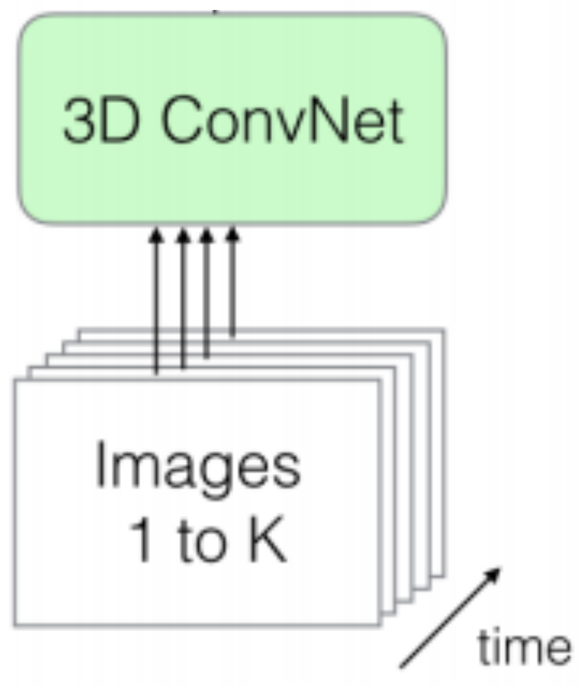
b) Video Model

"Attention is All You Need", Vaswani et al., NIPS 2017

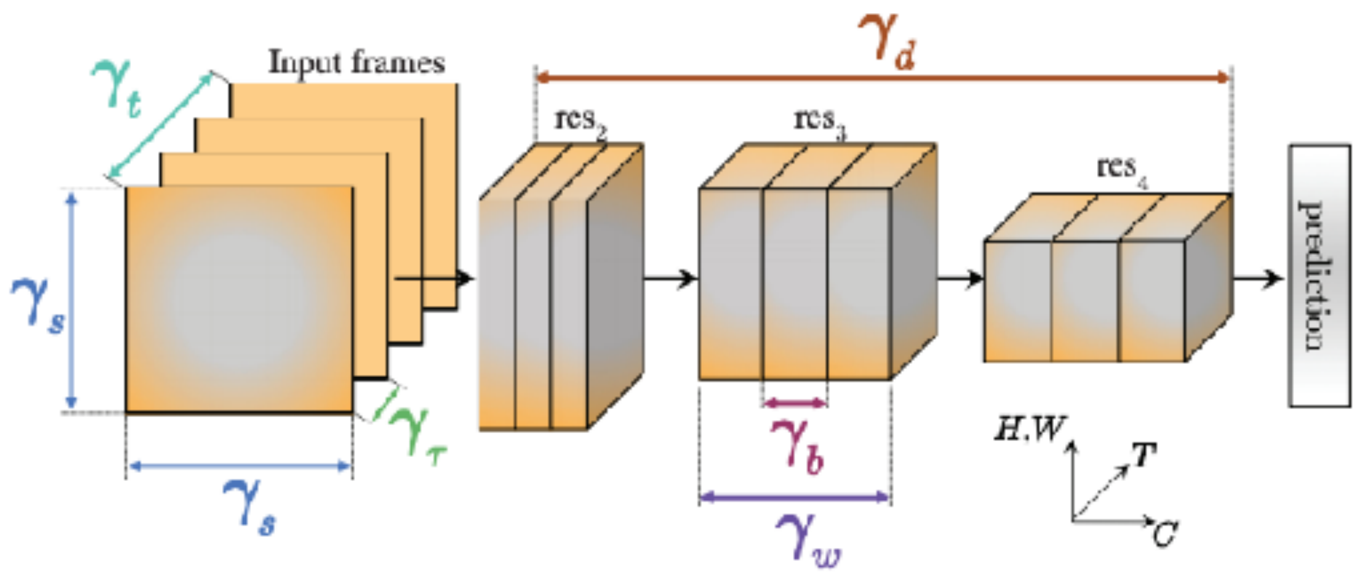
State-of-the-Art in Video Classification



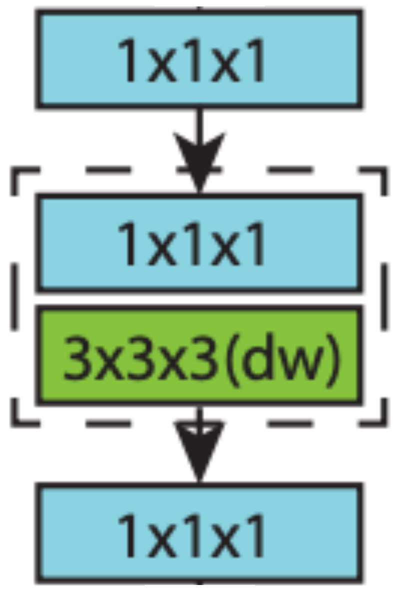
SlowFast Networks
[Feichtenhofer et al. 2019]



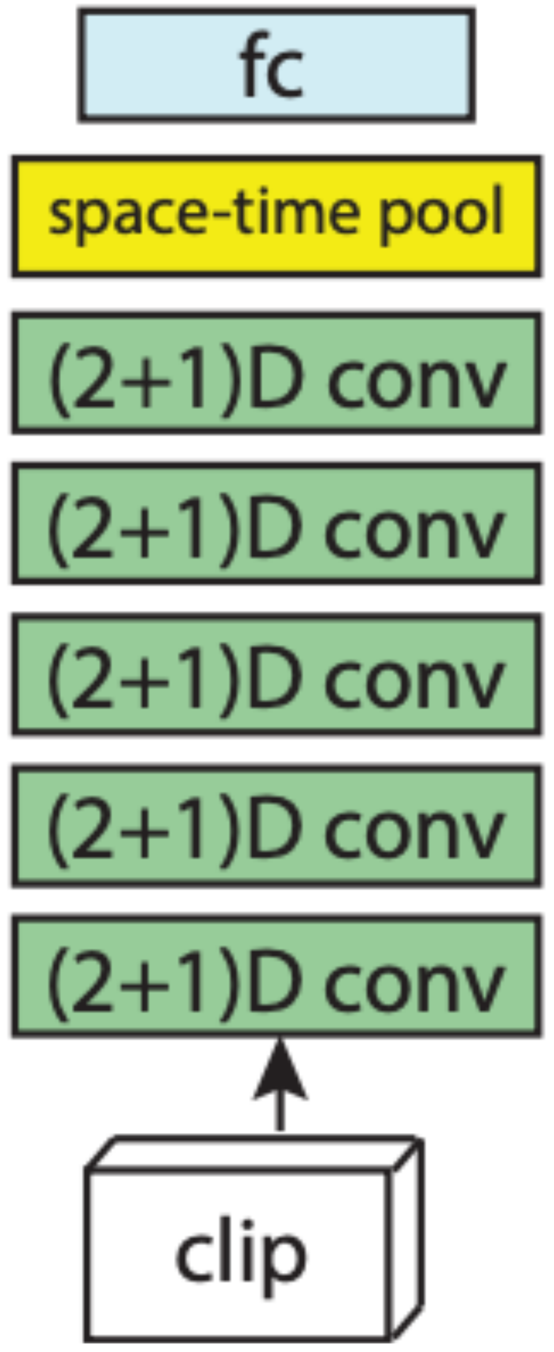
Inflated 3D Networks
[Carreira et al. 2018]



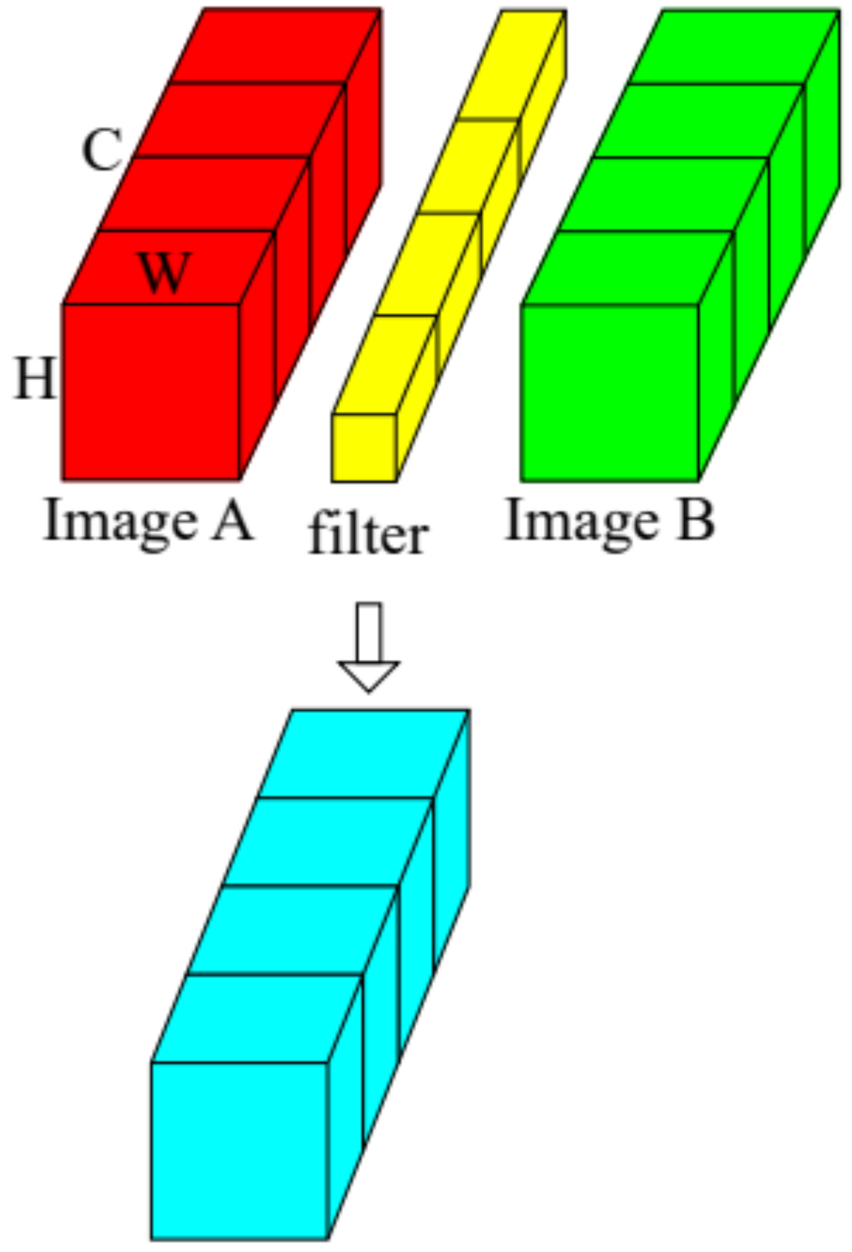
Expanded 3D Networks
[Feichtenhofer 2020]



Channel Separated Networks
[Tran et al. 2019]



R(2+1)D Networks
[Tran et al. 2018]



Correlation Networks
[Wang et al. 2020]

3D Convolutions vs Self-Attention

3D Convolutions:

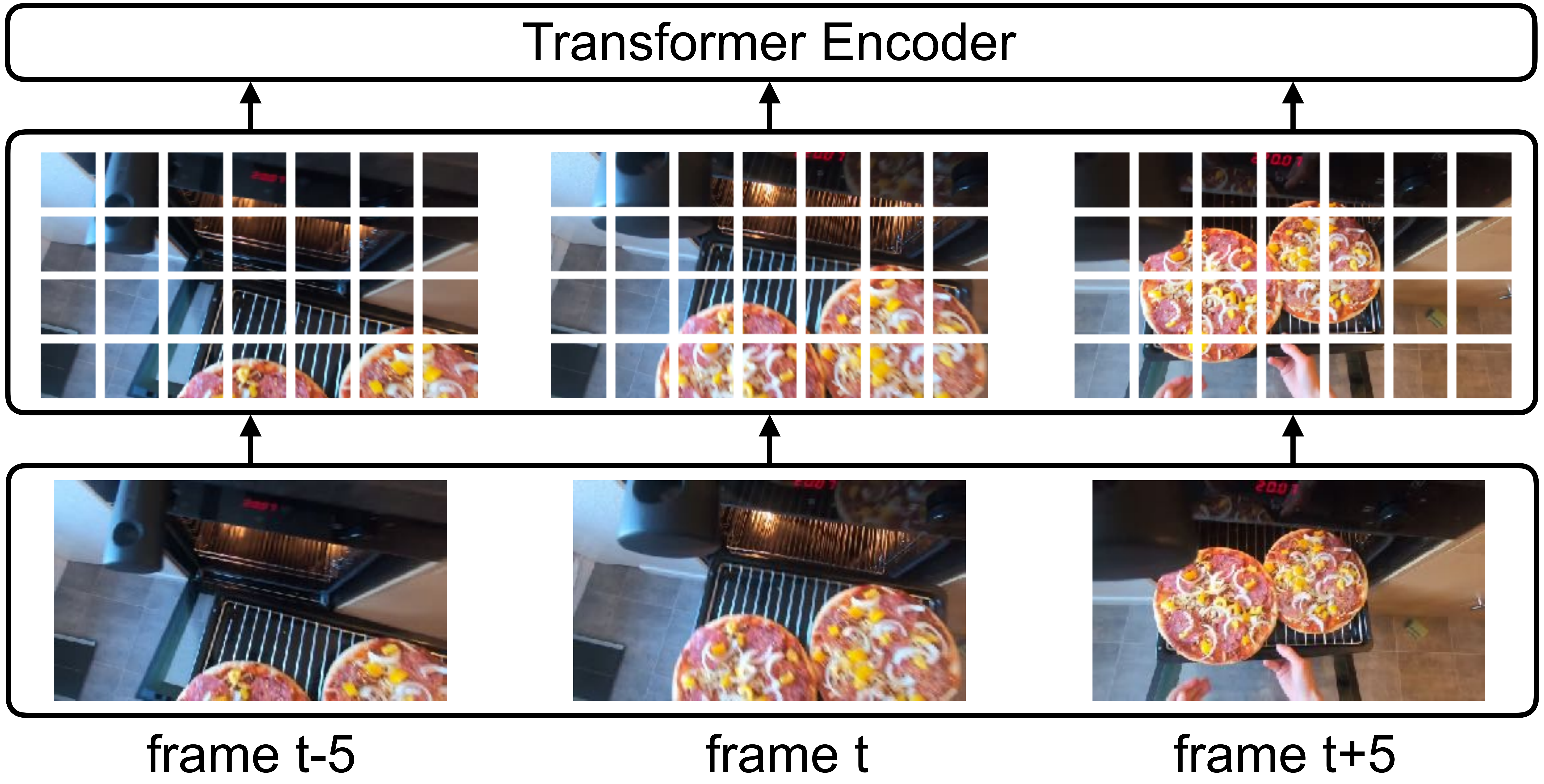
- Strong inductive bias.
- Captures short-range patterns.
- Difficult to scale.

Self-Attention:

- Fewer inductive biases.
- Can capture both short-range and long-range dependencies.
- Easier to scale model capacity.

Video Decomposition

- We decompose the video into a sequence of frame-level patches.

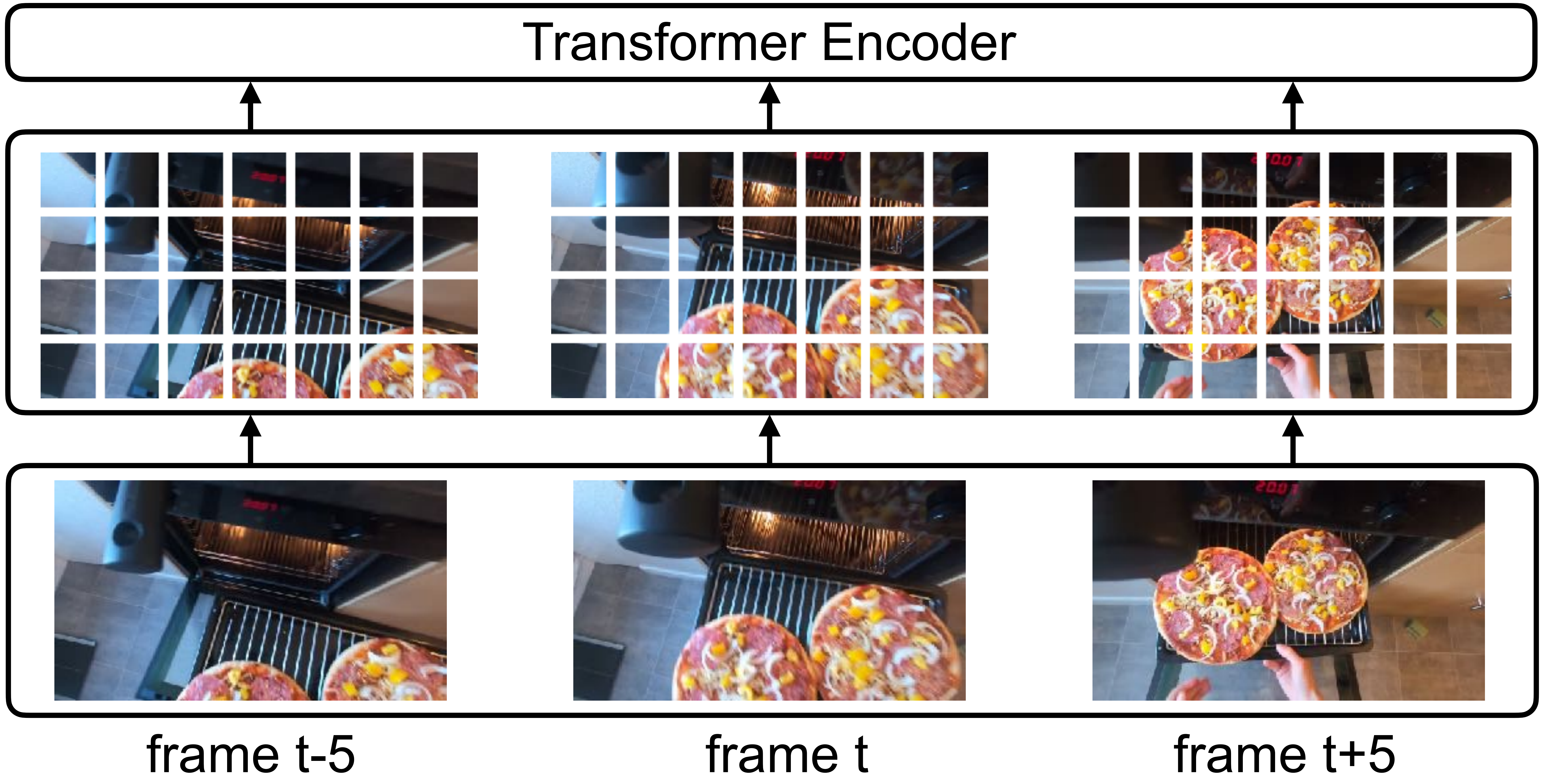


"An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale", Dosovitskiy et al., ICLR 2021

Video Decomposition

- We decompose the video into a sequence of frame-level patches.

Computing similarity for all pairs of patches is costly.

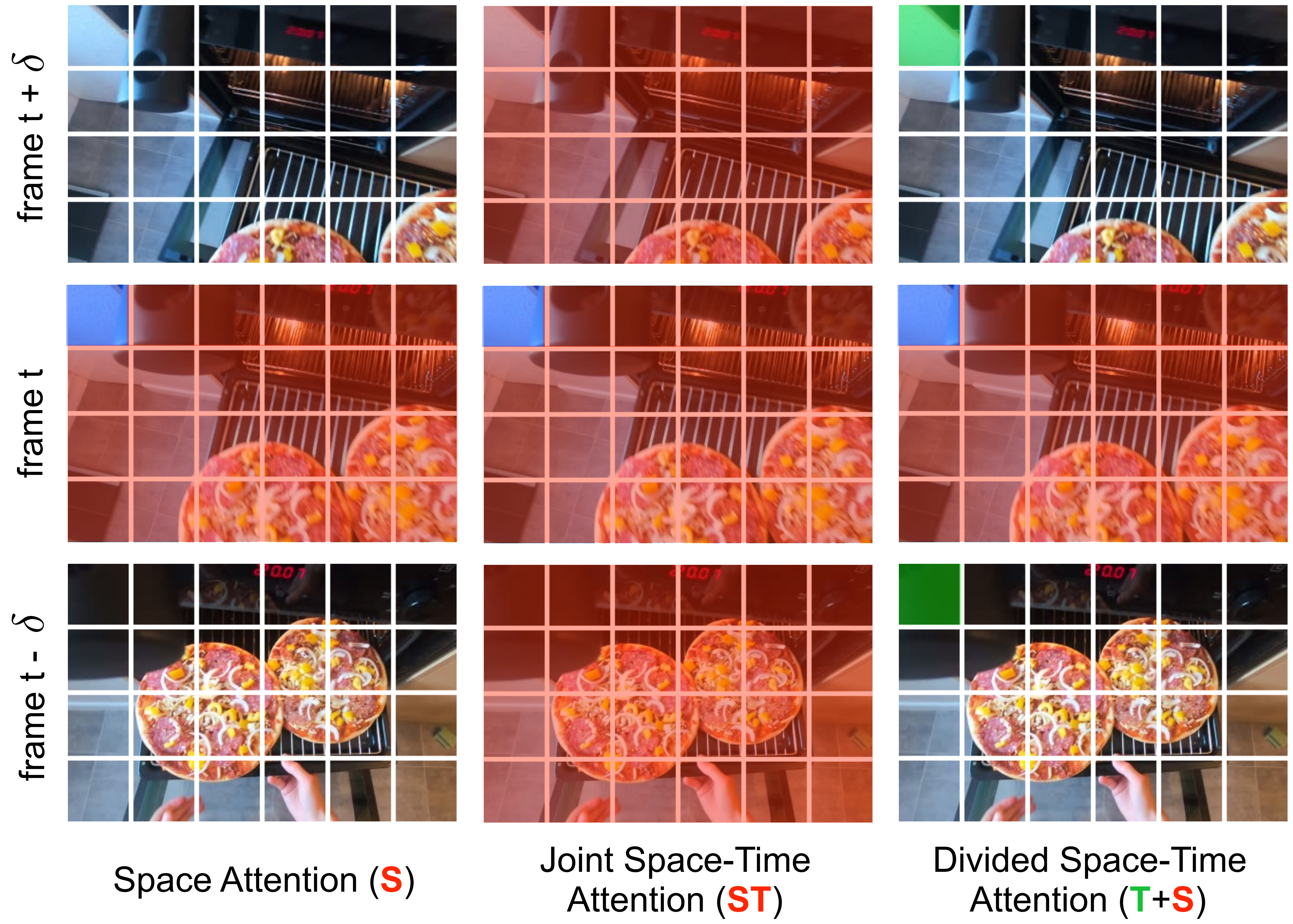


"An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale", Dosovitskiy et al., ICLR 2021

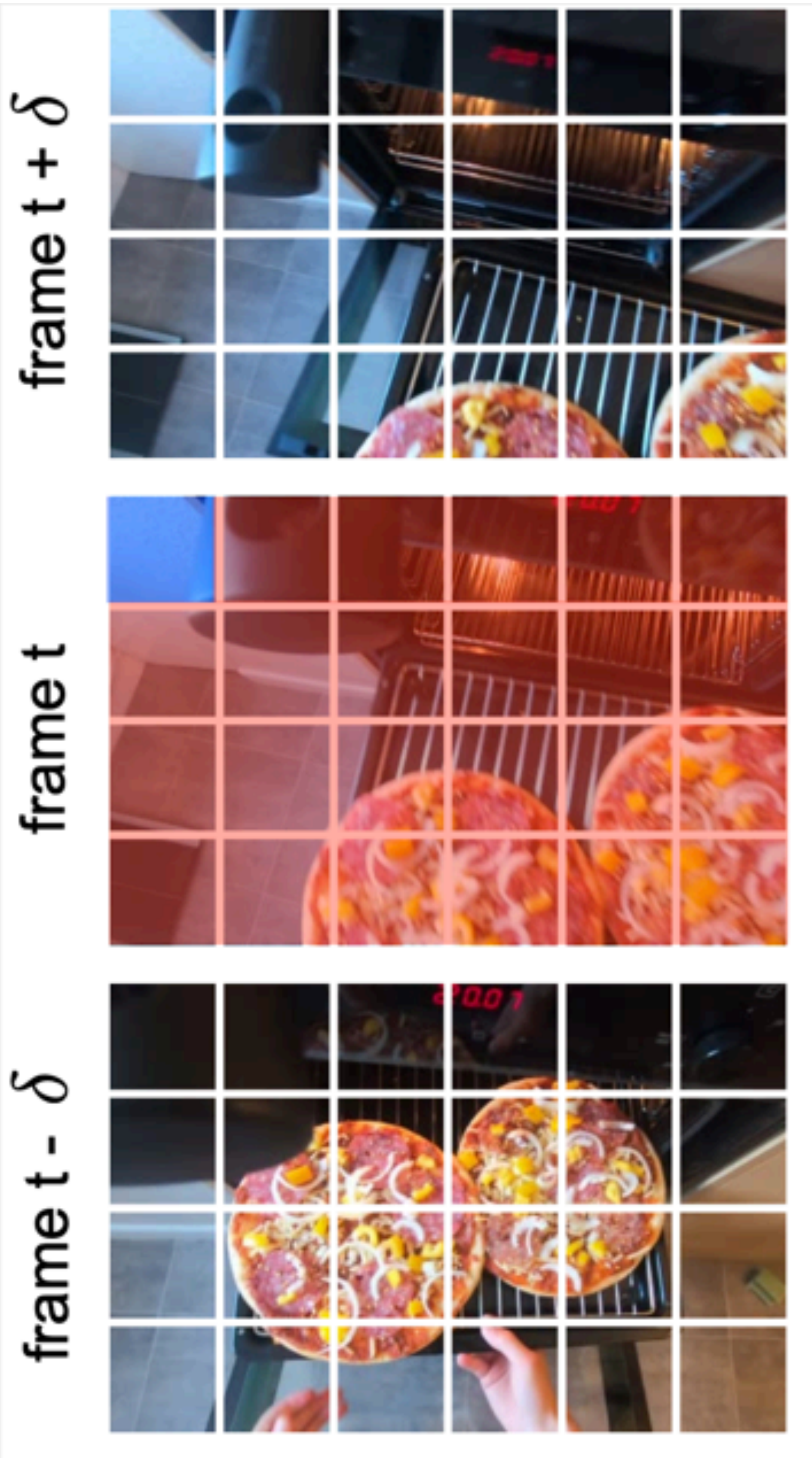
1. What is the right space-time self-attention pattern?

Space-Time Self-Attention

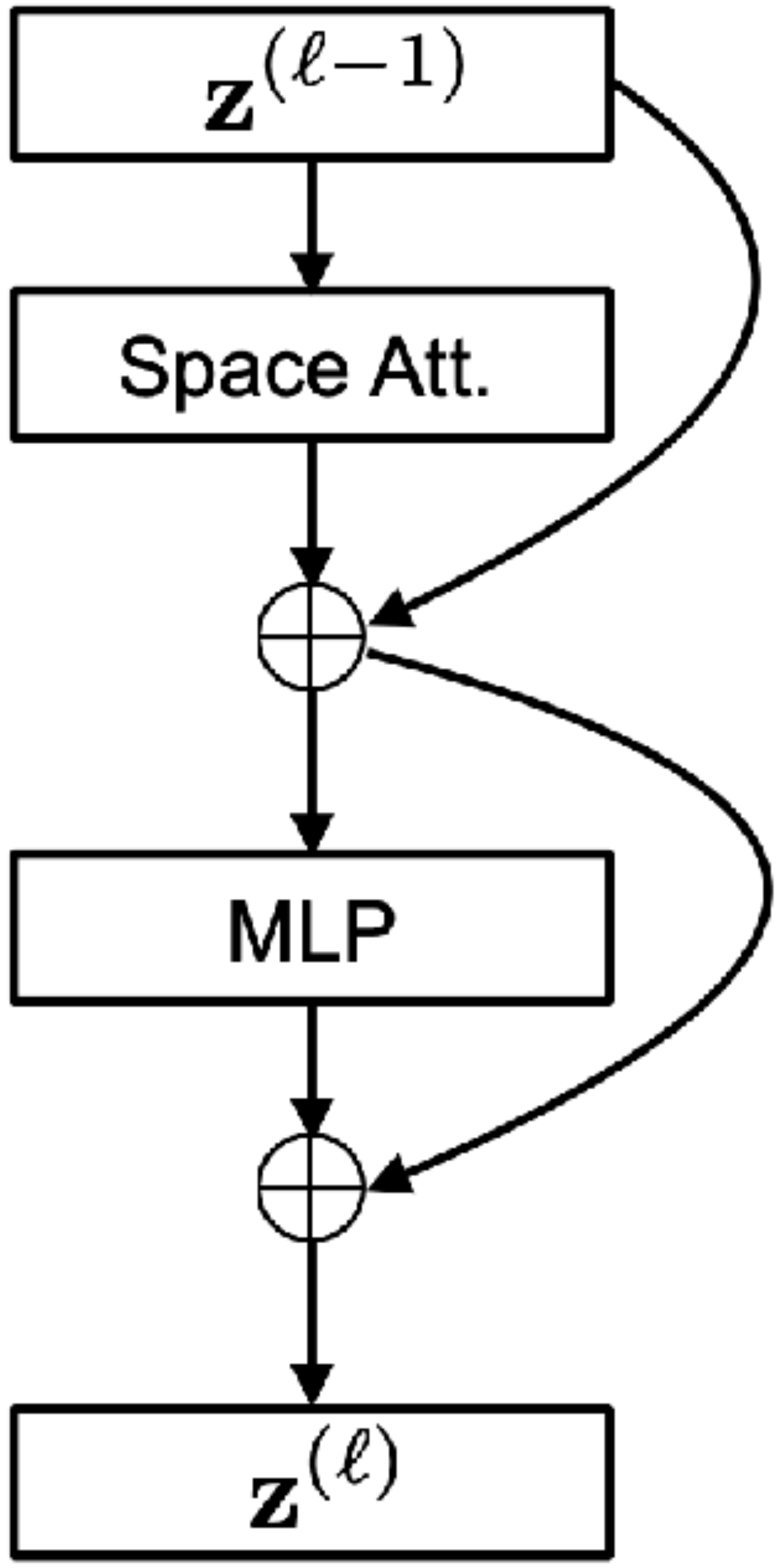
- We investigate several space-time self-attention schemes.



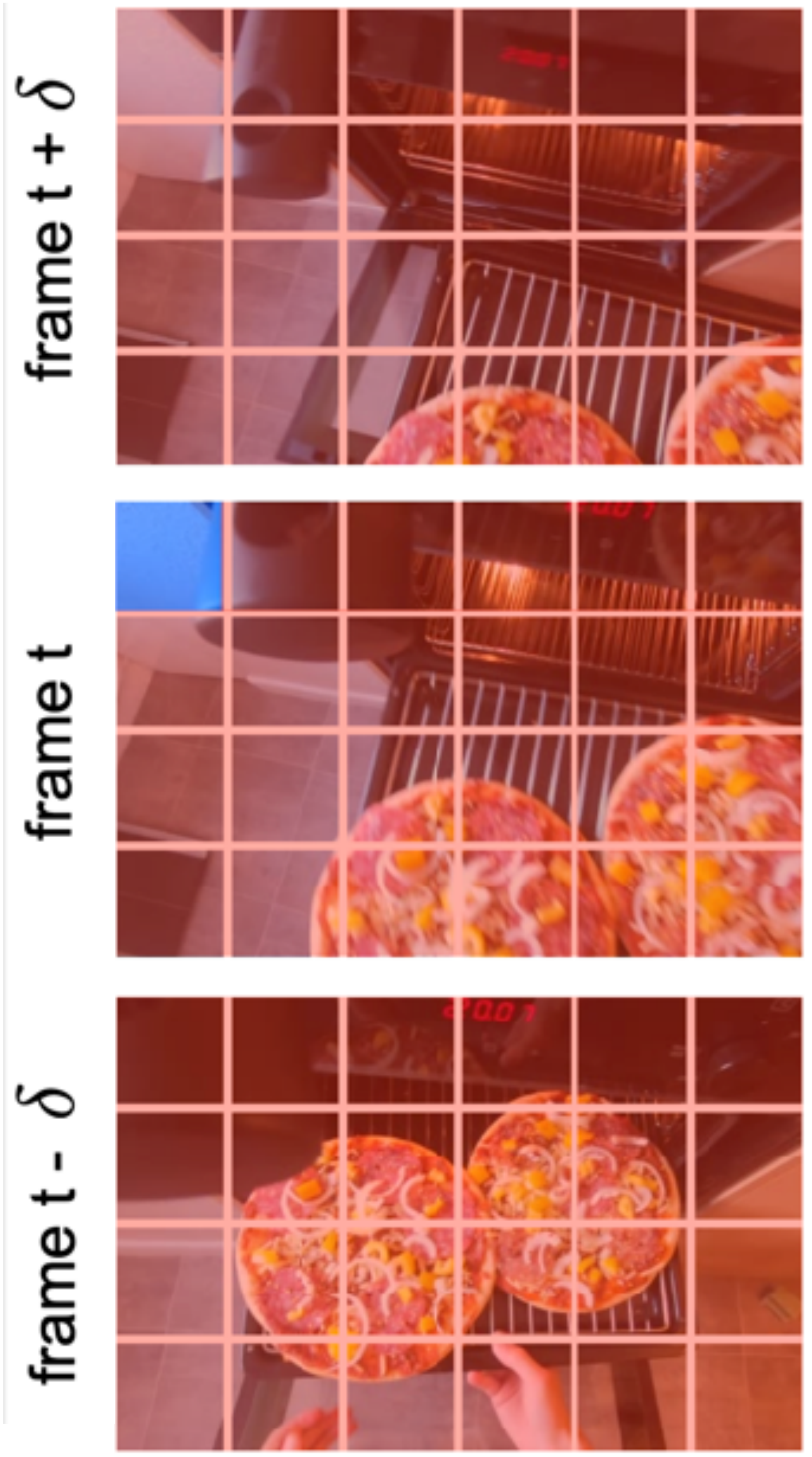
Spatial Self-Attention



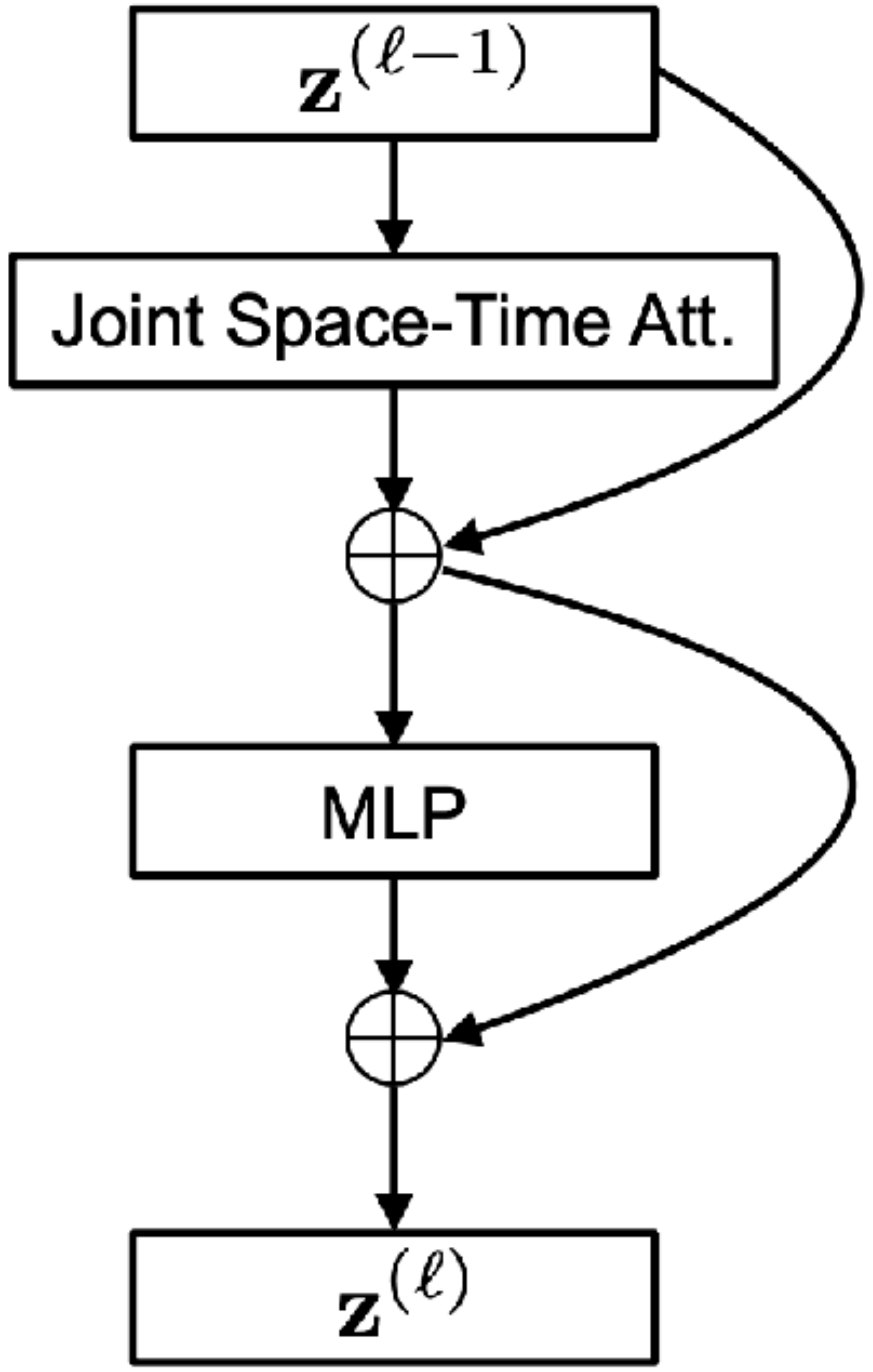
Space Attention (S)



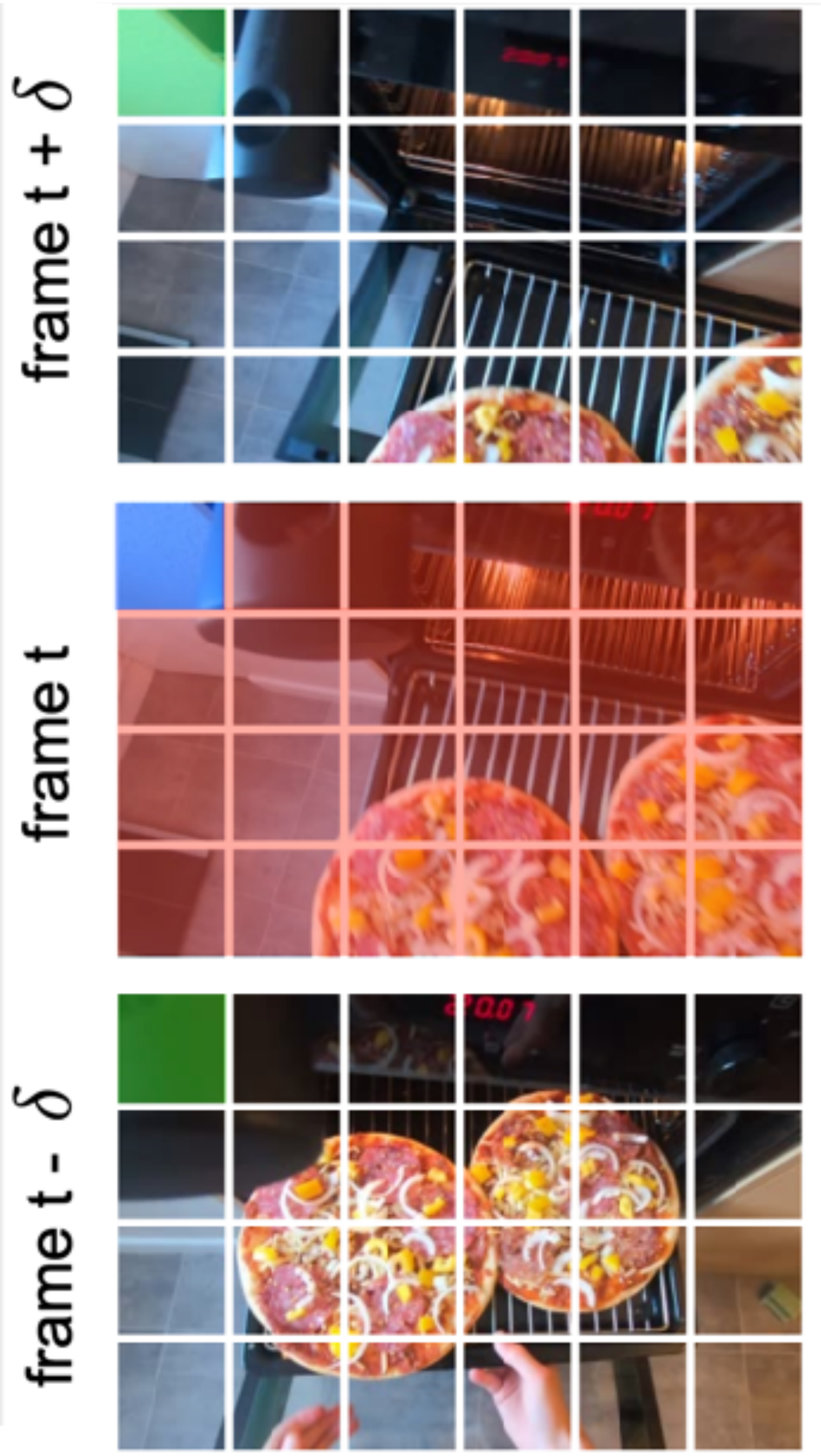
Joint Space-Time Self-Attention



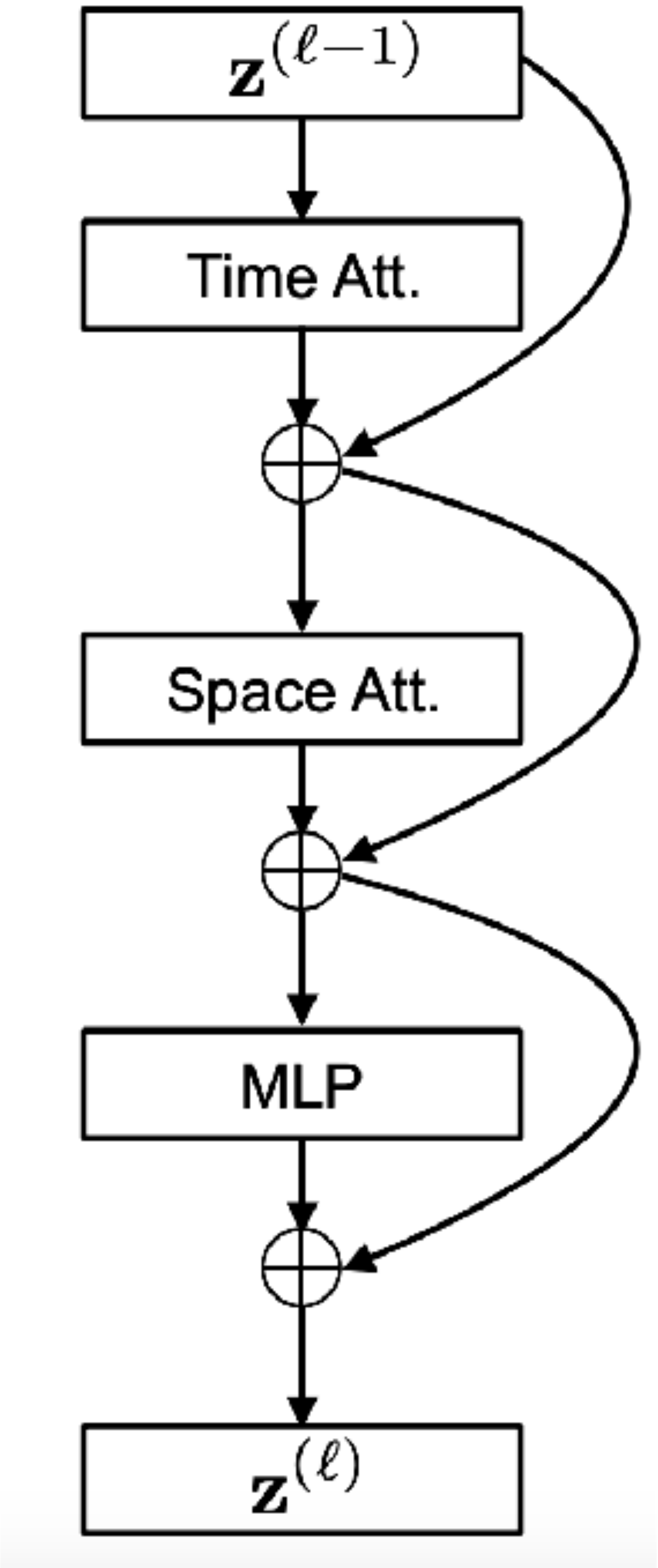
Joint Space-Time
Attention (**ST**)



Divided Space-Time Self-Attention



Divided Space-Time
Attention (T+S)



Analysis of Self-Attention Schemes

- Each space-time self-attention scheme is evaluated on Kinetics-400, and Something-Something-V2 datasets.

Attention	Pretraining	Params	K400	SSv2
Space	ImageNet-21K	85.9M	76.9	36.6
Joint Space-Time	ImageNet-21K	85.9M	77.4	58.5
Divided Space-Time	ImageNet-21K	121.4M	**78.0**	**59.5**

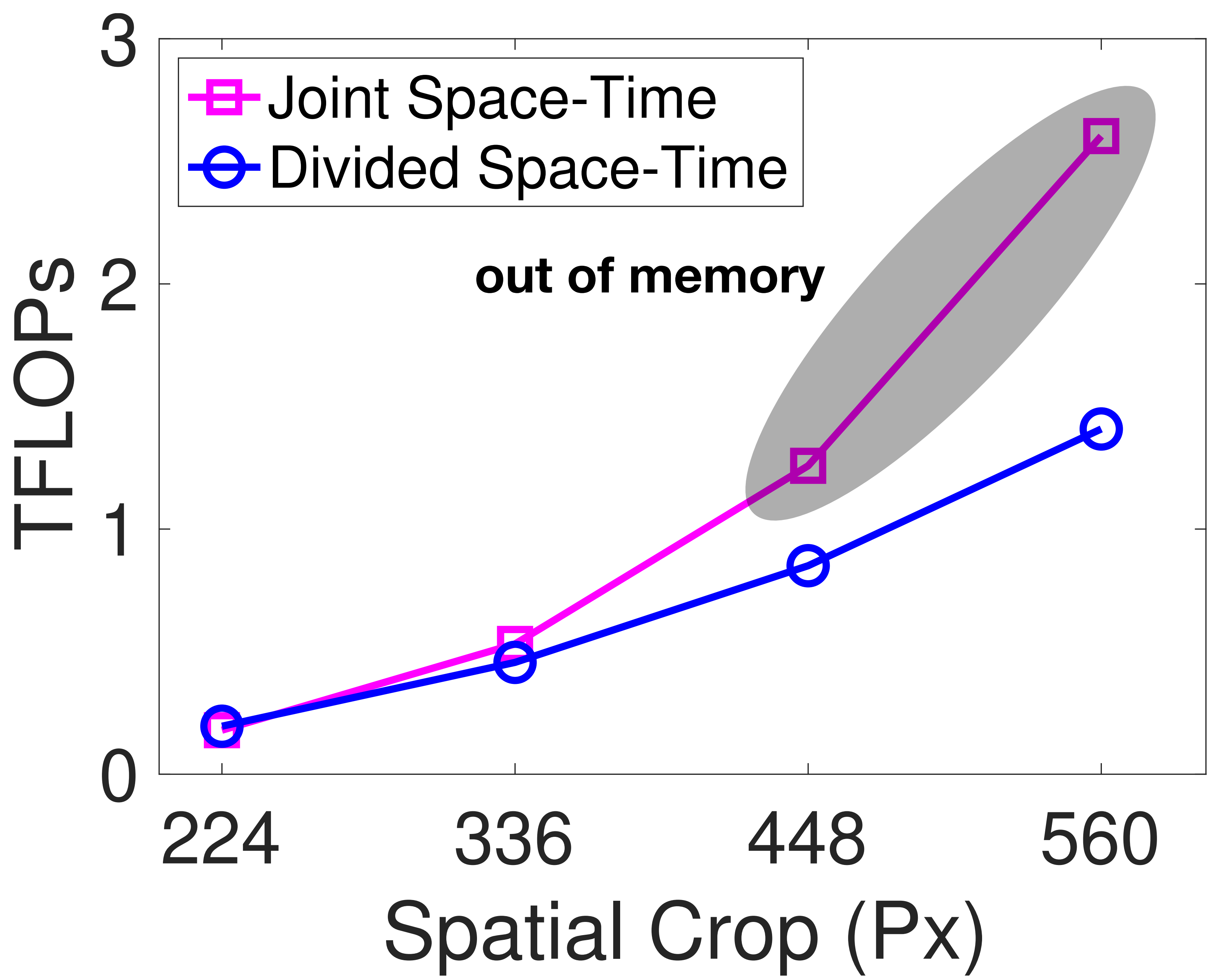
Analysis of Self-Attention Schemes

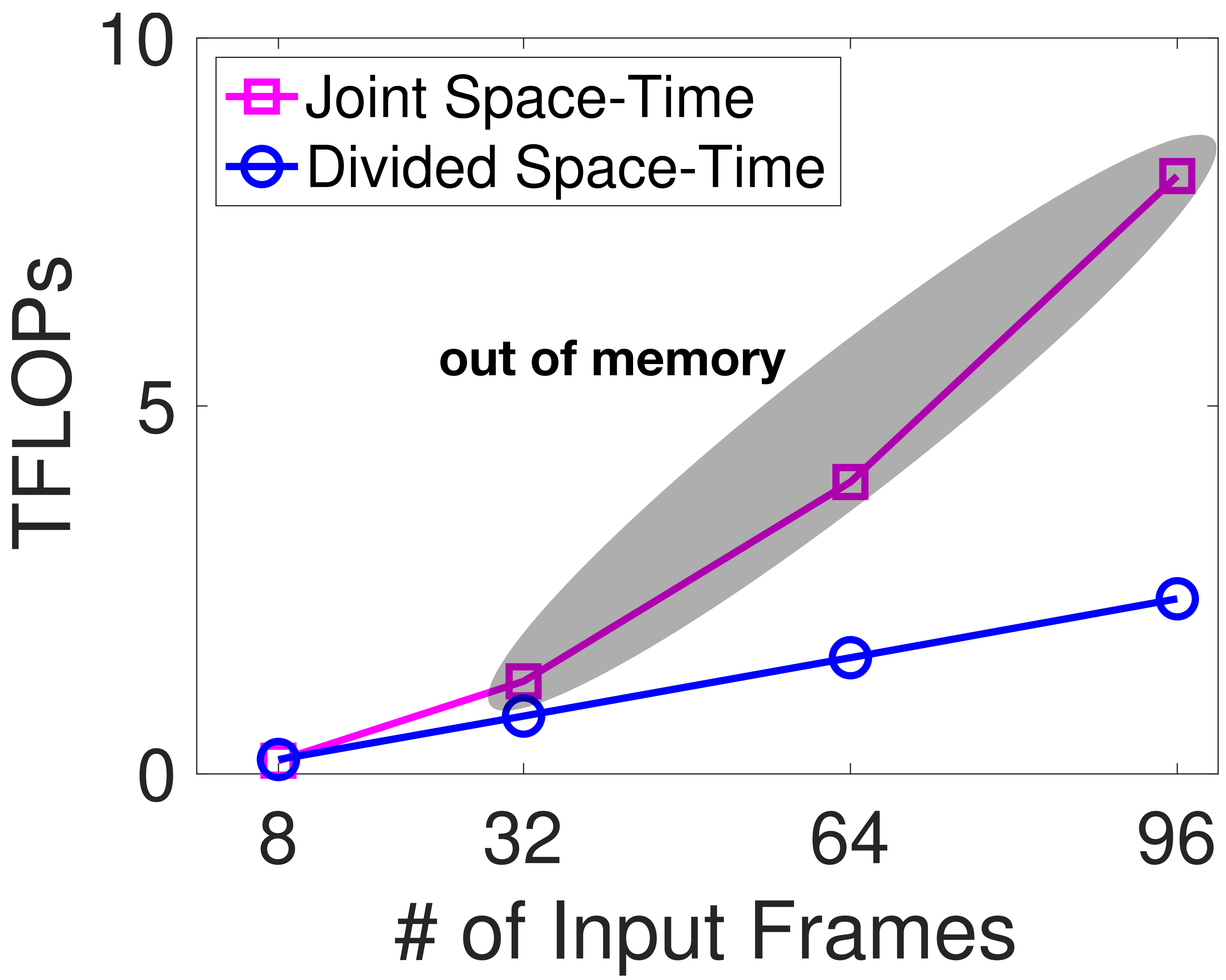
- Each space-time self-attention scheme is evaluated on Kinetics-400, and Something-Something-V2 datasets.

Attention	Pretraining	Params	K400	SSv2
Space	ImageNet-21K	85.9M	76.9	36.6
Joint Space-Time	ImageNet-21K	85.9M	77.4	58.5
Divided Space-Time	ImageNet-21K	121.4M	**78.0**	**59.5**

Analysis of Self-Attention Schemes

- As we increase the spatial resolution, or the video length, our proposed divided space-time attention leads to dramatic computational savings.





2. Is space-time attention better than 3D convolutions?

Comparison to 3D CNNs

- We investigate the distinguishing properties of TimeSformer compared to 3D CNNs.

Model	Pretrain	K400 Training Time (hours)	K400 Acc.	Inference TFLOPs	Params
I3D 8x8 R50	ImageNet-1K	444	71.0	1.11	28.0M
I3D 8x8 R50	ImageNet-1K	1440	73.4	1.11	28.0M
SlowFast R50	ImageNet-1K	448	70.0	1.97	34.6M
SlowFast R50	ImageNet-1K	3840	75.6	1.97	34.6M
SlowFast R50	N/A	6336	76.4	1.97	34.6M
TimeSformer	ImageNet-1K	**416**	75.8	**0.59**	121.4M
TimeSformer	ImageNet-21K	**416**	**78.0**	**0.59**	121.4M

Comparison to 3D CNNs

- We investigate the distinguishing properties of TimeSformer compared to 3D CNNs.

Model	Pretrain	K400 Training Time (hours)	K400 Acc.	Inference TFLOPs	Params
I3D 8x8 R50	ImageNet-1K	444	71.0	1.11	28.0M
I3D 8x8 R50	ImageNet-1K	1440	73.4	1.11	28.0M
SlowFast R50	ImageNet-1K	448	70.0	1.97	34.6M
SlowFast R50	ImageNet-1K	3840	75.6	1.97	34.6M
SlowFast R50	N/A	6336	76.4	1.97	34.6M
TimeSformer	ImageNet-1K	**416**	75.8	**0.59**	121.4M
TimeSformer	ImageNet-21K	**416**	**78.0**	**0.59**	121.4M

Comparison to 3D CNNs

- We investigate the distinguishing properties of TimeSformer compared to 3D CNNs.

Model	Pretrain	K400 Training Time (hours)	K400 Acc.	Inference TFLOPs	Params
I3D 8x8 R50	ImageNet-1K	444	71.0	1.11	28.0M
I3D 8x8 R50	ImageNet-1K	1440	73.4	1.11	28.0M
SlowFast R50	ImageNet-1K	448	70.0	1.97	34.6M
SlowFast R50	ImageNet-1K	3840	75.6	1.97	34.6M
SlowFast R50	N/A	6336	76.4	1.97	34.6M
TimeSformer	ImageNet-1K	**416**	75.8	**0.59**	121.4M
TimeSformer	ImageNet-21K	**416**	**78.0**	**0.59**	121.4M

Comparison to 3D CNNs

- We investigate the distinguishing properties of TimeSformer compared to 3D CNNs.

Model	Pretrain	K400 Training Time (hours)	K400 Acc.	Inference TFLOPs	Params
I3D 8x8 R50	ImageNet-1K	444	71.0	1.11	28.0M
I3D 8x8 R50	ImageNet-1K	1440	73.4	1.11	28.0M
SlowFast R50	ImageNet-1K	448	70.0	1.97	34.6M
SlowFast R50	ImageNet-1K	3840	75.6	1.97	34.6M
SlowFast R50	N/A	6336	76.4	1.97	34.6M
TimeSformer	ImageNet-1K	**416**	75.8	**0.59**	121.4M
TimeSformer	ImageNet-21K	**416**	**78.0**	**0.59**	121.4M

Comparison to 3D CNNs

- We investigate the distinguishing properties of TimeSformer compared to 3D CNNs.

Model	Pretrain	K400 Training Time (hours)	K400 Acc.	Inference TFLOPs	Params
I3D 8x8 R50	ImageNet-1K	444	71.0	1.11	28.0M
I3D 8x8 R50	ImageNet-1K	1440	73.4	1.11	28.0M
SlowFast R50	ImageNet-1K	448	70.0	1.97	34.6M
SlowFast R50	ImageNet-1K	3840	75.6	1.97	34.6M
SlowFast R50	N/A	6336	76.4	1.97	34.6M
TimeSformer	ImageNet-1K	**416**	75.8	**0.59**	121.4M
TimeSformer	ImageNet-21K	**416**	**78.0**	**0.59**	121.4M

Comparison to 3D CNNs

- We investigate the distinguishing properties of TimeSformer compared to 3D CNNs.

Model	Pretrain	K400 Training Time (hours)	K400 Acc.	Inference TFLOPs	Params
I3D 8x8 R50	ImageNet-1K	444	71.0	1.11	28.0M
I3D 8x8 R50	ImageNet-1K	1440	73.4	1.11	28.0M
SlowFast R50	ImageNet-1K	448	70.0	1.97	34.6M
SlowFast R50	ImageNet-1K	3840	75.6	1.97	34.6M
SlowFast R50	N/A	6336	76.4	1.97	34.6M
TimeSformer	ImageNet-1K	**416**	75.8	**0.59**	121.4M
TimeSformer	ImageNet-21K	**416**	**78.0**	**0.59**	121.4M

Comparison to 3D CNNs

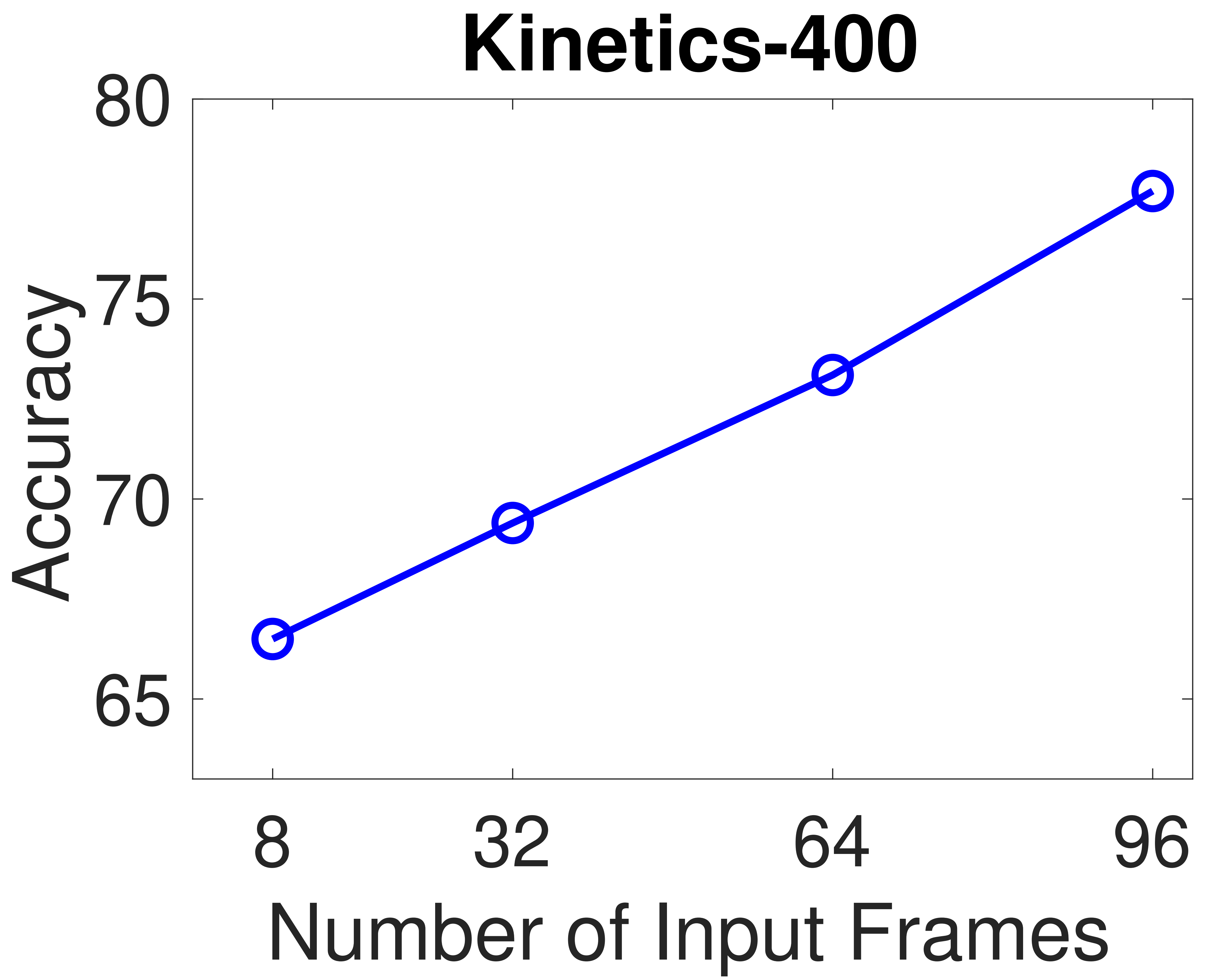
- We investigate the distinguishing properties of TimeSformer compared to 3D CNNs.

Model	Pretrain	K400 Training Time (hours)	K400 Acc.	Inference TFLOPs	Params
I3D 8x8 R50	ImageNet-1K	444	71.0	1.11	28.0M
I3D 8x8 R50	ImageNet-1K	1440	73.4	1.11	28.0M
SlowFast R50	ImageNet-1K	448	70.0	1.97	34.6M
SlowFast R50	ImageNet-1K	3840	75.6	1.97	34.6M
SlowFast R50	N/A	6336	76.4	1.97	34.6M
TimeSformer	ImageNet-1K	**416**	75.8	**0.59**	121.4M
TimeSformer	ImageNet-21K	**416**	**78.0**	**0.59**	121.4M

3. What is space-time attention particularly useful for?

Increasing the Video Length

- The scalability of our model allows it to operate on longer videos compared to most 3D CNNs.



Long-Term Video Modeling

- We evaluate our model's ability for long-term video modeling.

Key Details:

- **1059** long-term action categories (making breakfast, cleaning a house, etc).
- On average, each video is **~7min** long.
- **85K** training & **35K** testing videos.
- Performance is evaluated using a standard top-1 accuracy metric.

"Learning a Text-Video Embedding by Watching Hundred Million Narrated Video Clips", Miech et al., ICCV 2019

Long-Term Video Modeling

- “Single Clip Coverage” denotes the number of seconds spanned by a single clip.
- “# Test Clips” is the average number of clips needed to cover the entire input video during inference.

Method	# Input Frames	Frame Sampling Rate	Single Clip Coverage	# Test Clips	Top-1 Acc
SlowFast R101	8	1/32	8.5s	48	48.2
SlowFast R101	32	1/32	34.1s	12	50.8
SlowFast R101	64	1/32	68.3s	6	51.5
SlowFast R101	96	1/32	102.4s	4	51.2
TimeSformer	8	1/32	8.5s	48	56.0
TimeSformer	32	1/32	34.1s	12	59.2
TimeSformer	64	1/32	68.3s	6	60.2
TimeSformer	96	1/32	102.4s	4	62.1

Long-Term Video Modeling

- "Single Clip Coverage" denotes the number of seconds spanned by a single clip.
- "# Test Clips" is the average number of clips needed to cover the entire input video during inference.

Method	# Input Frames	Frame Sampling Rate	Single Clip Coverage	# Test Clips	Top-1 Acc
SlowFast R101	8	1/32	8.5s	48	48.2
SlowFast R101	32	1/32	34.1s	12	50.8
SlowFast R101	64	1/32	68.3s	6	51.5
SlowFast R101	96	1/32	102.4s	4	51.2
TimeSformer	8	1/32	8.5s	48	56.0
TimeSformer	32	1/32	34.1s	12	59.2
TimeSformer	64	1/32	68.3s	6	60.2
TimeSformer	96	1/32	102.4s	4	62.1

Long-Term Video Modeling

- "Single Clip Coverage" denotes the number of seconds spanned by a single clip.
- "# Test Clips" is the average number of clips needed to cover the entire input video during inference.

Method	# Input Frames	Frame Sampling Rate	Single Clip Coverage	# Test Clips	Top-1 Acc
SlowFast R101	8	1/32	8.5s	48	48.2
SlowFast R101	32	1/32	34.1s	12	50.8
SlowFast R101	64	1/32	68.3s	6	51.5
SlowFast R101	96	1/32	102.4s	4	51.2
TimeSformer	8	1/32	8.5s	48	56.0
TimeSformer	32	1/32	34.1s	12	59.2
TimeSformer	64	1/32	68.3s	6	60.2
TimeSformer	96	1/32	102.4s	4	62.1

4. Is space-time attention all you need for video understanding?

- Compared to modern 3D CNNs, TimeSformer has a larger learning capacity, and a comparable or even lower inference cost.

- 😊 Compared to modern 3D CNNs, TimeSformer has a larger learning capacity, and a comparable or even lower inference cost.
- 😊 Our method does not require a very long optimization schedule, and thus, it can be trained efficiently on video data.

- 😊 Compared to modern 3D CNNs, TimeSformer has a larger learning capacity, and a comparable or even lower inference cost.
- 😊 Our method does not require a very long optimization schedule, and thus, it can be trained efficiently on video data.
- 😊 TimeSformer can handle much longer videos, which makes it highly suitable for long-term video modeling.

- 😊 Compared to modern 3D CNNs, TimeSformer has a larger learning capacity, and a comparable or even lower inference cost.

- 😊 Our method does not require a very long optimization schedule, and thus, it can be trained efficiently on video data.

- 😊 TimeSformer can handle much longer videos, which makes it highly suitable for long-term video modeling.

- ☹️ Due to a large number of parameters, TimeSformer requires image-level pretraining.

- 😊 Compared to modern 3D CNNs, TimeSformer has a larger learning capacity, and a comparable or even lower inference cost.

- 😊 Our method does not require a very long optimization schedule, and thus, it can be trained efficiently on video data.

- 😊 TimeSformer can handle much longer videos, which makes it highly suitable for long-term video modeling.

- ☹️ Due to a large number of parameters, TimeSformer requires image-level pretraining.

- ☹️ Improvements are needed for learning more effective features on temporally heavy datasets (e.g. SSv2).

Discussion Questions

- Is space-time attention all you need for video understanding?

Discussion Questions

- Is space-time attention all you need for video understanding?
- Can TimeSformer recognize actions that involve fast-moving objects?