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

ICML 2021

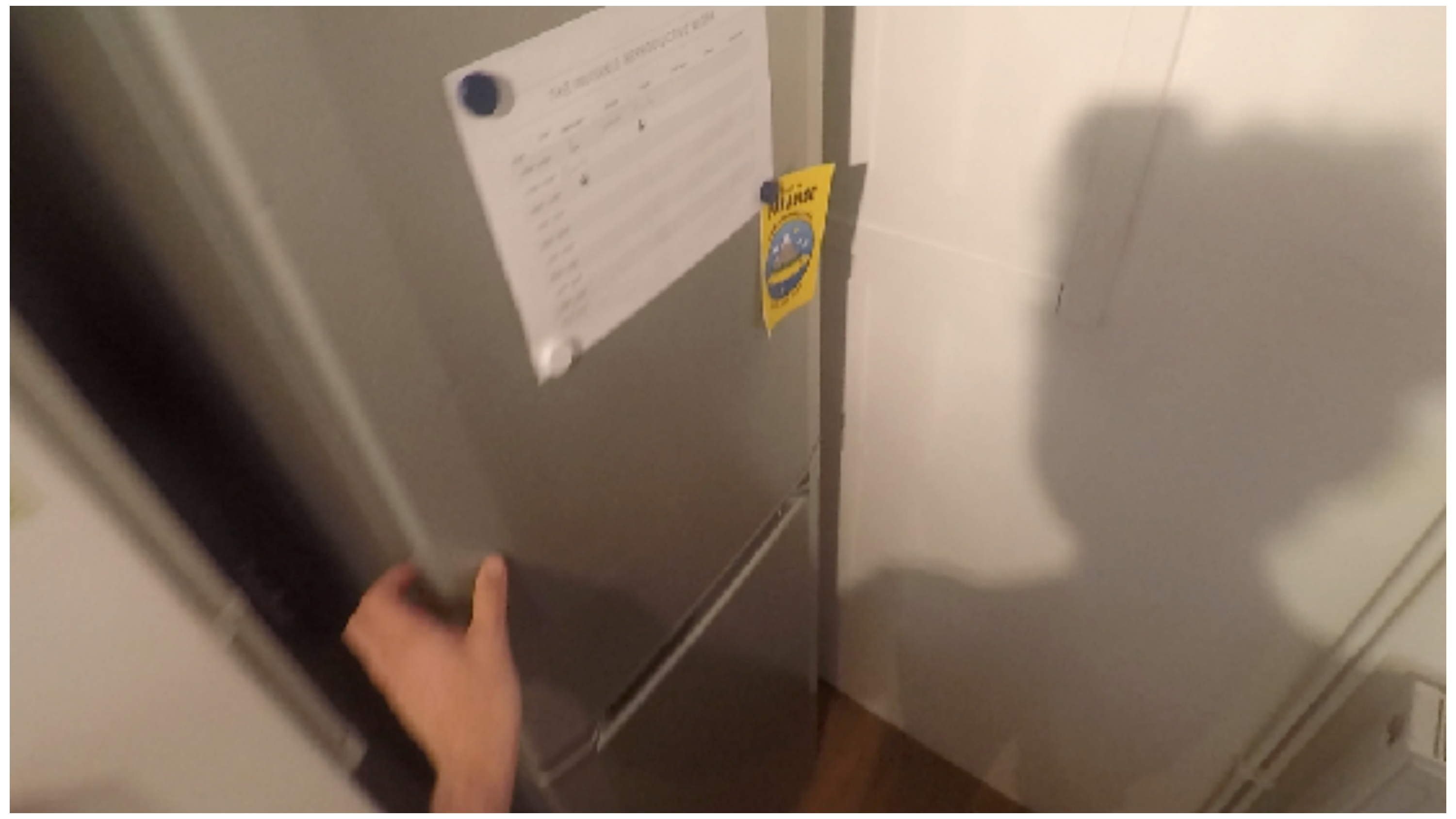
Gedas Bertasius, Heng Wang, Lorenzo Torresani

Video Classification

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

Cartwheeling

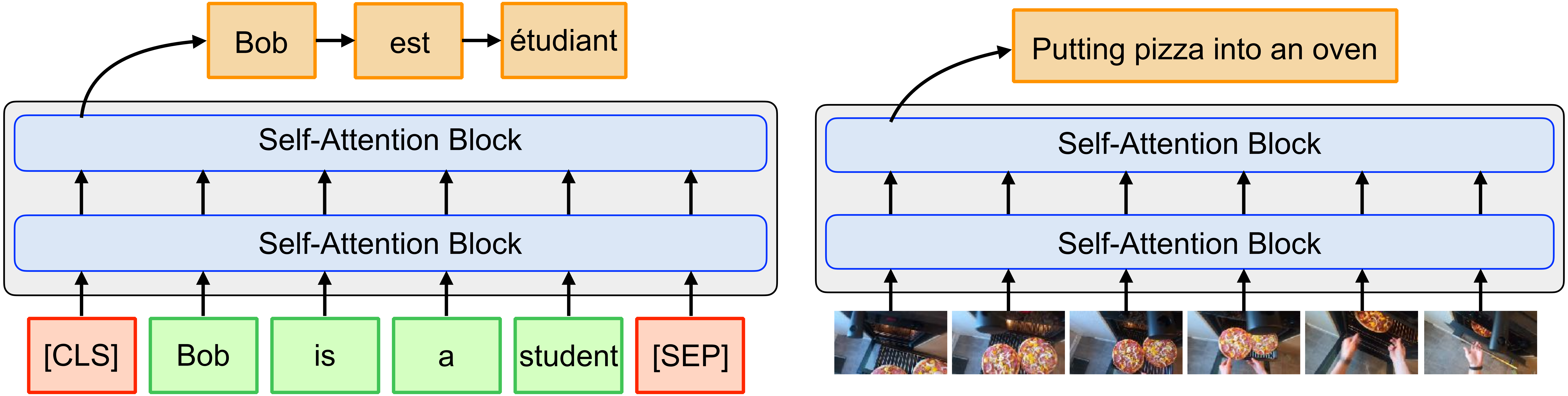
Braiding Hair



Opening a Fridge

Modern Language Models

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



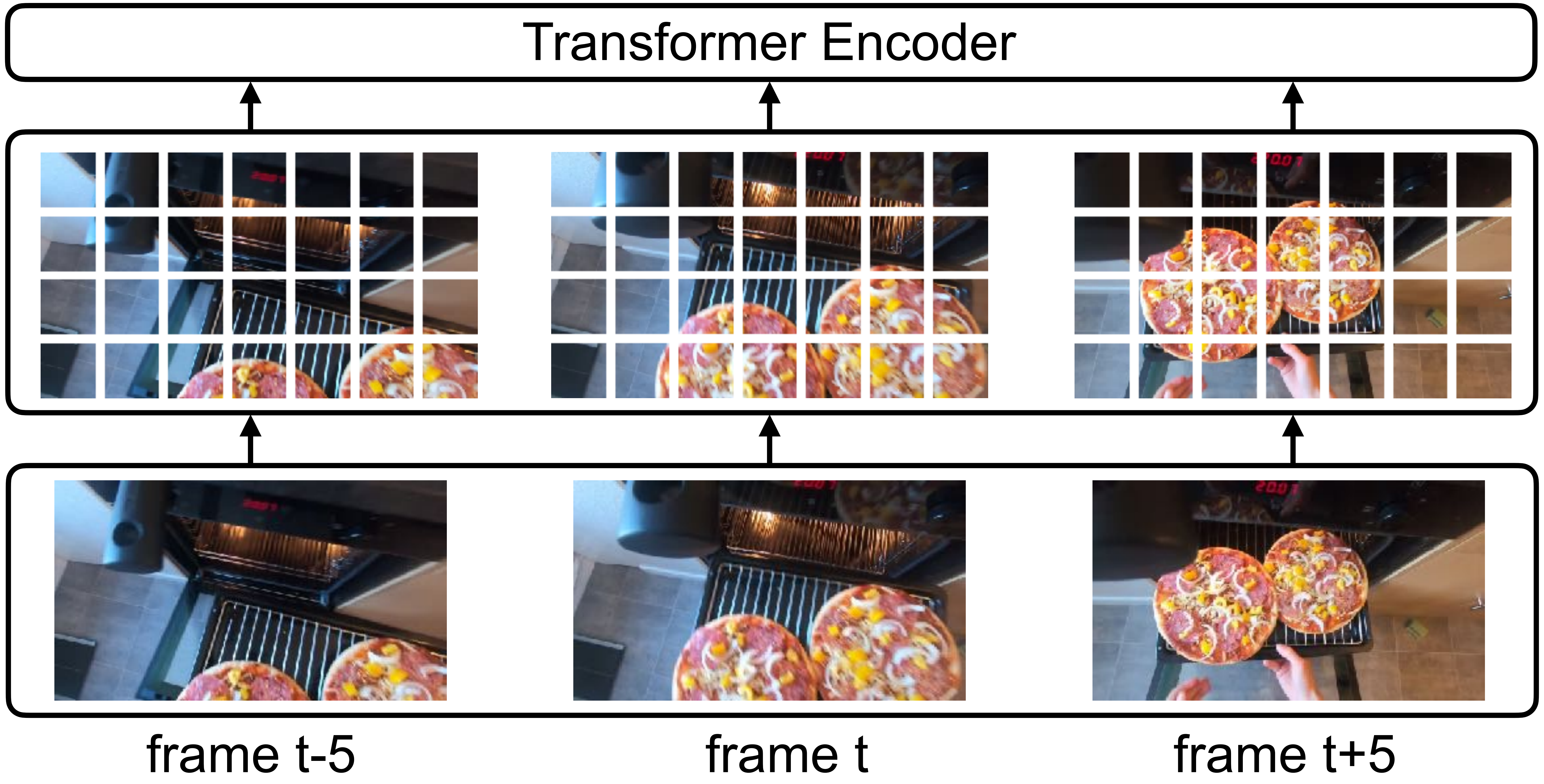
a) Language Model

b) Video Model

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

Video Decomposition

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

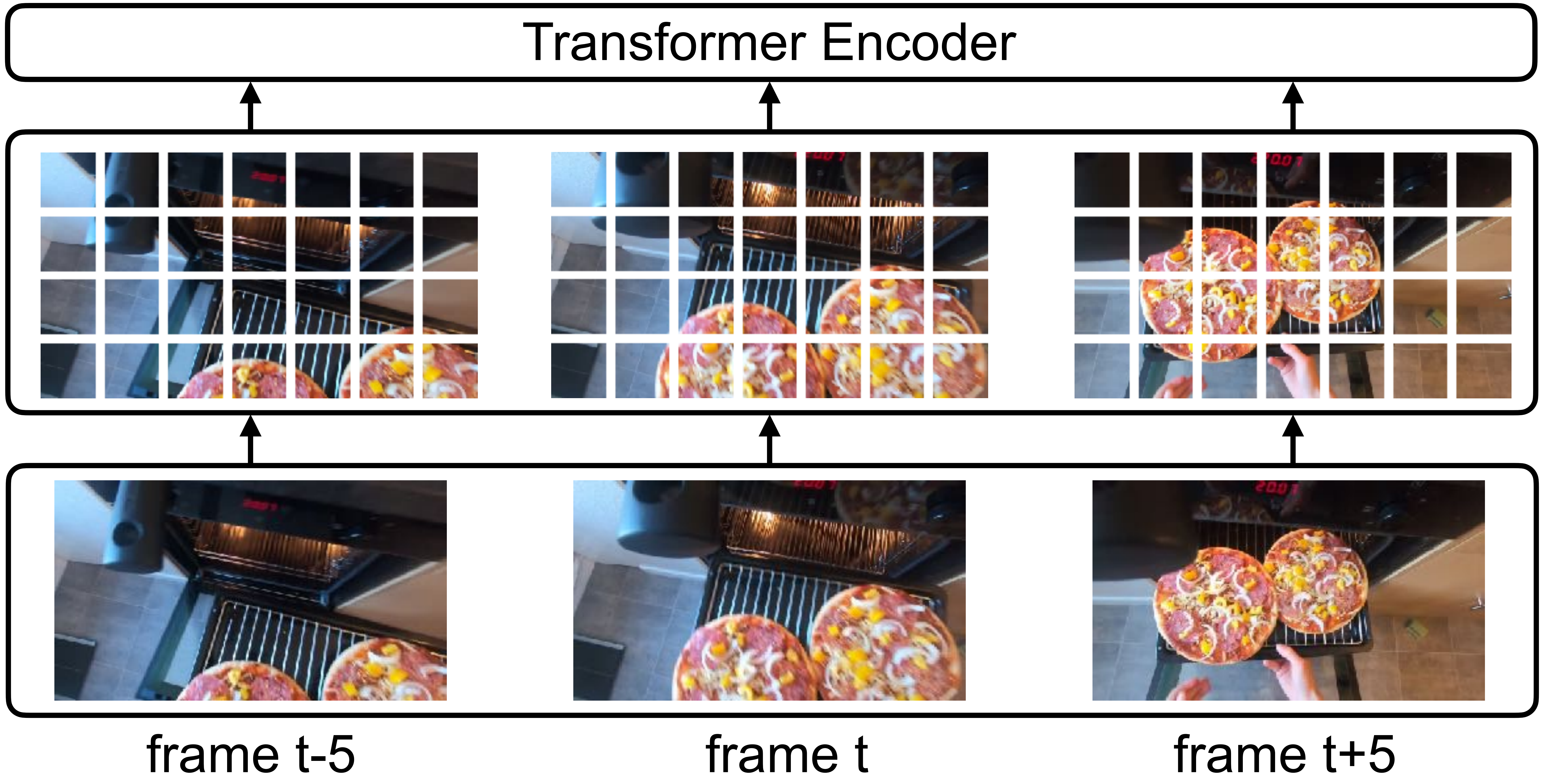


"An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale", Dosovitskiy et al., ICLR 2020

Video Decomposition

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

Computing similarity for all pairs of patches is costly.

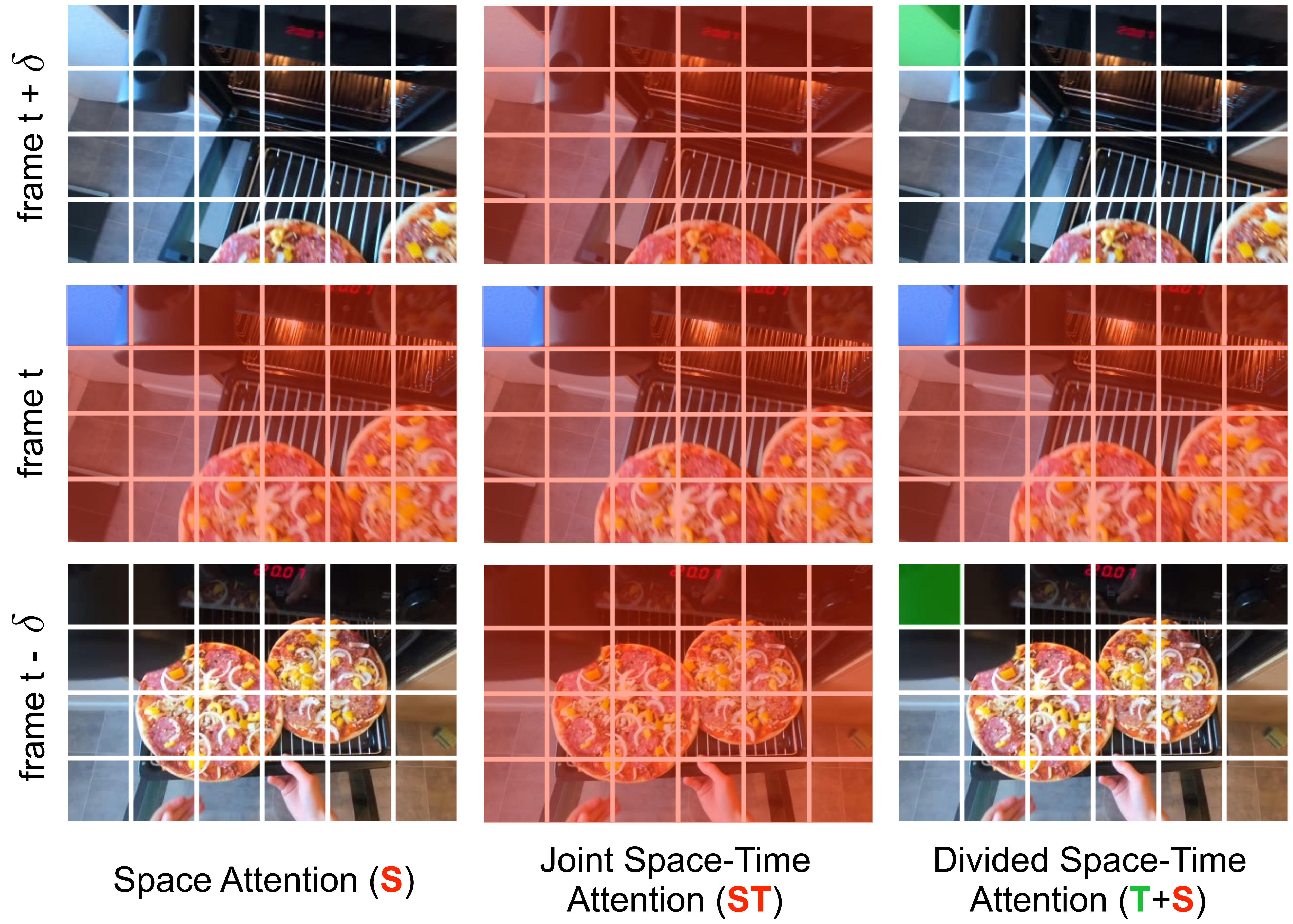


"An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale", Dosovitskiy et al., ICLR 2020

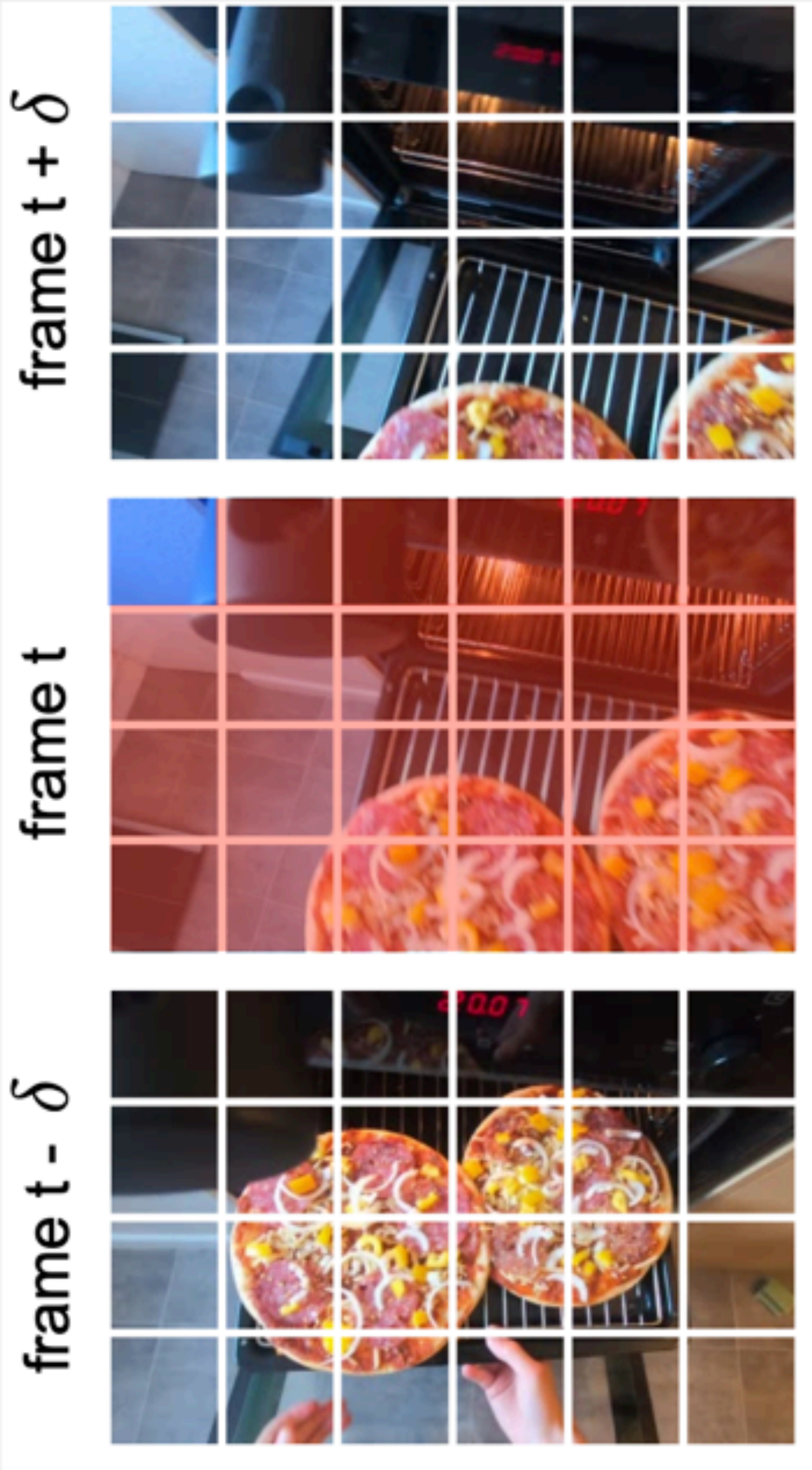
1. What is the right space-time self-attention pattern?

Space-Time Self-Attention

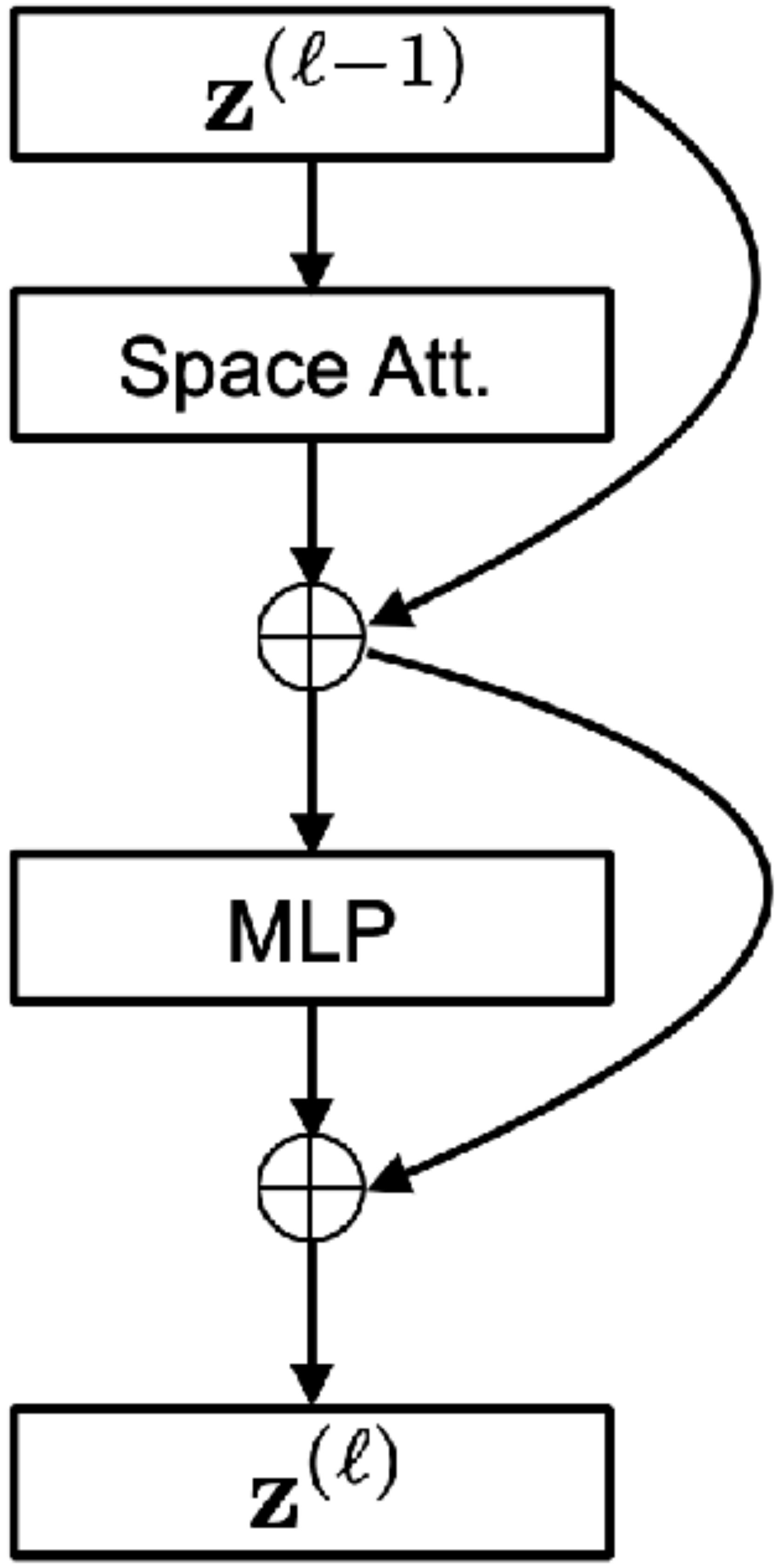
- We investigate several space-time self-attention schemes.



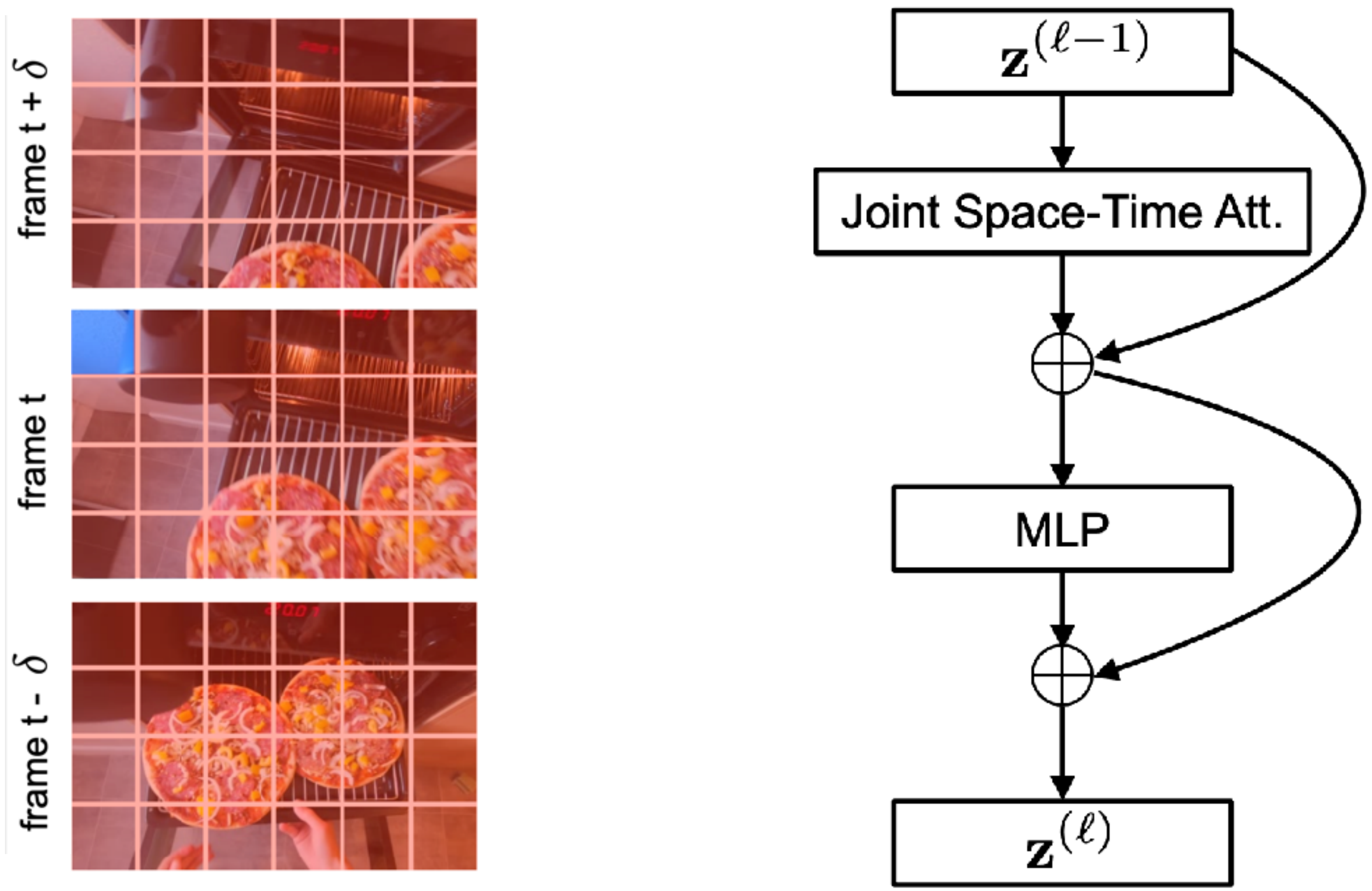
Spatial Self-Attention



Space Attention (S)

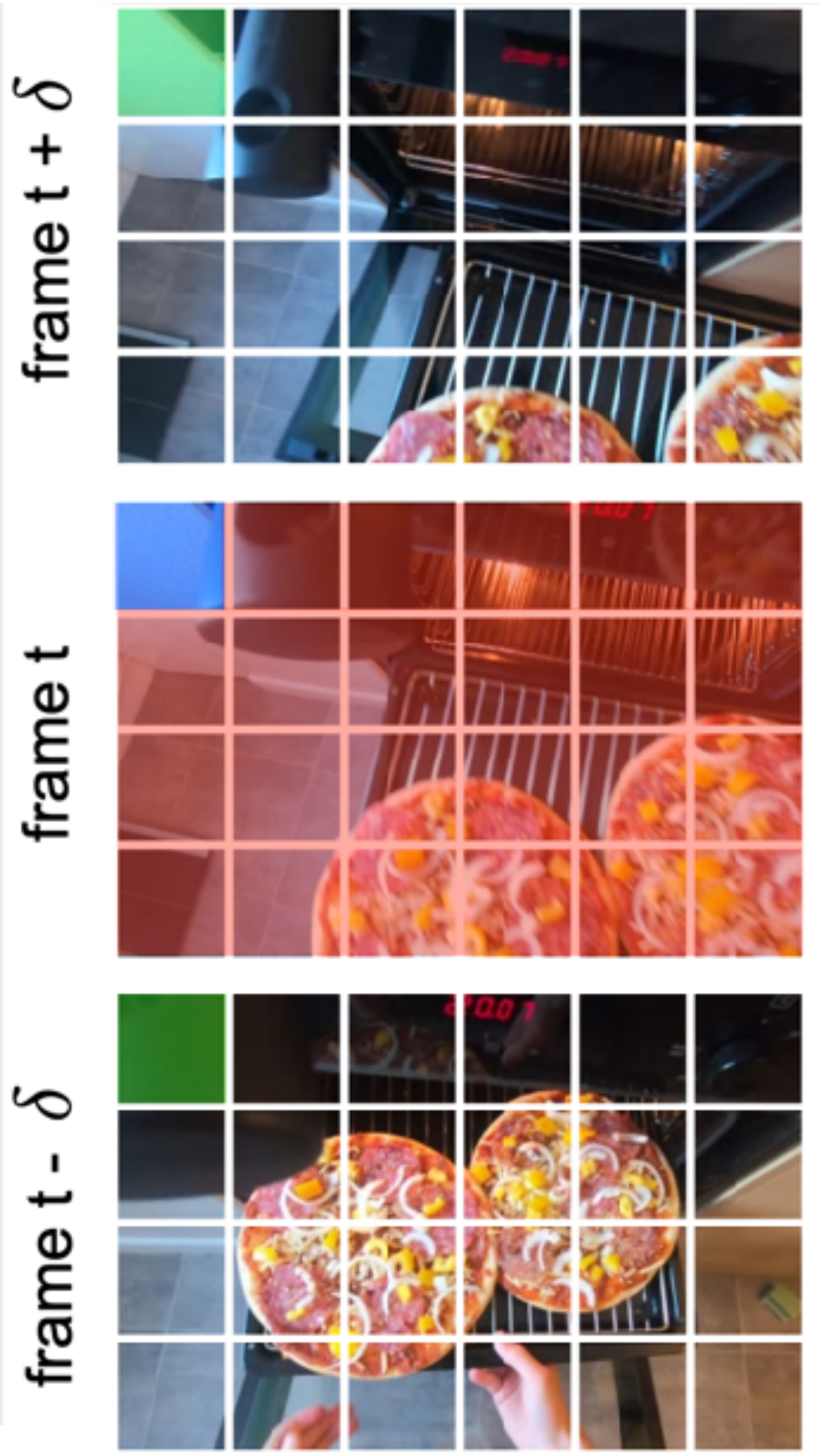


Joint Space-Time Self-Attention

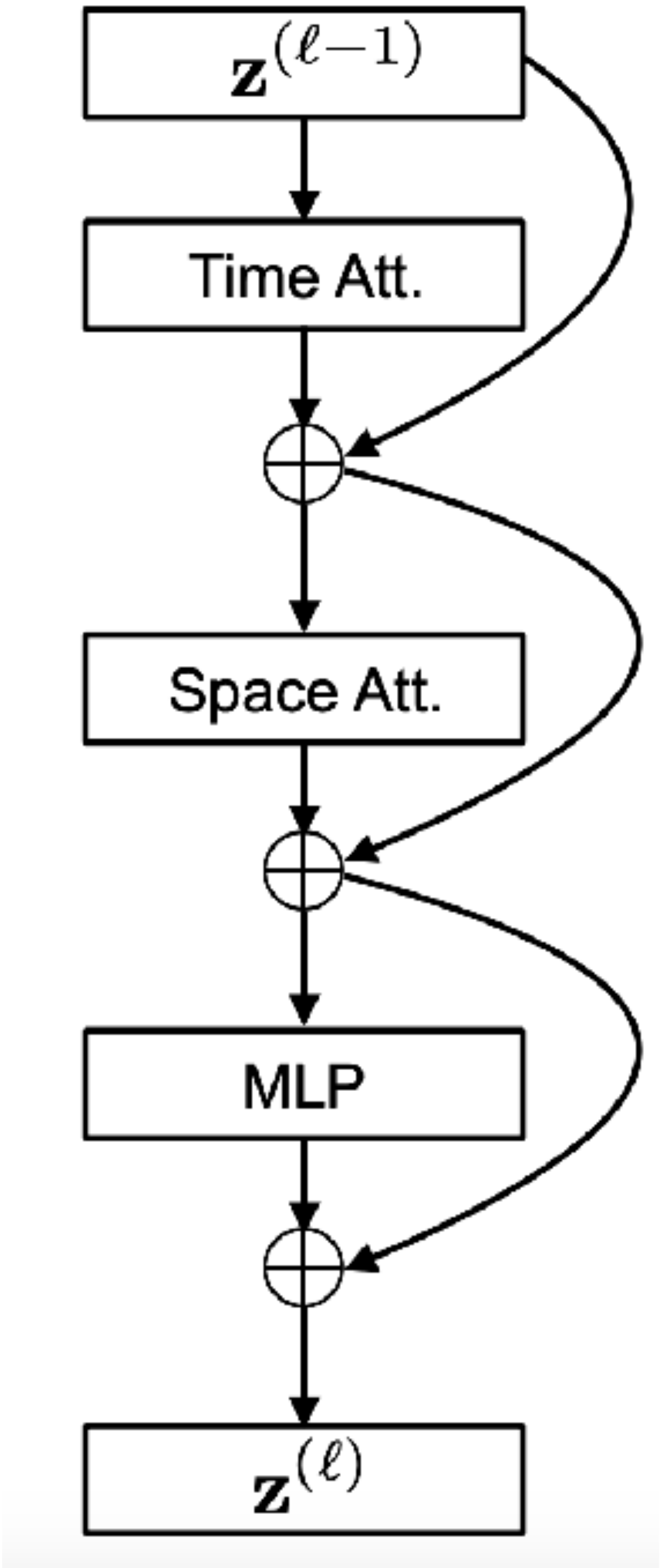


Joint Space-Time
Attention (**ST**)

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

- 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**

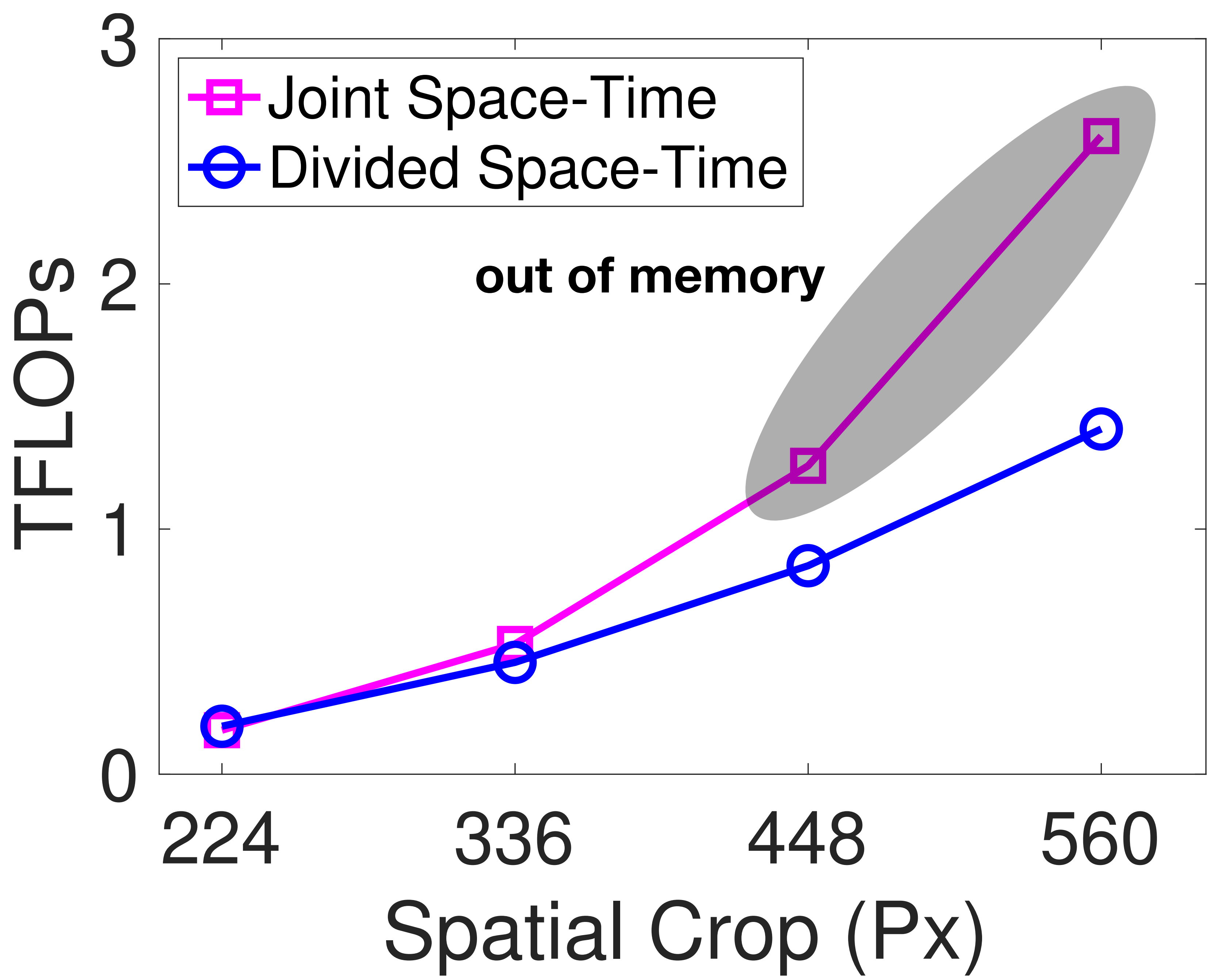
Analysis of Self-Attention Schemes

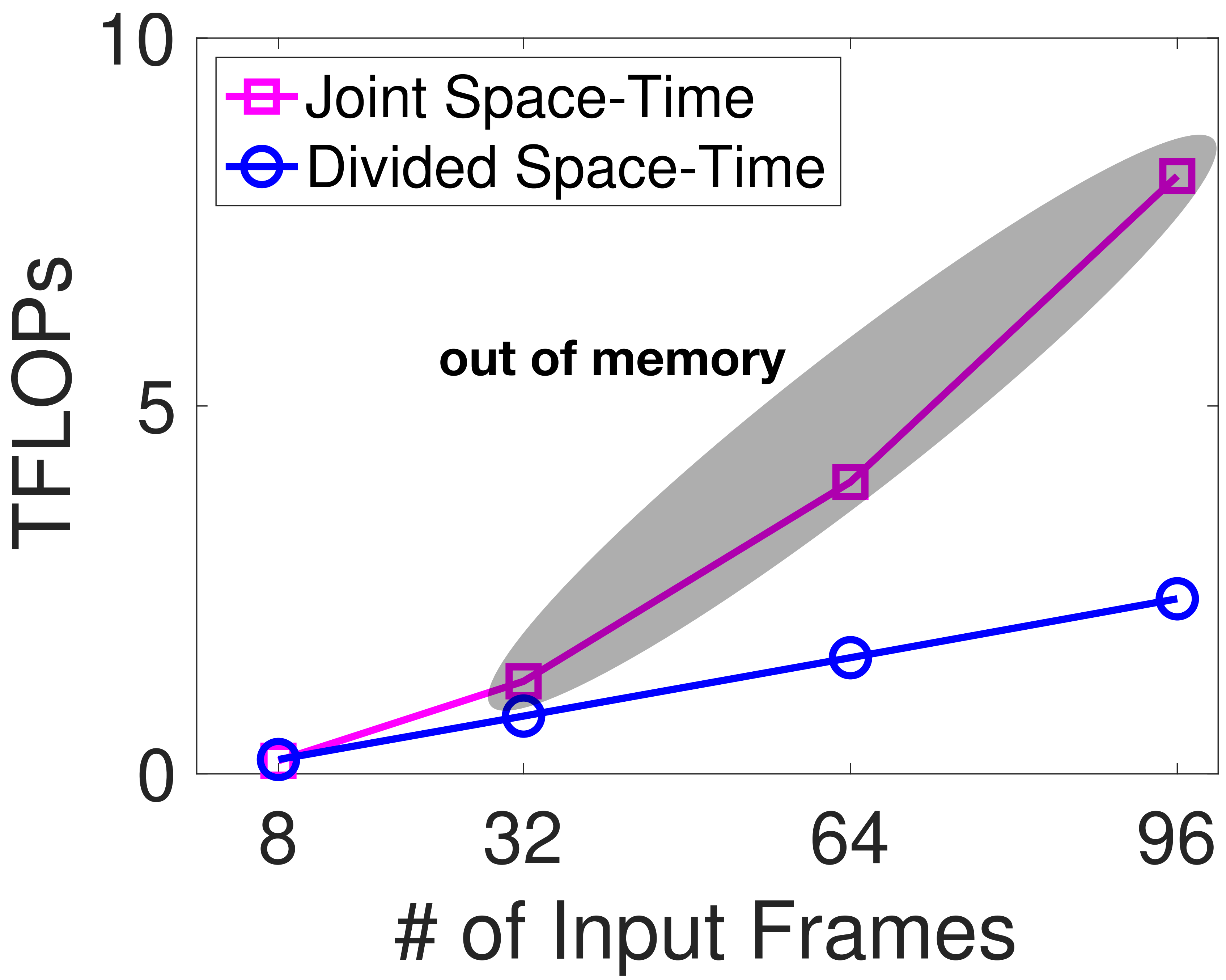
- Each space-time self-attention scheme is evaluated on Kinetics-400, and Something-Something-V2 datasets.

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

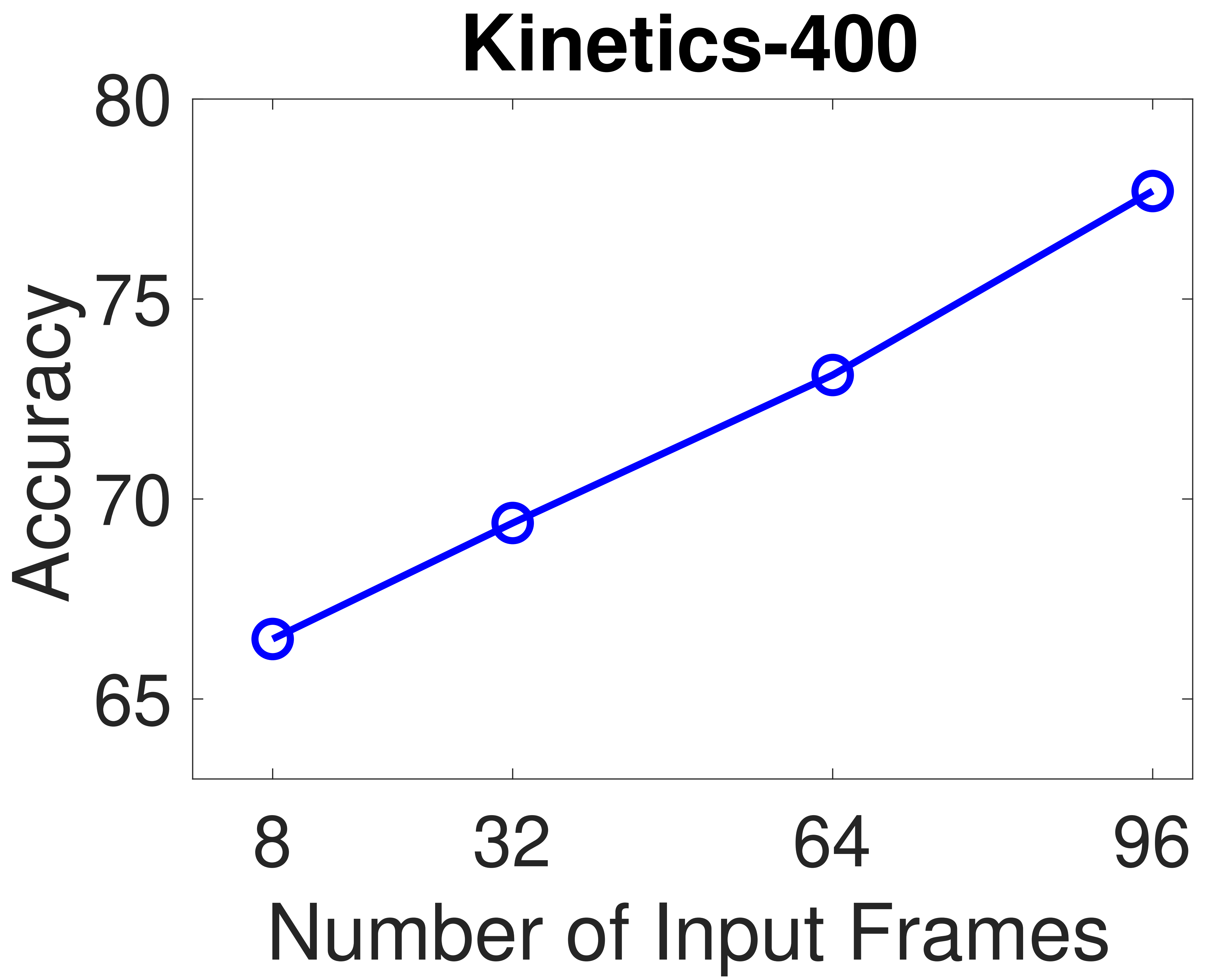
- We investigate the distinguishing properties of TimeSformer compared to 3D CNNs.

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.

Method	# Input Frames	Single Clip Coverage	Top-1 Acc
SlowFast	8	8.5s	48.2
SlowFast	32	34.1s	50.8
SlowFast	64	68.3s	51.5
SlowFast	96	102.4s	51.2
TimeSformer	8	8.5s	56.8
TimeSformer	32	34.1s	61.2
TimeSformer	64	68.3s	62.2
TimeSformer	96	102.4s	**62.6**

Long-Term Video Modeling

- “Single Clip Coverage” denotes the number of seconds spanned by a single clip.

Method	# Input Frames	Single Clip Coverage	Top-1 Acc
SlowFast	8	8.5s	48.2
SlowFast	32	34.1s	50.8
SlowFast	64	68.3s	51.5
SlowFast	96	102.4s	51.2
TimeSformer	8	8.5s	56.8
TimeSformer	32	34.1s	61.2
TimeSformer	64	68.3s	62.2
TimeSformer	96	102.4s	**62.6**

Long-Term Video Modeling

- "Single Clip Coverage" denotes the number of seconds spanned by a single clip.

Method	# Input Frames	Single Clip Coverage	Top-1 Acc
SlowFast	8	8.5s	48.2
SlowFast	32	34.1s	50.8
SlowFast	64	68.3s	51.5
SlowFast	96	102.4s	51.2
TimeSformer	8	8.5s	56.8
TimeSformer	32	34.1s	61.2
TimeSformer	64	68.3s	62.2
TimeSformer	96	102.4s	**62.6**

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

1. Can TimeSformer recognize actions that involve fast-moving objects?

2. Why does TimeSformer struggle with temporally-heavy datasets such as SSv2? How can we improve it?

3. What is the main reason that divided attention can outperform joint attention?

4. How would the performance change if we swapped the order of time and space attention in each block?

5. Why does the accuracy suddenly drop when the spatial crop side reaches 560 pixels?

6. Why does using the larger ImageNet-21K compared to the ImageNet-1K results in better performance on the K400 dataset but a similar performance on the SSv2 dataset?

7. What are the main advantages of video transformers over 3D CNNs (if any)?

8. Are the comparisons with 3D CNNs fair (given the varying parameter counts)?

9. What are the potential advantages of combining CNNs with Transformers for video recognition?

10. Will transformers replace convolution-based methods for video understanding? Why or why not?

11. How would this approach work for capturing longer range temporal dependencies (10min or more)?

Discussion Questions

1. Can TimeSformer recognize actions that involve fast-moving objects?

Discussion Questions

2. Why does TimeSformer struggle with temporally-heavy datasets such as SSv2? How can we improve it?

Discussion Questions

3. What is the main reason that divided attention can outperform joint attention?

Discussion Questions

4. How would the performance change if we swapped the order of time and space attention in each block?

Discussion Questions

5. Why does the accuracy suddenly drop when the spatial crop side reaches 560 pixels?

Discussion Questions

6. Why does using the larger ImageNet-21K compared to the ImageNet-1K results in better performance on the K400 dataset but a similar performance on the SSv2 dataset?

Discussion Questions

7. What are the main advantages of video transformers over 3D CNNs (if any)?

Discussion Questions

8. Are the comparisons with 3D CNNs fair (given the varying parameter counts)?

Discussion Questions

9. What are the potential advantages of combining CNNs with Transformers for video recognition?

Discussion Questions

10. Will transformers replace convolution-based methods for video understanding? Why or why not?

Discussion Questions

11. How would this approach work for capturing longer range temporal dependencies (10min or more)?