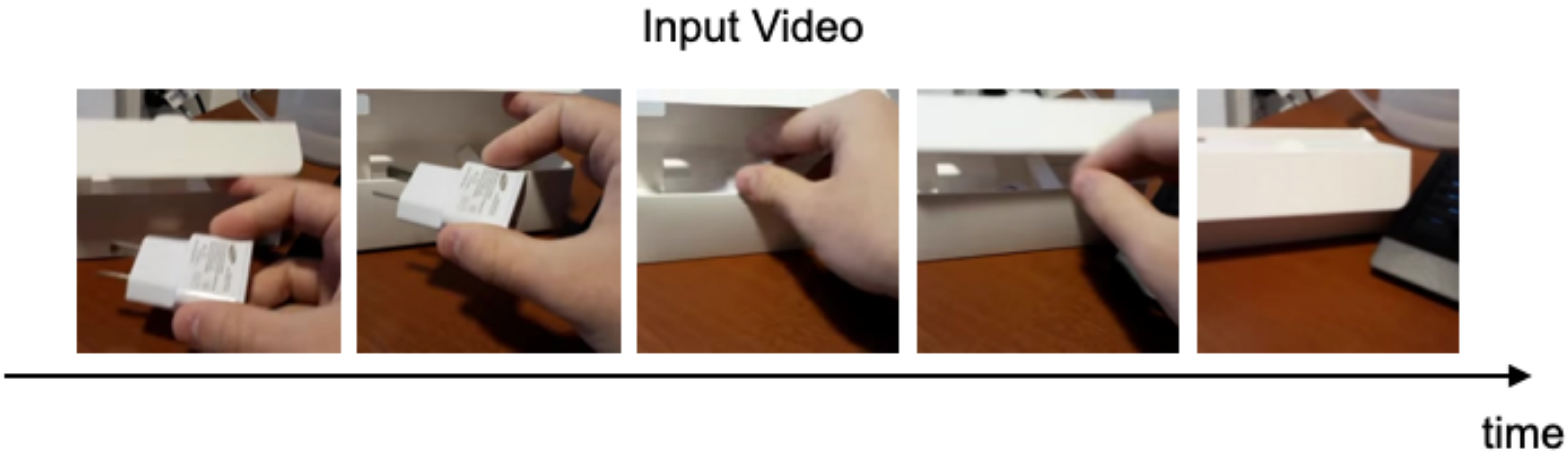
Keeping Your Eye on the Ball: Trajectory Attention in Video Transformers

NeurIPS 2021

Mandela Patrick, Dylan Campbell, Yuki M. Asano, Ishan Misra, Florian Metze, Christoph Feichtenhofer, Andrea Vedaldi, João F. Henriques

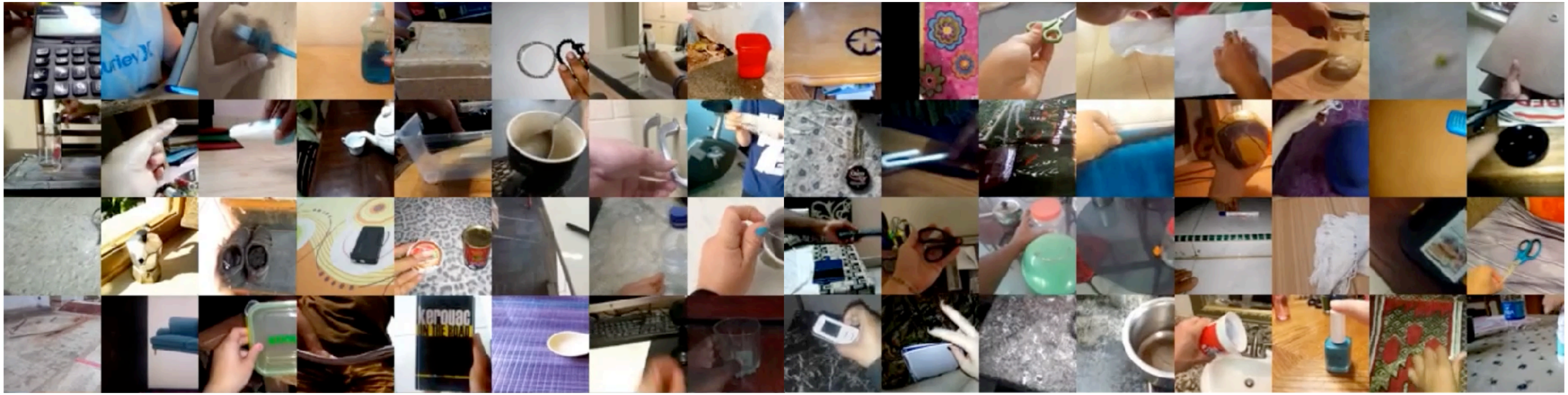
Motivation

- Motion cues often provide relevant information for recognizing a person's actions.



Something-Something Dataset

- ~200K videos of "temporally heavy" human actions (e.g., Covering [something] with [something], Letting [something] roll along a flat surface, etc.).



Something-Something Dataset

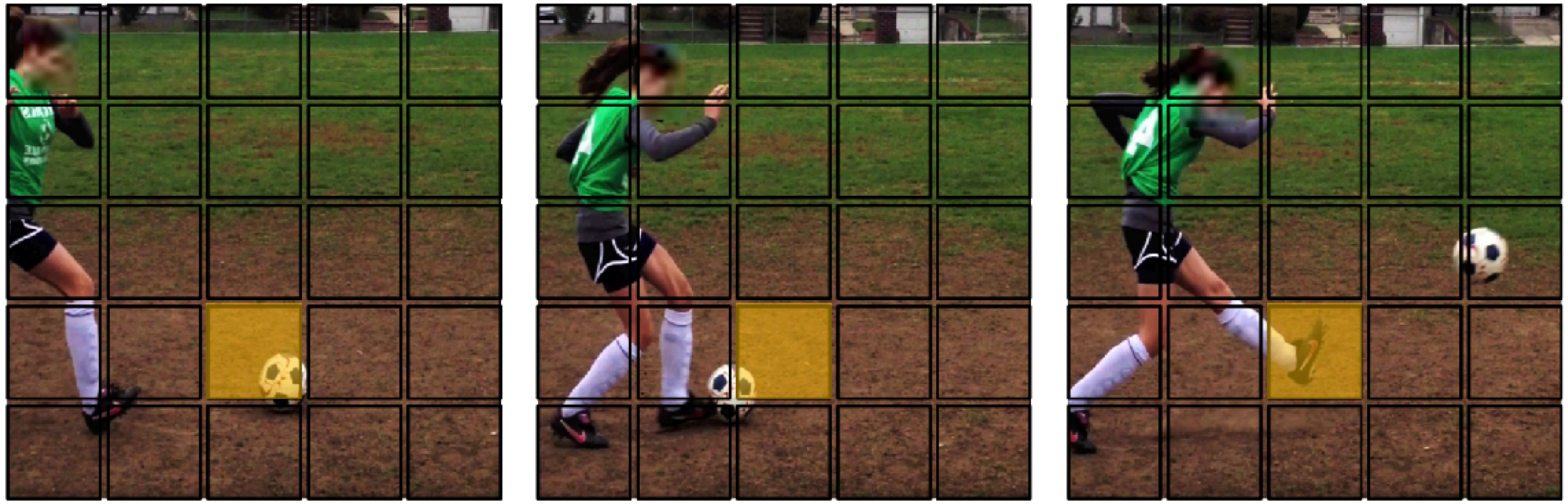
- ~200K videos of “temporally heavy” human actions (e.g., Covering [something] with [something], Letting [something] roll along a flat surface, etc.).

Method	SSv2
SlowFast (Feichtenhofer et al., 2019b)	61.7
TSM (Lin et al., 2019)	63.4
STM (Jiang et al., 2019)	64.2
MSNet (Kwon et al., 2020)	64.7
TEA (Li et al., 2020b)	65.1
bLVNet (Fan et al., 2019)	**65.2**
TimeSformer	59.5

TimeSformer underperforms on temporally heavy datasets like Something-Something-v2

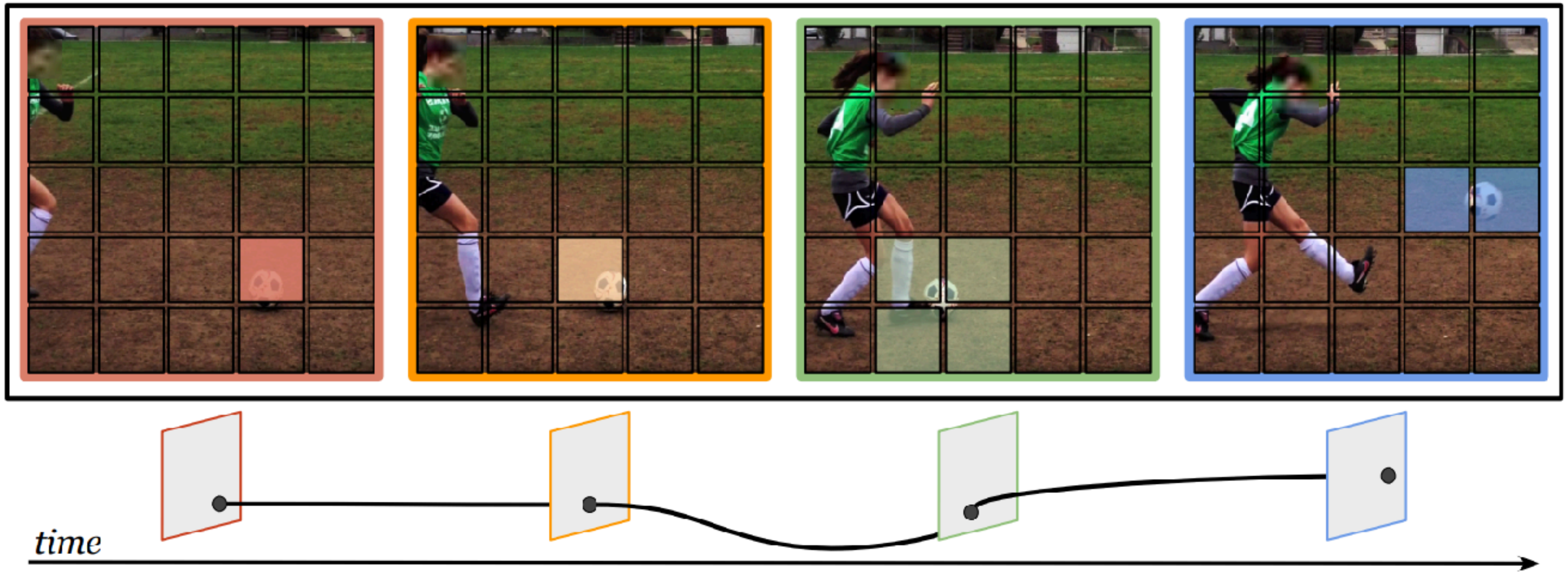
Motivation

- The authors argue that TimeSformer's divided space-time attention cannot capture motion trajectories of objects across time.



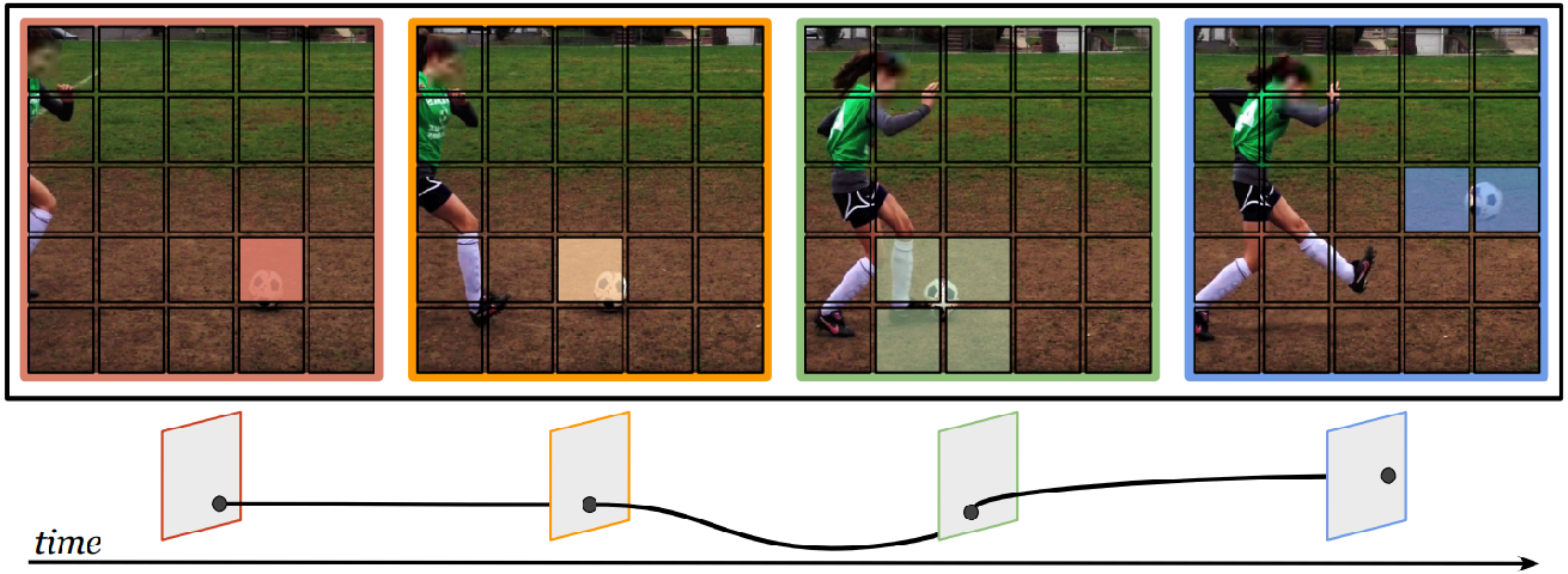
Trajectory Attention

- Instead, the authors propose trajectory attention, a mechanism for aggregating information along implicitly determined motion paths.



Trajectory Attention

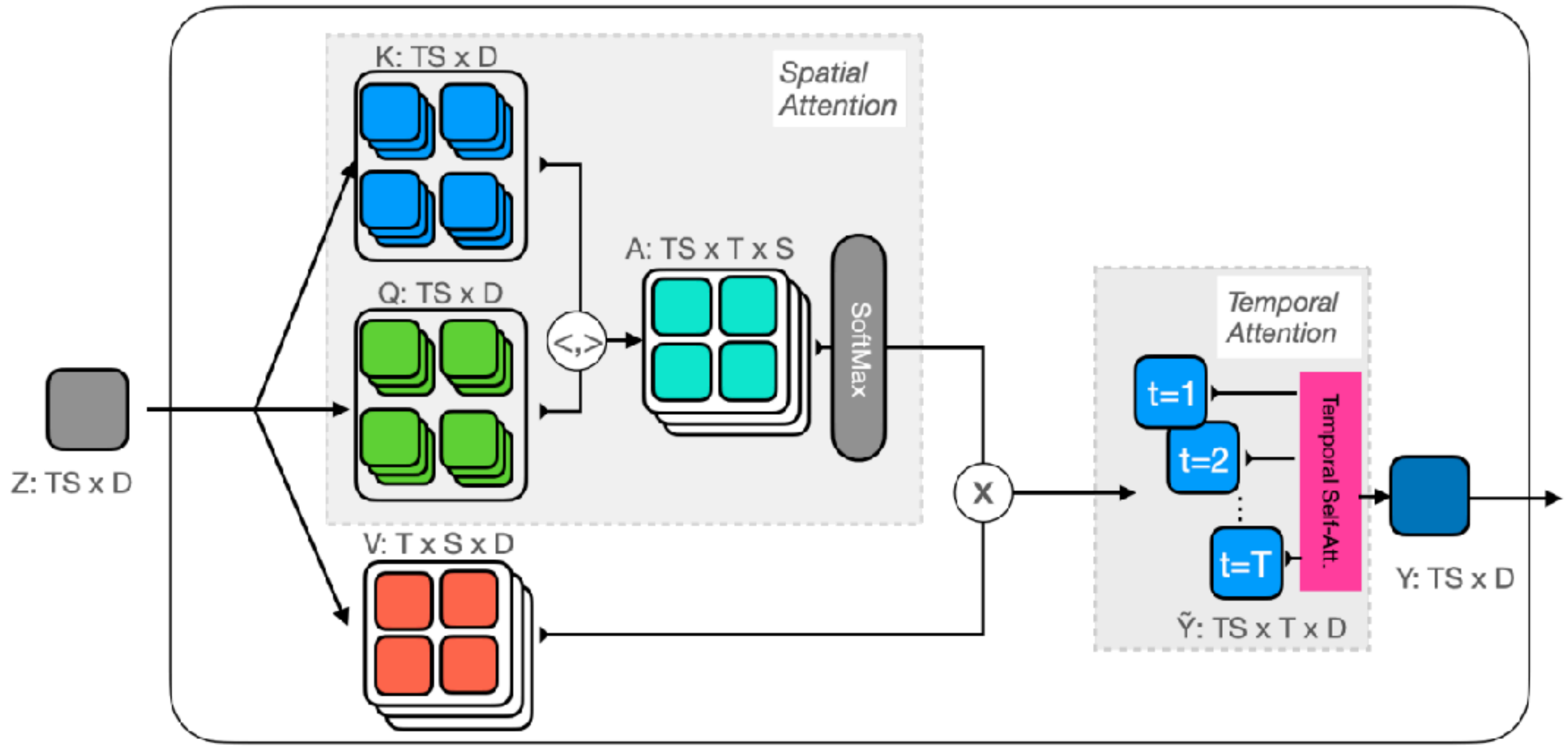
- Instead, the authors propose trajectory attention, a mechanism for aggregating information along implicitly determined motion paths.



Without using optical flow or any other explicit motion-based data.

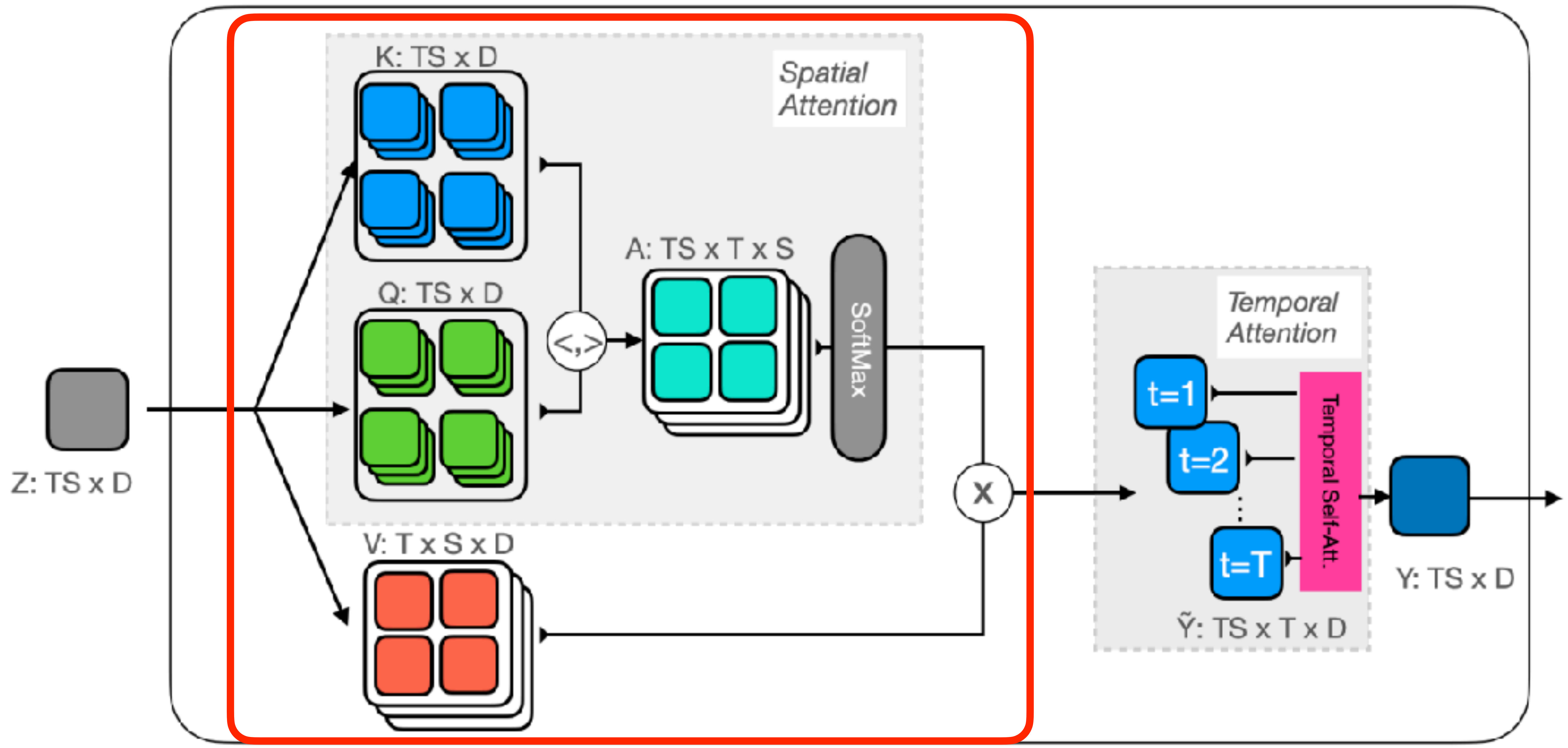
Technical Approach

- The proposed trajectory attention operation takes as input a TSxD spatiotemporal tensor where T is the number of frames and S depicts the spatial dimension, and D is the feature dimensionality.



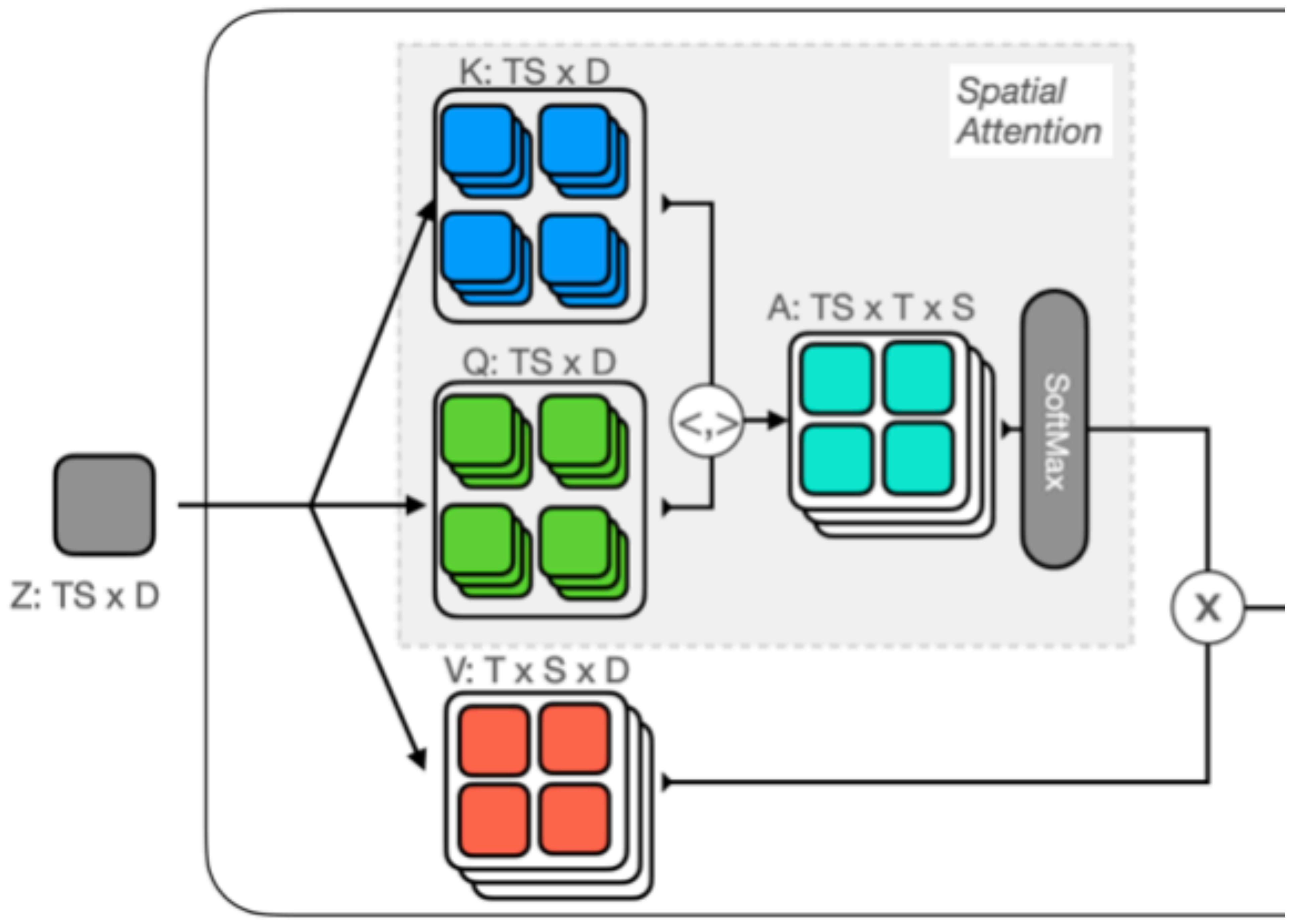
Technical Approach

- The proposed trajectory attention operation takes as input a TSxD spatiotemporal tensor where T is the number of frames and S depicts the spatial dimension, and D is the feature dimensionality.



Technical Approach

- Compared to prior approaches, the attention is performed along trajectories.



This is a standard quadratic self-attention over spatiotemporal feature volume.

Trajectory Attention

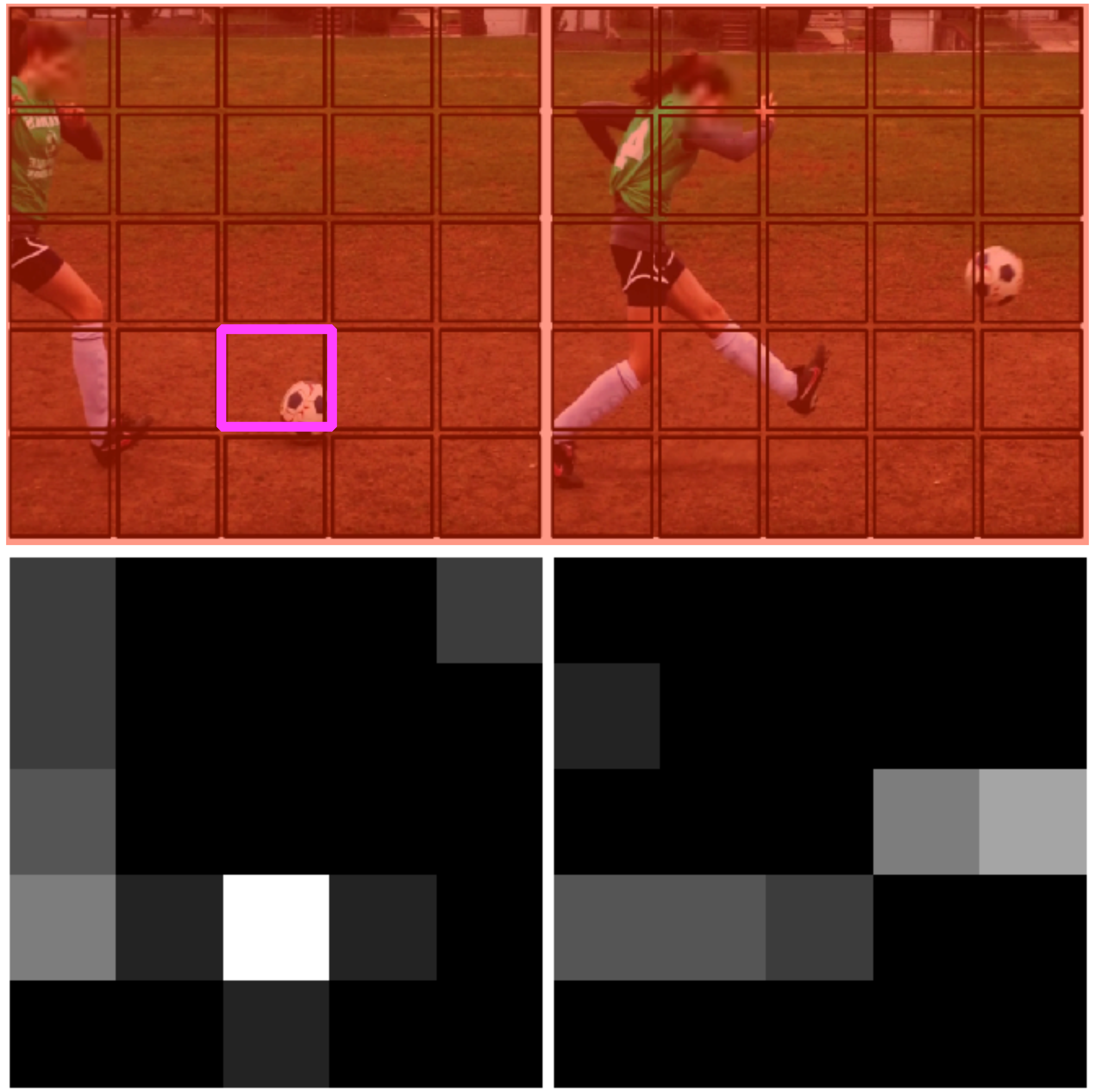
- In practice, the trajectory attention is implemented using standard joint space-time attention.

The reference patch

Trajectory Attention

- In practice, the trajectory attention is implemented using standard joint space-time attention.

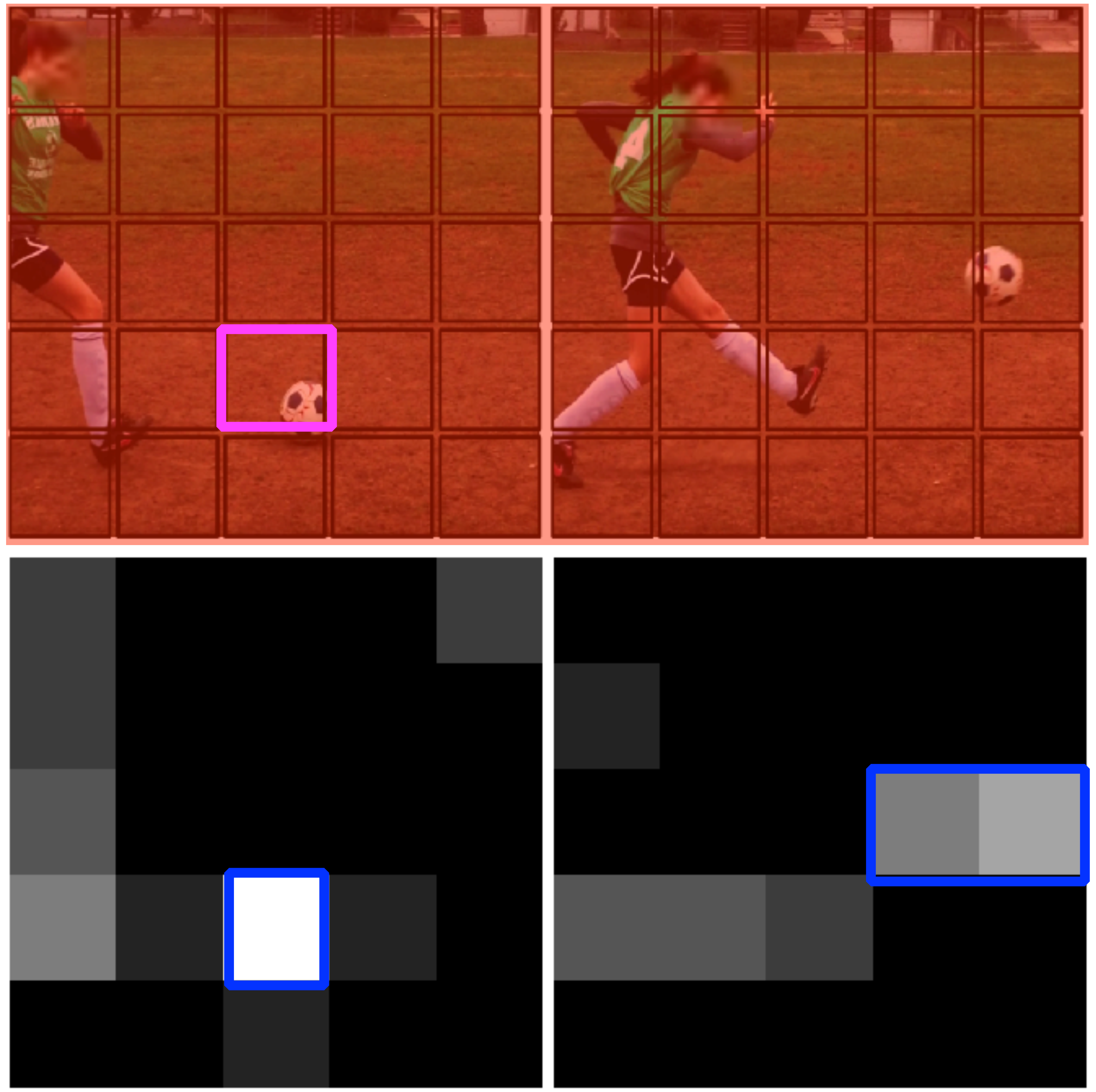
The reference patch



Trajectory Attention

- In practice, the trajectory attention is implemented using standard joint space-time attention.

The reference patch



Attention-driven discovery of the trajectory of the ball.

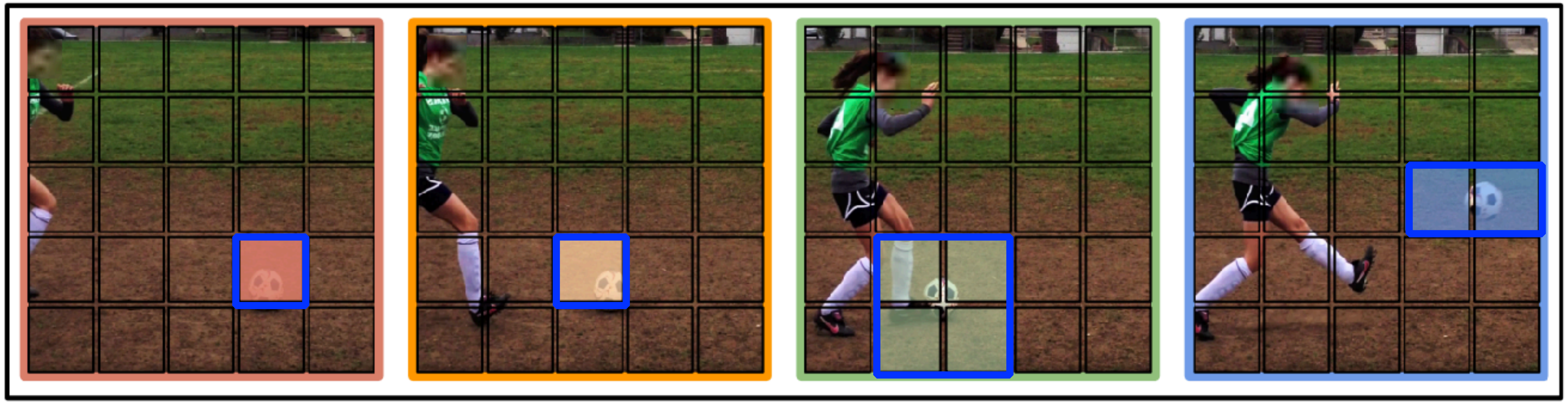
Trajectory Attention

- Instead, the authors propose trajectory attention, a mechanism for aggregating information along implicitly determined motion paths.

The representation for the reference patch …

Trajectory Attention

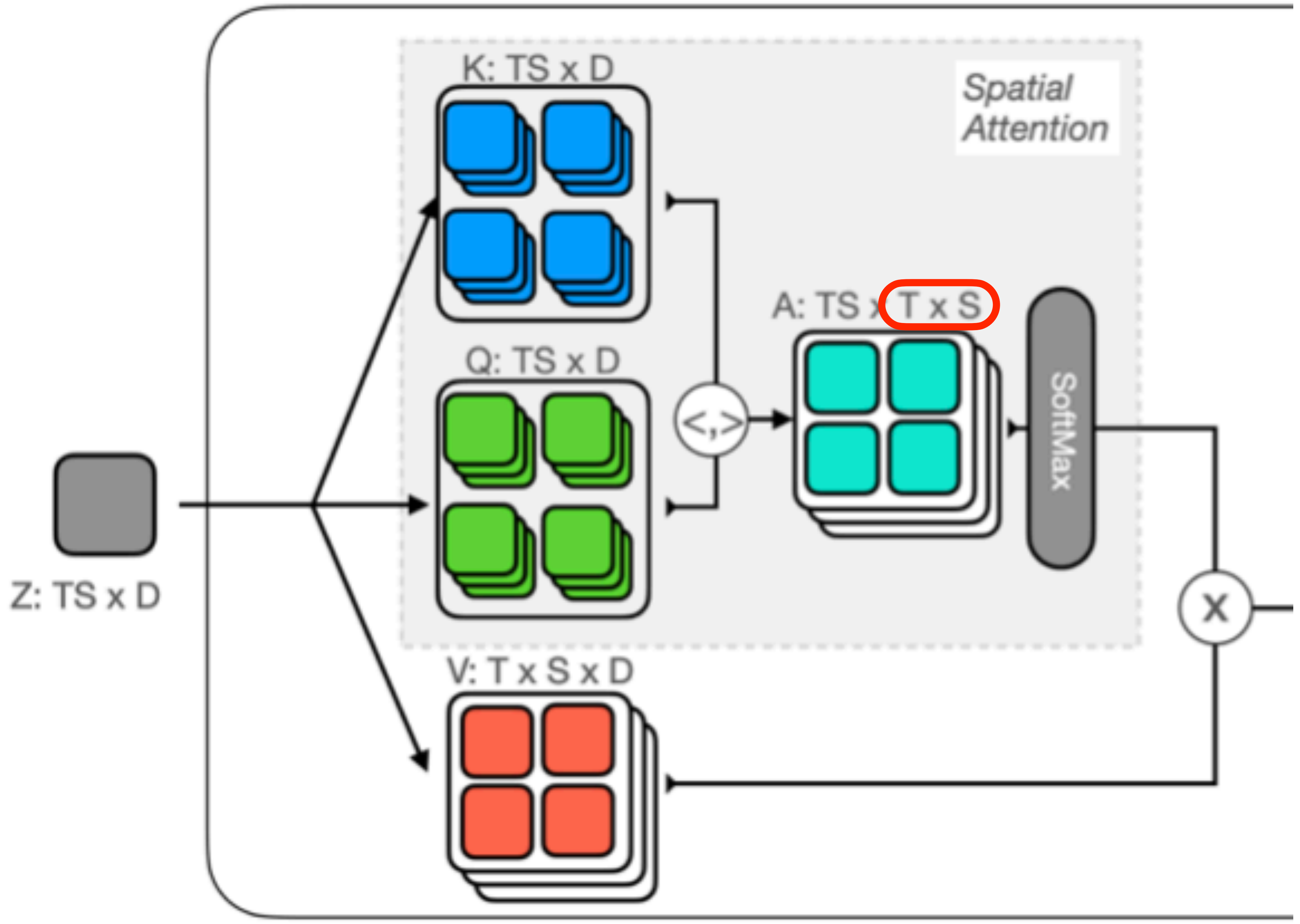
- Instead, the authors propose trajectory attention, a mechanism for aggregating information along implicitly determined motion paths.



… is computed as a weighted average
of patches along the trajectory.

Technical Approach

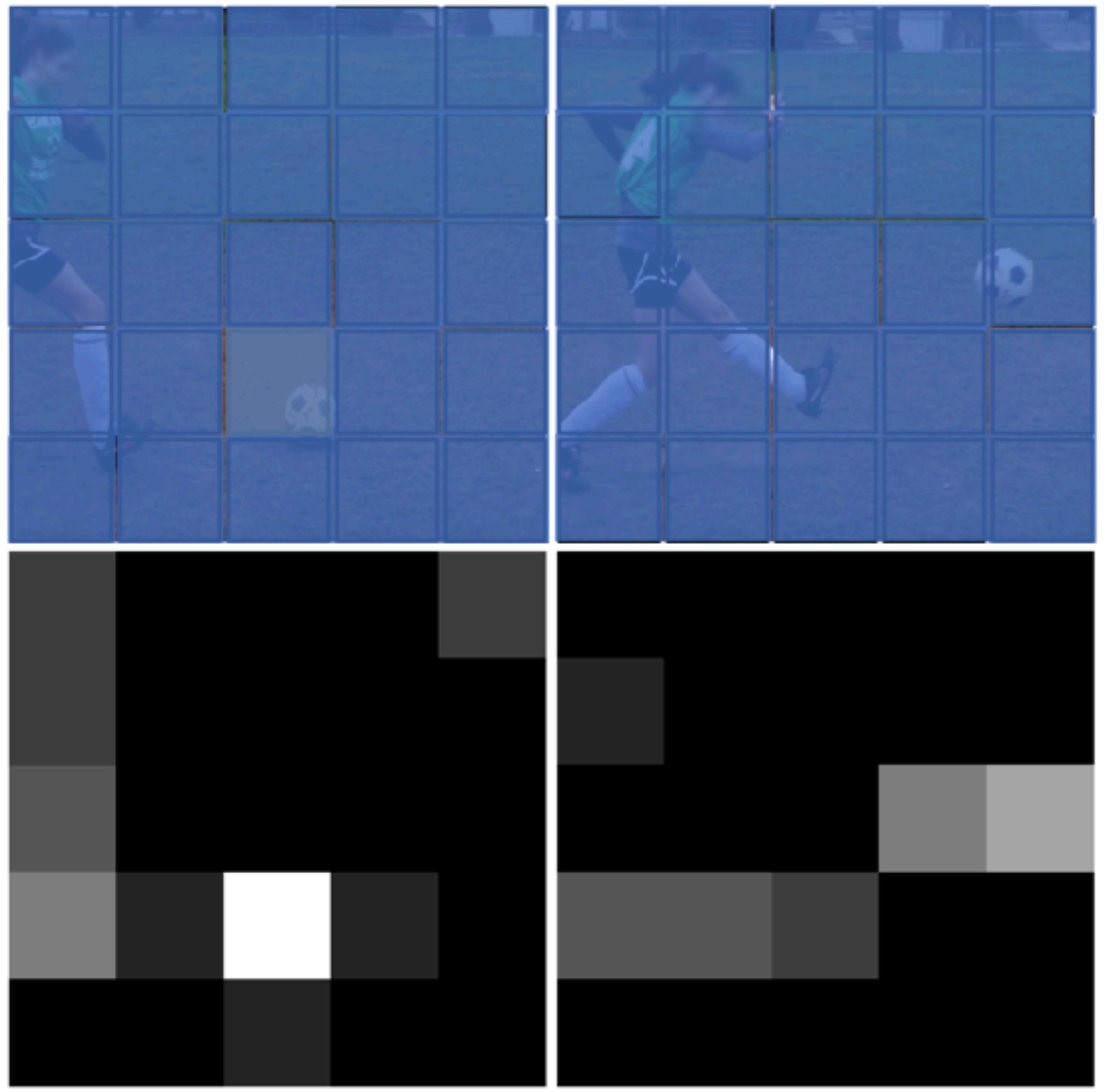
- Compared to prior approaches, the attention is performed along trajectories.



The only difference is in the normalization step.

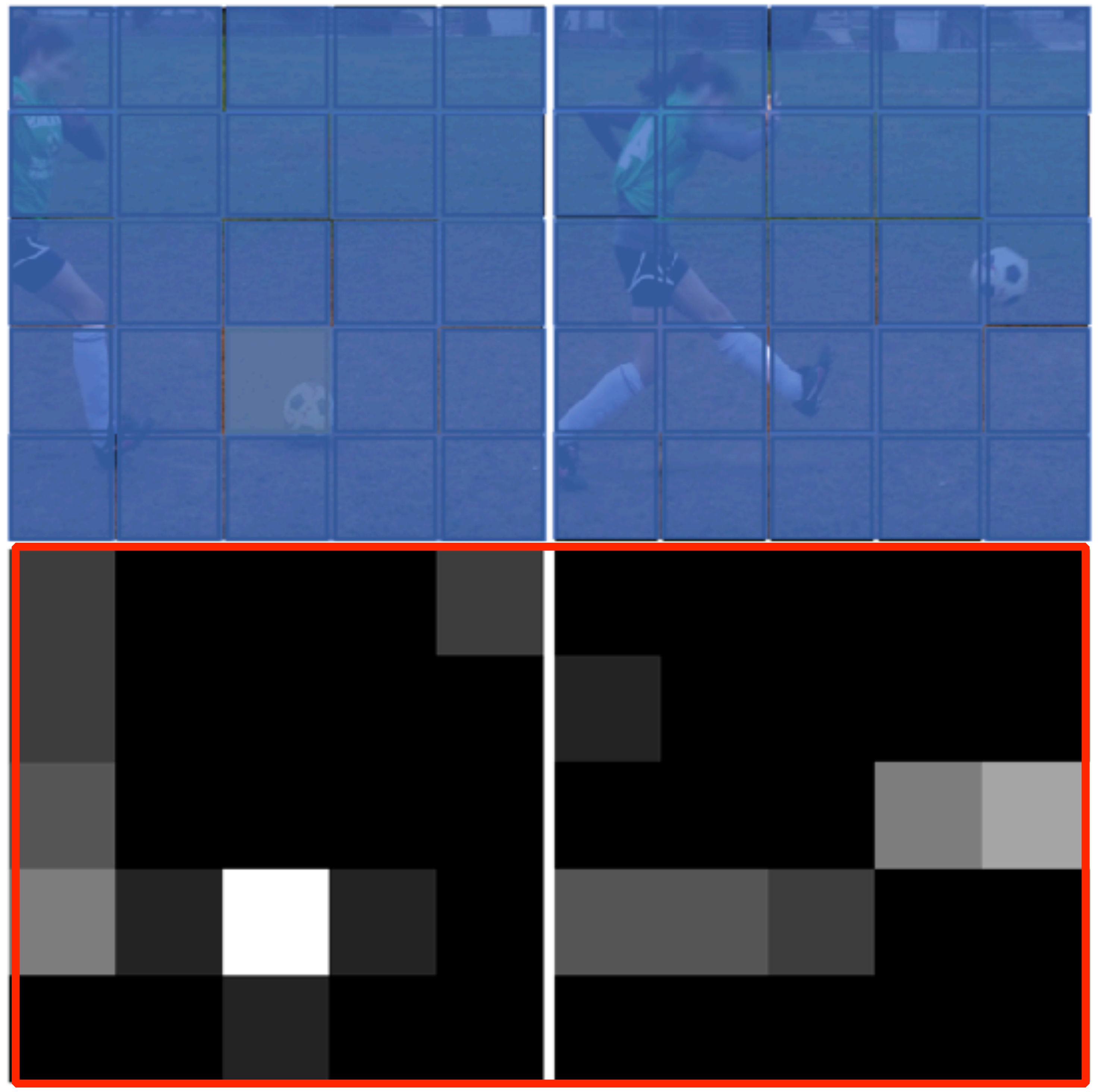
Technical Approach

- In standard spatiotemporal self-attention, the normalization is done across the entire spatiotemporal volume.



Technical Approach

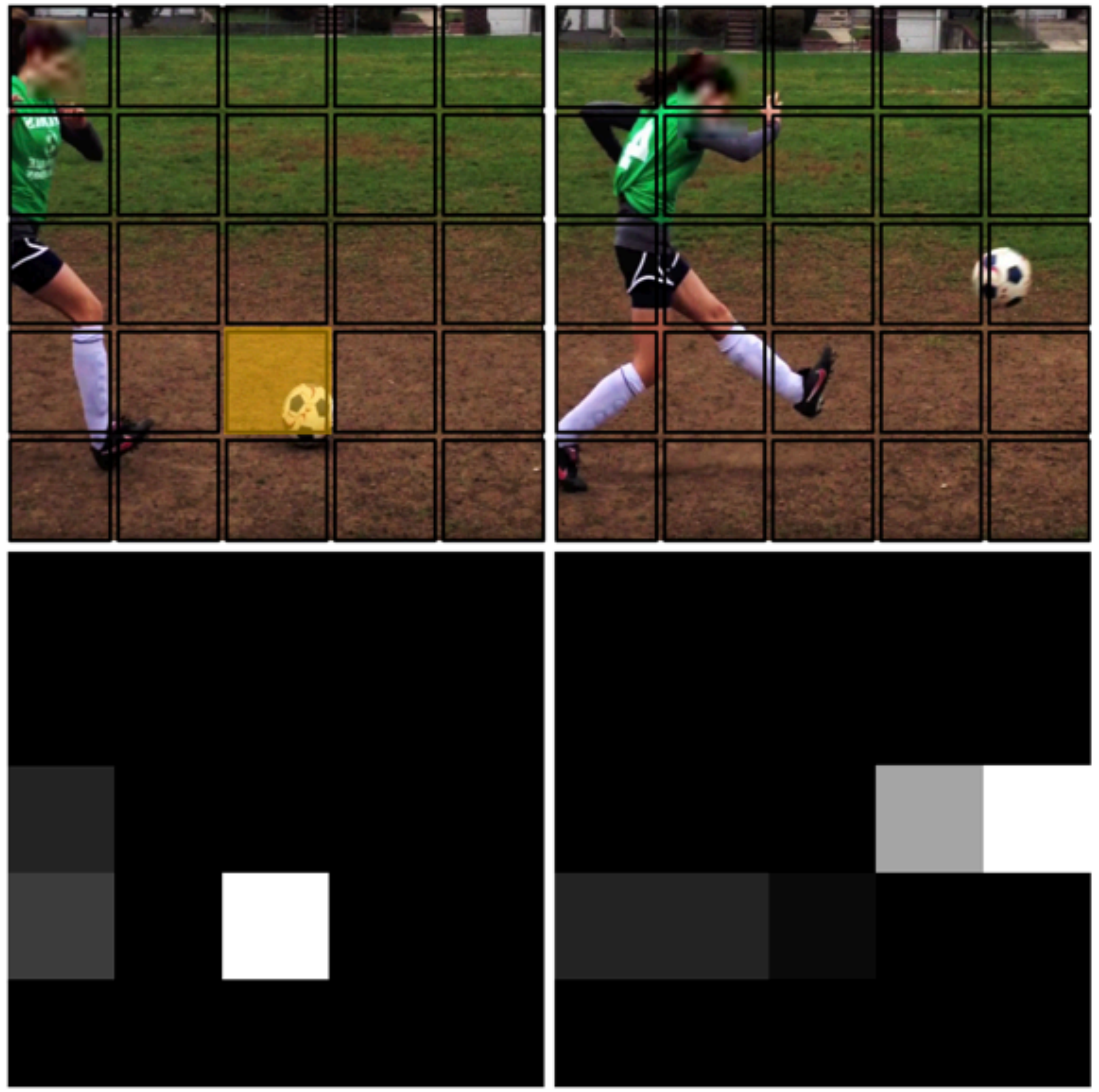
- In standard spatiotemporal self-attention, the normalization is done across the entire spatiotemporal volume.



All the attention values across the volume have to sum up to 1.

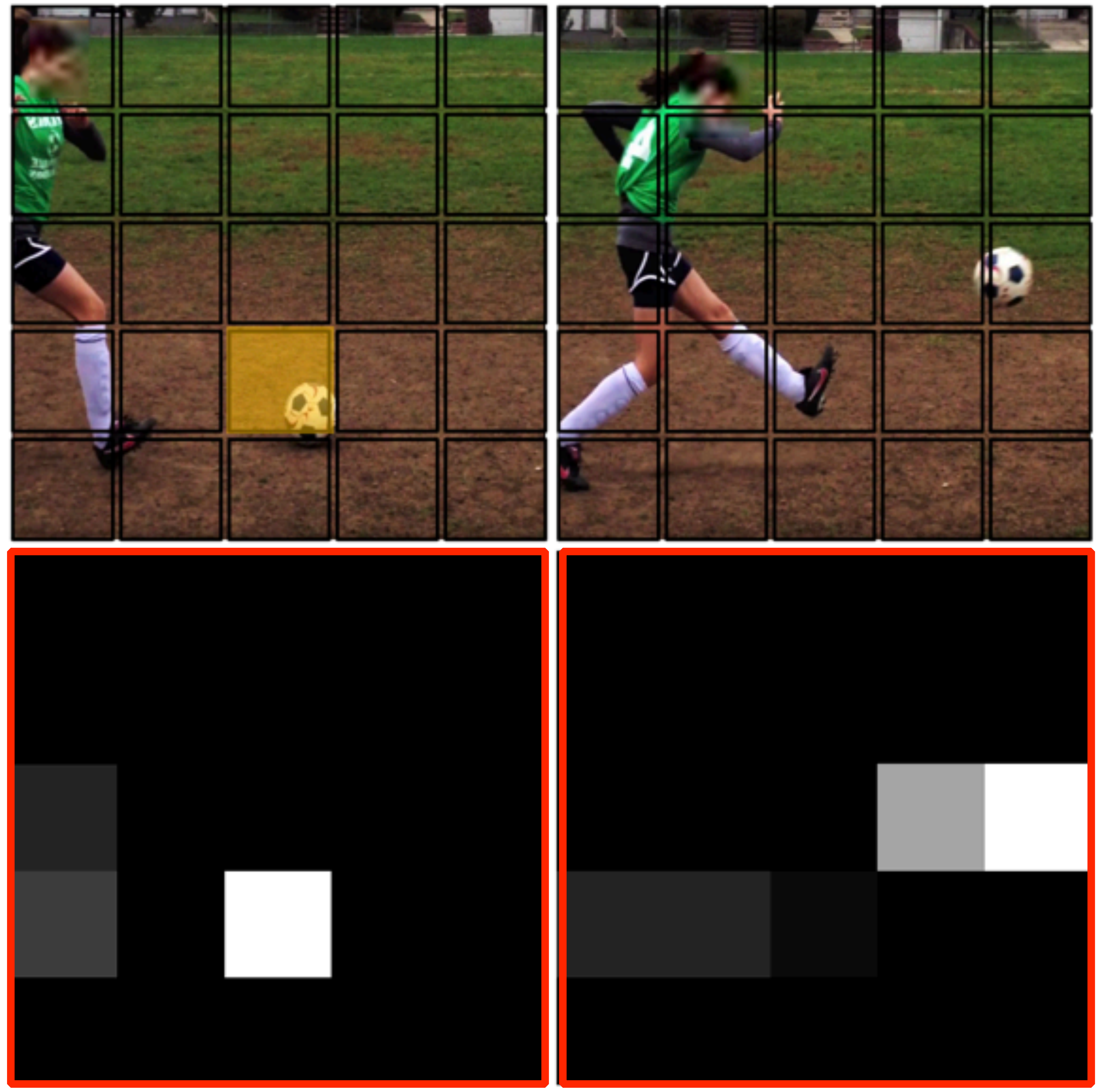
Technical Approach

- The authors propose to use per-frame softmax normalization.



Technical Approach

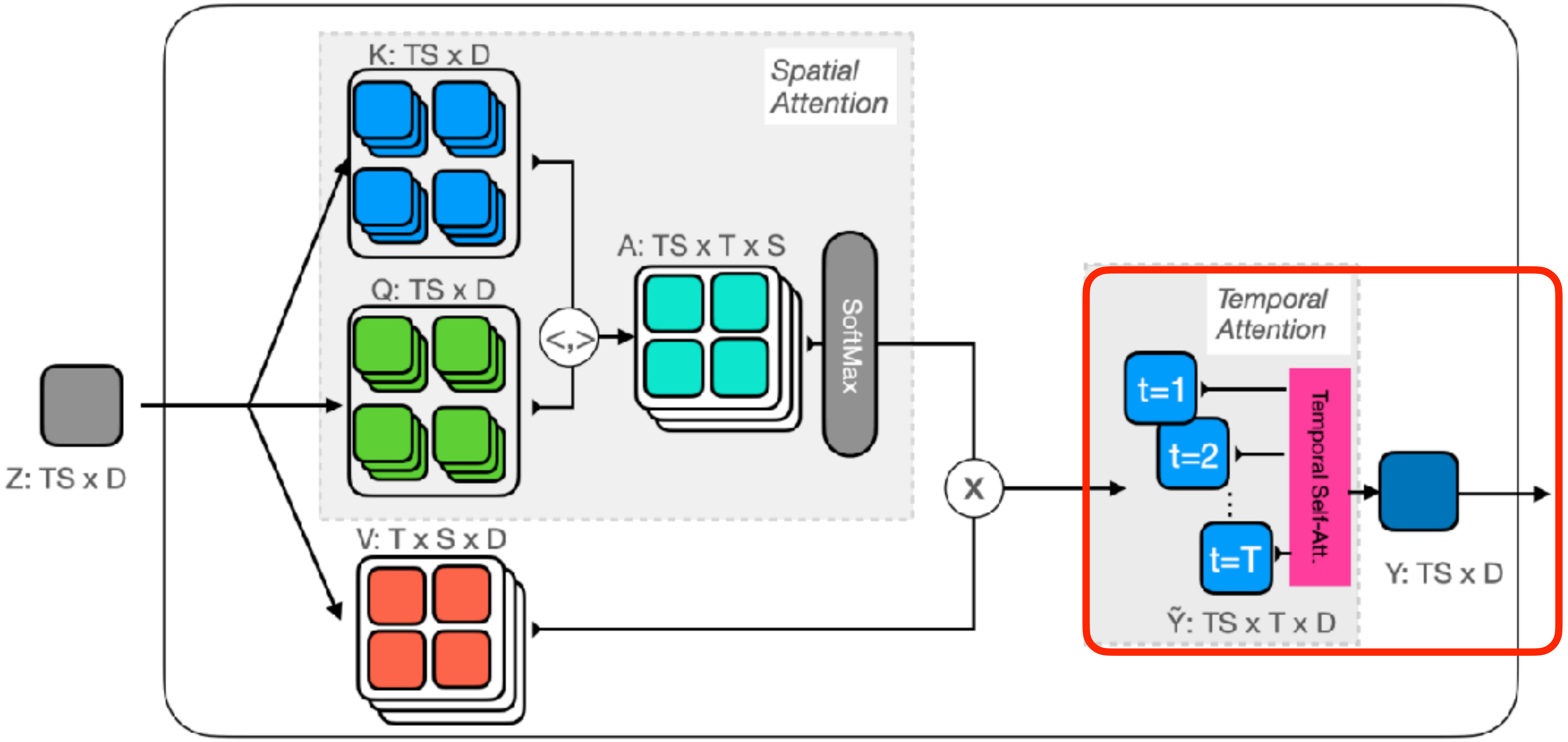
- The authors propose to use per-frame softmax normalization.



The values inside each frame have to sum up to 1.

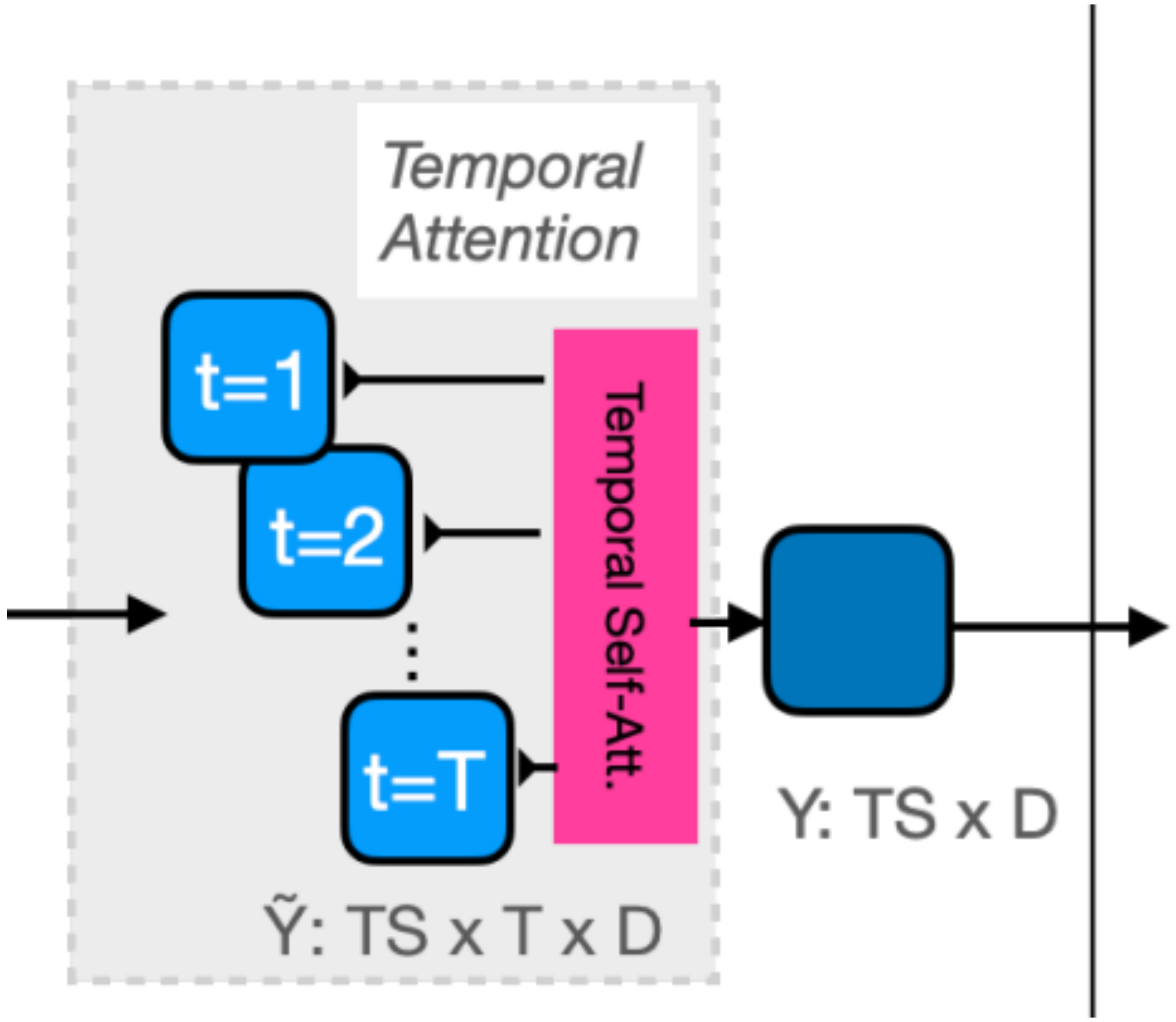
Technical Approach

- Once the trajectories are computed, the authors pool them across time to reason about intra-frame information/connections.



Technical Approach

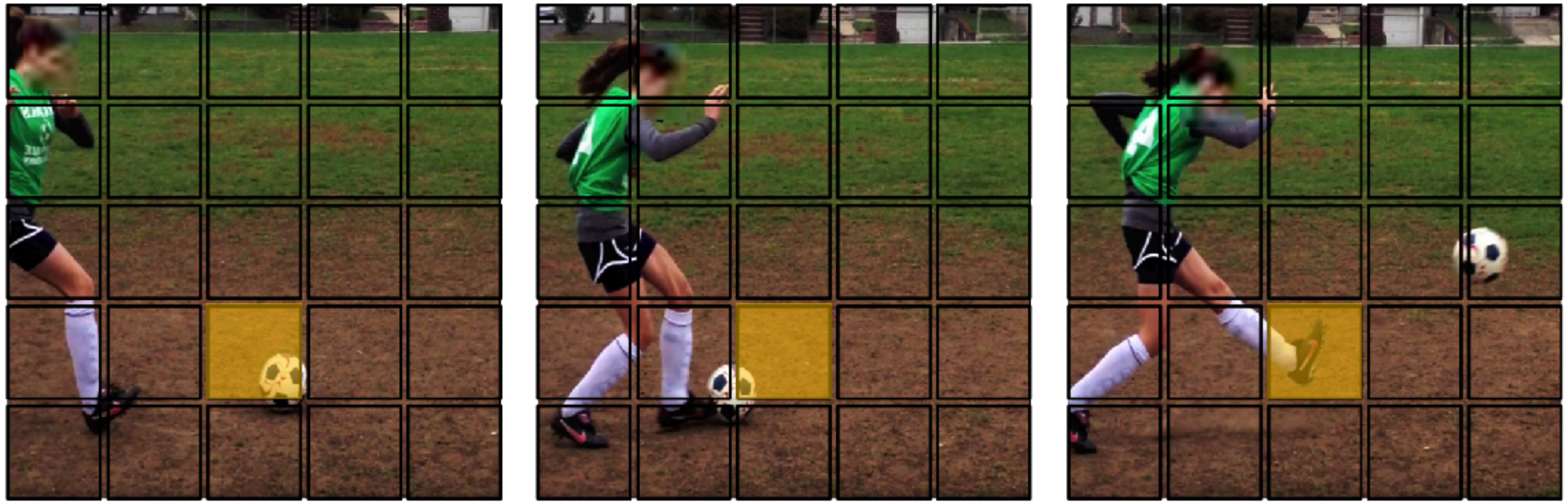
- Once the trajectories are computed, the authors pool them across time to reason about intra-frame information/connections.



Standard temporal attention from the divided space-time attention block.

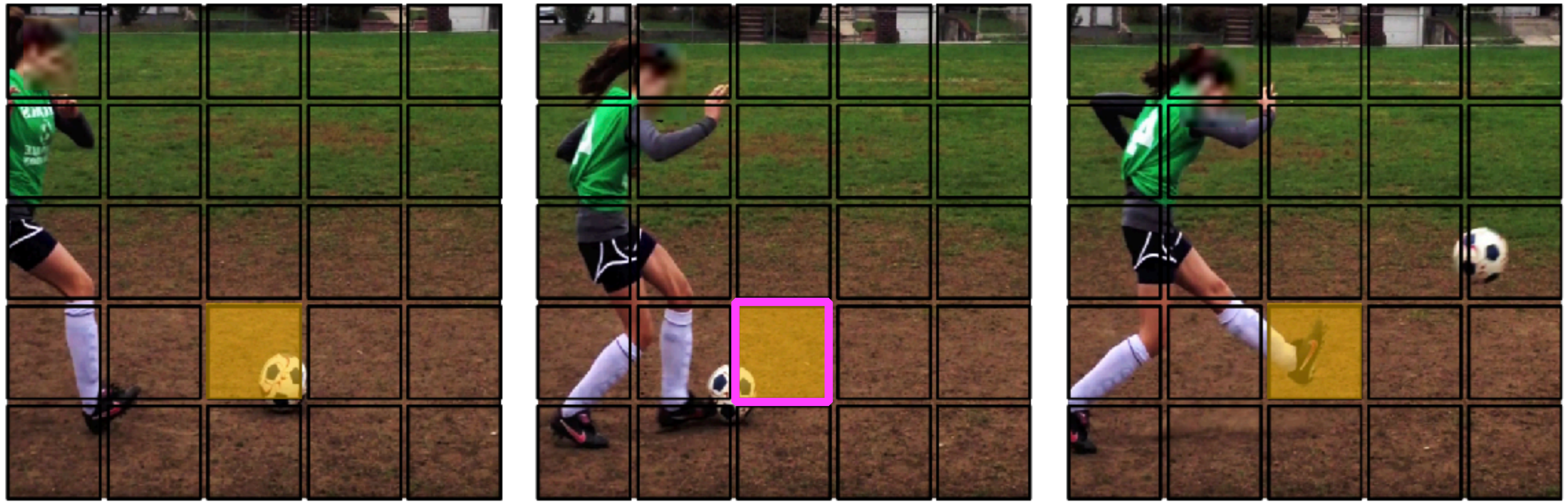
Temporal Attention

- For each query patch, the attention is applied at the same spatial location but across different frames.



Temporal Attention

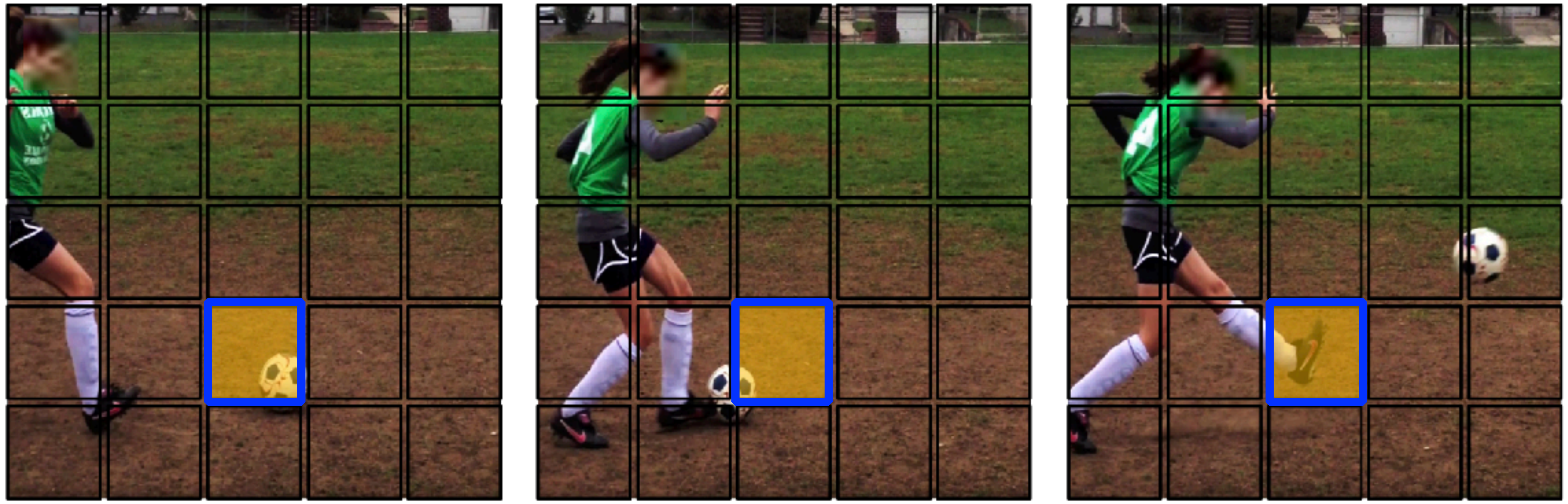
- For each query patch, the attention is applied at the same spatial location but across different frames.



The representation for the reference patch …

Temporal Attention

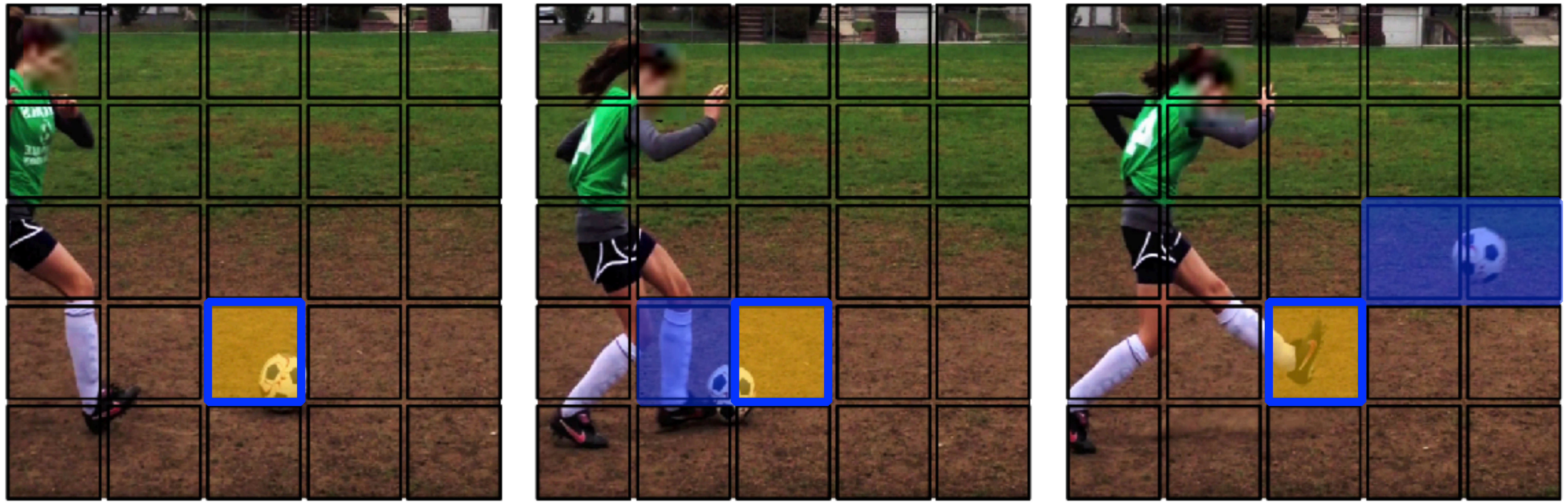
- For each query patch, the attention is applied at the same spatial location but across different frames.



… is computed as a weighted average of patches at the same spatial but different temporal locations

Temporal Attention

- For each query patch, the attention is applied at the same spatial location but across different frames.



Even though these patches might not be aligned, they might now incorporate relevant trajectory information (i.e., after the first attention step).

Joint Space-Time Attention

- Computing pairwise similarities between every single pair of patches in a video can be very computationally costly.
- Stacking two distinct attention layers on top of each other is even more costly.

Computational Cost

- The proposed trajectory attention is a lot more computationally expensive than (1) joint space-time attention, and (2) divided space-time attention.

Attention	GFLOPS
Joint Space-Time	180.6
Divided Space-Time	185.8
Trajectory	369.5

Computational Cost

- The proposed trajectory attention is a lot more computationally expensive than (1) joint space-time attention, and (2) divided space-time attention.

Attention	GFLOPS
Joint Space-Time	180.6
Divided Space-Time	185.8
Trajectory	369.5

Completely contrary to the motivation of the paper!

Approximating Attention

- The idea is similar to standard matrix factorization / low-rank decomposition methods.

Algorithm 1 Orthoformer (proposed) attention

1: $\mathbf{P} \leftarrow \text{MostOrthogonalSubset}(\mathbf{Q}, \mathbf{K}, R)$
2: $\mathbf{\Omega}_1 = \mathcal{S}(\mathbf{Q}^\top \mathbf{P} / \sqrt{D})$
3: $\mathbf{\Omega}_2 = \mathcal{S}(\mathbf{P}^\top \mathbf{K} / \sqrt{D})$
4: $\mathbf{Y} = \mathbf{\Omega}_1 (\mathbf{\Omega}_2 \mathbf{V})$

Approximating Attention

- The idea is similar to standard matrix factorization / low-rank decomposition methods.

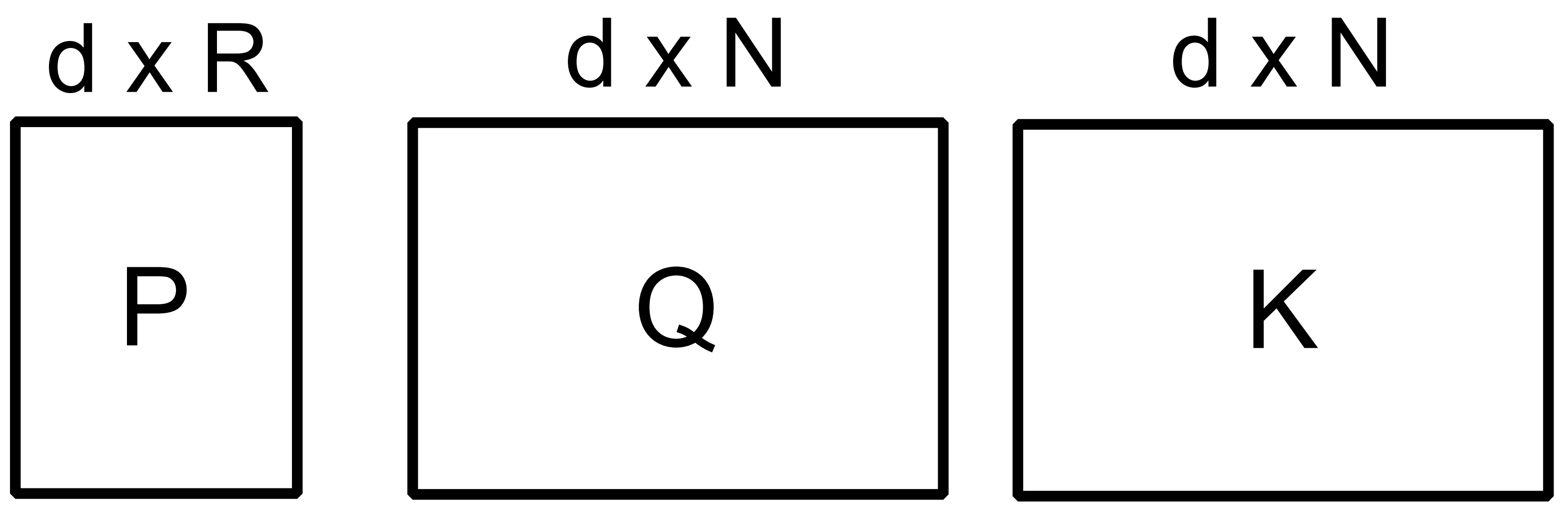
Algorithm 1 Orthoformer (proposed) attention

1: $\mathbf{P} \leftarrow \text{MostOrthogonalSubset}(\mathbf{Q}, \mathbf{K}, R)$

2: $\mathbf{\Omega}_1 = \mathcal{S}(\mathbf{Q}^\top \mathbf{P} / \sqrt{D})$

3: $\mathbf{\Omega}_2 = \mathcal{S}(\mathbf{P}^\top \mathbf{K} / \sqrt{D})$

4: $\mathbf{Y} = \mathbf{\Omega}_1 (\mathbf{\Omega}_2 \mathbf{V})$



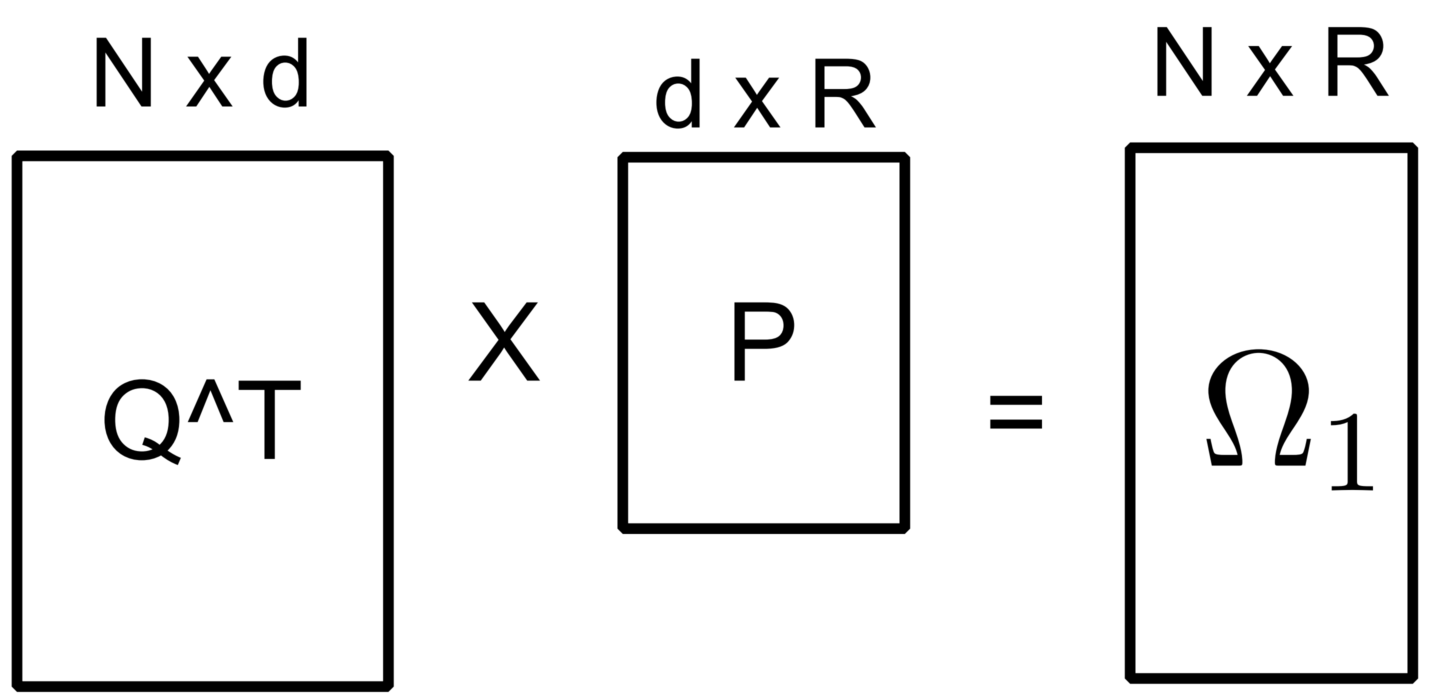
d - feature dimensionality
N - the number of original tokens
R - the number of prototypes (R << N)

Approximating Attention

- The idea is similar to standard matrix factorization / low-rank decomposition methods.

Algorithm 1 Orthoformer (proposed) attention

1: $\mathbf{P} \leftarrow \mathrm{MostOrthogonalSubset}(\mathbf{Q}, \mathbf{K}, R)$
2: $\mathbf{\Omega}_1 = \mathcal{S}(\mathbf{Q}^\top \mathbf{P}/\sqrt{D})$
3: $\mathbf{\Omega}_2 = \mathcal{S}(\mathbf{P}^\top \mathbf{K}/\sqrt{D})$
4: $\mathbf{Y} = \mathbf{\Omega}_1(\mathbf{\Omega}_2 \mathbf{V})$



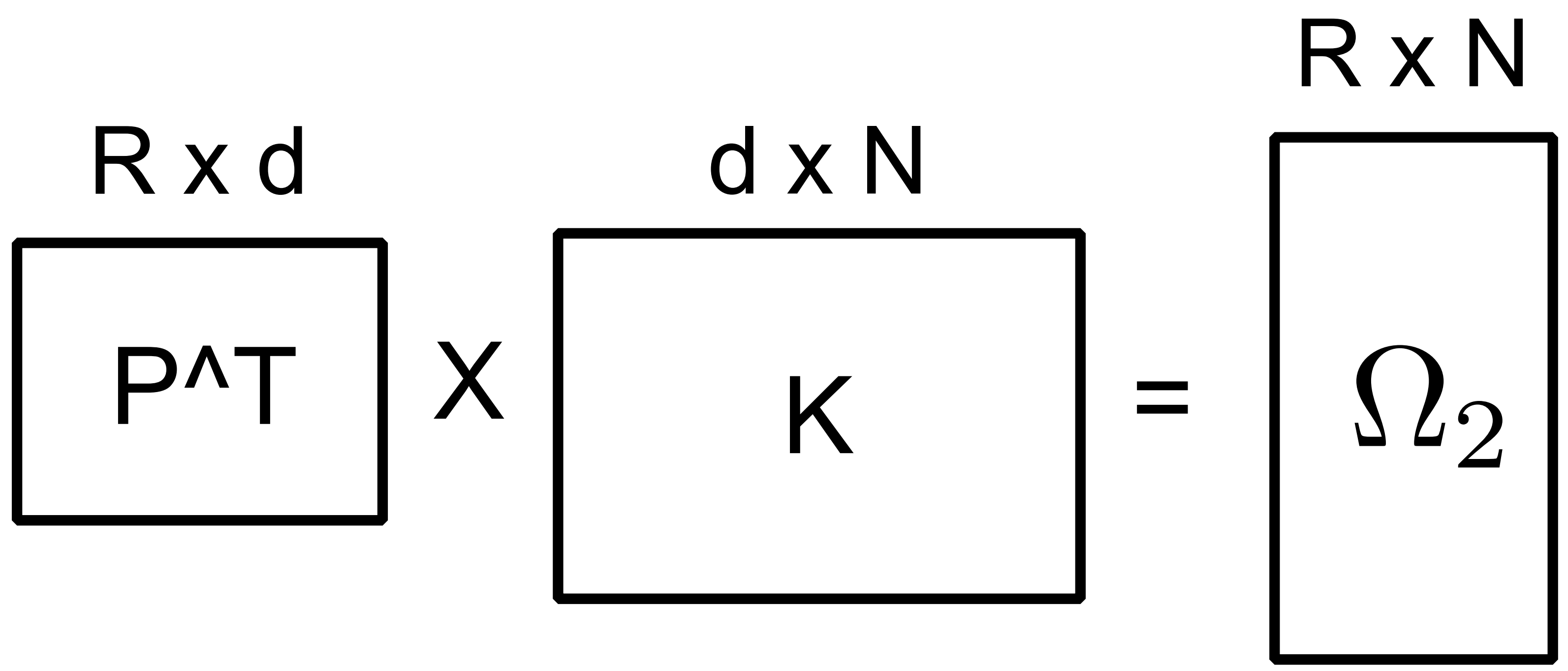
Projecting the queries to the query prototypes.

Approximating Attention

- The idea is similar to standard matrix factorization / low-rank decomposition methods.

Algorithm 1 Orthoformer (proposed) attention

1: $\mathbf{P} \leftarrow \text{MostOrthogonalSubset}(\mathbf{Q}, \mathbf{K}, R)$
2: $\mathbf{\Omega}_1 = \mathcal{S}(\mathbf{Q}^\top \mathbf{P}/\sqrt{D})$
3: $\mathbf{\Omega}_2 = \mathcal{S}(\mathbf{P}^\top \mathbf{K}/\sqrt{D})$
4: $\mathbf{Y} = \mathbf{\Omega}_1(\mathbf{\Omega}_2 \mathbf{V})$



Projecting the keys to the key prototypes.

Approximating Attention

- The idea is similar to standard matrix factorization / low-rank decomposition methods.

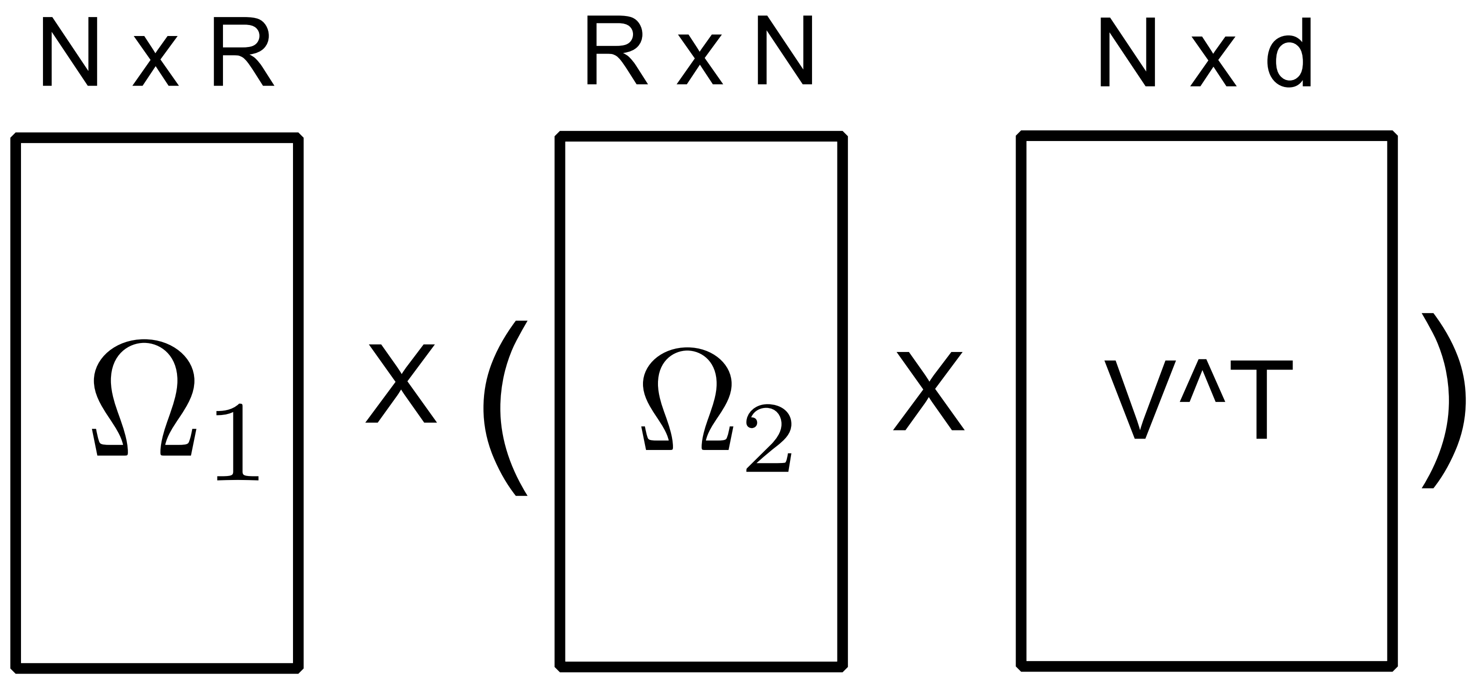
Algorithm 1 Orthoformer (proposed) attention

$\mathbf{P} \leftarrow \mathrm{MostOrthogonalSubset}(\mathbf{Q}, \mathbf{K}, R)$

$\mathbf{\Omega}_1 = \mathcal{S}(\mathbf{Q}^\top \mathbf{P} / \sqrt{D})$

$\mathbf{\Omega}_2 = \mathcal{S}(\mathbf{P}^\top \mathbf{K} / \sqrt{D})$

$\boxed{\mathbf{Y} = \mathbf{\Omega}_1(\mathbf{\Omega}_2 \mathbf{V})}$

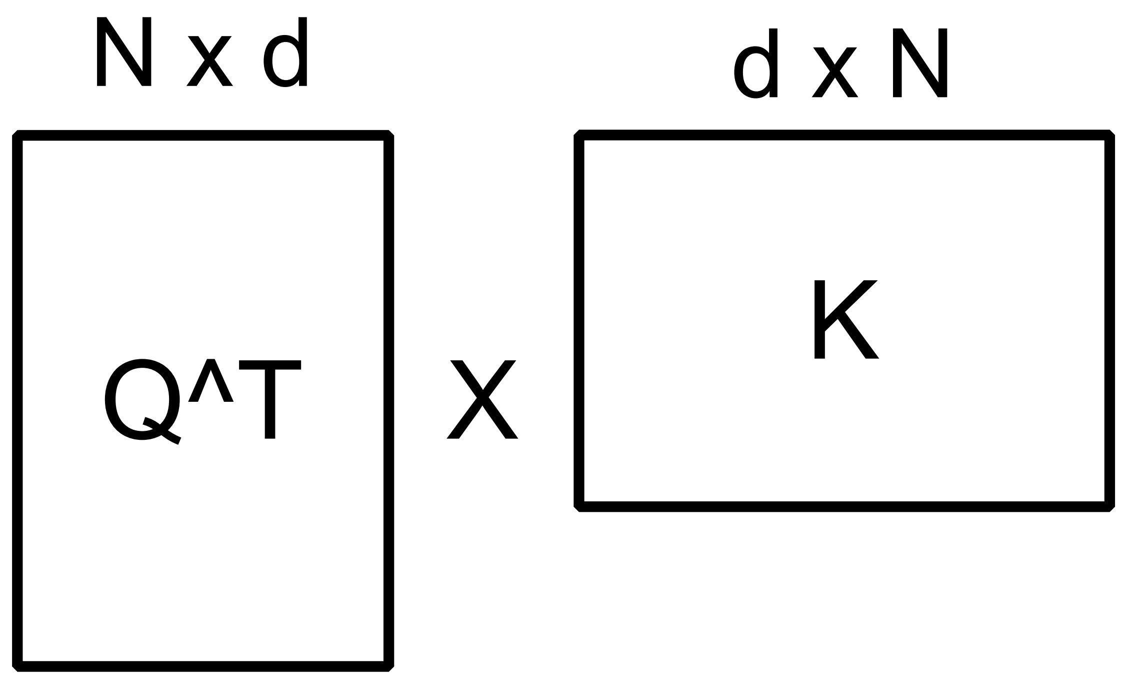


Approximating Attention

- The idea is similar to standard matrix factorization / low-rank decomposition methods.

Algorithm 1 Orthoformer (proposed) attention

1: $\mathbf{P} \leftarrow \text{MostOrthogonalSubset}(\mathbf{Q}, \mathbf{K}, R)$
2: $\mathbf{\Omega}_1 = \mathcal{S}(\mathbf{Q}^\top \mathbf{P}/\sqrt{D})$
3: $\mathbf{\Omega}_2 = \mathcal{S}(\mathbf{P}^\top \mathbf{K}/\sqrt{D})$
4: $\mathbf{Y} = \mathbf{\Omega}_1(\mathbf{\Omega}_2 \mathbf{V})$



Such approximation eliminates the need to do N^2 comparisons where N=ST can be a very a large number.

Approximation Ablations

- The results are evaluated on the Kinetics-400 and Something-Something-V2 action recognition datasets (using top-1 accuracy).

Attention	Approx.	Mem.	K-400	SSv2
Trajectory (E)	N/A	7.4	**79.7**	**66.5**
Trajectory (A)	Performer	5.1	72.9	52.7
	Nyströmformer	3.8	**77.5**	**64.0**
	Orthoformer	3.6	**77.5**	63.8

Approximation Ablations

- The results are evaluated on the Kinetics-400 and Something-Something-V2 action recognition datasets (using top-1 accuracy).

Attention	Approx.	Mem.	K-400	SSv2
Trajectory (E)	N/A	7.4	**79.7**	**66.5**
Trajectory (A)	Performer	5.1	72.9	52.7
	Nyströmformer	3.8	**77.5**	**64.0**
	Orthoformer	3.6	**77.5**	63.8

Substantial drop in performance.

Approximation Ablations

- The results are evaluated on the Kinetics-400 and Something-Something-V2 action recognition datasets (using top-1 accuracy).

Attention	Approx.	Mem.	K-400	SSv2
Trajectory (E)	N/A	7.4	**79.7**	**66.5**
Trajectory (A)	Performer	5.1	72.9	52.7
	Nyströmformer	3.8	**77.5**	**64.0**
	Orthoformer	3.6	**77.5**	63.8

Where are the computational cost metrics (i.e., GFLOPS) ???

Comparison to Prior Work

- The results are evaluated on Kinetics-400 (using top-1 accuracy).

(b) **Kinetics-400**

Method	Pretrain	Top-1	Top-5	GFLOPs×views
I3D [12]	IN-1K	72.1	89.3	108×N/A
R(2+1)D [82]	-	72.0	90.0	152×5×23
S3D-G [94]	IN-1K	74.7	93.4	142.8×N/A
X3D-XL [26]	-	79.1	93.9	48.4×3×10
SlowFast [27]	-	79.8	93.9	234×3×10
VTN [56]	IN-21K	78.6	93.7	4218×1×1
VidTr-L [49]	IN-21K	79.1	93.9	392×3×10
Tformer-L[8]	IN-21K	80.7	94.7	2380×3×1
MViT-B [24]	-	81.2	95.1	455×3×3
ViViT-L [3]	IN-21K	**81.3**	94.7	3992×3×4
Mformer	IN-21K	79.7	94.2	369.5×3×10
Mformer-L	IN-21K	80.2	94.8	1185.1×3×10
Mformer-HR	IN-21K	81.1	**95.2**	958.8×3×10

Comparison to Prior Work

- The results are evaluated on Kinetics-400 (using top-1 accuracy).

(b) **Kinetics-400**

Method	Pretrain	Top-1	Top-5	GFLOPs×views
I3D [12]	IN-1K	72.1	89.3	108×N/A
R(2+1)D [82]	-	72.0	90.0	152×5×23
S3D-G [94]	IN-1K	74.7	93.4	142.8×N/A
X3D-XL [26]	-	79.1	93.9	48.4×3×10
SlowFast [27]	-	79.8	93.9	234×3×10
VTN [56]	IN-21K	78.6	93.7	4218×1×1
VidTr-L [49]	IN-21K	79.1	93.9	392×3×10
Tformer-L[8]	IN-21K	80.7	94.7	2380×3×1
MViT-B [24]	-	81.2	95.1	455×3×3
ViViT-L [3]	IN-21K	**81.3**	94.7	3992×3×4
Mformer	IN-21K	79.7	94.2	369.5×3×10
Mformer-L	IN-21K	80.2	94.8	1185.1×3×10
Mformer-HR	IN-21K	81.1	**95.2**	958.8×3×10

Good results but the computational cost is huge

Comparison to Prior Work

- The results are evaluated on SSv2 (using top-1 accuracy).

(a) **Something–Something V2**

Model	Pretrain	Top-1	Top-5	GFLOPs ×views
SlowFast [27]	K-400	61.7	-	65.7×3×1
TSM [51]	K-400	63.4	88.5	62.4×3×2
STM [36]	IN-1K	64.2	89.8	66.5×3×10
MSNet [44]	IN-1K	64.7	89.4	67×1×1
TEA [50]	IN-1K	65.1	-	70×3×10
bLVNet [25]	IN-1K	65.2	90.3	128.6×3×10
VidTr-L [49]	IN-21K+K-400	60.2	-	351×3×10
Tformer-L [8]	IN-21K	62.5	-	1703×3×1
ViViT-L [3]	IN-21K+K-400	65.4	89.8	3992×4×3
MViT-B [24]	K-400	67.1	90.8	170×3×1
Mformer	IN-21K+K-400	66.5	90.1	369.5×3×1
Mformer-L	IN-21K+K-400	**68.1**	**91.2**	1185.1×3×1
Mformer-HR	IN-21K+K-400	67.1	90.6	958.8×3×1

Comparison to Prior Work

- The results are evaluated on SSv2 (using top-1 accuracy).

(a) **Something–Something V2**

Model	Pretrain	Top-1	Top-5	GFLOPs ×views
SlowFast [27]	K-400	61.7	-	65.7×3×1
TSM [51]	K-400	63.4	88.5	62.4×3×2
STM [36]	IN-1K	64.2	89.8	66.5×3×10
MSNet [44]	IN-1K	64.7	89.4	67×1×1
TEA [50]	IN-1K	65.1	-	70×3×10
bLVNet [25]	IN-1K	65.2	90.3	128.6×3×10
VidTr-L [49]	IN-21K+K-400	60.2	-	351×3×10
Tformer-L [8]	IN-21K	62.5	-	1703×3×1
ViViT-L [3]	IN-21K+K-400	65.4	89.8	3992×4×3
MViT-B [24]	K-400	67.1	90.8	170×3×1
Mformer	IN-21K+K-400	66.5	90.1	369.5×3×1
Mformer-L	IN-21K+K-400	**68.1**	**91.2**	1185.1×3×1
Mformer-HR	IN-21K+K-400	67.1	90.6	958.8×3×1

Strong quantitative results

Ablation on Normalization

- Comparisons between spatial and spatiotemporal normalization schemes.

Norm_S	Norm_{ST}	GFLOPS	K-400	SSv2
✗	✓	369.5	77.2	60.9
✓	✗	369.5	**79.7**	**66.5**

Ablation on Normalization

- Comparisons between spatial and spatiotemporal normalization schemes.

Norm_S	Norm_{ST}	GFLOPS	K-400	SSv2
✗	✓	369.5	77.2	60.9
✓	✗	369.5	**79.7**	**66.5**

Spatial attention normalization works much better than spatiotemporal normalization.

Discussion Questions

- Why is the attention approximation scheme introduced but not used in the final variants of the approach?

Discussion Questions

- Why is the attention approximation scheme introduced but not used in the final variants of the approach?
- Why is the computational cost of the approximated model never reported in GFLOPs?

Discussion Questions

- Why is the attention approximation scheme introduced but not used in the final variants of the approach?
- Why is the computational cost of the approximated model never reported in GFLOPs?
- Based on the experiments, can we confidently say that the trajectory attention is the way to go?

Discussion Questions

- Why is the attention approximation scheme introduced but not used in the final variants of the approach?
- Why is the computational cost of the approximated model never reported in GFLOPs?
- Based on the experiments, can we confidently say that the trajectory attention is the way to go?
- Why is the spatial normalization scheme so much more effective than the spatiotemporal one?

Summary

- Chaotic and poorly crafted paper.
- The proposed approach leads to better results but at a large computational cost.
- Technical contributions are somewhat incremental (i.e., the trajectory attention combines two standard attention schemes & changes normalization).
- Quantitative improvements on temporally-heavy datasets (i.e., SSv2) are impressive.