UNC | COLLEGE OF ARTS AND SCIENCES
Computer Science

Understanding Complex Human Activities in Long Videos

Gedas Bertasius

EPIC @ CVPR’22

06-20-2022

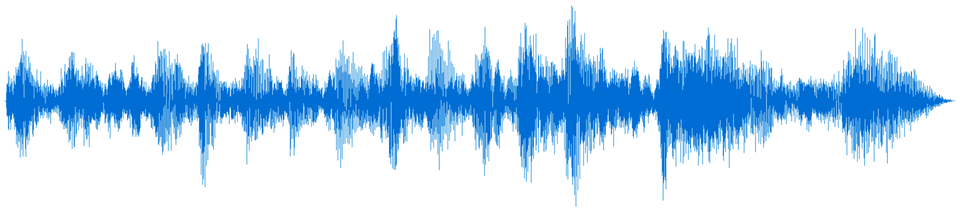
Motivation

- The ultimate goal of my research is to build computer vision models for understanding human behavior.

Motivation

- To develop a personalized AI assistant, we need systems that can reason about (1) multiple modalities (2) over long temporal extent.

i) Multimodal Perception

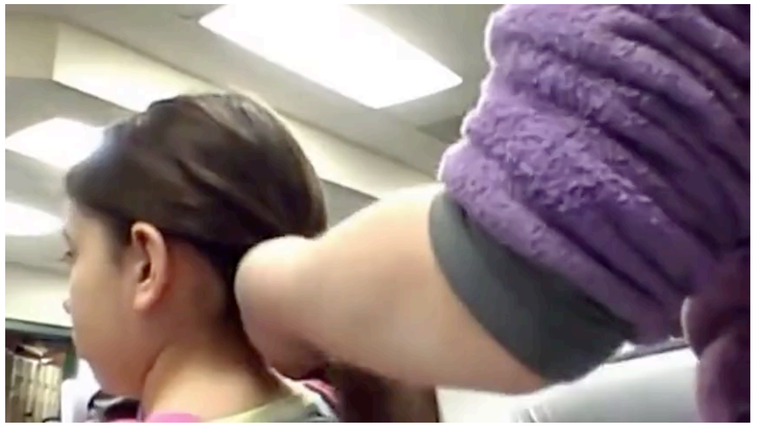


ii) Long-range Video Understanding

Motivation

- Instead, most modern video models operate on manually trimmed videos that span only a few seconds.

Cartwheeling



Braiding Hair

Learning To Recognize Procedural Activities with Distant Supervision

CVPR 2022


Xudong Lin


Fabio Petroni


Gedas Bertasius


Marcus Rohrbach


Shih-Fu Chang

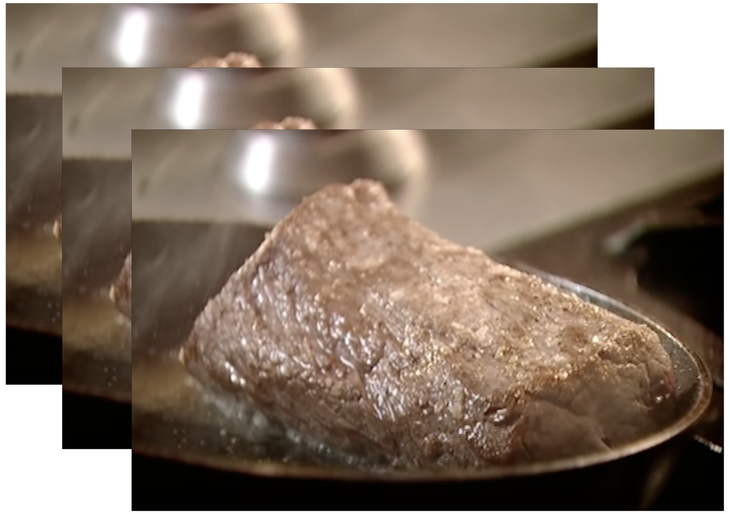

Lorenzo Torresani

Motivation

- Imagine that you are preparing a complex meal (e.g., Beef Wellington).

Motivation

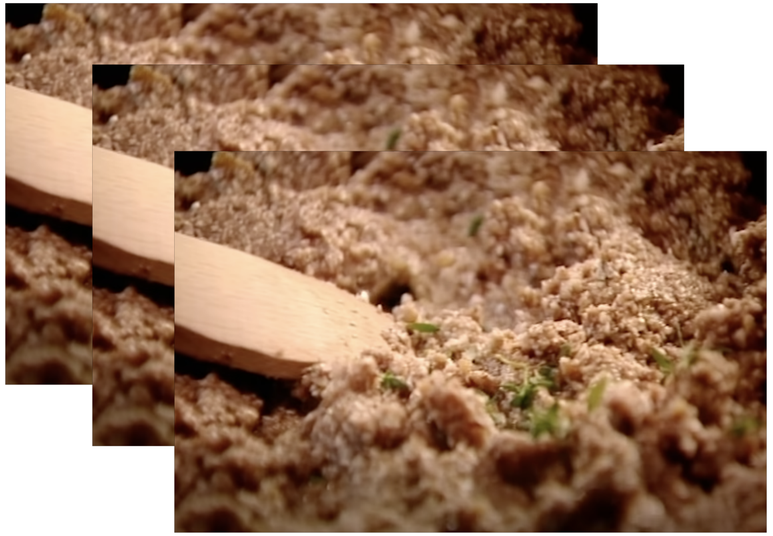
- Imagine that you are preparing a complex meal (e.g., Beef Wellington).



Step 1: Searing the fillet mignon.

Motivation

- Imagine that you are preparing a complex meal (e.g., Beef Wellington).

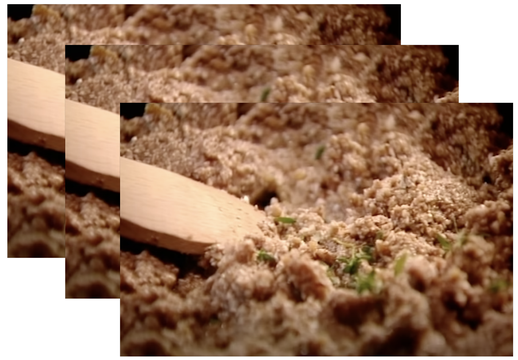


Step 2: Frying the mushrooms until they are dried.

Motivation

- Imagine that you are preparing a complex meal (e.g., Beef Wellington).

Step 1



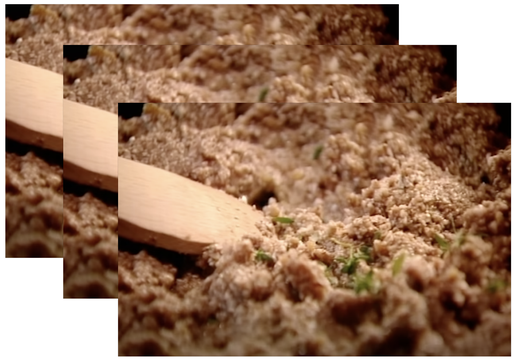
Step 2

Umm, what is the next step?

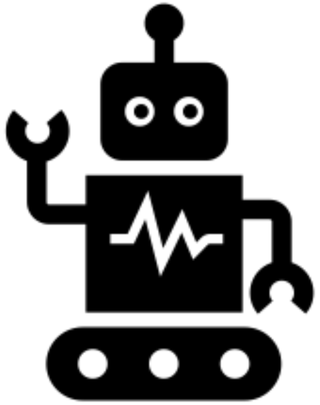
Motivation

- Imagine that you are preparing a complex meal (e.g., Beef Wellington).

Step 1



Step 2

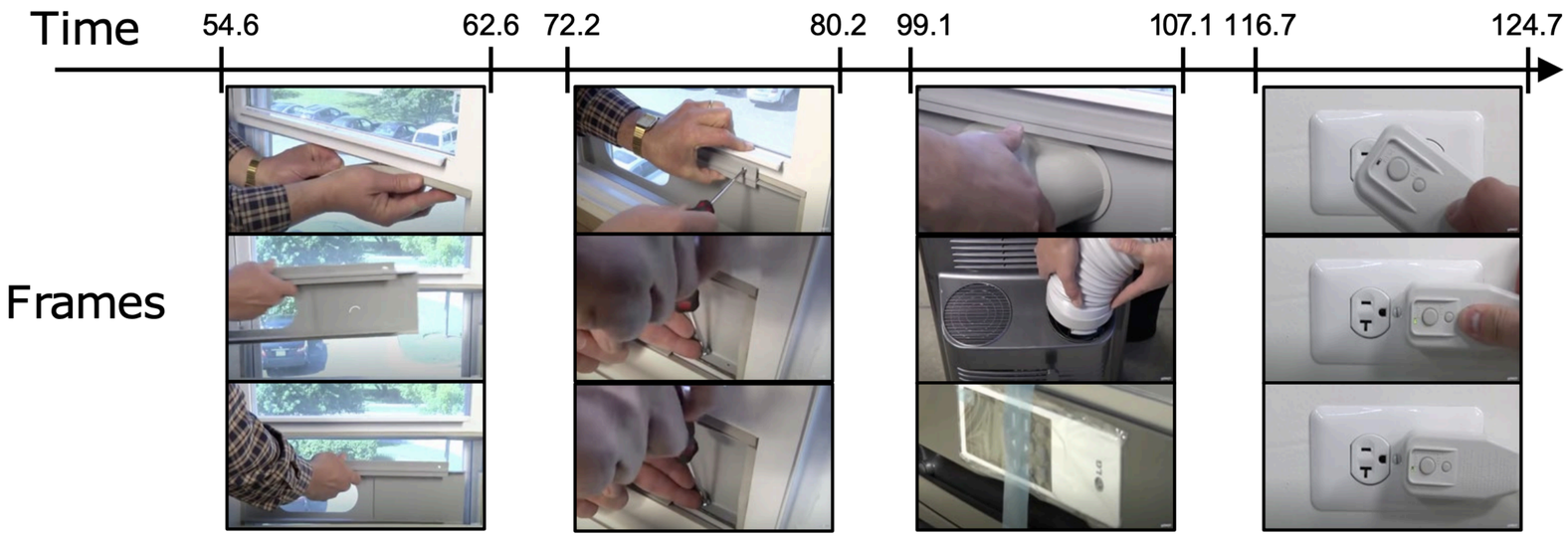


I see you have done:
Step 1: Searing the fillet mignon to brown.
Step 2: Frying the mushrooms until they are dried.

You are preparing Beef Wellington! The next step is:
Step 3: Shingle the prosciutto on the plastic wrap and spread the mushrooms over the prosciutto.

Motivation

- Manually annotating individual steps in a long procedural video is costly and not scalable.



Prior Work

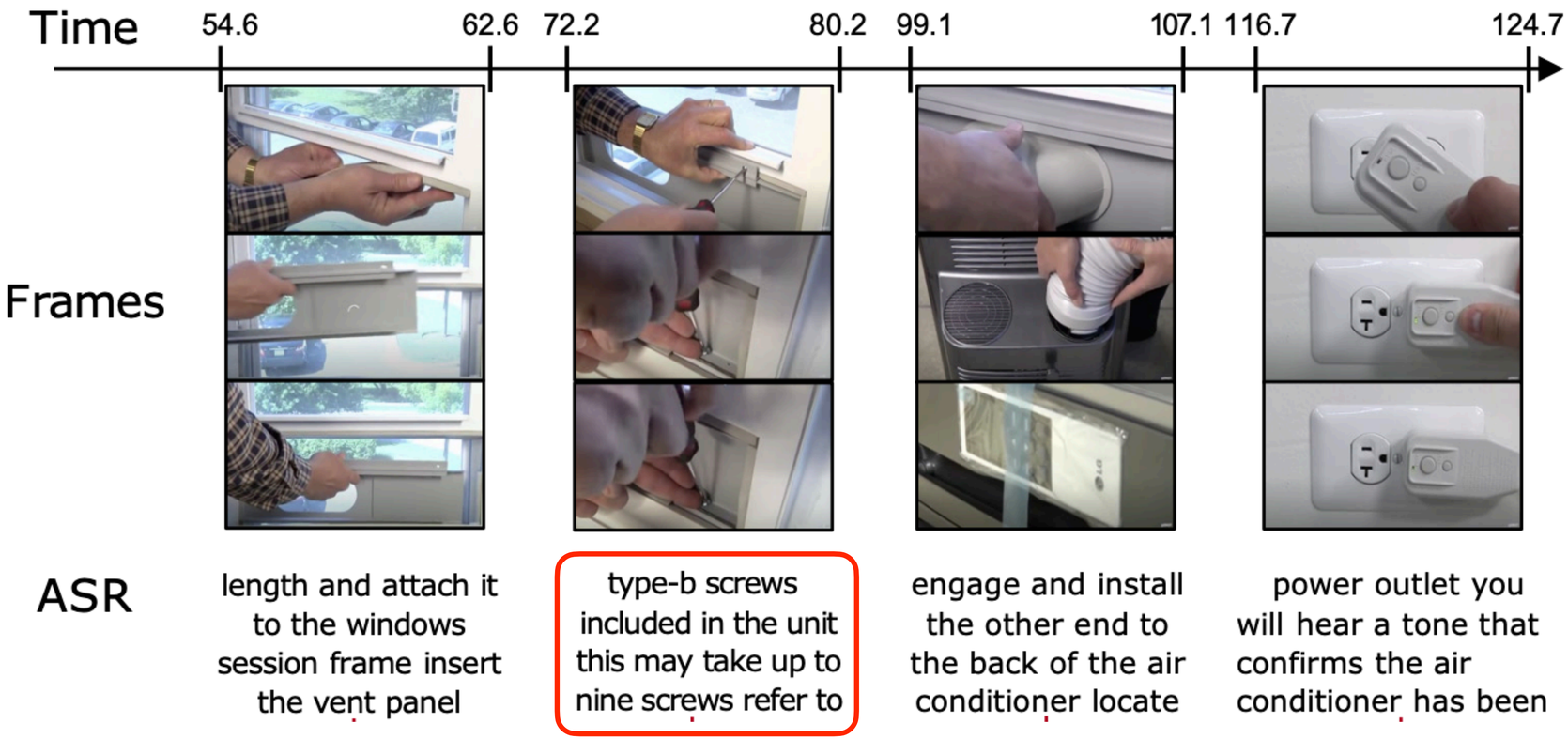
- Learning a video representation from the automatically transcribed speech narrations in instructional videos.

[1] Miech et al. “HowTo100M: Learning a Text-Video Embedding by Watching Hundred Million Narrated Video Clips,” ICCV 2019

[2] Miech et al. “End-to-End Learning of Visual Representations from Uncurated Instructional Videos,” CVPR 2020

Noise in ASR

- Automatically transcribed speech narrations can be very noisy.

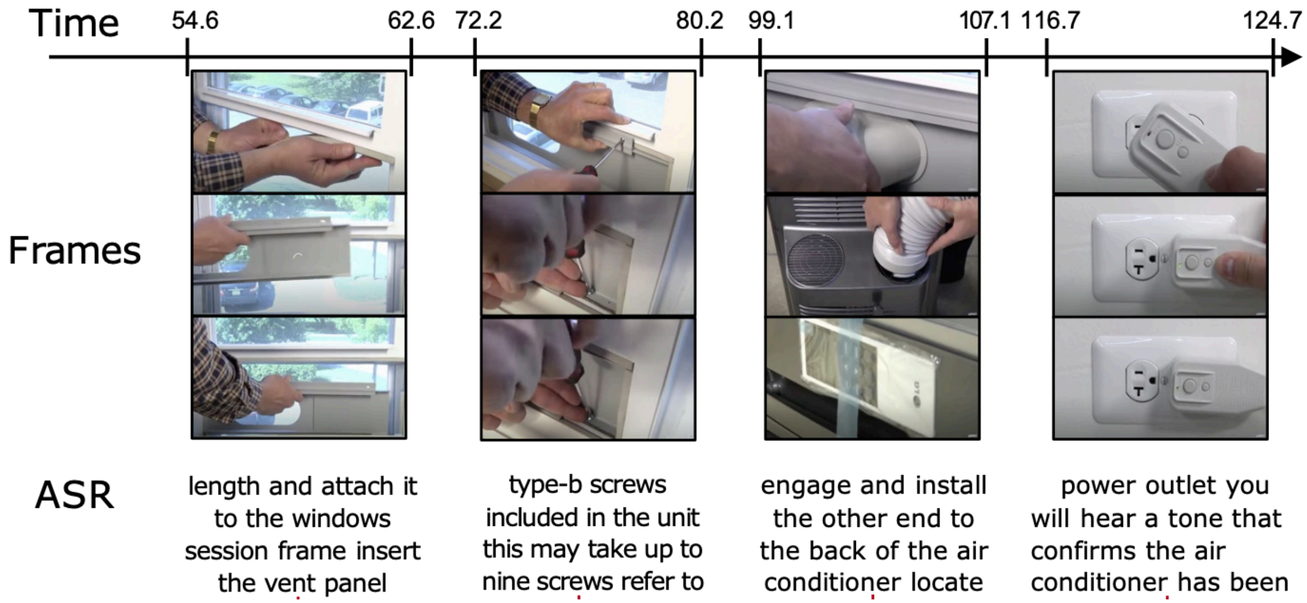


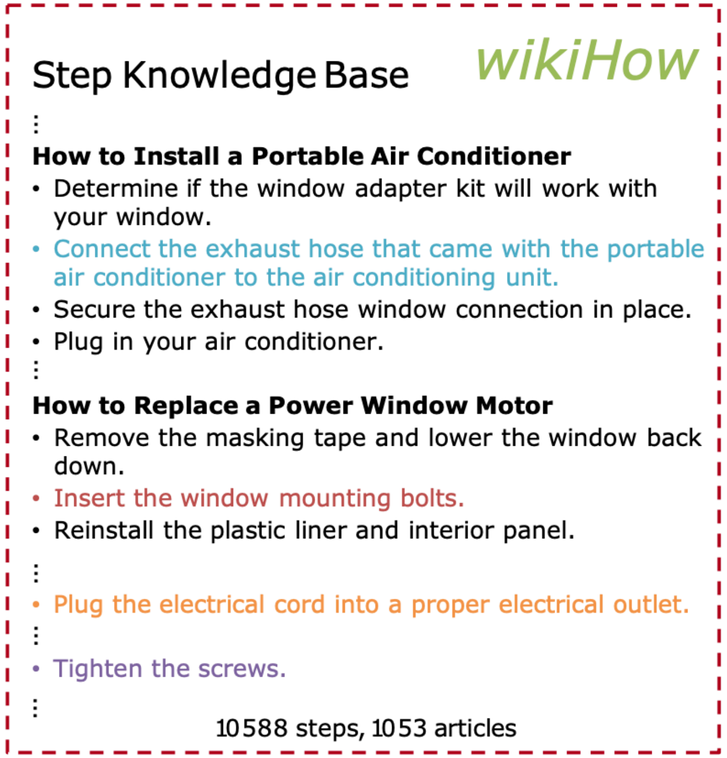
WikiHow Knowledge Base

- WikiHow is a crowdsourced multimedia repository containing over 230,000 “how-to” articles.

Distant Supervision

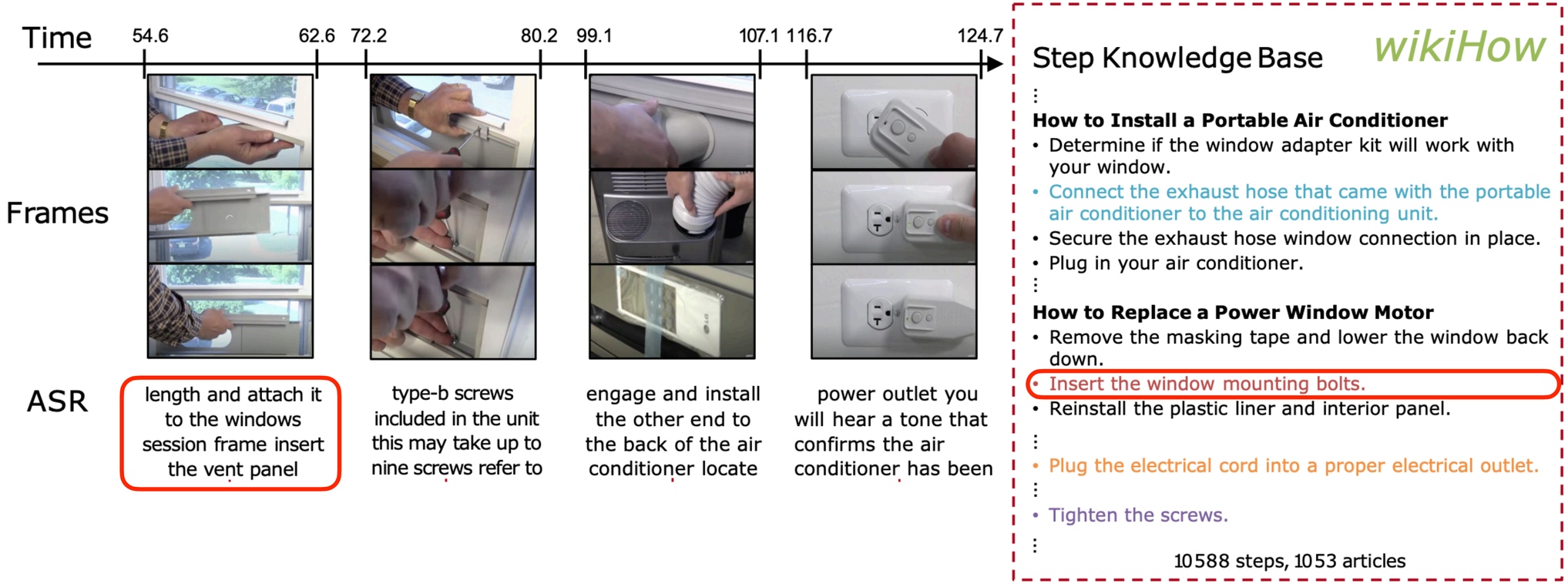
- To address this issue, we leverage the distant supervision of a textual WikiHow knowledge base.





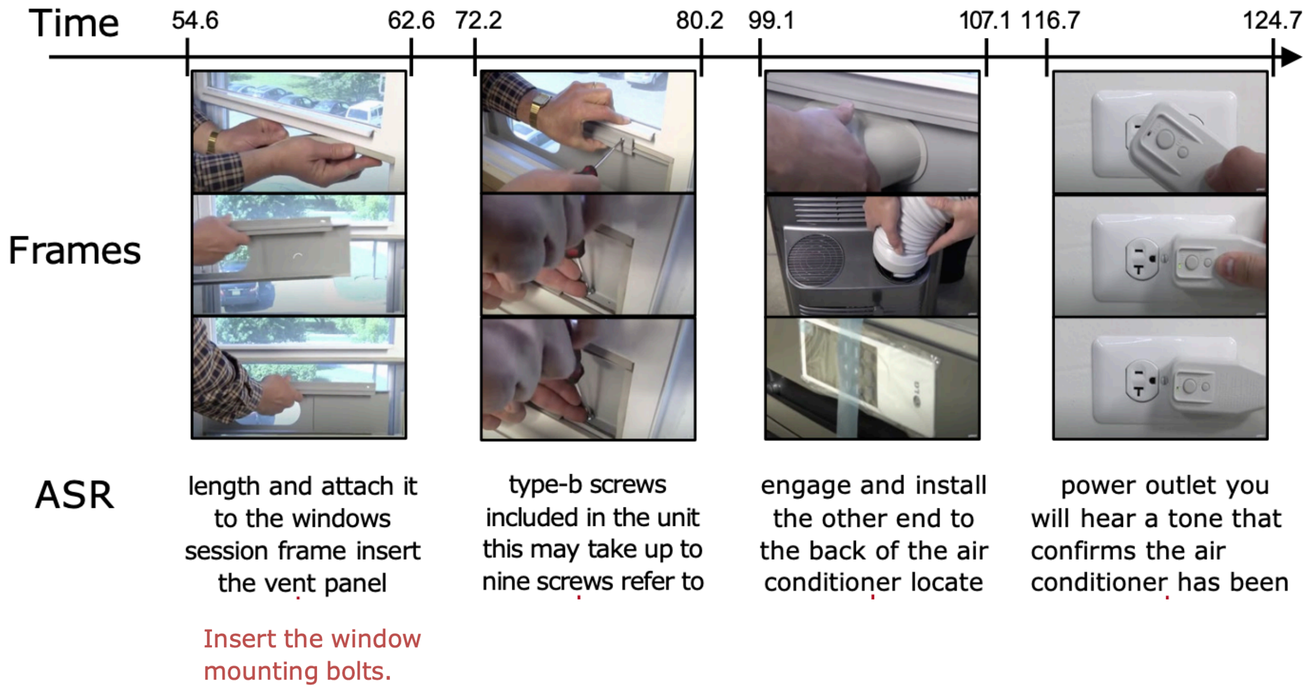
Distant Supervision

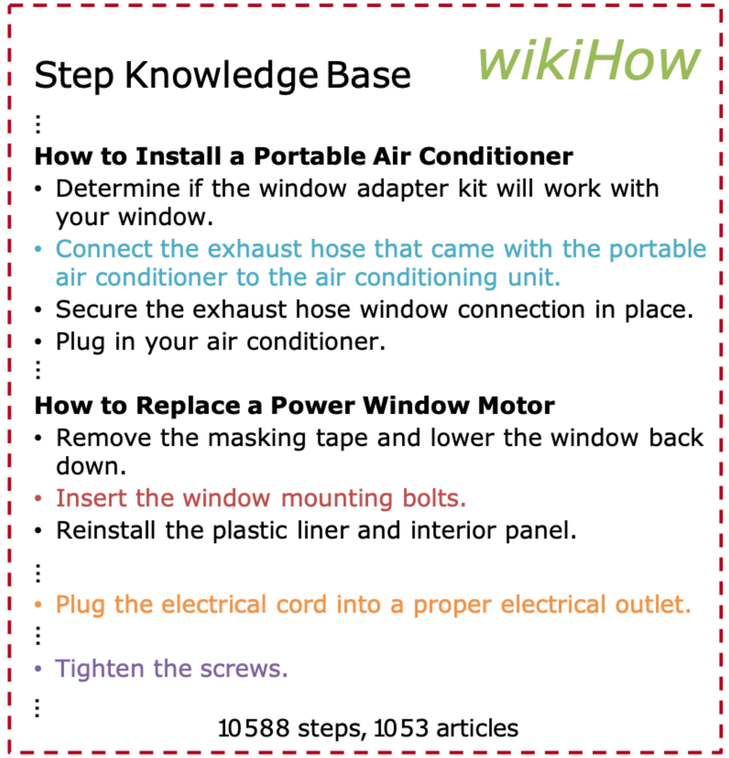
- To address this issue, we leverage the distant supervision of a textual WikiHow knowledge base.



Distant Supervision

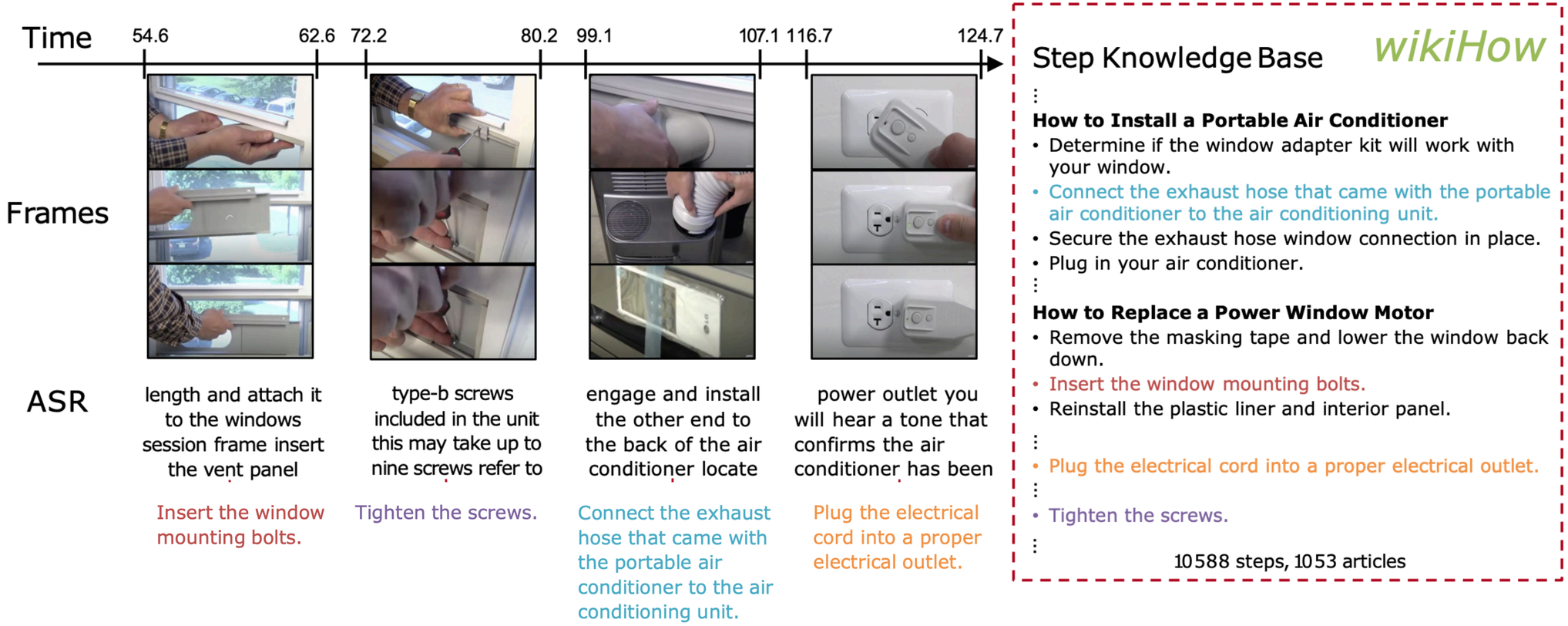
- To address this issue, we leverage the distant supervision of a textual WikiHow knowledge base.





Distant Supervision

- To address this issue, we leverage the distant supervision of a textual WikiHow knowledge base.



Matching ASR to Step Descriptions

- To assign a pseudo step-description label to each video segment, we use a textual similarity measure.

$$P(y_s^{(t)}|x_l) \approx \frac{\exp\left(\mathcal{S}(a_l, y_s^{(t)})\right)}{\sum_{t,s} \exp\left(\mathcal{S}(a_l, y_s^{(t)})\right)}.$$

Matching ASR to Step Descriptions

- To assign a pseudo step-description label to each video segment, we use a textual similarity measure.

Language-based description of step s for task t.

A video segment l from an instructional video x.

$$P(y_s^{(t)}|x_l) \approx \frac{\exp\left(\mathcal{S}(a_l, y_s^{(t)})\right)}{\sum_{t,s} \exp\left(\mathcal{S}(a_l, y_s^{(t)})\right)}.$$

Matching ASR to Step Descriptions

- To assign a pseudo step-description label to each video segment, we use a textual similarity measure.

ASR associated with a video segment l.

Language-based description of step s for task t.

Normalized dot product similarity.

$$P(y_s^{(t)}|x_l) \approx \frac{\exp\left(\mathcal{S}(a_l, y_s^{(t)})\right)}{\sum_{t,s} \exp\left(\mathcal{S}(a_l, y_s^{(t)})\right)}.$$

Matching ASR to Step Descriptions

- To assign a pseudo step-description label to each video segment, we use a textual similarity measure.

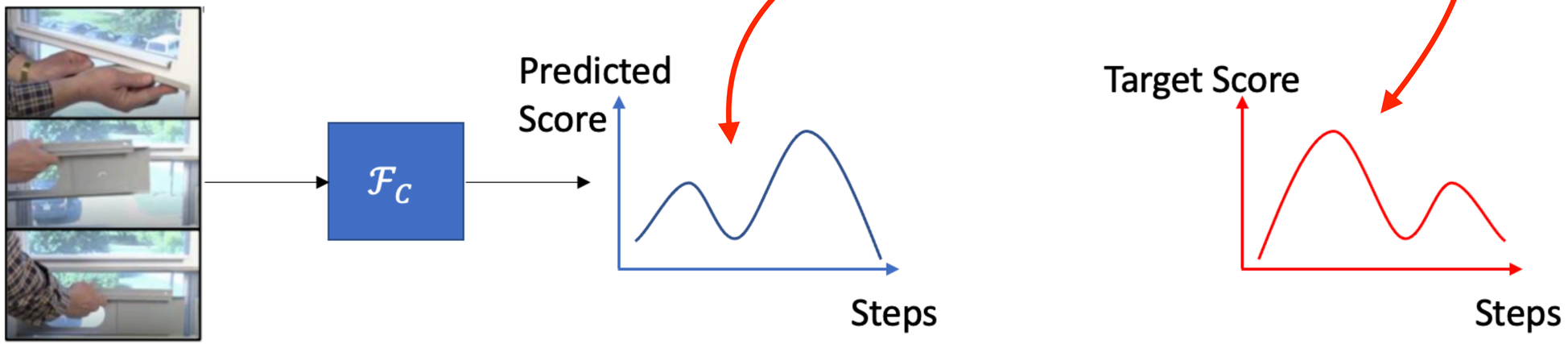
$$P(y_s^{(t)}|x_l) \approx \frac{\exp\left(\mathcal{S}(a_l, y_s^{(t)})\right)}{\sum_{t,s} \exp\left(\mathcal{S}(a_l, y_s^{(t)})\right)}.$$

Normalization over similarities w.r.t. all step descriptions for all tasks.

Learning Step Embeddings

We minimize the KL-Divergence between the predicted distribution and target distributions.

$$\min_{\theta} \sum_{t,s} P(y_s^{(t)}|x_l) \log \frac{P(y_s^{(t)}|x_l)}{[\mathcal{F}_C(x_l;\theta)]_{(t,s)}}.$$



Experimental Results

- Pretraining is done on HowTo100M.

a) HowTo100M

- Evaluation is done on COIN, Breakfast and EPIC-Kitchens.

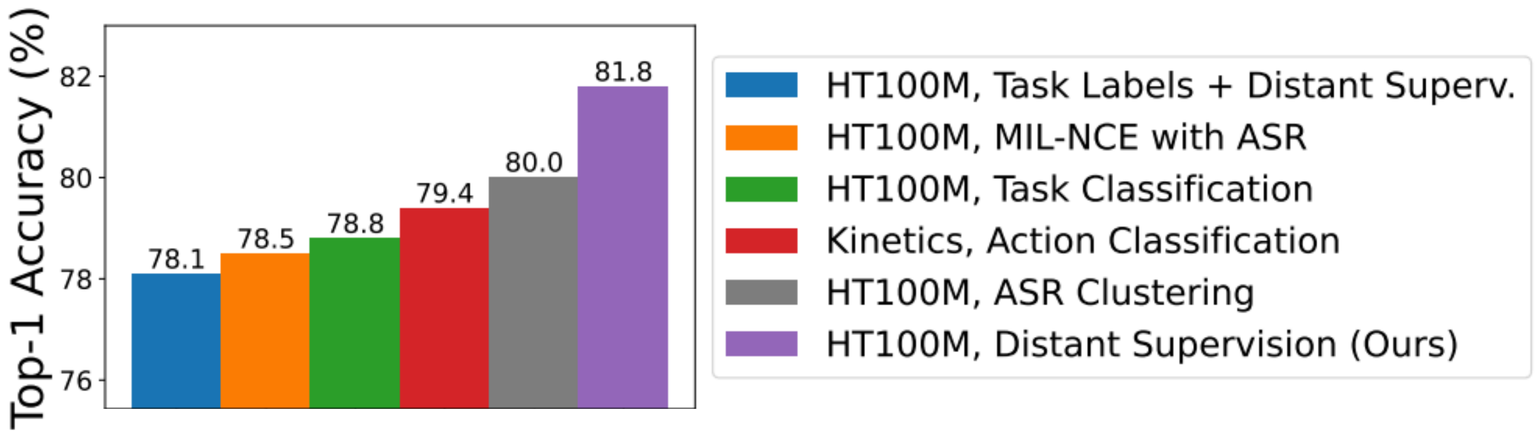
b) COIN

c) Breakfast

d) EPIC-Kitchens

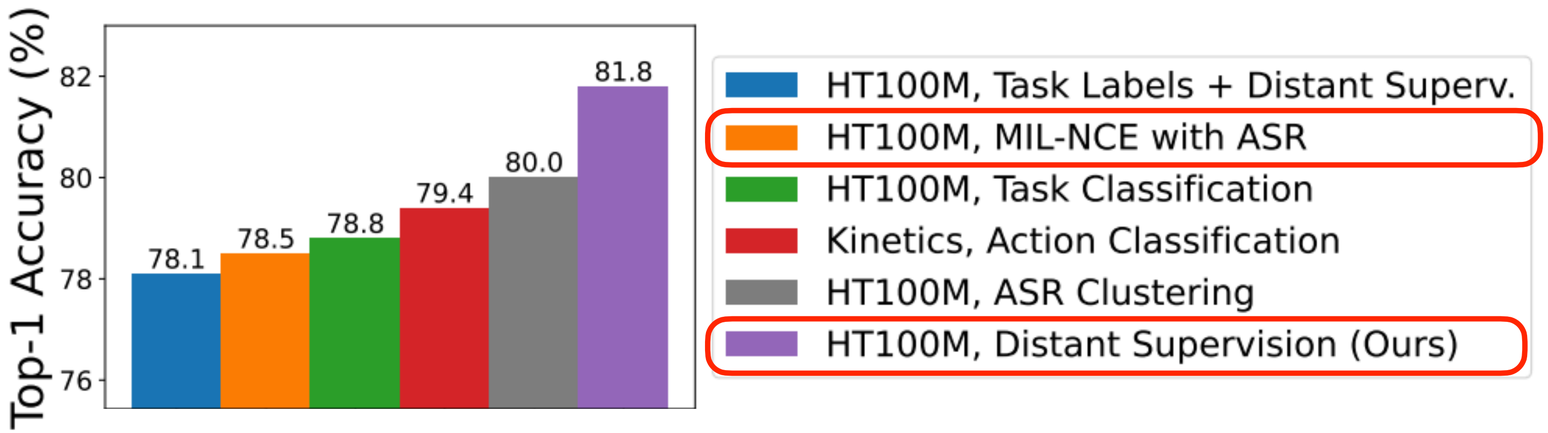
Experimental Results

- Comparing different forms of supervision on COIN procedural activity classification task.



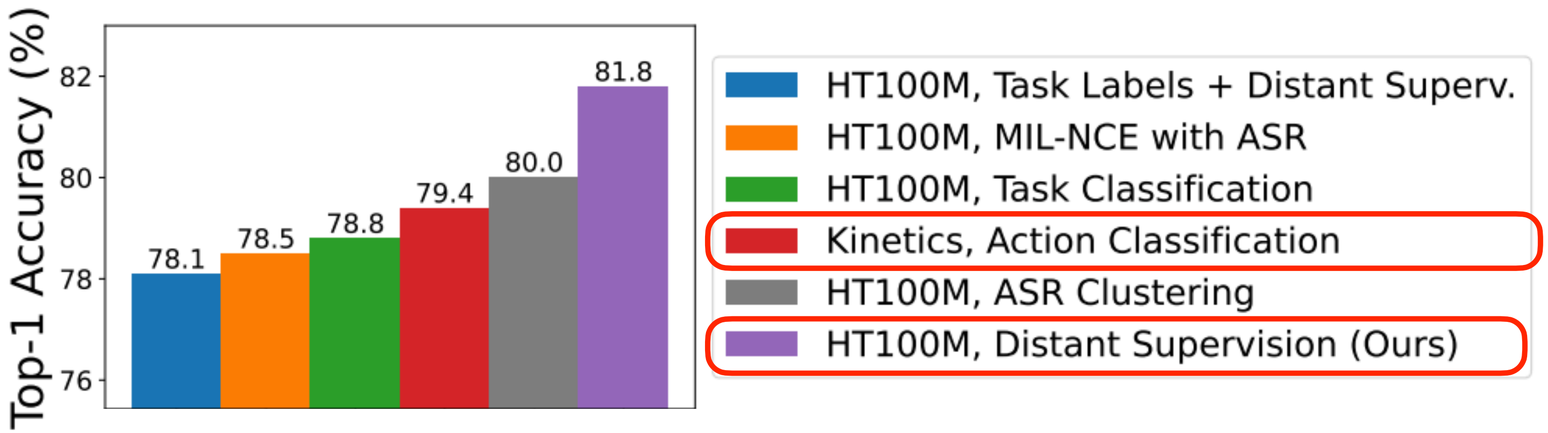
Experimental Results

- Comparing different forms of supervision on COIN procedural activity classification task.



Experimental Results

- Comparing different forms of supervision on COIN procedural activity classification task.



Experimental Results

- Comparison to the state-of-the-art on 4 downstream recognition tasks.

Segment Model	Pretraining Supervision	Pretraining Dataset	Linear Acc (%)
TSN (RGB+Flow) [57]	Supervised: action labels	Kinetics	36.5*
S3D [39]	Unsupervised: MIL-NCE on ASR	HT100M	37.5*
ClipBERT [34]	Supervised: captions	COCO + Visual Genome	30.8
VideoCLIP [65]	Unsupervised: NCE on ASR	HT100M	39.4
SlowFast [17]	Supervised: action labels	Kinetics	32.9
TimeSformer [8]	Supervised: action labels	Kinetics	48.3
TimeSformer [8]	Unsupervised: k-means on ASR	HT100M	46.5
TimeSformer	**Unsupervised: distant supervision (ours)**	HT100M	**54.1**

a) Step classification on COIN

Long-term Model	Segment Model	Pretraining Supervision	Pretraining Dataset	Acc (%)
Timeception [25]	3D-ResNet [62]	Supervised: action labels	Kinetics	71.3
VideoGraph [26]	I3D [11]	Supervised: action labels	Kinetics	69.5
GHRM [69]	I3D [11]	Supervised: action labels	Kinetics	75.5
Basic Transformer	S3D [39]	Unsupervised: MIL-NCE on ASR	HT100M	74.4
Basic Transformer	SlowFast [17]	Supervised: action labels	Kinetics	76.1
Basic Transformer	TimeSformer [8]	Supervised: action labels	Kinetics	81.1
Basic Transformer	TimeSformer [8]	Unsupervised: k-means on ASR	HT100M	81.4
Basic Transformer	**TimeSformer**	**Unsupervised: distant supervision (ours)**	HT100M	**88.7**
Transformer w/ KB Transfer	**TimeSformer**	**Unsupervised: distant supervision (ours)**	HT100M	**89.9**

b) Task classification on Breakfast

Long-term Model	Segment Model	Pretraining Supervision	Pretraining Dataset	Acc (%)
Basic Transformer	S3D [39]	Unsupervised: MIL-NCE on ASR	HT100M	28.1
Basic Transformer	SlowFast [17]	Supervised: action labels	Kinetics	25.6
Basic Transformer	TimeSformer [8]	Supervised: action labels	Kinetics	34.7
Basic Transformer	TimeSformer [8]	Unsupervised: k-means on ASR	HT100M	34.0
Basic Transformer	**TimeSformer**	**Unsupervised: distant supervision (ours)**	HT100M	**38.2**
Transformer w/ KB Transfer	**TimeSformer**	**Unsupervised: distant supervision (ours)**	HT100M	**39.4**

c) Step forecasting on COIN

Segment Model	Pretraining Supervision	Pretraining Dataset	Action (%)	Verb (%)	Noun (%)
TSN [61]	-	-	33.2	60.2	46.0
TRN [68]	-	-	35.3	65.9	45.4
TBN [29]	-	-	36.7	66.0	47.2
MoViNet [30]	-	-	**47.7**	**72.2**	57.3
TSM [36]	Supervised: action labels	Kinetics	38.3	67.9	49.0
SlowFast [17]	Supervised: action labels	Kinetics	38.5	65.6	50.0
ViViT-L [6]	Supervised: action labels	Kinetics	44.0	66.4	56.8
TimeSformer [8]	Supervised: action labels	Kinetics	42.3	66.6	54.4
TimeSformer	**Unsupervised: distant supervision (ours)**	HT100M	44.4	67.1	**58.1**

d) Egocentric video classification on EPIC-Kitchens

Take-Home Message

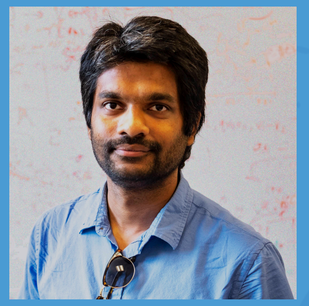
- Distant supervision from an external textual knowledge base enables learning a strong visual representation for procedural activity recognition.

UNC | COLLEGE OF ARTS AND SCIENCES
Computer Science

ECLIPSE: Efficient Long-range Video Retrieval using Sight and Sound

Yan-Bo Lin

Jie Lei

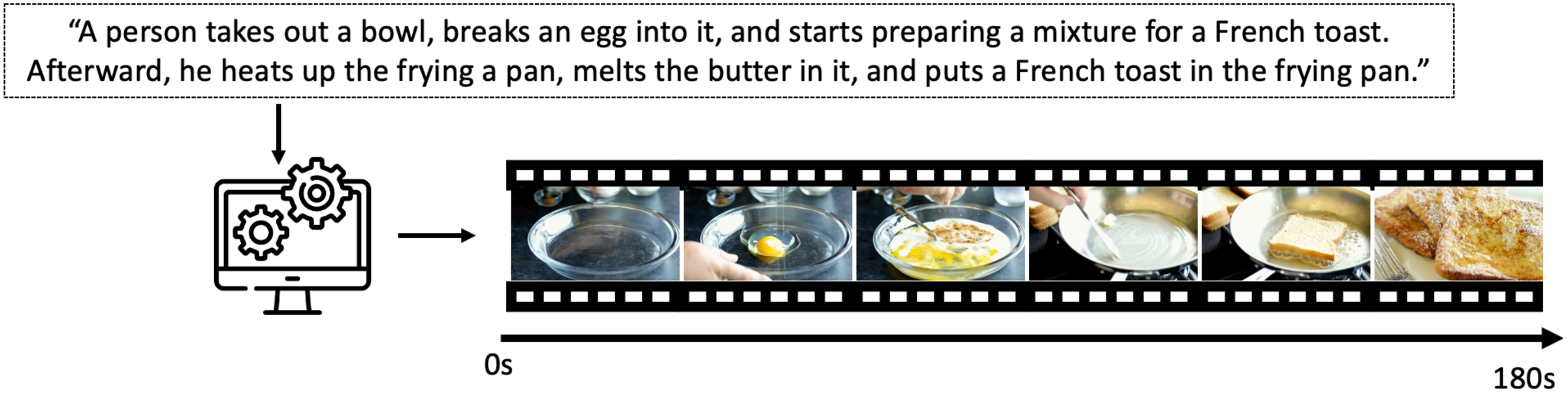


Mohit Bansal

Gedas Bertasius

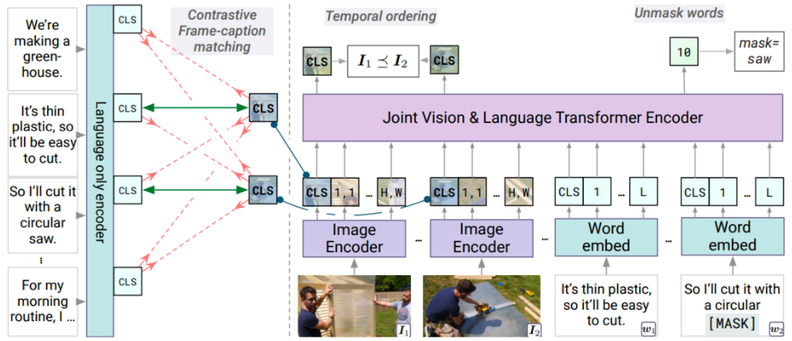
Problem Overview

- Given a textual query describing a several minute-long video, we aim to retrieve the correct video.

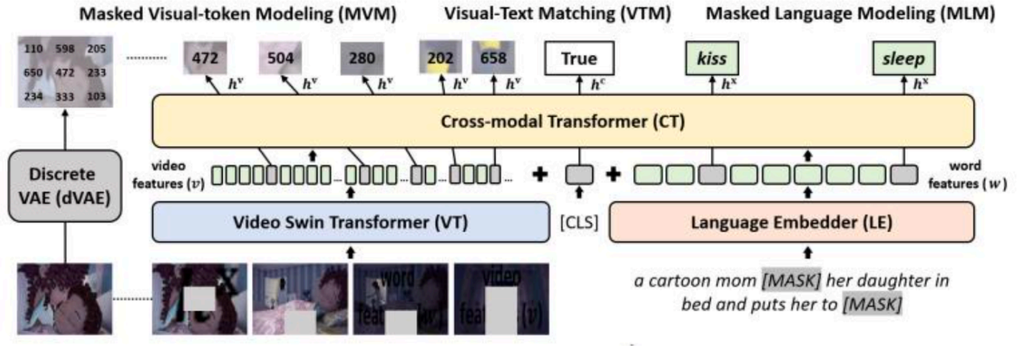


Prior Work

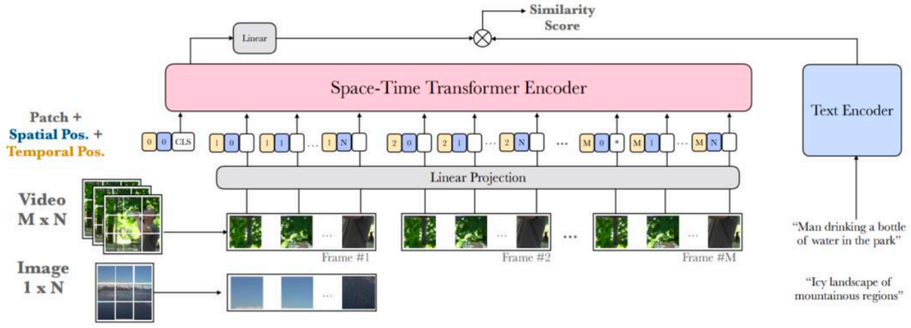
- Most prior text-to-video retrieval approaches focus on short-range videos (~5-15s in length).



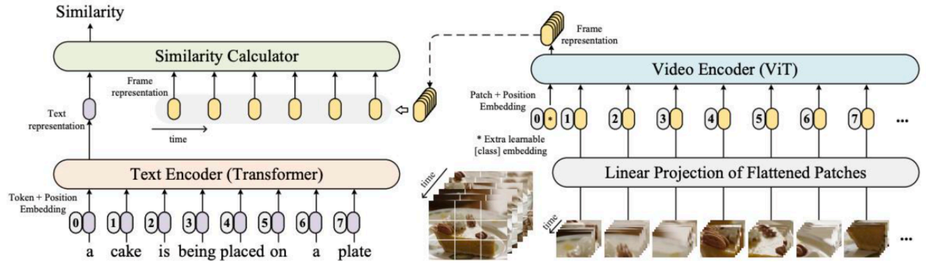
Zellers, Rowan, et al. "MERLOT: Multimodal Neural Script Knowledge Models." NeurIPS'21.



Fu et al. "VIOLET : End-to-End Video-Language Transformers with Masked Visual-token Modeling" arXiv'21



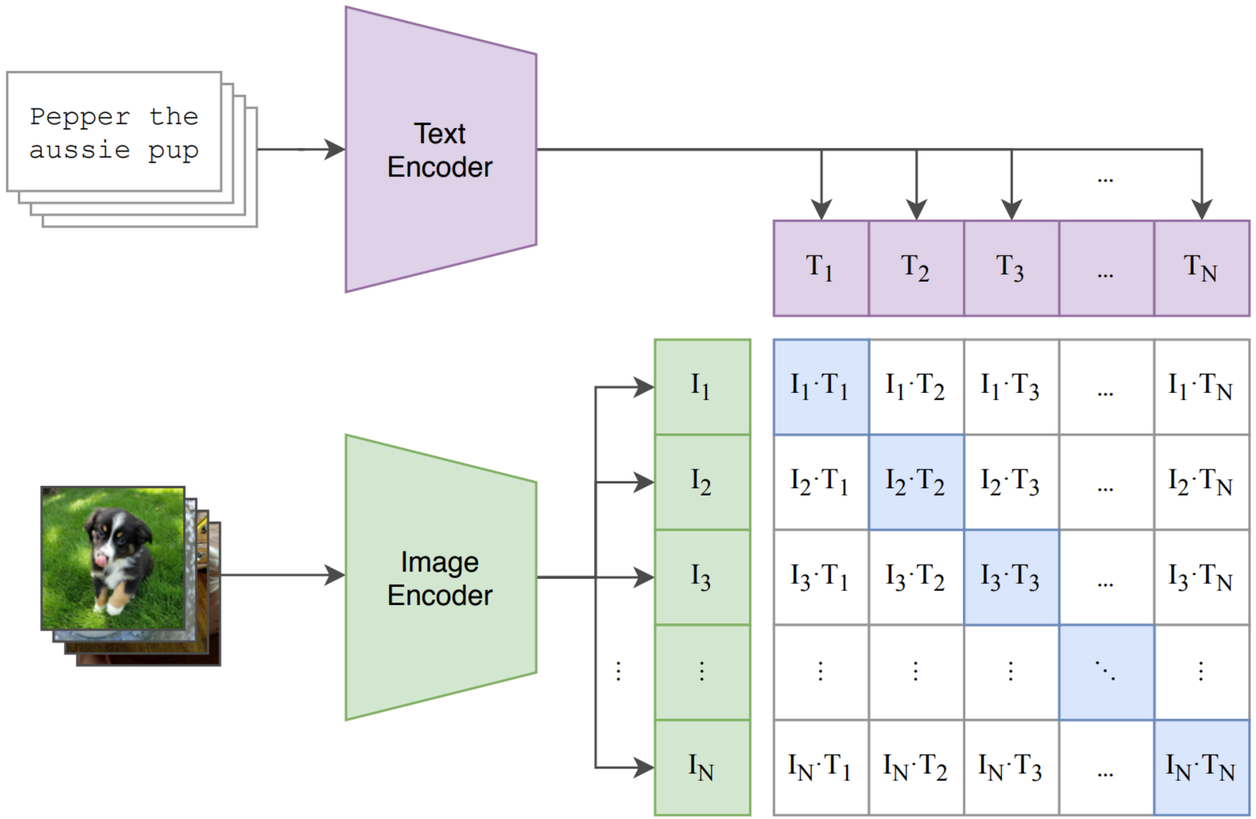
Bain, Max, et al. "Frozen in time: A joint video and image encoder for end-to-end retrieval." ICCV'21.



Luo et al. "CLIP4Clip: An Empirical Study of CLIP for End to End Video Clip Retrieval" arXiv'21.

Prior Work

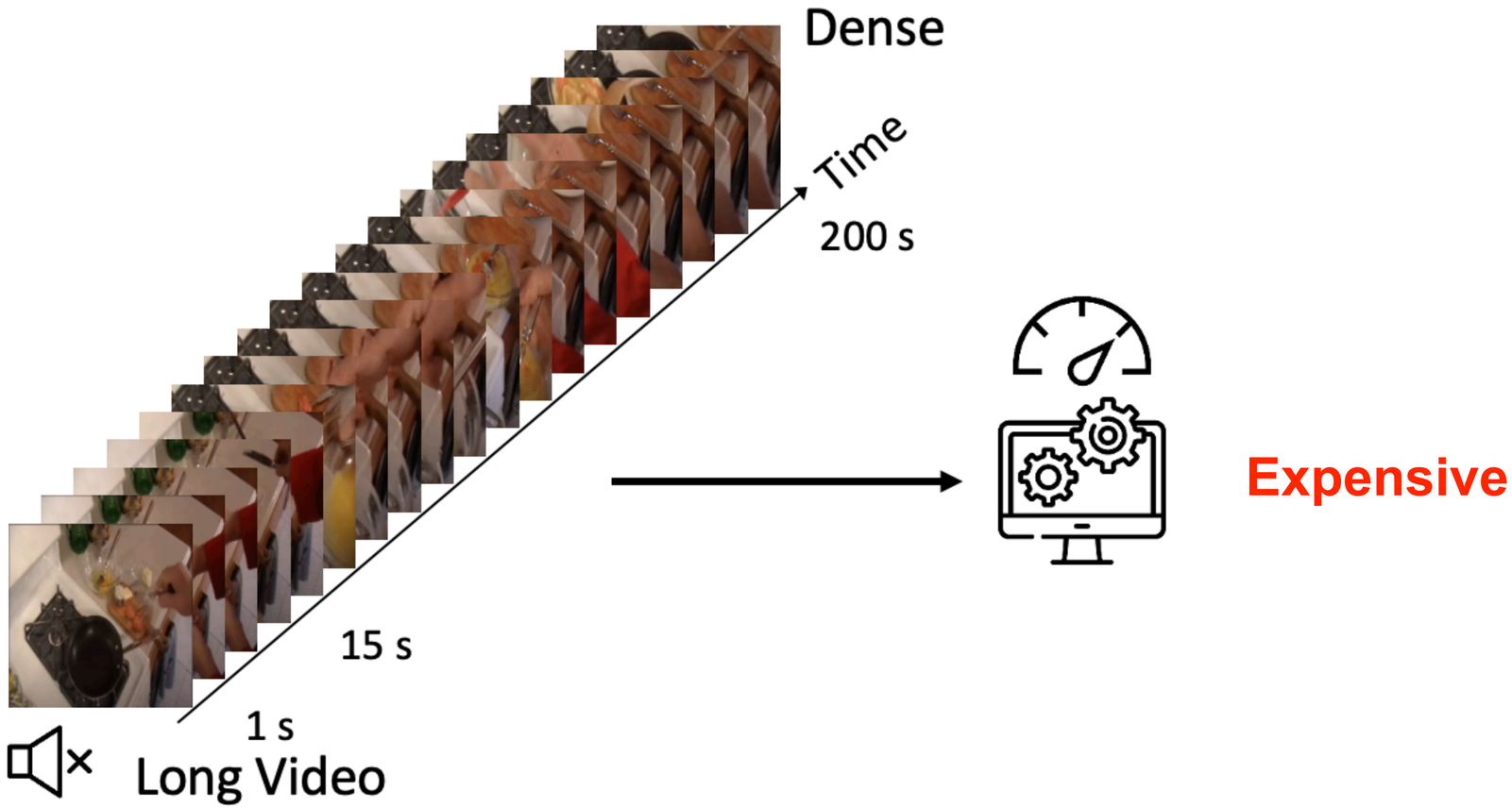
- CLIP is one of the most widely adopted vision-and-language models.



Radford et al. “Learning Transferable Visual Models From Natural Language Supervision,” ICML 2021

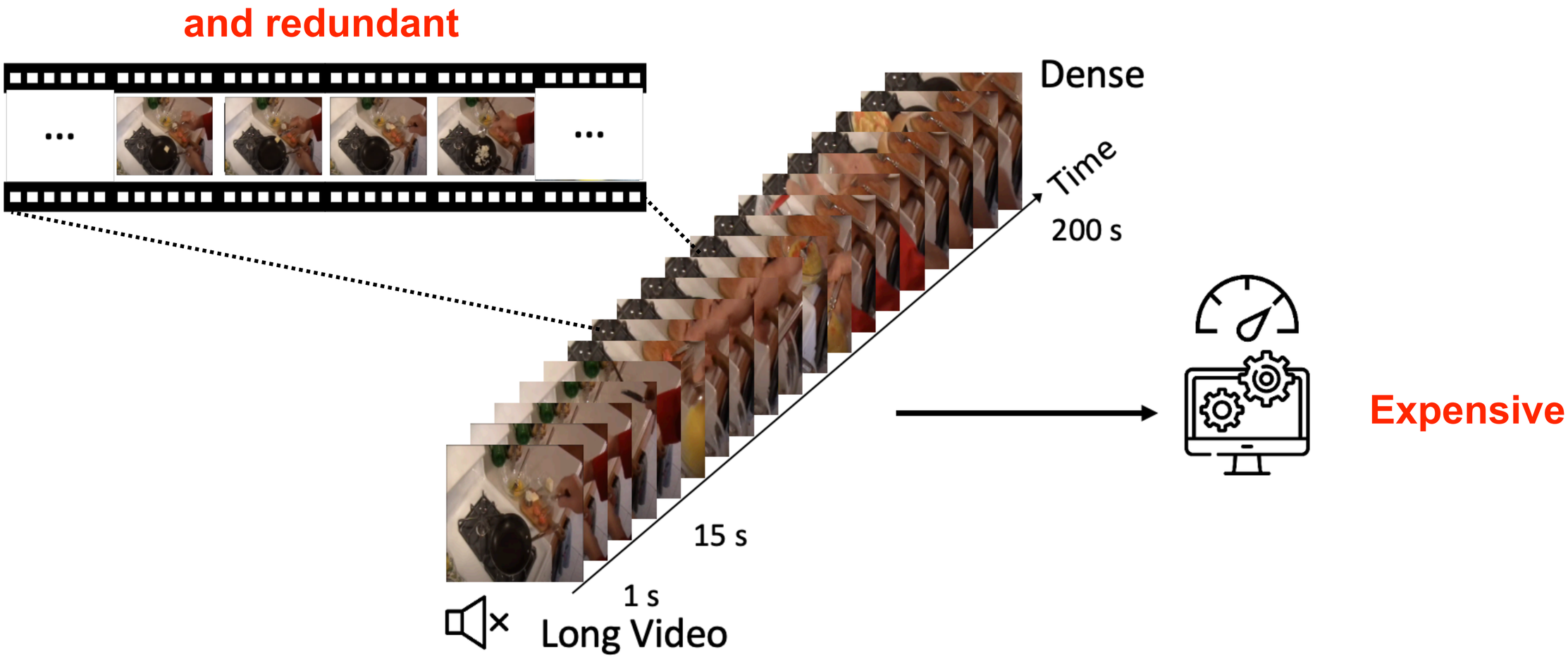
Motivation

- Long videos are costly to process and they also have high informational redundancy.



Motivation

- Long videos are costly to process and they also have high informational redundancy.



Motivation

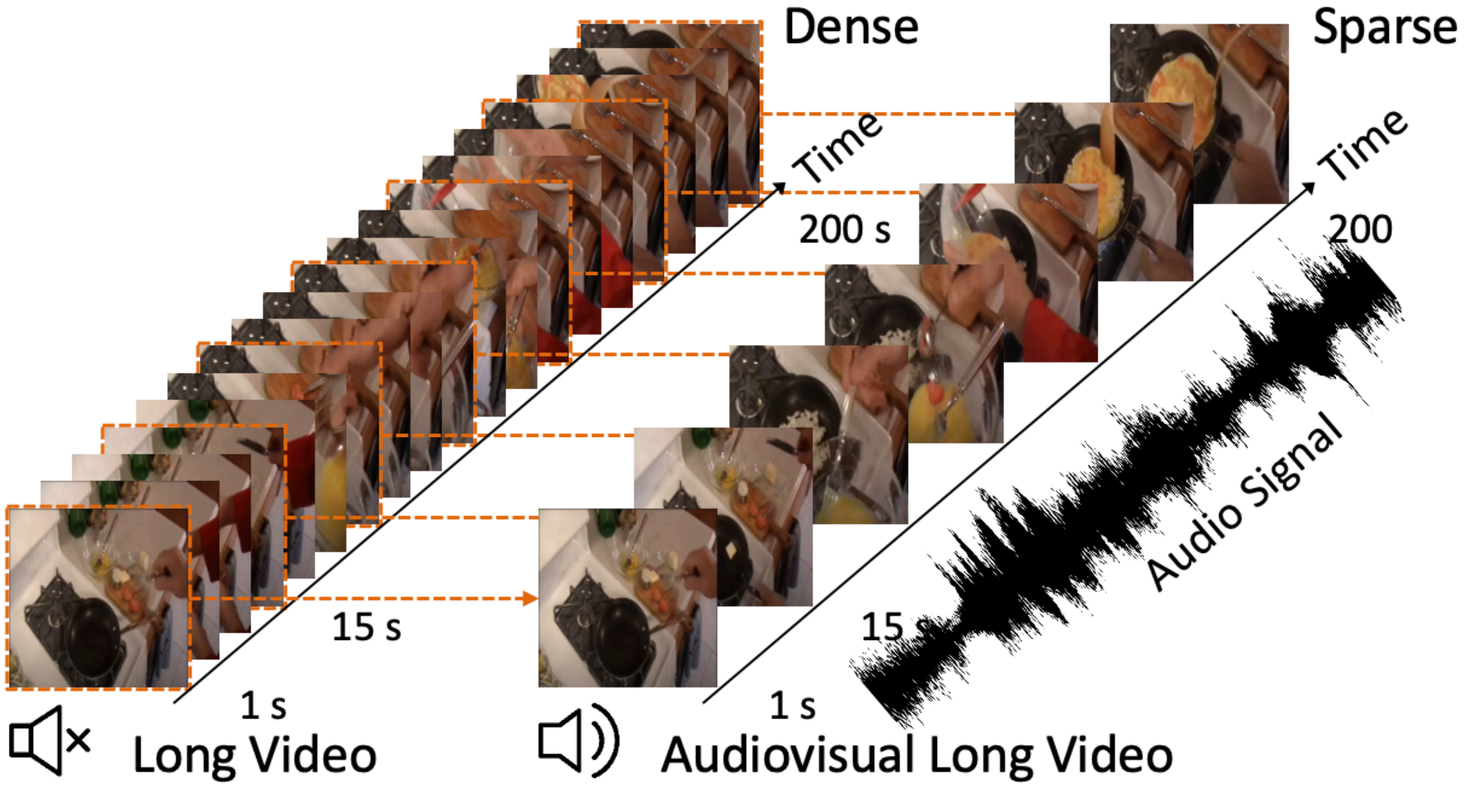
Audio is **compact** and **cheap to process**.

Motivation

- Audio effectively captures temporal dynamics in videos.

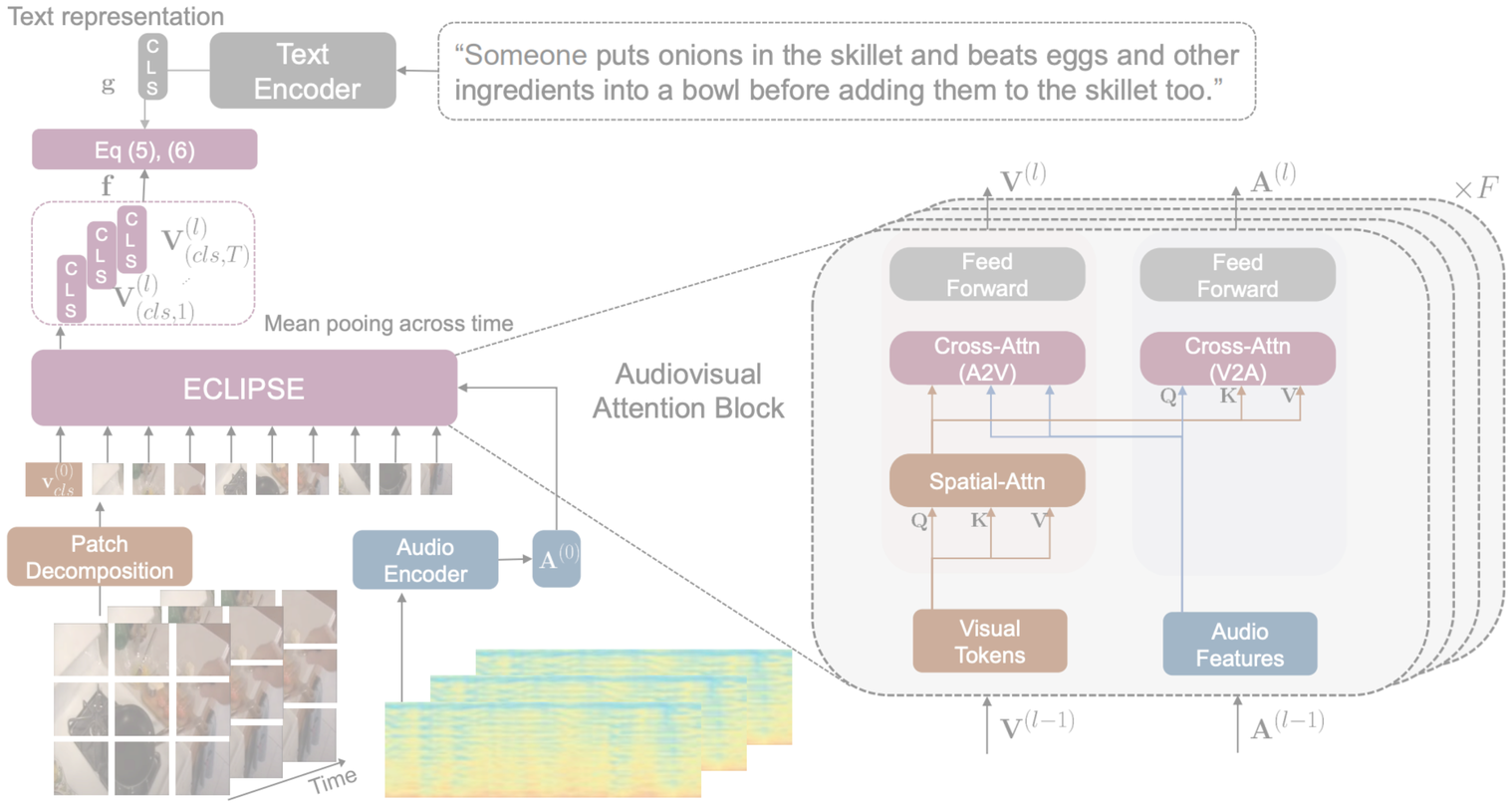
Our Approach

- Our proposed ECLIPSE model (**E**fficient **CLIP** with **S**ound **E**ncoding) adapts CLIP to audiovisual retrieval of long videos.



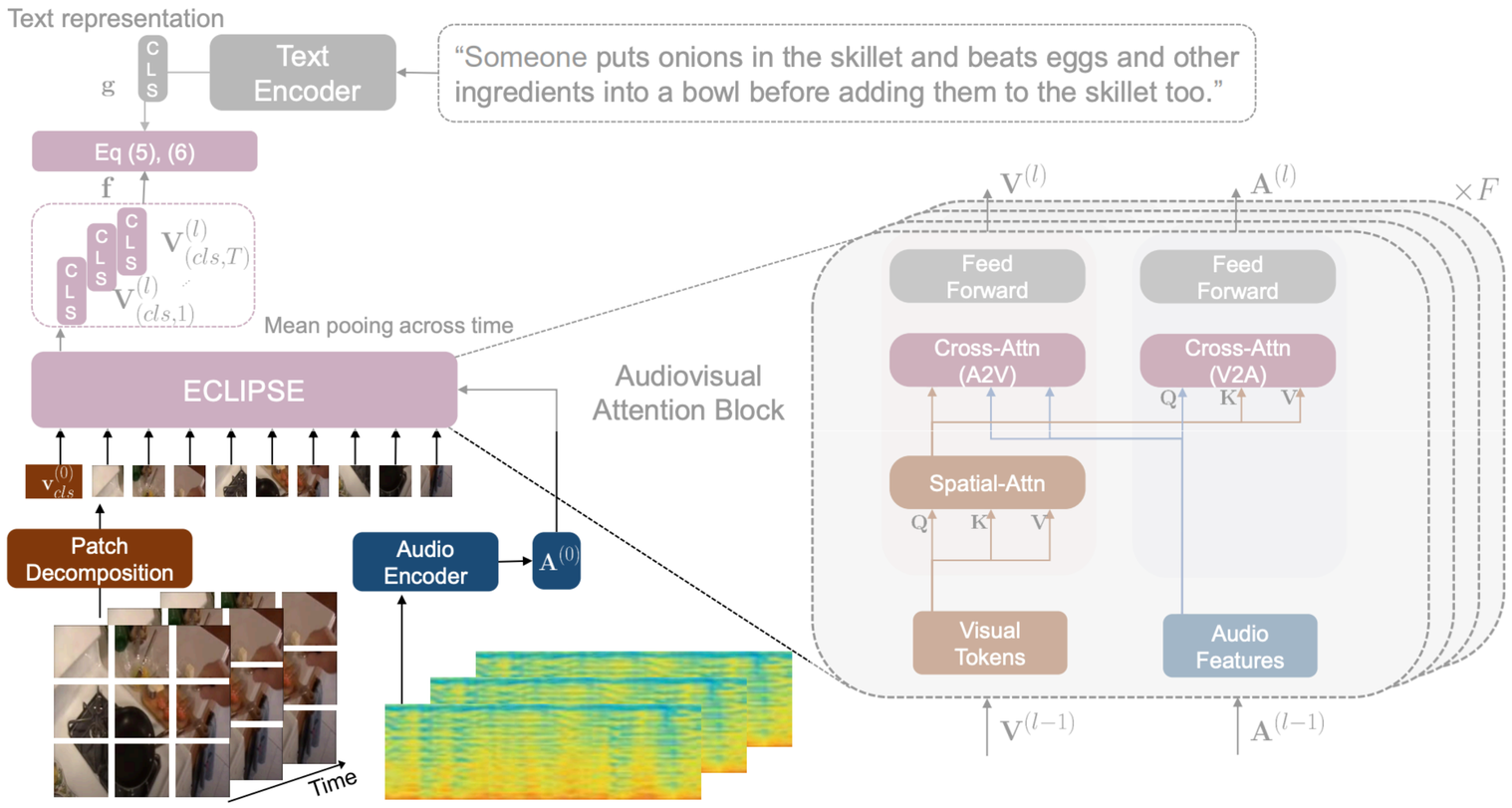
Architecture

- We adapt CLIP to long-range text-to-video retrieval by adding an efficient audiovisual attention block into the Transformer architecture.



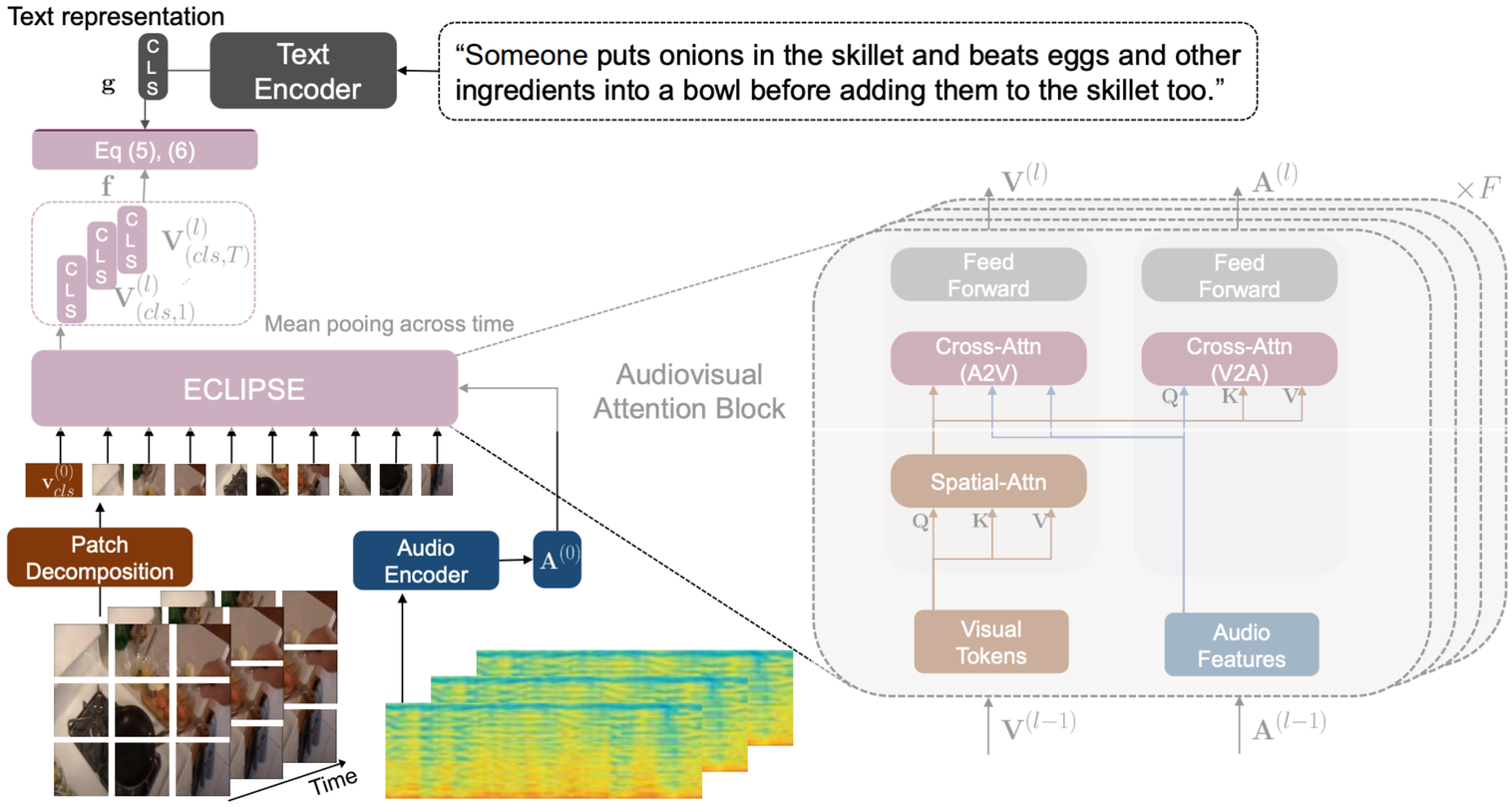
Architecture

- We adapt CLIP to long-range text-to-video retrieval by adding an efficient audiovisual attention block into the Transformer architecture.



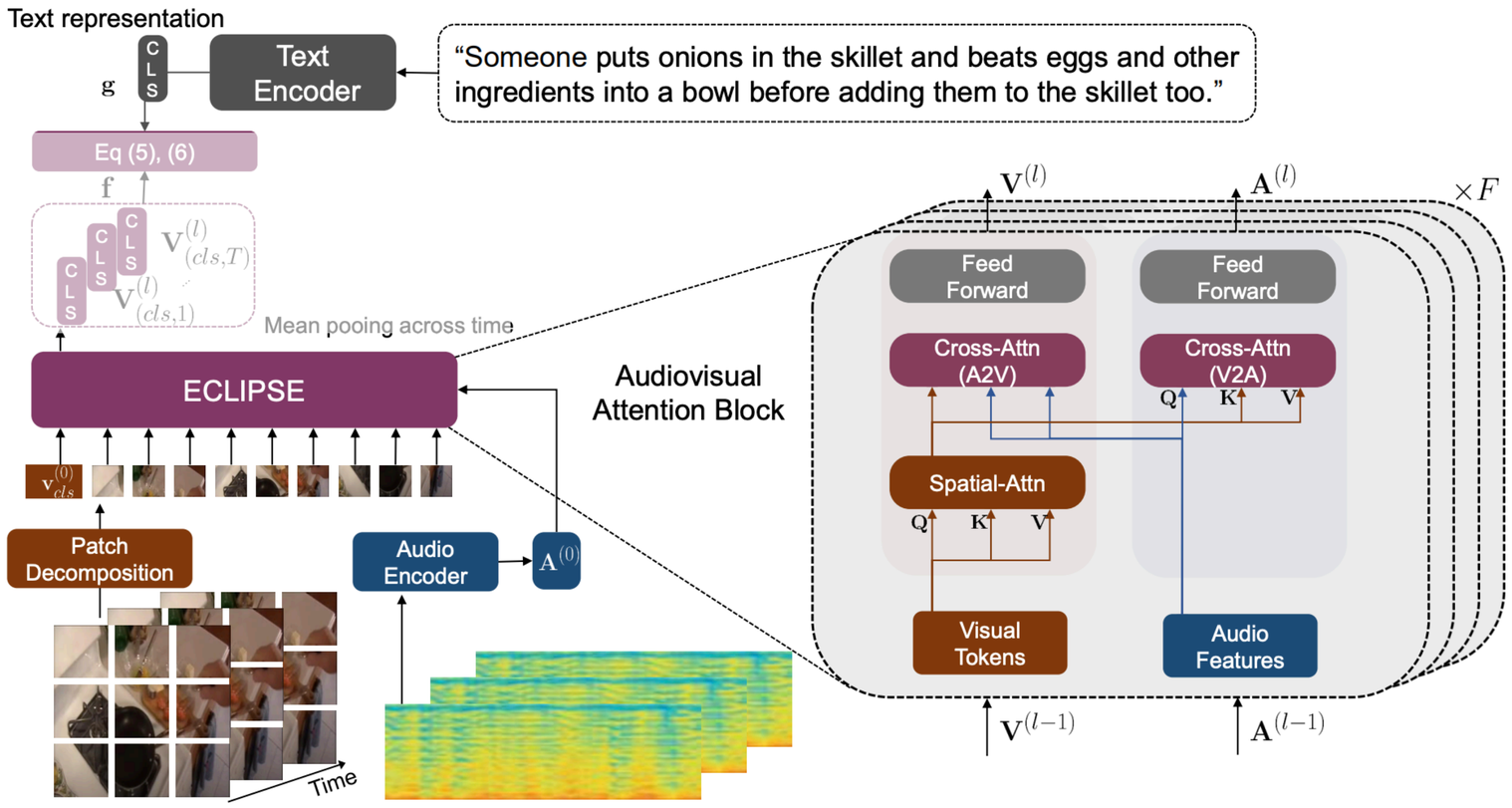
Architecture

- We adapt CLIP to long-range text-to-video retrieval by adding an efficient audiovisual attention block into the Transformer architecture.



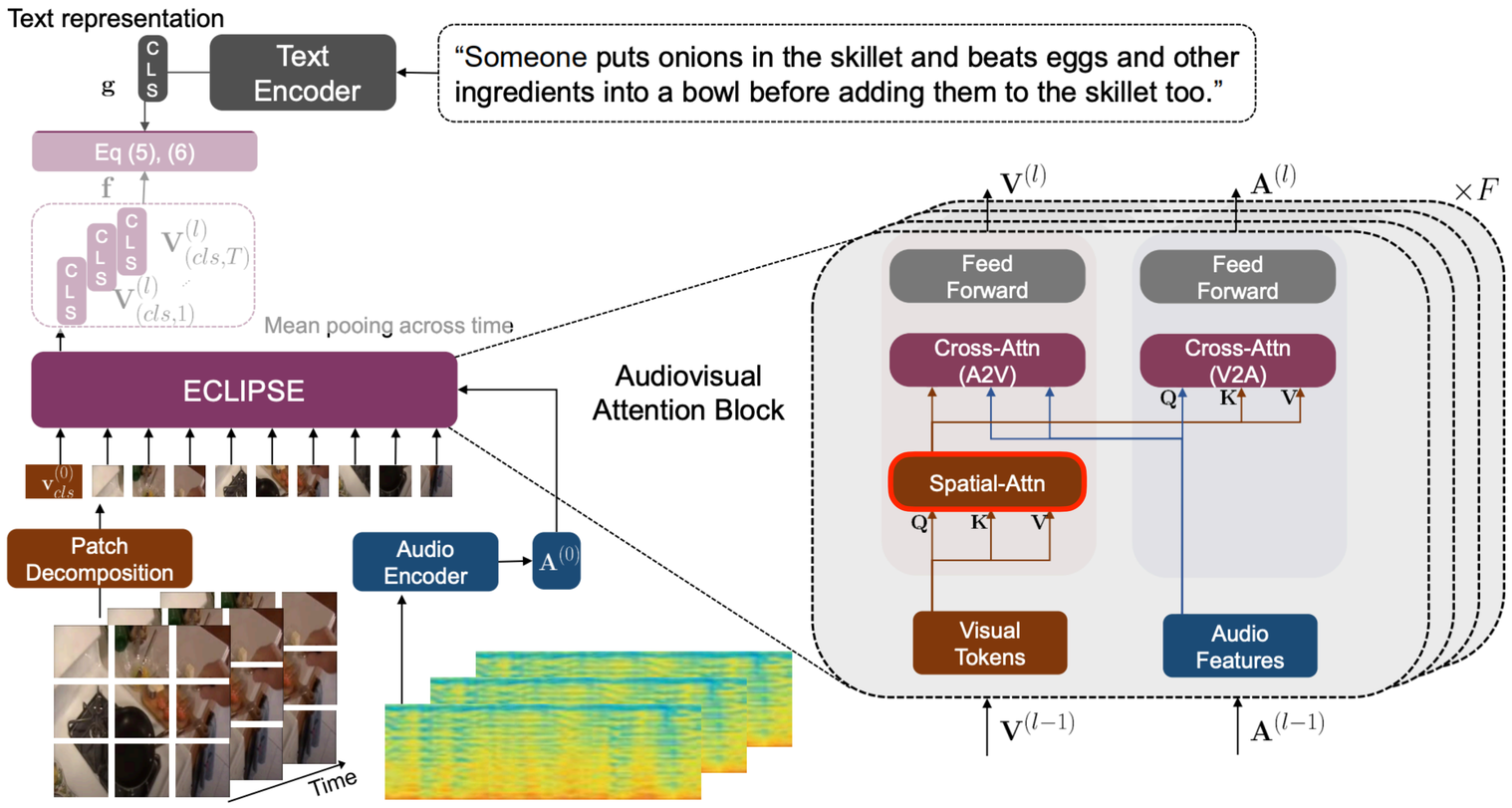
Architecture

- We adapt CLIP to long-range text-to-video retrieval by adding an efficient audiovisual attention block into the Transformer architecture.



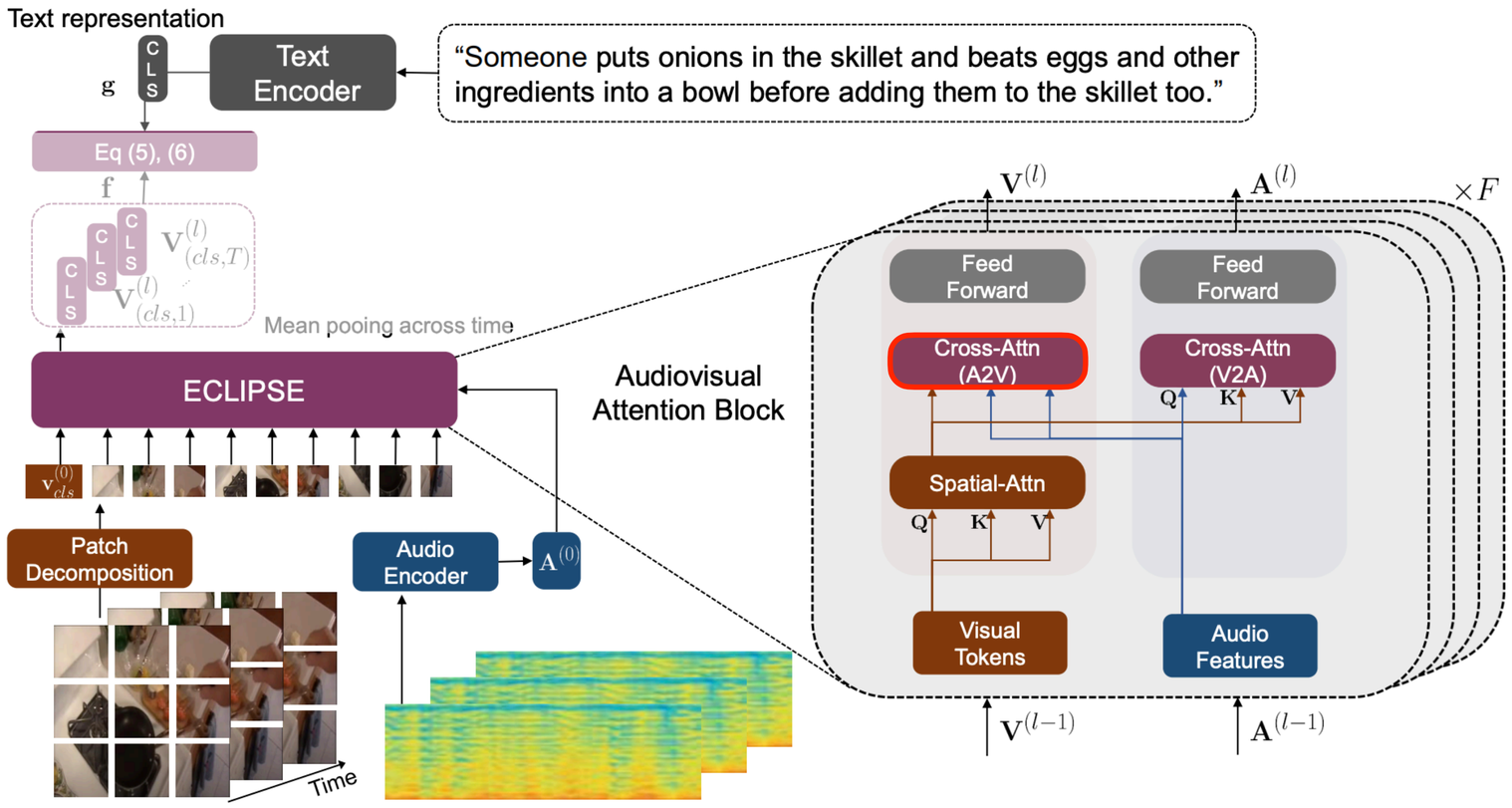
Architecture

- We adapt CLIP to long-range text-to-video retrieval by adding an efficient audiovisual attention block into the Transformer architecture.



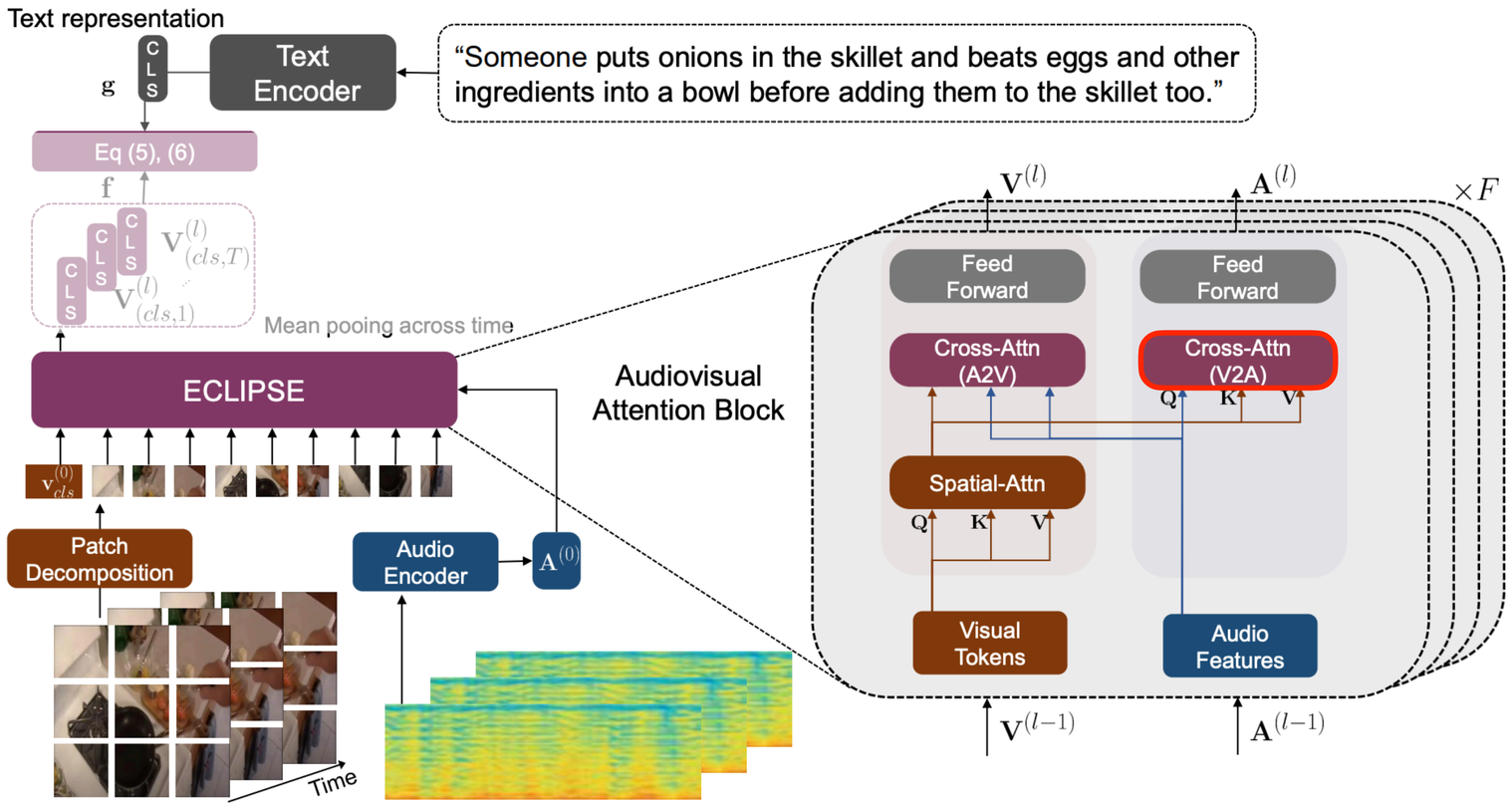
Architecture

- We adapt CLIP to long-range text-to-video retrieval by adding an efficient audiovisual attention block into the Transformer architecture.



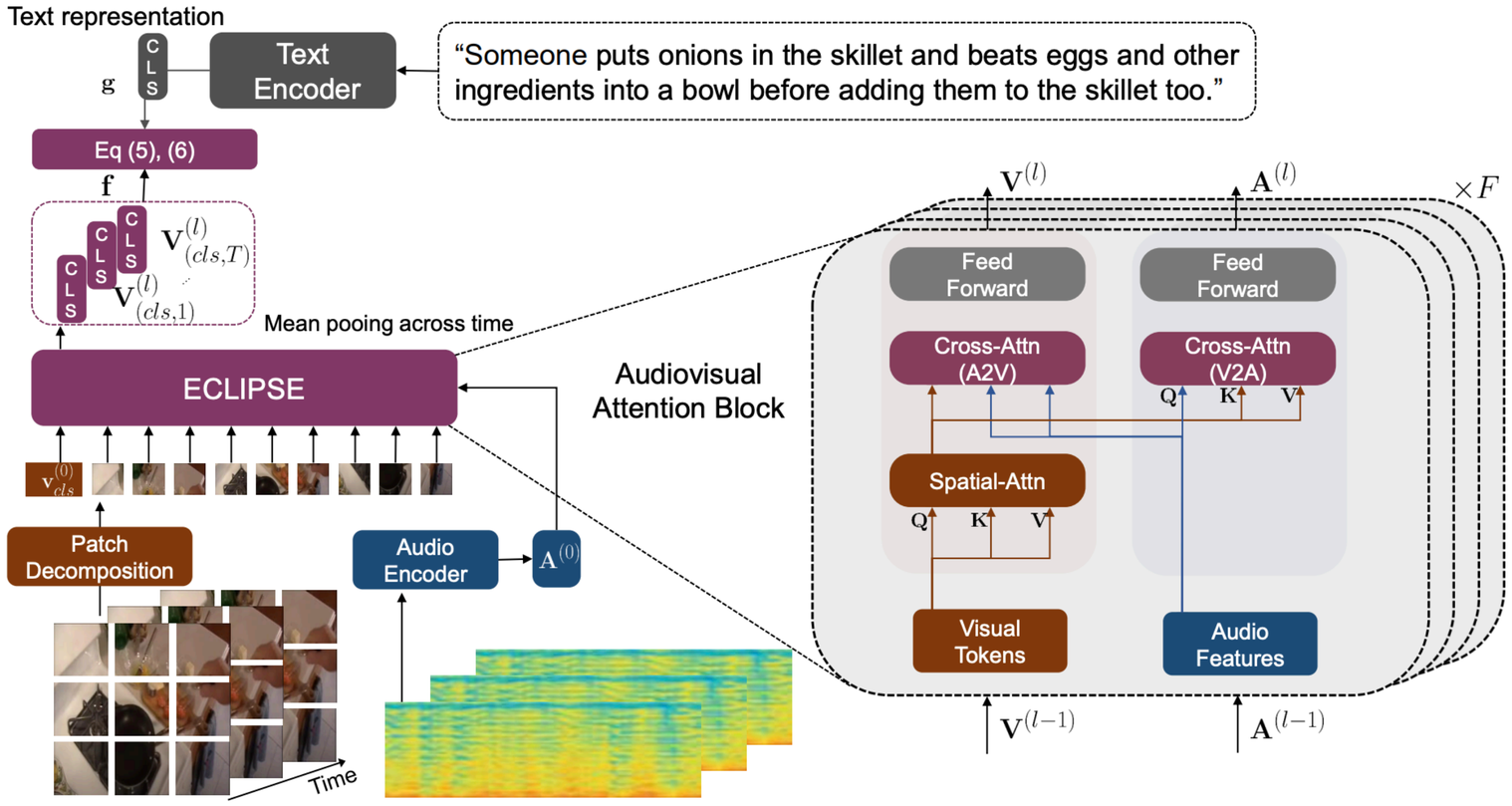
Architecture

- We adapt CLIP to long-range text-to-video retrieval by adding an efficient audiovisual attention block into the Transformer architecture.



Architecture

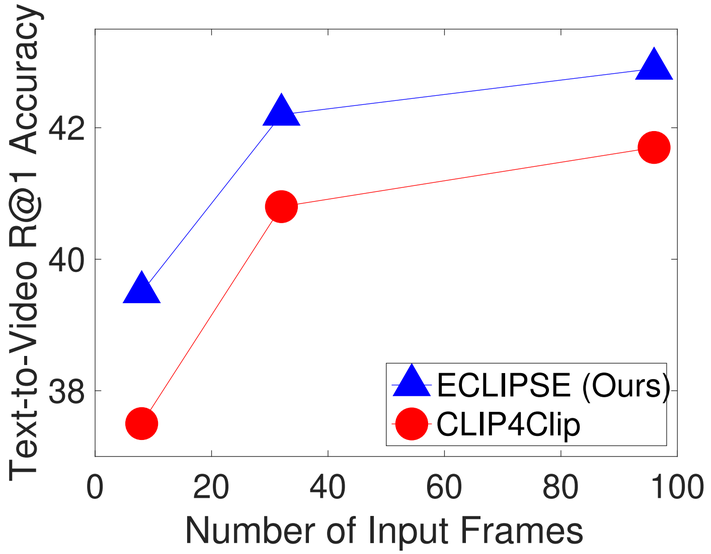
- We adapt CLIP to long-range text-to-video retrieval by adding an efficient audiovisual attention block into the Transformer architecture.



Results on ActivityNet

- Our method outperforms CLIP4Clip while using 3x fewer frames.

Method	Num. Frames	Inference GFLOPs ↓	GPU Mem. (in MB) ↓	Samples per Sec. ↑	T2V R@1 ↑
CLIP4Clip	96	1251	24,802	17.39	41.7
ECLIPSE	32	**827**	**10,637**	**50.93**	**42.3**



Results on Other Datasets

- Our method outperforms other methods while also being more efficient.

Method	Frames	QVH [25]	DiDeMo [26]	YC2 [27]	Charades [28]	GFLOPs
Frozen-in-Time [3]	32	55.0	34.6	32.2	11.9	1426
CLIP4Clip [16]	96	70.2	42.5	37.6	13.9	1251
ECLIPSE	32	**70.8**	**44.2**	**38.5**	**15.7**	**827**

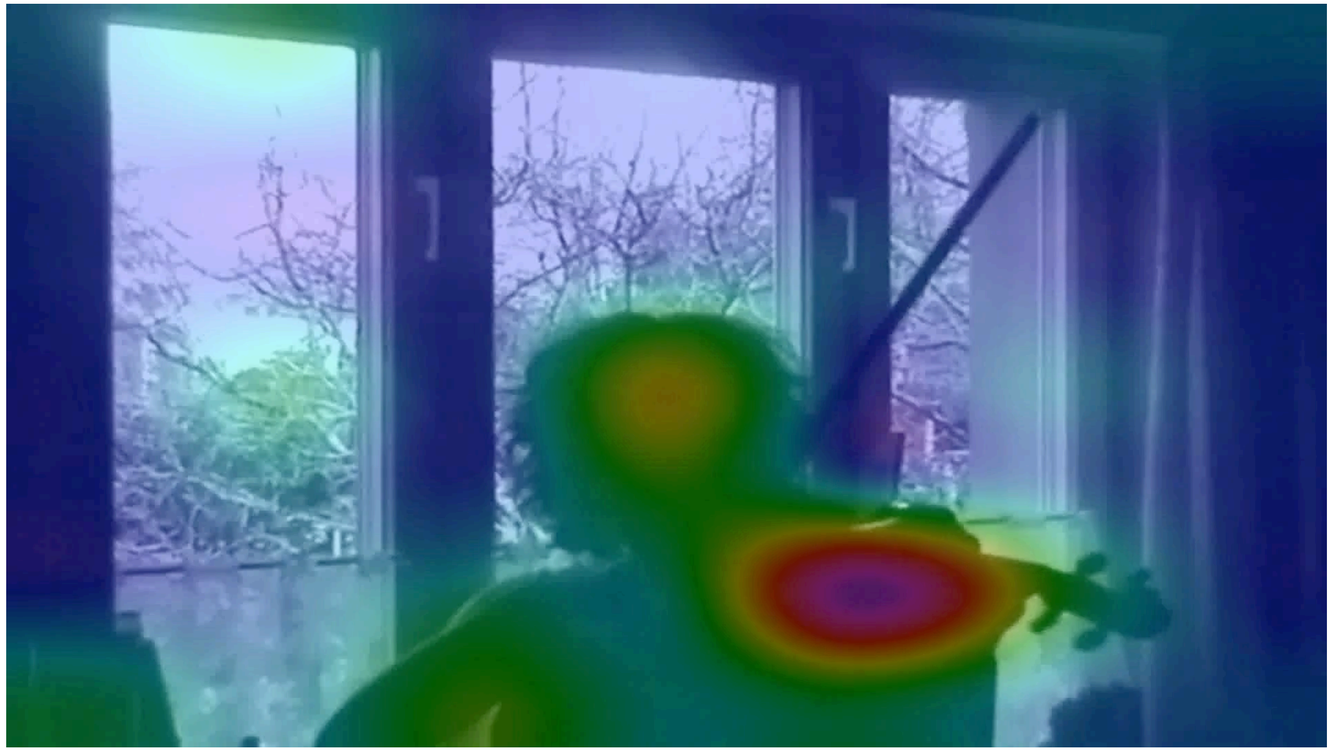
Results on Other Datasets

- Our method outperforms other methods while also being more efficient.

Method	Frames	QVH [25]	DiDeMo [26]	YC2 [27]	Charades [28]	GFLOPs
Frozen-in-Time [3]	32	55.0	34.6	32.2	11.9	1426
CLIP4Clip [16]	96	70.2	42.5	37.6	13.9	1251
ECLIPSE	32	**70.8**	**44.2**	**38.5**	**15.7**	**827**

Sound Localization

Our method learns to localize sounds even though it was not explicitly trained to do so.



Take-Home Message

- Audio can be used to replace redundant and costly to process video parts for efficient long-range video retrieval.

UNC | COLLEGE OF ARTS AND SCIENCES
Computer Science

Long Movie Clip Classification with State-Space Video Models

Md Mohaiminul Islam

Gedas Bertasius

Motivation

- Applying Transformers to long video sequences is costly.

14*14*60 = 11760 tokens

Motivation

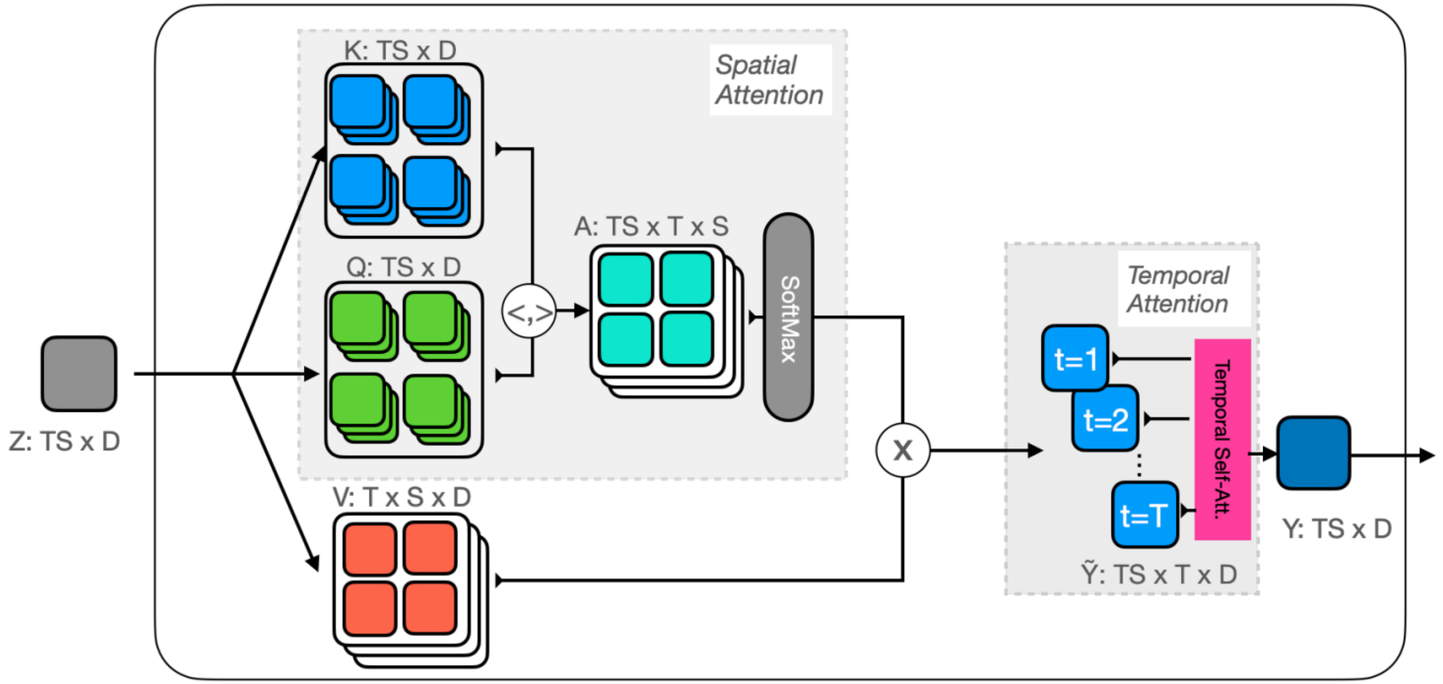
- Applying Transformers to long video sequences is costly.

14*14*60 = 11760 tokens

1.38 Million pair-wise comparison for self-attention operation

Efficient Attention for Video

- Efficient attention schemes typically lead to a substantial drop in accuracy.



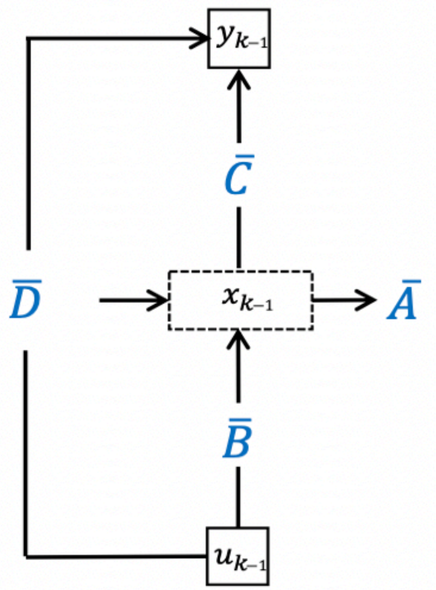
Patrick et al. “Keeping Your Eye on the Ball: Trajectory Attention in Video Transformers,” NeurIPS 2021

Structured State-Space Model

- Structured state-space models can model long-range dependences.

Recurrent Computation

$$x_k = Ax_{k-1} + Bu_k$$
$$y_k = Cx_k + Du_k$$



No need to compute similarities between all pairs of tokens.

Gu et al. “Efficiently Modeling Long Sequences with Structured State Spaces,” ICLR 2022

Long Range Arena

- Structured State-space (S4) model outperforms all prior methods on every task from the Long Range Arena benchmark.

Model	ListOps	Text	Retrieval	Image	Pathfinder	Path-X	Avg
Transformer	36.37	64.27	57.46	42.44	71.40	✗	53.66
Reformer	37.27	56.10	53.40	38.07	68.50	✗	50.56
BigBird	36.05	64.02	59.29	40.83	74.87	✗	54.17
Linear Trans.	16.13	65.90	53.09	42.34	75.30	✗	50.46
Performer	18.01	65.40	53.82	42.77	77.05	✗	51.18
FNet	35.33	65.11	59.61	38.67	77.80	✗	54.42
Nyströmformer	37.15	65.52	79.56	41.58	70.94	✗	57.46
Luna-256	37.25	64.57	79.29	47.38	77.72	✗	59.37
S4	**58.35**	**76.02**	**87.09**	**87.26**	**86.05**	**88.10**	**80.48**

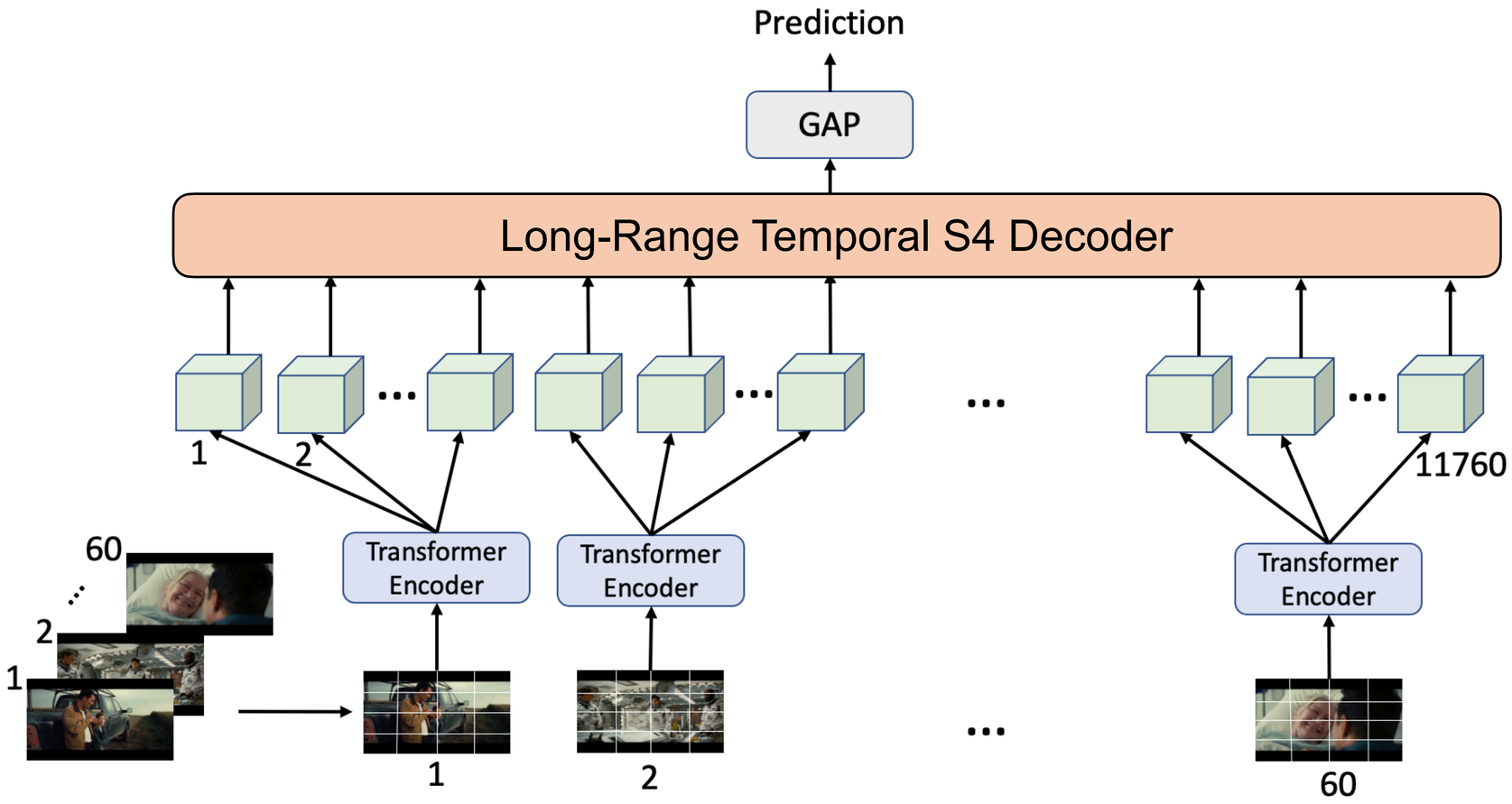
Theoretical Runtime

- Compared to self-attention, structured state-space operation is much more efficient.

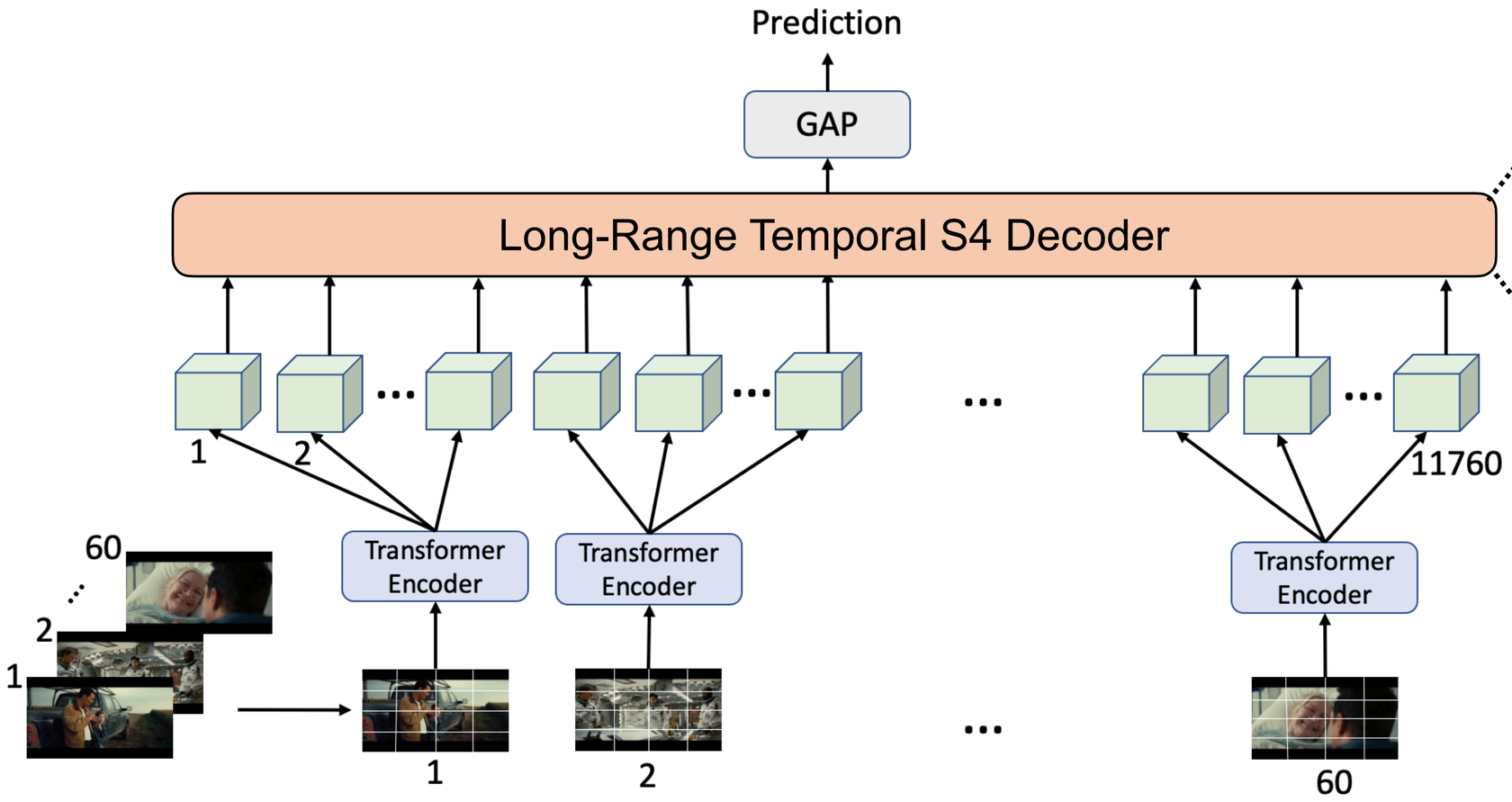
	Attention	State-Space
Memory	$B(L^2 + HL)$	**BLH**

L: Sequence length, B: Batch size, H: Hidden dimension

The ViS4mer Model



The ViS4mer Model



Long-range dependencies can be modeled using a convolution

$$y = \overline{K} * u$$

$$\overline{K} = (\bar{C}\bar{A}^{k}\bar{B})$$

Computational Efficiency

- ViS4mer requires significantly less GPU memory than the Transformer-based models.

Method	GPU Memory Usage (GB)
Transformer	41.38
ViS4mer (**Ours**)	**5.15**

Results on COIN and Breakfast

- ViS4mer achieves competitive results on long-range procedural activity classification datasets.

Model	Pretraning Dataset	Pretraining Samples	Accuracy(↑)
VideoGraph [24]	Kinetics	306K	69.50
Timeception [23]	Kinetics	306K	71.30
GHRM [57]	Kinetics	306K	75.50
Distant Supervision [32]	HowTo100M	**136M**	**89.90**
ViS4mer	Kinetics	495K	88.17

a) Breakfast

Model	Pretraning Dataset	Pretraining Samples	Accuracy(↑)
TSN [44]	Kinetics	306K	73.40
Distant Supervision [32]	HowTo100M	**136M**	**90.00**
ViS4mer	Kinetics	495K	88.41

b) COIN

Comparison with Efficient Attention Schemes

- ViS4mer outperforms other efficient attention schemes on 8 out of 9 movie understanding tasks from the Long Video Understanding (LVU) benchmark.

	Content (↑)			Metadata (↑)				User (↓)			
	Relation	Speak	Scene	Director	Genre	Writer	Year	Like	Views	Sam./s (↑)	Mem (↓)
Self-attention	52.38	37.31	62.79	56.07	52.70	42.26	39.16	0.31	3.83	1.88	41.38
Performer	50.00	38.80	60.46	58.87	49.45	48.21	41.25	0.31	3.93	4.67	5.93
Orthoformer	50.00	39.30	66.27	55.14	**55.79**	47.02	43.35	0.29	3.86	4.85	5.56
State-space	**57.14**	**40.79**	**67.44**	**62.61**	54.71	**48.8**	**44.75**	**0.26**	**3.63**	**4.95**	**5.15**

Take-Home Message

- Combining the strengths of self-attention and structured state-space allows us to build a memory-efficient long-range video classification model.

Questions?