HERMES: temporal-coHERent long-forM understanding with Episodes and Semantics

ICCV 2025

HERMES: temporal-co**HER**ent long-for**M** understanding with **E**pisodes and **S**emantics

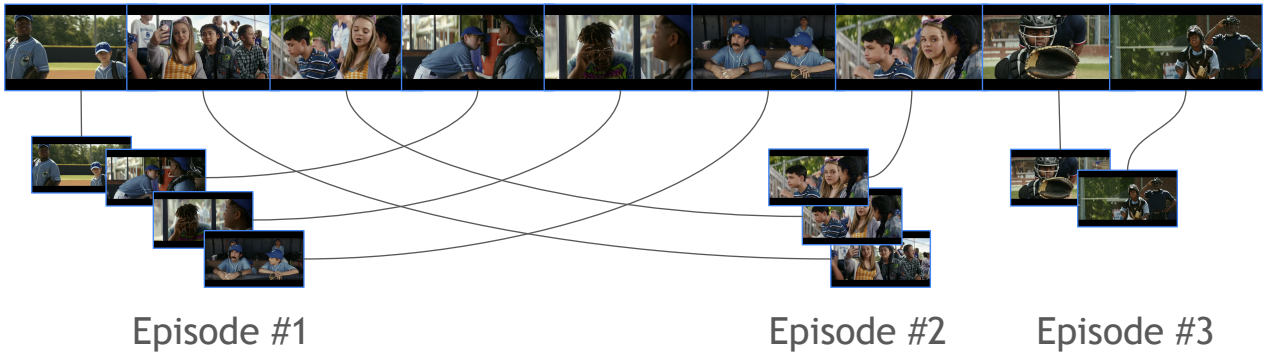
In Video Understanding, we have always traded accuracy for efficiency...

Not anymore.

Introducing **HERMES**.

HERMES: temporal-coHERent long-forM understanding with Episodes and Semantics

Given a video such as the one below, **HERMES** does two things:



1st, it combines similar scenes into **contiguous episodes**

2nd, it **identifies key semantics** scattered throughout the video

HERMES: temporal-coHERent long-forM understanding with Episodes and Semantics

Given a video such as the one below, **HERMES** does two things:

1. 1st, it combines similar scenes into **contiguous episodes**
2. 2nd, it **identifies key semantics** scattered throughout the video
3. It then **combines the Episodes and the Semantics**

HERMES achieves SOTA results in several benchmarks while being much faster

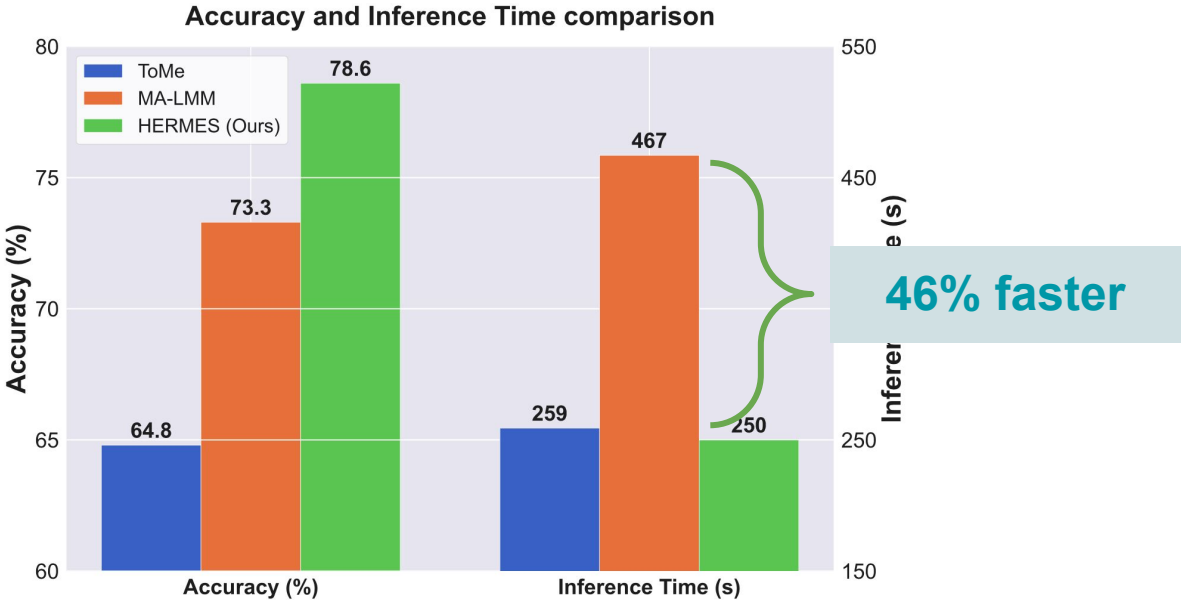
Model	LVU								Breakfast	COIN
	Content			Metadata				Avg		
	Relation	Speak	Scene	Director	Genre	Writer	Year			
FACT [21]	-	-	-	-	-	-	-	-	86.1	-
Obj. Transformer [43]	53.1	39.4	56.9	52.1	54.6	34.5	39.1	47.1	-	-
VIS4mer [15]	57.1	40.8	67.4	62.6	54.7	48.8	44.8	53.7	88.2	88.4
TranS4mer [16]	59.5	39.2	70.9	63.9	55.9	46.9	45.5	54.5	90.3	89.2
S5 [41]	67.1	42.1	73.5	67.3	65.4	51.3	48.0	59.2	90.7	90.8
Movies2Scenes [5]	**71.2**	42.2	68.2	70.9	57.8	55.9	53.7	60.0	-	-
MA-LMM [14]	58.2	44.8	80.3	74.6	61.0	70.4	51.9	63.0	93.0	93.2
HERMES (Ours)	67.6	**47.5**	**90.0**	**82.6**	**69.5**	**77.2**	**57.7**	**70.3**	**95.2**	**93.5**

+7.3% Accurate

Model	Acc.	Time	Mem. (GB)
LongVA (7B)	54.11	1	42.5
+ ECO	54.19	0.700 (-30%)	**22.9**
+ SeTR	**54.56**	0.726 (-27%)	32.7

46% less GPU Usage

HERMES achieves **SOTA** results in several benchmarks while being **much faster**



HERMES is open-source, feel free to use it yourself

Code

Project Page