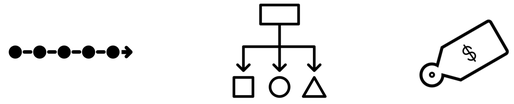
Learning Video Representations from Large Language Models

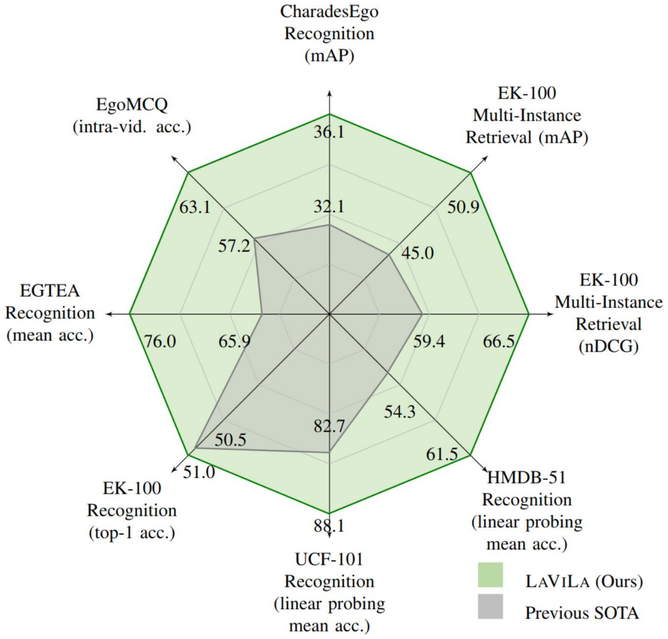
Yue Zhao[1,2], Ishan Misra[2], Philipp Krähenbühl[1], Rohit Girdhar[2]
[1]UT Austin [2]Meta AI

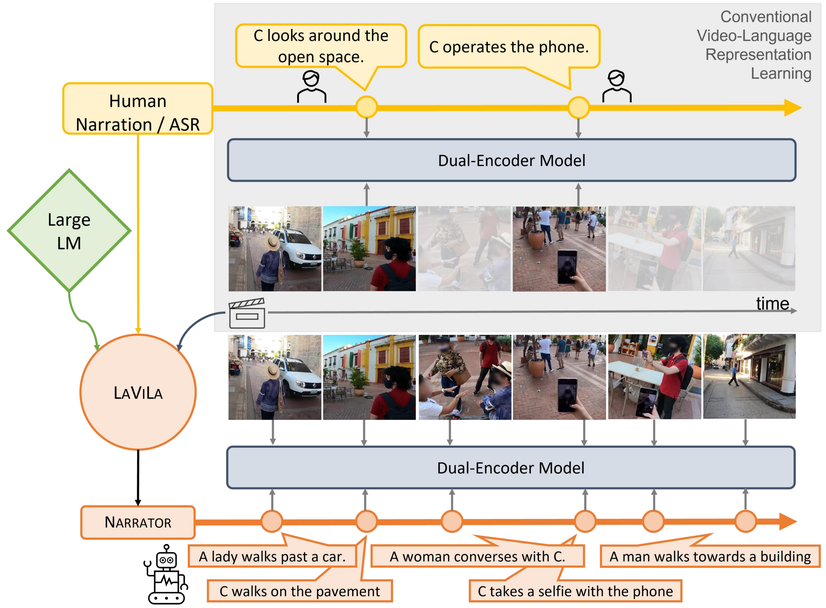
JUNE 18-22, 2023 CVPR VANCOUVER, CANADA

Our Approach: LAVILA

- Goal: Learn Video-Text Representation through Pre-training
- Language-model augmented Video-Language pre-training
  - LLM + visual condition => Narrators (video captioning)

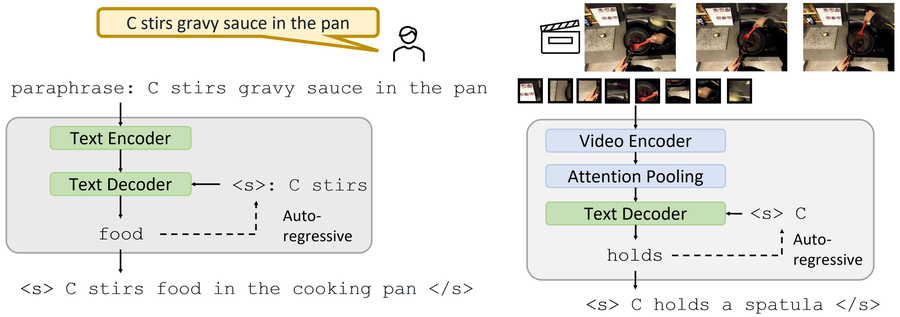




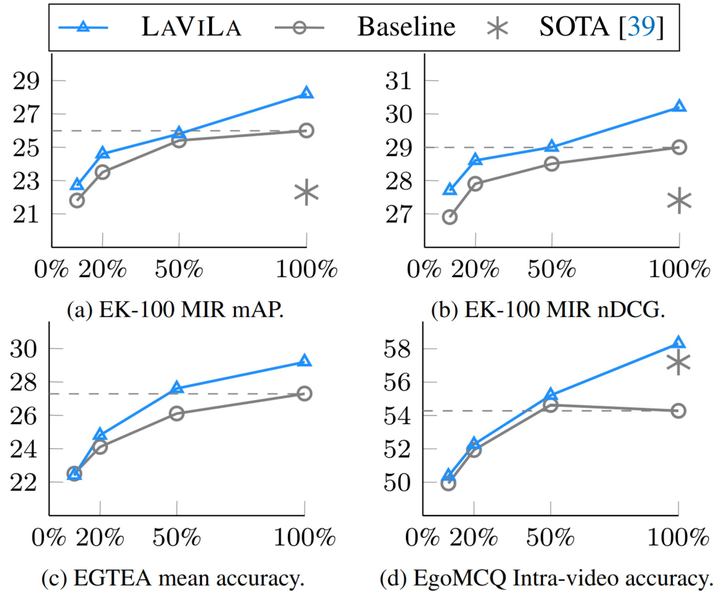


Language supervisions

- REPHRASER: narration -> rephrased narration
- NARRATOR: video clip -> pseudo-captioned narration



Data Scaling



Model Scaling