



Better and Longer Video Understanding

WENHAO CHAI

25 Aug, 2025

Wenhao Chai

Wenhao Chai is a first-year Ph.D. Student in Computer Science at Princeton University, working with Prof. Zhuang Liu. He received his master's degree at University of Washington and bachelor's degree at Zhejiang University. He previously studied at Stanford University, working with Prof. Christopher D. Manning. He has internship at Pika Labs and Microsoft Research Asia. His research spans a wide range of topics in machine learning and computer vision.



Overview

- 01 What is Video Understanding?
- 02 **SAMURAI**: Before LLM
- 03 **MovieChat**: Long-form Video
- 04 **VDC**: Detailed Video Captioning
- 05 **VideoMMLU**: Lecture Understanding
- 06 **AuroraLong**: Bring RNN Back



PRINCETON
UNIVERSITY

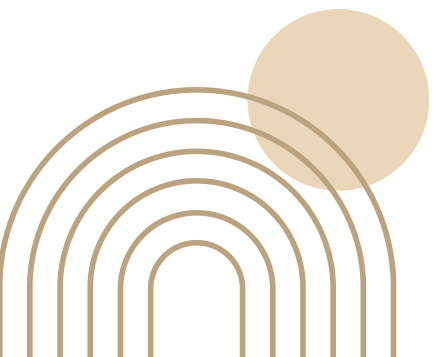


What is Video Understanding

Video is a time-ordered stream of images, typically 24–60 frames per second, so it carries far more information than a single image: motion, causality, and temporal context.

Video understanding is broadly useful across domains: sports analytics, autonomous driving and drones, robotics, healthcare (surgery/endoscopy), security, retail/ads, media recommendation, education, and AR/VR.

In the past, a typical video understanding model could only handle a single task, such as detecting anomalies, classifying human actions, or tracking object motion.



SAMURAI



One of our work on traditional video understanding

SAMURAI: Adapting Segment Anything Model for Zero-Shot Visual Tracking with Motion-Aware Memory

- **7,000 Stars on GitHub in 180 days**

The most popular tracking project ever.

- **State-of-the-Art Performance**

LaSOT, GOT-10k, TrackingNet, NFS, OTB100, etc.

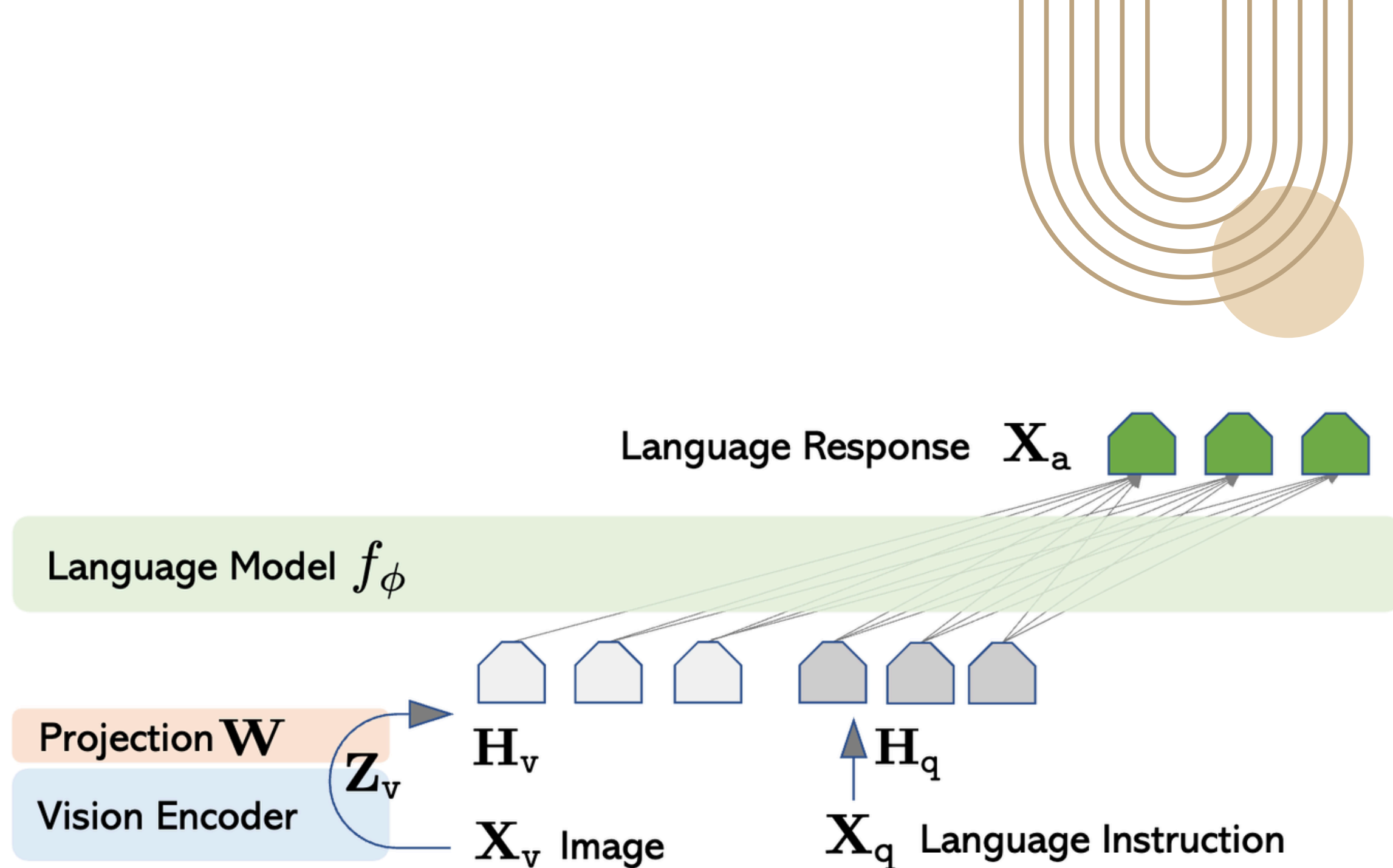
- **50 Citations in 180 days**


- **What's next for video understanding?**
with LLM!

Video LLMs

How We Connect?

- Connect ViT and LLM
- Adapt from Image LLMs
- Handle longer sequences
- May need more compute
- But less data





Some of our work on LLM-based video understanding

Long ● **MovieChat: From Dense Token to Sparse Memory for Long Video Understanding** Computer Vision and Pattern Recognition (CVPR), 2024
MovieChat+: Question-aware Sparse Memory for Long Video Question Answering IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI), 2025

Detail ● **AuroraCap: Efficient, Performant Video Detailed Captioning and a New Benchmark** International Conference on Learning Representations (ICLR), 2025

Knowledge ● **Video-MMLU: A Massive Multi-Discipline Lecture Understanding Benchmark** International Conference on Computer Vision (ICCV) Workshop @ Findings

Efficient ● **AuroraLong: Bringing RNNs Back to Efficient Open-Ended Video Understanding** International Conference on Computer Vision (ICCV), 2025

MovieChat

First ever video understanding system that can take over 10,000 frames as input.

Long-form Video

hours / 10,000 frames

Vision Encoder

frame / clip level

Short-term Memory

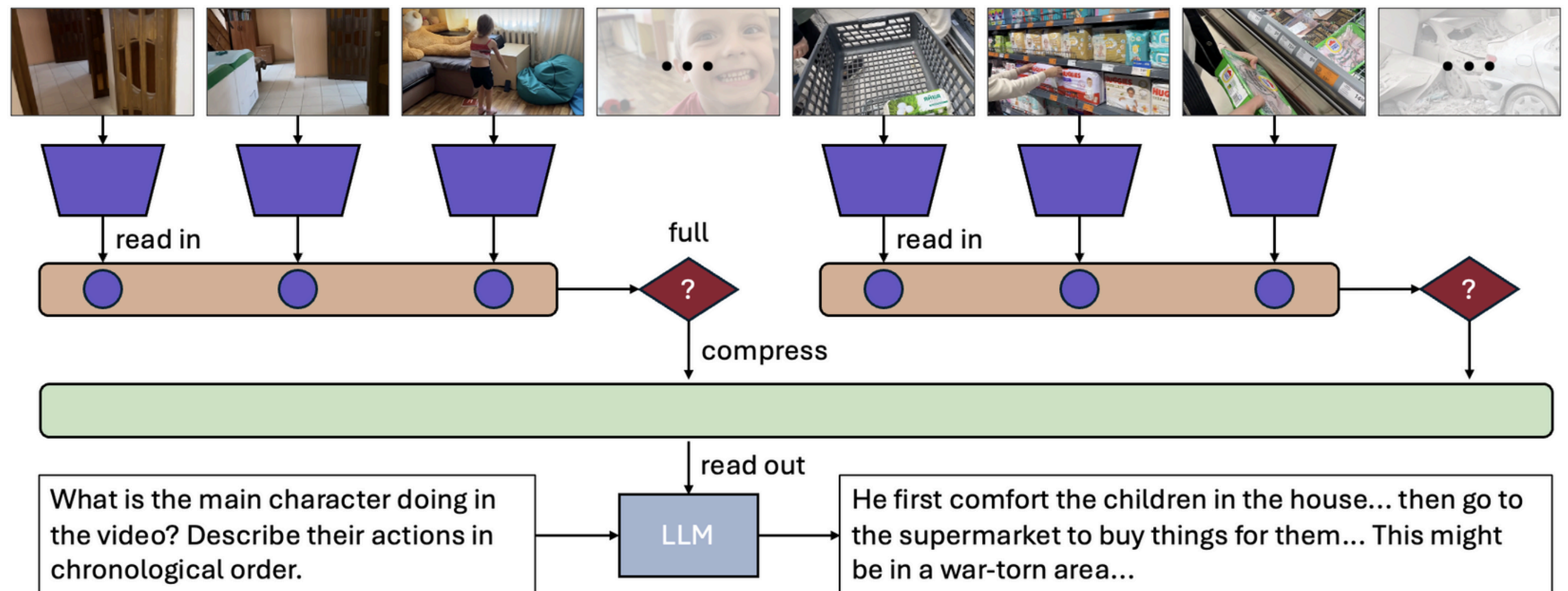
limited stack

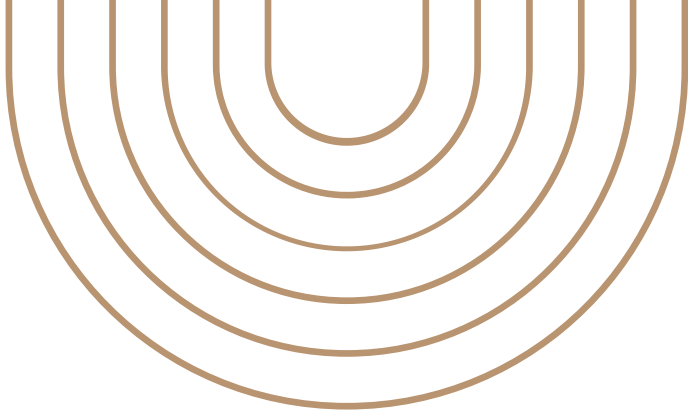
Long-term Memory

unlimited set

LLM Reasoning

text question and answer





MovieChat

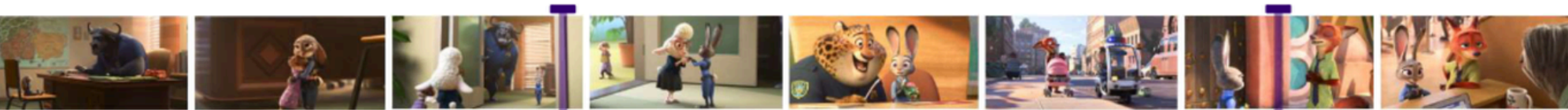
First ever long-form video understanding challenge hosted in CVPR 2024.

LOVEU@CVPR'24

Home · Program · **Track1** · Track2

Track 1: Long-Term Video Question Answering

- This track aims at encouraging our participants to advance long-term video understanding system.
- The competition is based on [MovieChat](#) test set only.
- Top 3 winners will be mentioned at the workshop and formally recognized.



Q#1 at 3930

Q#2 at 11130

Question #1

What is Chief Bogo doing?

Answer #1

In the video, the character Chief Bogo is seen talking to someone on a laptop while sitting at a table in a dark room. Chief Bogo is a cartoon character wearing a brown outfit.

Question #2

Who is this rabbit named Judy talking to?

Answer #2

In the video, the rabbit police named Judy is talking to the person in a movie theater, who is a fox.

Question #3

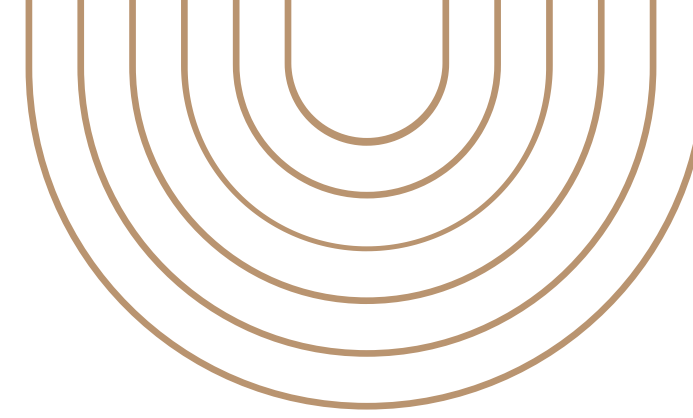
Can you describe this video?

Answer #3

The video shows various scenes of an animation with different animals in a restaurant or office environment. At the beginning, we see an animated cartoon rabbit and a deer talking to each other in a restaurant. Next, there is a cartoon bunny talking to a cat at a desk, followed by a scene of a cartoon deer and a mouse talking to each other. The video then continues with the animated cartoon dog and cat sitting at a table in a restaurant, and another scene where a man and a dog are in a room with a computer on a table. There are more scenes of the same characters later on, including one with a bunny and a fox sitting at a desk, and another with a bunny and a dog talking to each other in the background. Finally, there are several more scenes with different animals, including a fox, a mouse, and a cat, all of which are in a different environment.

Model with Link	Comment	Breakpoint Acc	Global Acc
Video-LLaMA	End-to-end	39.1	51.7
VideoChat	End-to-end	46.1	57.8
TimeChat	CoT, ICL, train on MovieChat	46.1	73.8
VideoChatGPT	End-to-end	48.0	47.6
MovieChat (baseline)	End-to-end	48.3	62.3
MovieChat+ (baseline)	End-to-end	49.6	71.2
Long-LLaVA	Eng-to-end	54.0	69.6
Long-LLaVA + Video-RAG	Eng-to-end	54.5	72.9
Streaming Long Video	Train on MovieChat	54.9	90.4
DrVideo	RAG	56.7	93.1
ReWind	End-to-end	57.2	87.6
HERMES	Train on MovieChat	57.3	78.6
Flash-VStream	Train on MovieChat	59.6	96.0
MM-Screenplayer	RAG	68.8	87.5

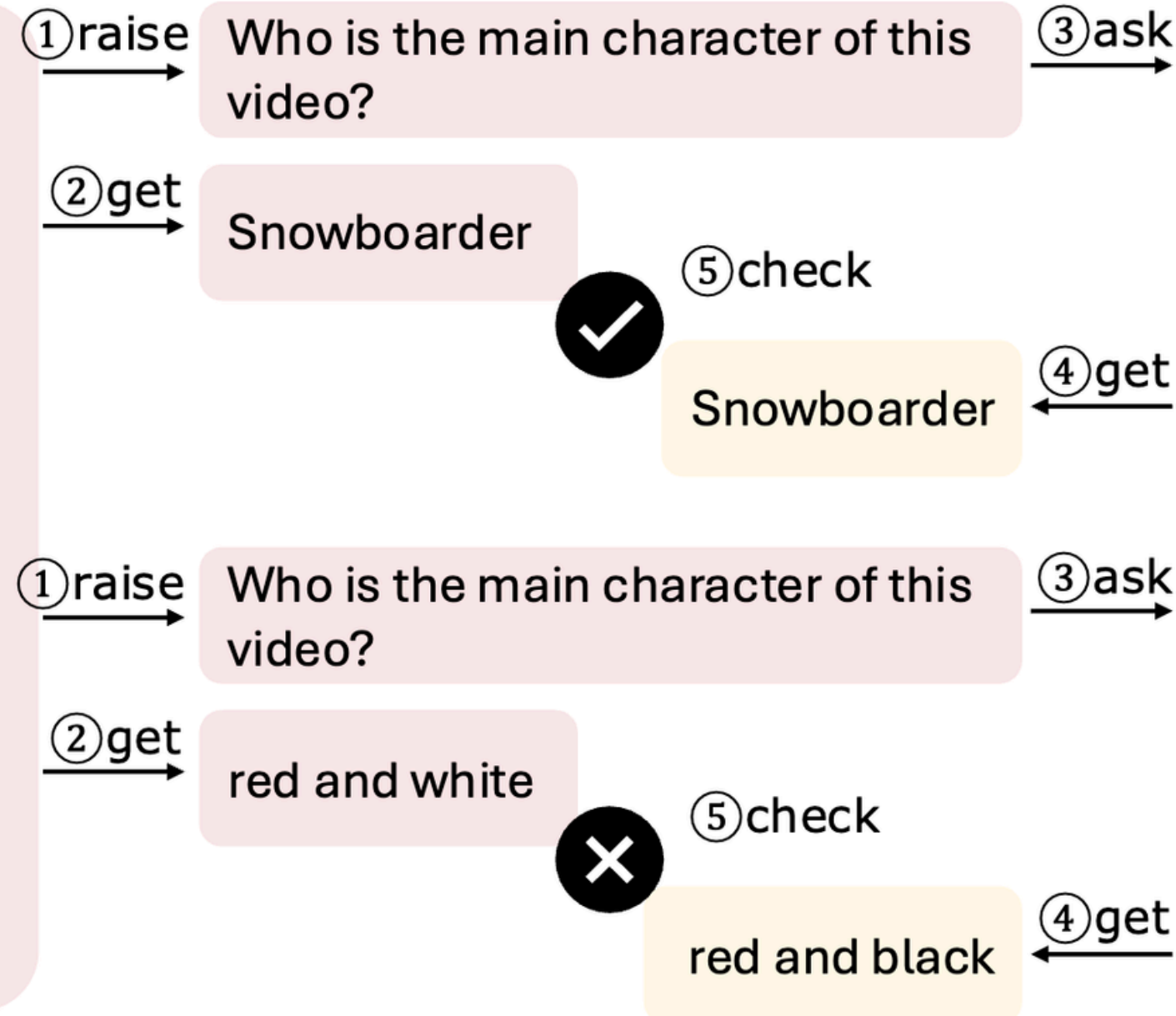
VDC



First ever evaluation system for detailed video captioning.

GT caption

The video showcases an exhilarating moment as a snowboarder soars through the air, executing a stunning trick. Dressed in a bold red and white jacket, black pants, and a protective helmet. The backdrop to this action-packed scene is a breathtaking snowy mountain landscape. The mountain's peak is visible in the distance. The overall composition of the video suggests a high-speed descent down the mountain ...



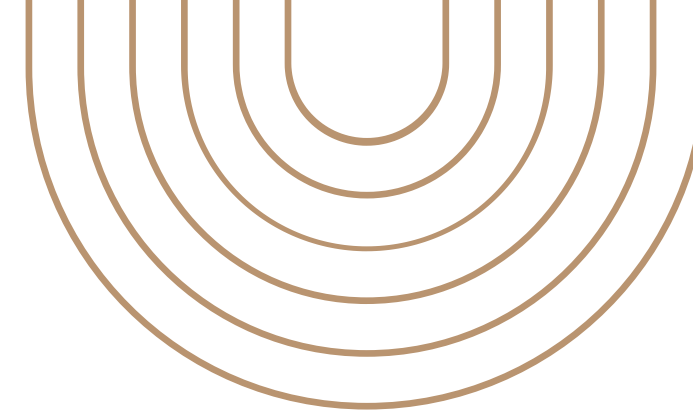
generated caption

The video captures a thrilling moment of a snowboarder in mid-air, performing an impressive trick. The snowboarder, clad in a vibrant red and black jacket, black pants, and a protective helmet. The snowboarder is holding onto a rope with one hand, suggesting that they are being pulled up the mountain by a snowmobile, a common practice in snowboarding to gain speed and momentum ...

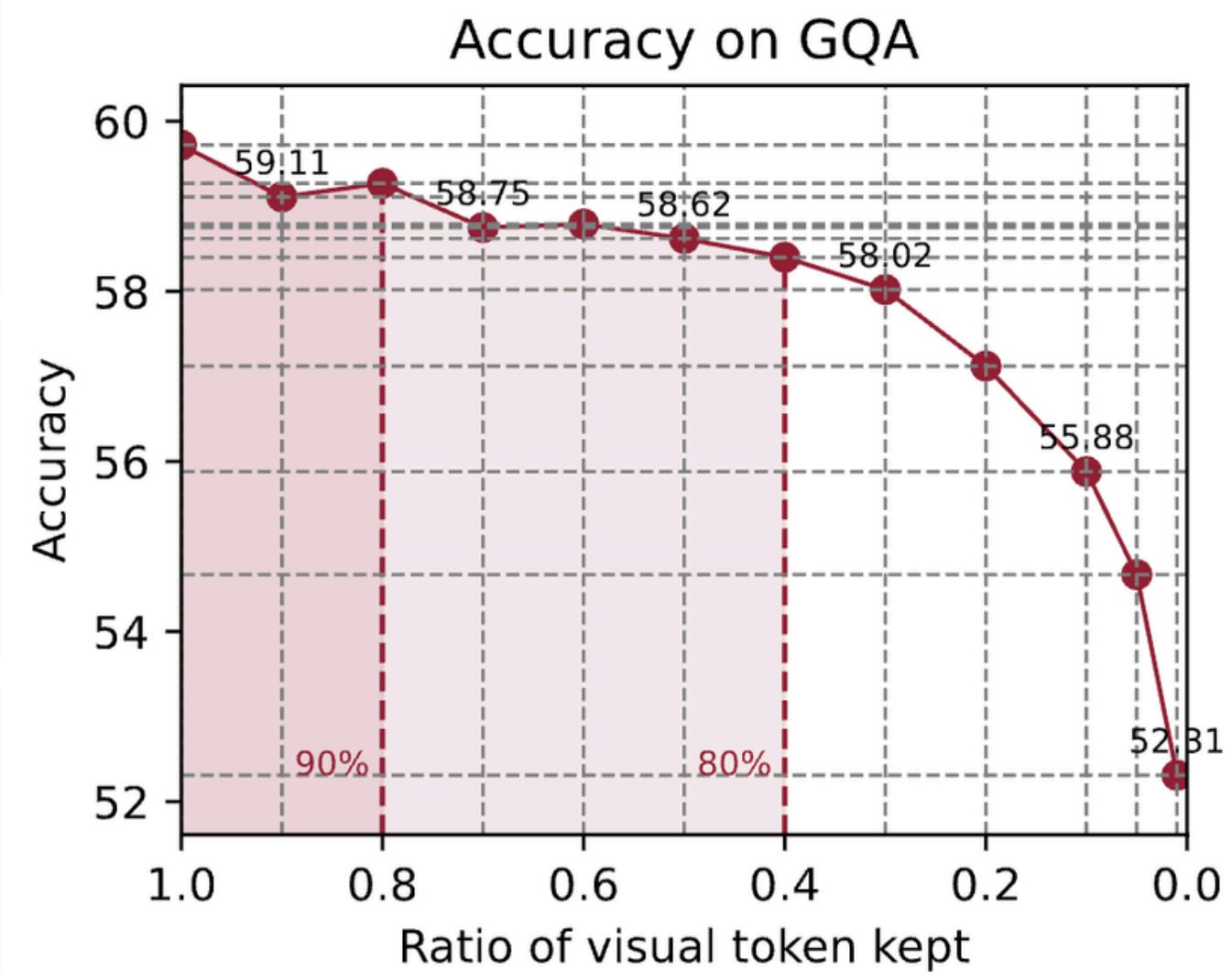
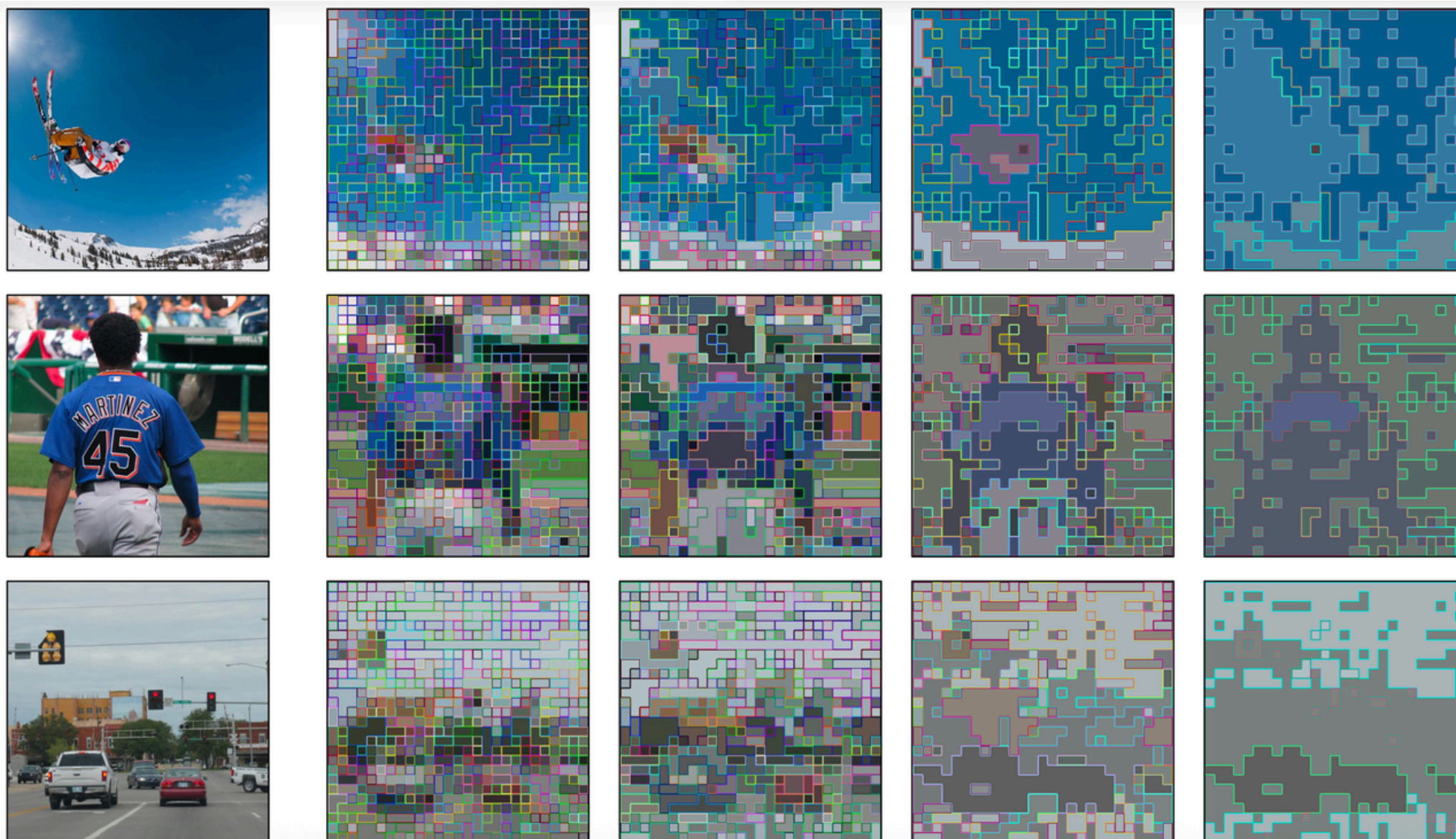
VDC with AuroraCap

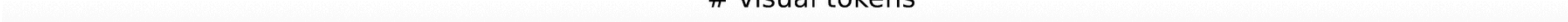


AuroraCap

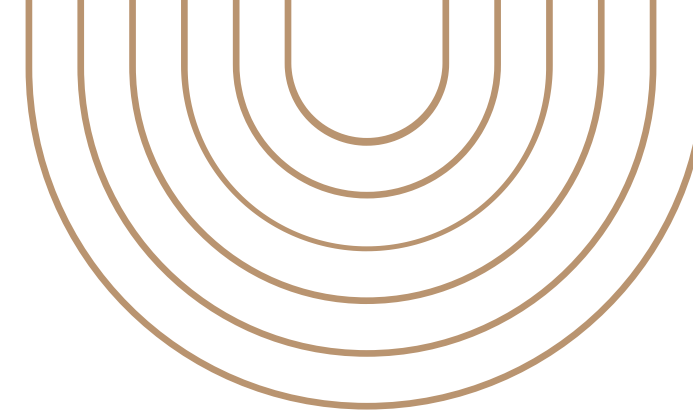


5% tokens but 90% performace.





VideoMMLU



Video-MMLU

A Massive Multi-Discipline Lecture Understanding Benchmark

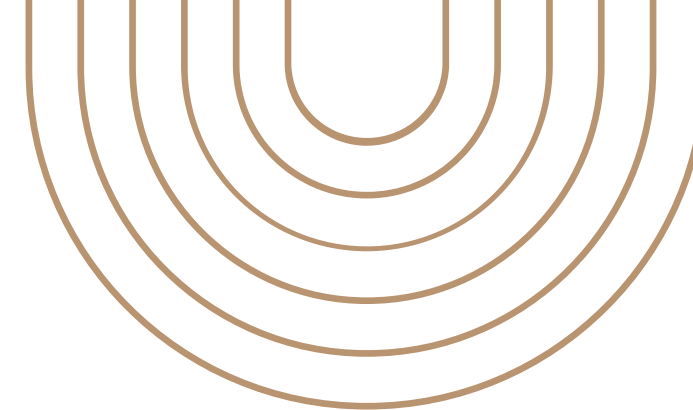
From Law of Sines
 $\sin(2\theta) = 2 \sin(\theta) \cos(\theta)$

Substitution Rule:
 $\int f(g(x))g'(x) dx = \int f(u)du = F(u) + C$

Example: $\int \frac{e^{2x}}{1+e^{2x}} dx$
 $u = 1 + e^{2x}$
 $du = 2e^{2x}dx$
 $f(u) = \frac{1}{u}$
 $F(u) = \ln|u|$

$a^3 - b^3$

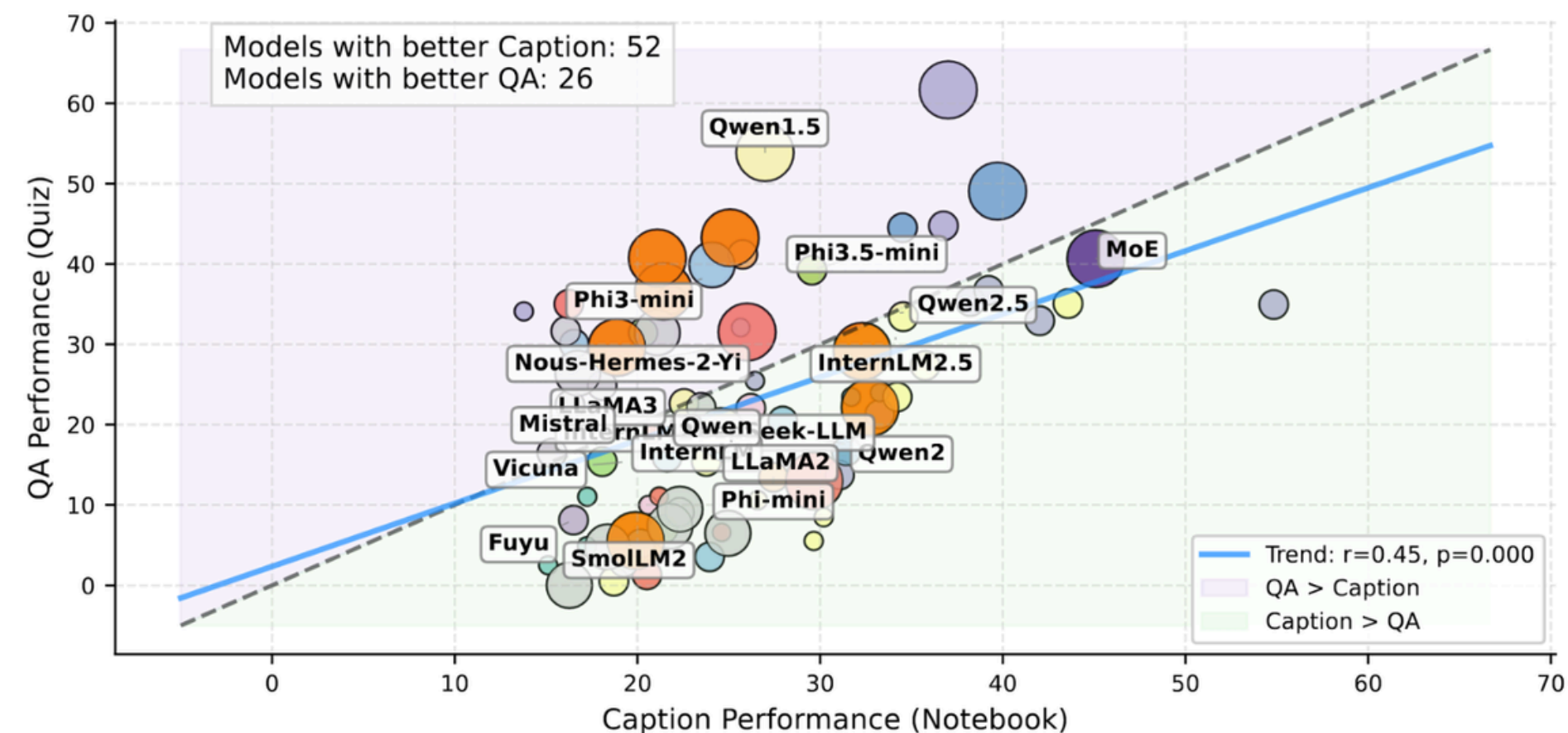
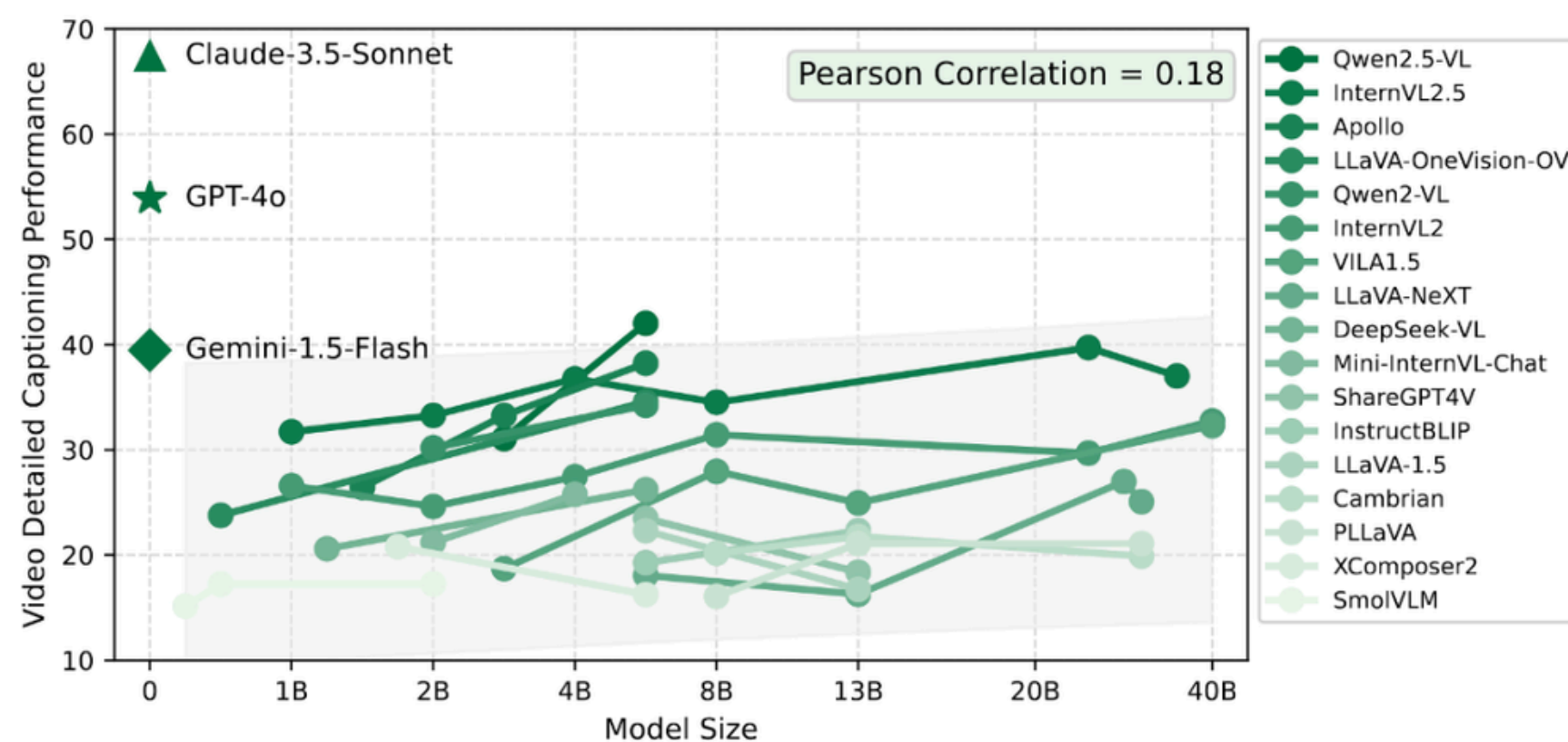
VideoMMLU



Can video LLMs really understand real-world lectures? NOT YET.

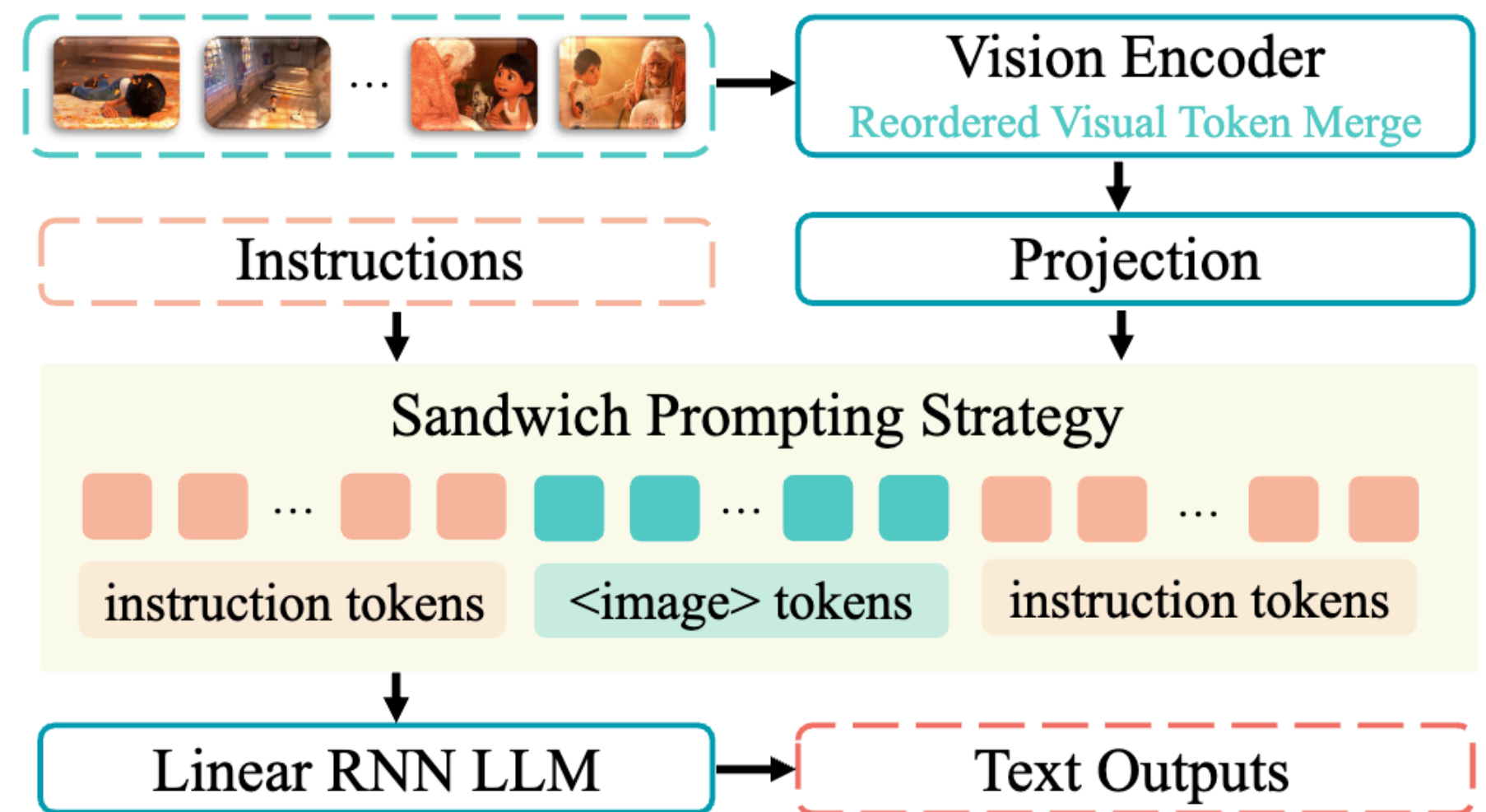
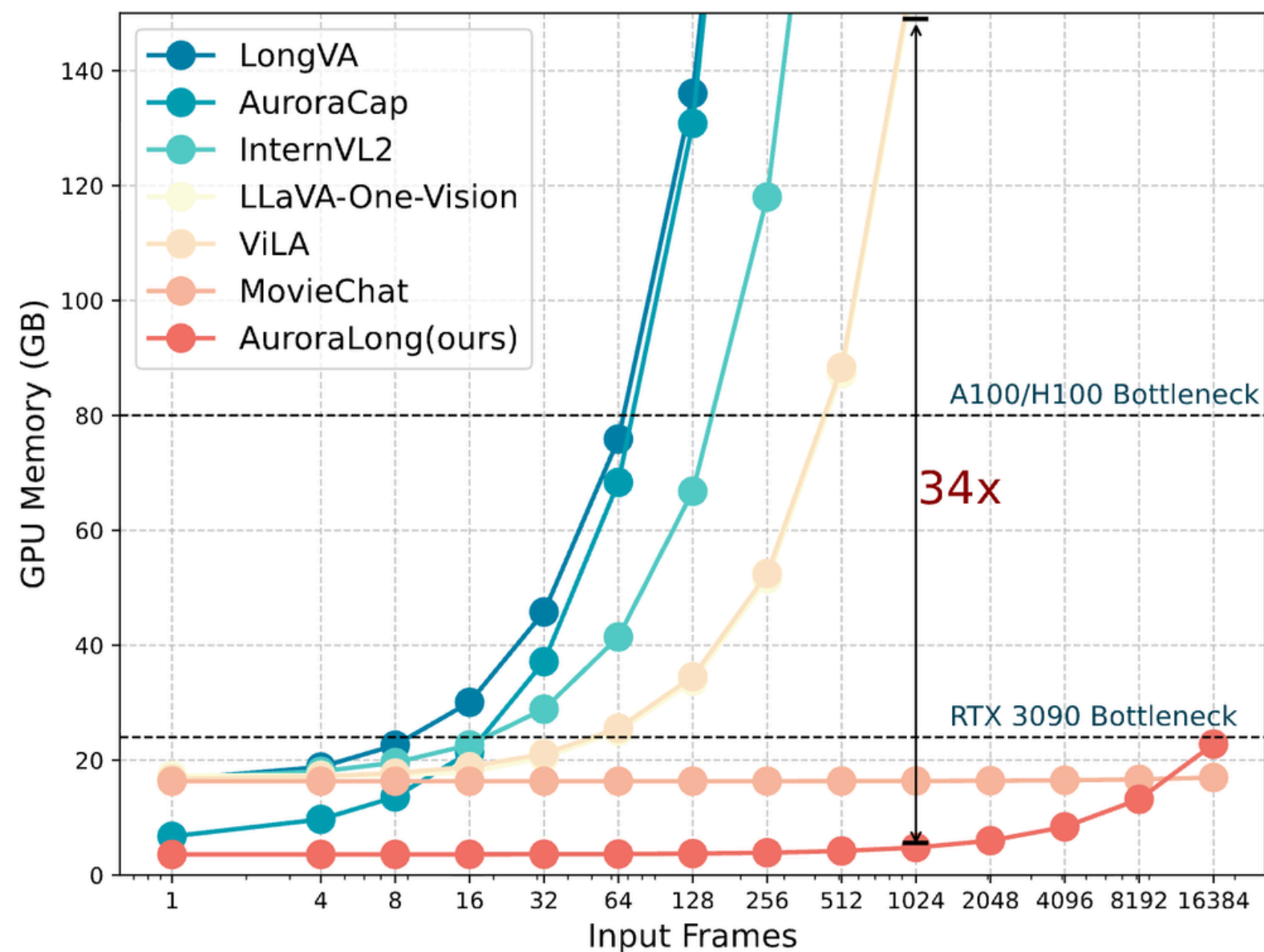
Finding 1. Large scale LMMs do not show clear advantages over smaller ones.

Although LMM scaling laws suggest significant performance improvements with increased model size, this trend is less pronounced in Video-MMLU. Model size shows a stronger correlation with performance in video QA compared to video captioning, implying reasoning benefits more from scaling.



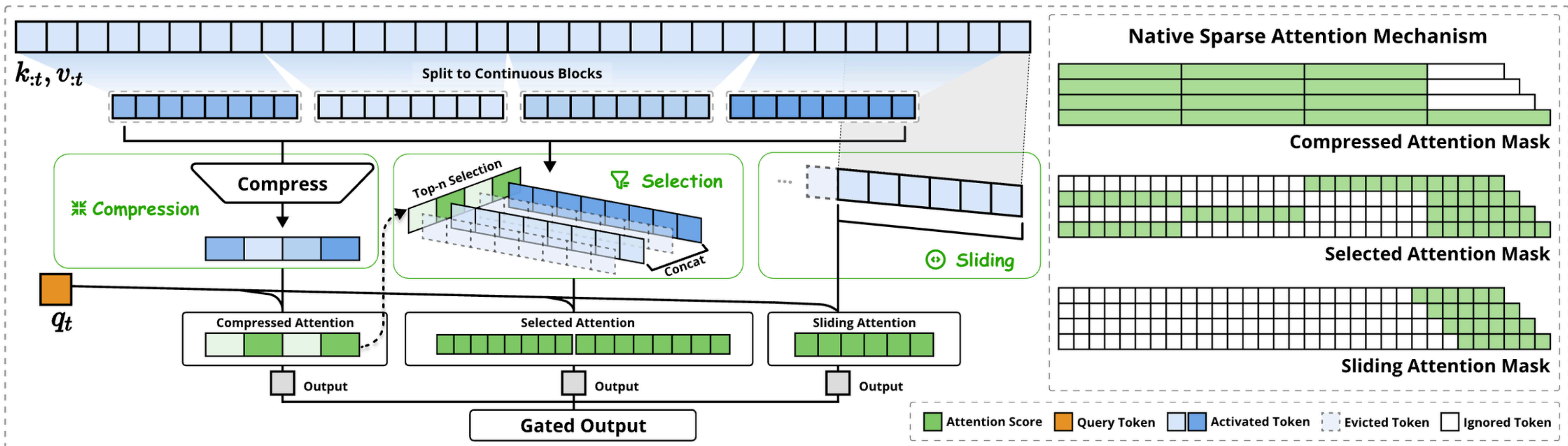
AuroraLong

RNN can beat Transformers in Video Understanding.



Something Ongoing

Can Sparse Attention performs well in Video Understanding? (Figure: DeepSeek NSA)





THANK YOU

Wenhao Chai