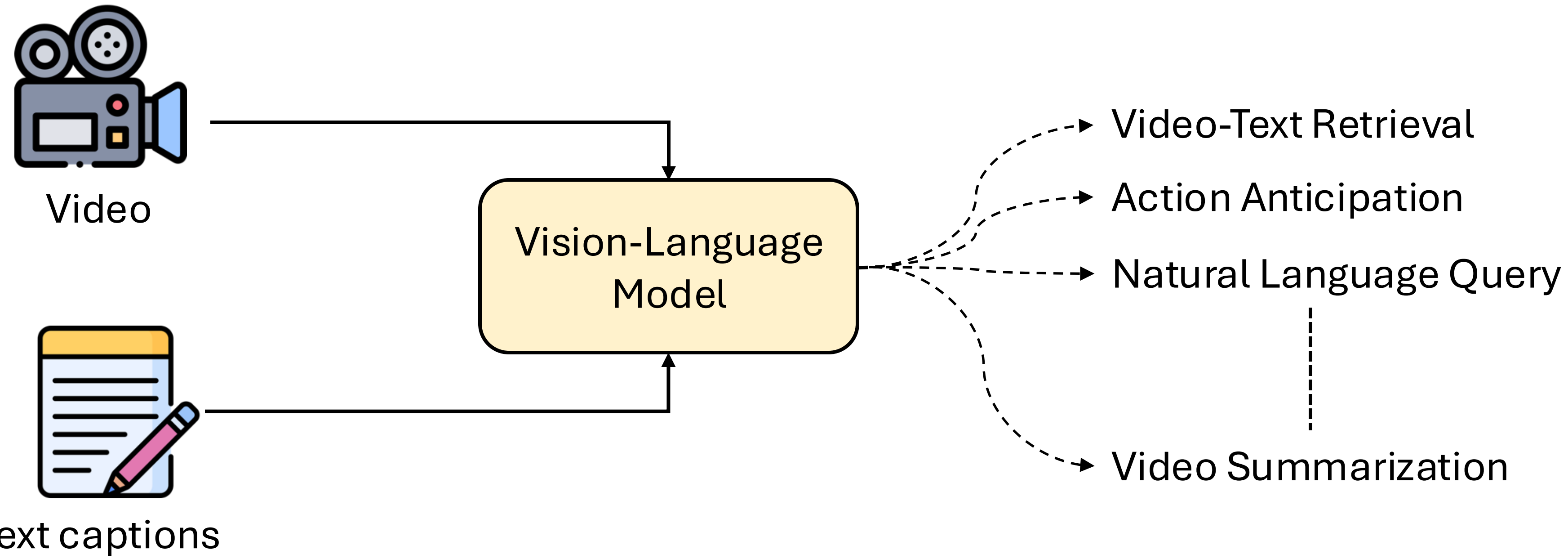
Mamba Fusion: Learning Actions Through Questioning

Zhikang Dong*, **Apoorva Beedu***, Jason Sheinkopf, Irfan Essa

Georgia Institute of Technology

https://github.com/Dongzhikang/MambaVL

Vision-Language Models

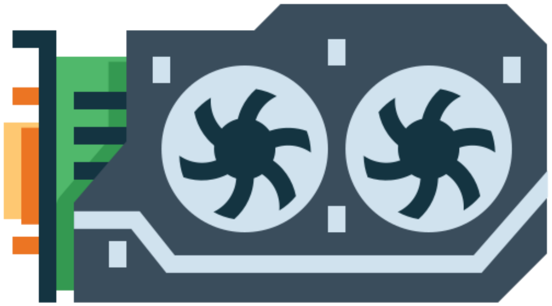


- Captions describing the video contents are used for training.
- Transformer-based architectures for both video and text encoding and fusion.

Pramanick, Shraman, et al. "Egovlpv2: Egocentric video-language pre-training with fusion in the backbone." Proceedings of the IEEE/CVF International Conference on Computer Vision. 2023.
Zhao, Yue, et al. "Learning video representations from large language models." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2023.

Challenges Faced by Transformers

Quadratic computational complexity



High GPU memory usage

Difficulty capturing long-term dependencies

Selective State Space Models like Mamba are a solution!

Gu, Albert, and Tri Dao. "Mamba: Linear-time sequence modeling with selective state spaces." arXiv preprint arXiv:2312.00752 (2023).

Captions for Action Recognition?

- Captions should not contain the actions for end-to-end training!
- LLM based caption generation does not provide the right context!

You are an expert at caption generation. Describe the action “Open Door” without using the words “Open” or “Door”. Provide 10 examples.

1. Grasp the handle, apply pressure, and shift the barrier aside to reveal an entryway.
2. Turn the knob, push or pull, and make way for movement through the passage.
3. Unlatch the panel, move it out of the way, and step through the opening.
4. Push against the wooden slab, allowing space to pass through.
5. Rotate the handle, displace the obstruction, and access the next area.
6. Release the latch, shift the divider, and enter the adjacent space.
7. Pull on the frame, creating an opening to step forward.
8. Apply force to the entryway’s barrier, making room to pass.
9. Press against the surface, moving it aside to clear the way.
10. Twist the lever, slide or swing the partition, and proceed through.

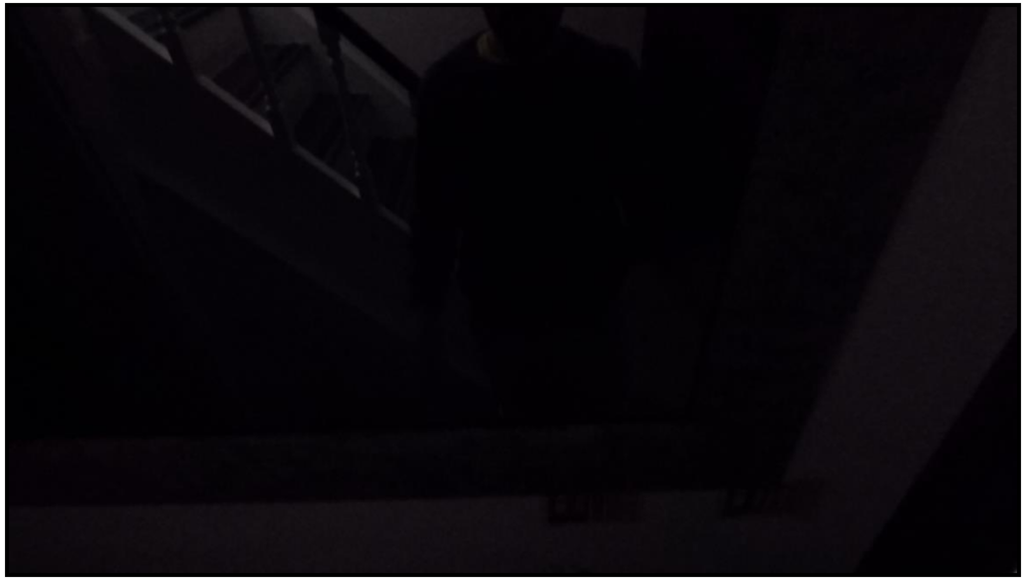
Captions aren’t suitable for action recognition!

Question-Answers are the Solution?

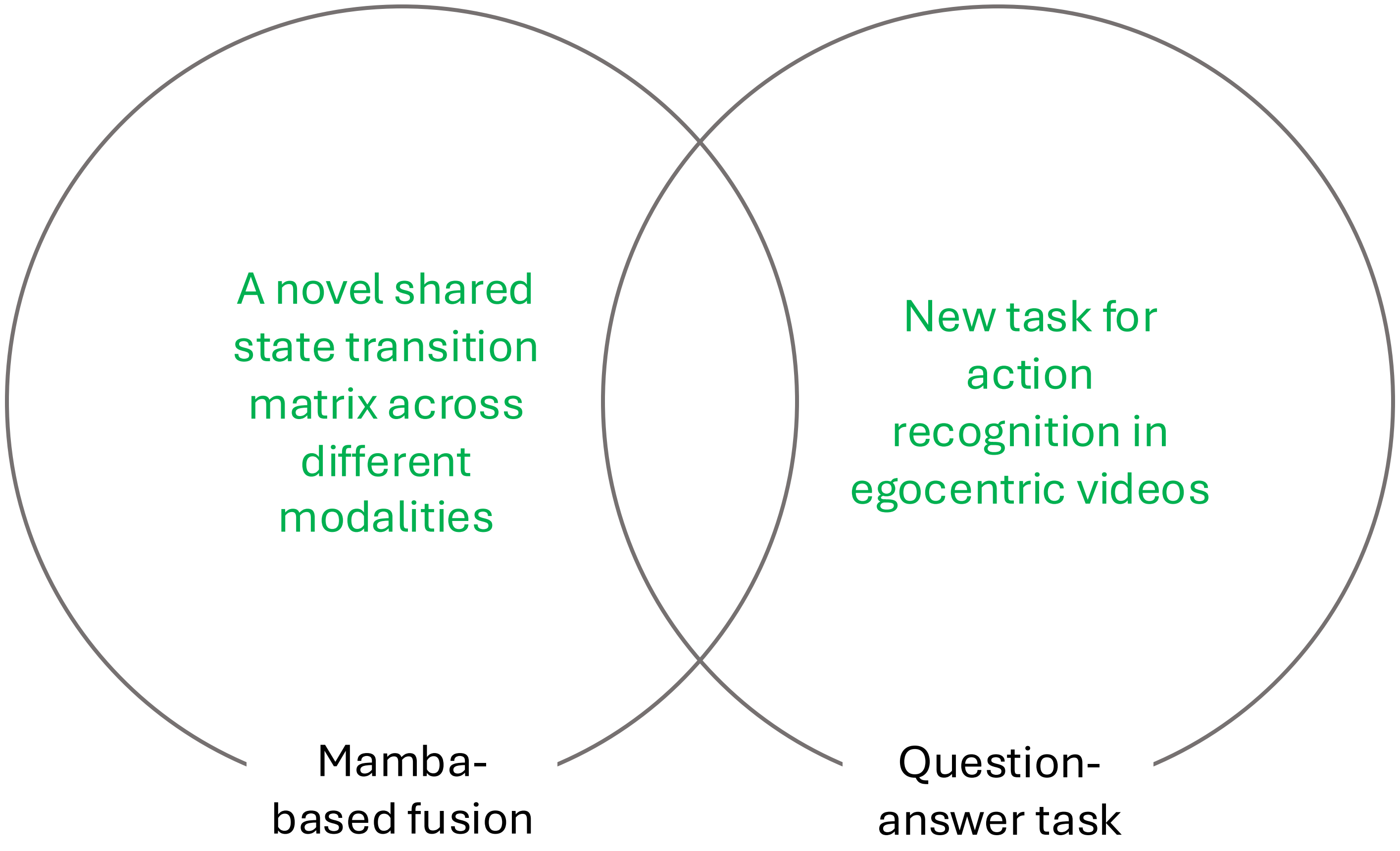
- Motivation: Question facilitate deeper reasoning than simple captions.
- Each action represented by two questions

  “What action is performed on the door by pulling on the handle with your hand?” (Answer: Open)

  “What object is being opened by pulling on the handle with your hand?” (Answer: Door)



MambaVL



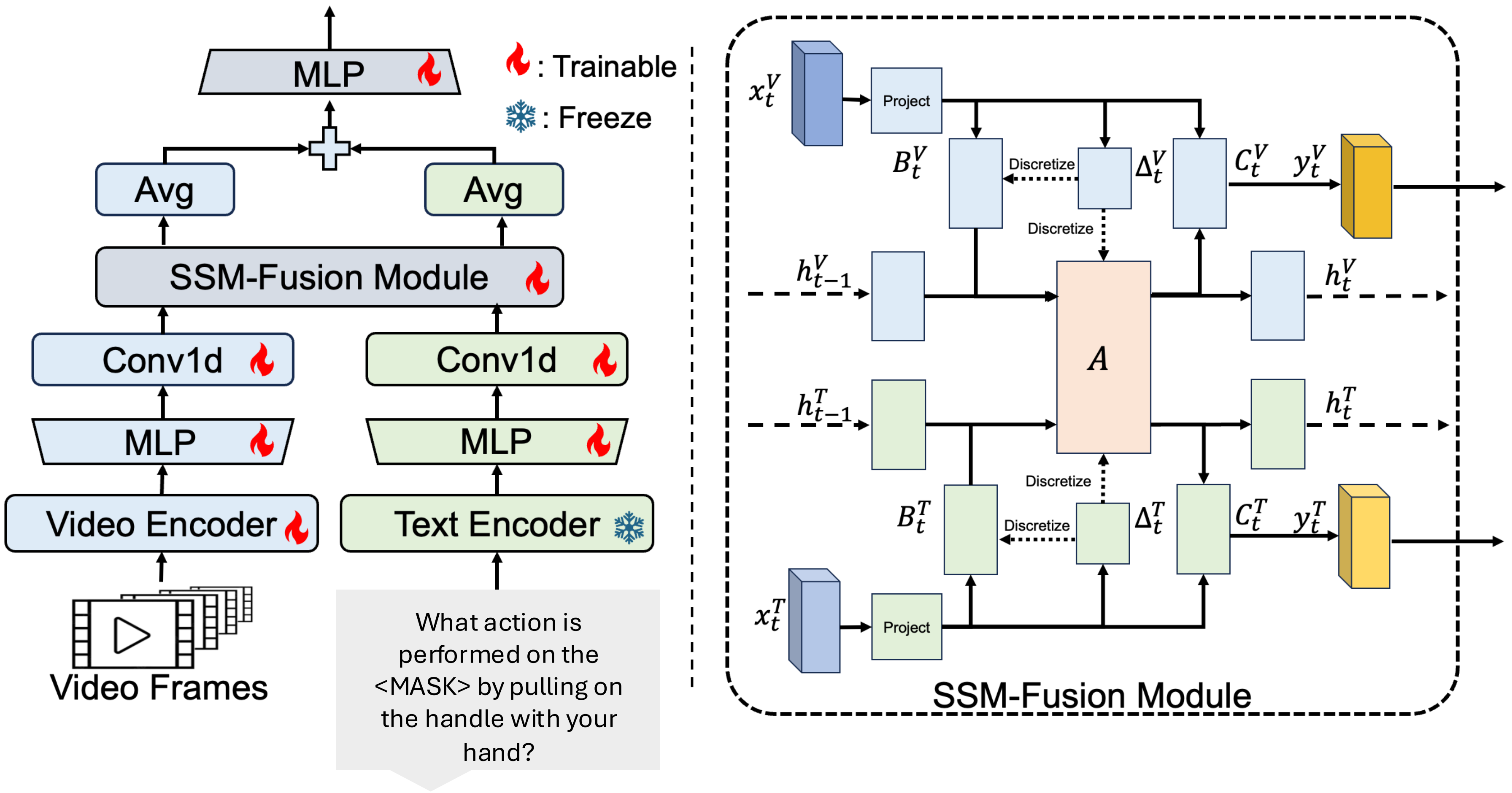
A novel shared state transition matrix across different modalities
New task for action recognition in egocentric videos
Mamba-based fusion
Question-answer task

Generating Question-Answer Pairs

You are provided with a noun, a verb, and a description of an egocentric video. Generate two questions based on this description:
1. Formulate a question that incorporates the given verb, asking which object (specified by the provided noun) is involved in the described action. The correct answer to this question should be the noun.
2. Construct a question that includes the given noun, inquiring about the action (specified by the provided verb) that is performed on this object. The correct answer to this question should be the verb.
Ensure that your questions are diverse but adhere to these guidelines. Separate your questions with a '\n' without using numbers or additional punctuation.
Some Examples:
Example 1:
Noun: 'dog', Verb: 'run', Description: 'A dog runs across the park.'
Questions:
What animal is running across the park?
Which activity is the dog performing in the park?
Example 2:
Noun: 'car', Verb: 'park', Description: 'A car is parked beside the road.'
Questions:
Identify the object that is parked beside the road?
What is the car doing beside the road?

Given noun: {noun}, verb: {verb}, description: {gpt_desc}. Your generated questions:

MambaVL Structure



MambaVL Structure

Algorithm 1 SSM-Fusion Module

Require: $\mathbf{x}^V$: $(\mathrm{B}, \mathrm{F}, \mathrm{D})$, $\mathbf{x}^T$: $(\mathrm{B}, \mathrm{L}, \mathrm{D})$
Ensure: $\mathbf{y}^V$: $(\mathrm{B}, \mathrm{F}, \mathrm{D})$, $\mathbf{y}^T$: $(\mathrm{B}, \mathrm{L}, \mathrm{D})$

$\boldsymbol{A}$: $(\mathrm{D}, \mathrm{N}) \leftarrow$ Parameter
$\boldsymbol{B}^V$: $(\mathrm{B}, \mathrm{F}, \mathrm{N}) \leftarrow \mathrm{Linear}^V_{\mathrm{B}}(\mathbf{x}^V)$
$\boldsymbol{B}^T$: $(\mathrm{B}, \mathrm{L}, \mathrm{N}) \leftarrow \mathrm{Linear}^T_{\mathrm{B}}(\mathbf{x}^T)$
$\boldsymbol{C}^V$: $(\mathrm{B}, \mathrm{F}, \mathrm{N}) \leftarrow \mathrm{Linear}^V_{\mathrm{C}}(\mathbf{x}^V)$
$\boldsymbol{C}^T$: $(\mathrm{B}, \mathrm{L}, \mathrm{N}) \leftarrow \mathrm{Linear}^T_{\mathrm{C}}(\mathbf{x}^T)$
Δ^V: $(\mathrm{B}, \mathrm{F}, \mathrm{D}) \leftarrow \mathrm{Softplus}(\mathrm{Parameter} + \mathrm{Linear}^V_{\Delta}(\mathbf{x}^V))$
Δ^T: $(\mathrm{B}, \mathrm{L}, \mathrm{D}) \leftarrow \mathrm{Softplus}(\mathrm{Parameter} + \mathrm{Linear}^T_{\Delta}(\mathbf{x}^T))$
$\overline{\boldsymbol{A}^V}, \overline{\boldsymbol{B}^V}$: $(\mathrm{B}, \mathrm{F}, \mathrm{D}, \mathrm{N}) \leftarrow \mathrm{Discretize}(\Delta^V, \boldsymbol{A}, \boldsymbol{B}^V)$
$\overline{\boldsymbol{A}^T}, \overline{\boldsymbol{B}^T}$: $(\mathrm{B}, \mathrm{L}, \mathrm{D}, \mathrm{N}) \leftarrow \mathrm{Discretize}(\Delta^T, \boldsymbol{A}, \boldsymbol{B}^T)$
$\mathbf{y}^{\mathbf{V}}$: $(\mathrm{B}, \mathrm{F}, \mathrm{D}) \leftarrow \mathrm{SSM}(\overline{\boldsymbol{A}^V}, \overline{\boldsymbol{B}^V}, C^V)$
$\mathbf{y}^{\mathbf{T}}$: $(\mathrm{B}, \mathrm{L}, \mathrm{D}) \leftarrow \mathrm{SSM}(\overline{\boldsymbol{A}^T}, \overline{\boldsymbol{B}^T}, C^T)$
return $\mathbf{y}^V$ and $\mathbf{y}^T$

Results

Model(Backbone)	Pretrain data	Verb	Noun	Action
MeMViT (24x3)	K600	71.4	60.3	48.4
Omnivore (swin-B)	IN-(21k+1k)+K400+SUN	69.5	61.7	49.9
MeMViT (16x4)	K400	70.6	58.5	46.2
ORViT (MF-HR)	IN-21k+K400	68.4	58.7	45.7
MambaVL (ORViT)	IN-21k+K400	**69.1**	**63.9**	**48.6**
AVION (VIT-B)	WIT + Ego4D	70.0	59.8	**49.1**
LaViLa (TSF-B)	WIT + Ego4D	69.0	58.4	46.9
MambaVL (ViT-B)	WIT + Ego4D	**70.9**	**61.1**	**49.1**
AVION (ViT-L)	WIT + Ego4D	73.0	65.4	54.4
LaViLa (TSF-L)	WIT + Ego4D	72.0	62.9	51.0
MambaVL (ViT-L)	WIT + Ego4D	**74.3**	**67.1**	**55.0**

Action Recognition: across model-sizes, MambaVL is stronger.

Method	Pretrain data	Overall		
		Verb	Noun	Action
AVT+ [34]	IN21K + EPIC boxes	28.2	32.0	15.9
MeMVIT (32x3) [20]	K700	**32.2**	**37.0**	17.7
MeMViT (16x4) [20]	K400	**32.8**	33.2	15.1
AFFT [35]	IN-21K	22.8	34.6	18.5
ORViT-MF [31]	IN-21k+K400	26.9	34.2	23.3
MambaVL (ORViT)	IN-21k+K400	29.1	35.1	**23.9**

Action Anticipation: MambaVL performs better than base model ORViT

1. AVION - Zhao, Yue, and Philipp Krähenbühl. "Training a large video model on a single machine in a day." *arXiv preprint arXiv:2309.16669* (2023).
2. ORViT - Herzig, Roei, et al. "Object-region video transformers." Proceedings of the ieee/cvf conference on computer vision and pattern recognition. 2022.
3. LaViLa - Zhao, Yue, et al. "Learning video representations from large language models." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2023.

Results

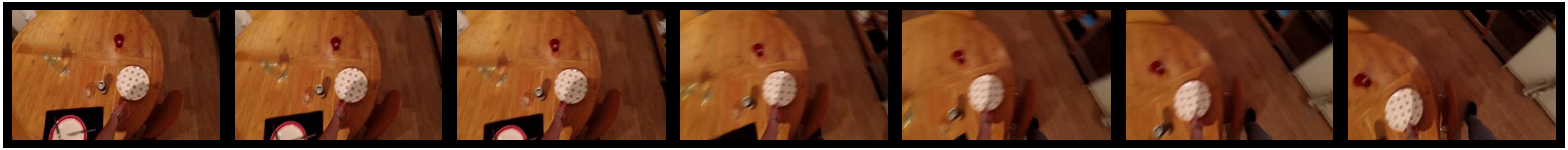
Model	GFLOPS	Params
ORViT	405.0	148M
ORViT + Transformer Fusion	413.5	242M
MambaVL	413.0	157M

Model comparison by GFLOPs and parameter count

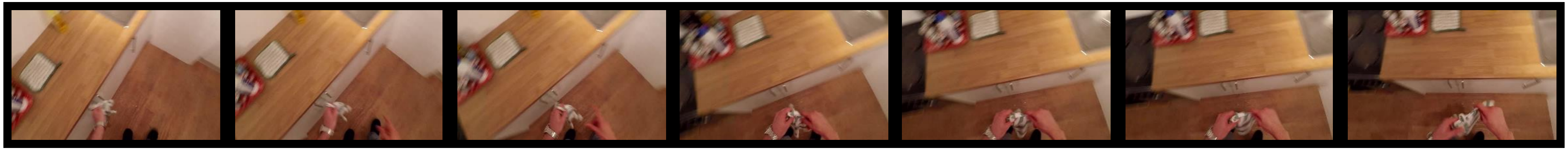
Fusion Method	Verb	Noun	Action
MLP	62.8	51.6	39.6
Transformer (6x4)	62.9	51.9	40.0
Transformer (12x12)	62.5	51.8	39.5
MambaVL	69.1	63.9	48.6

Comparison between different fusion methods

Qualitative Results



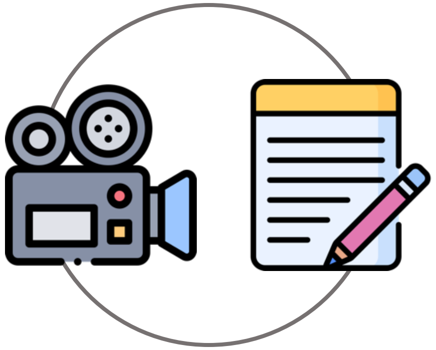
GT: Take Plate
Predicted: Take Plate



GT: Dry Hands
Predicted: Dry Hands

Advantages of MambaVL

Mamba: Linear complexity for long-range sequence modeling



Effective cross-modal information sharing

Flexible integration with existing methods

New Task! Question-Answering for Action Recognition

Thank you for watching!

Mamba Fusion: Learning Actions Through Questioning

Zhikang Dong*, Apoorva Beedu*, Jason Sheinkopf, Irfan Essa

For any questions, please reach out to:
Apoorva Beedu or Zhikang Dong
abeedu3@gatech.edu, zdong3@gatech.edu

https://github.com/Dongzhikang/MambaVL