Unifying Specialized Visual Encoders for Video Language Models

Jihoon Chung*,1, Tyler Zhu*,1, Max Gonzalez Saez-Diez[1], Juan Carlos Niebles[2], Honglu Zhou[2], Olga Russakovsky[1]

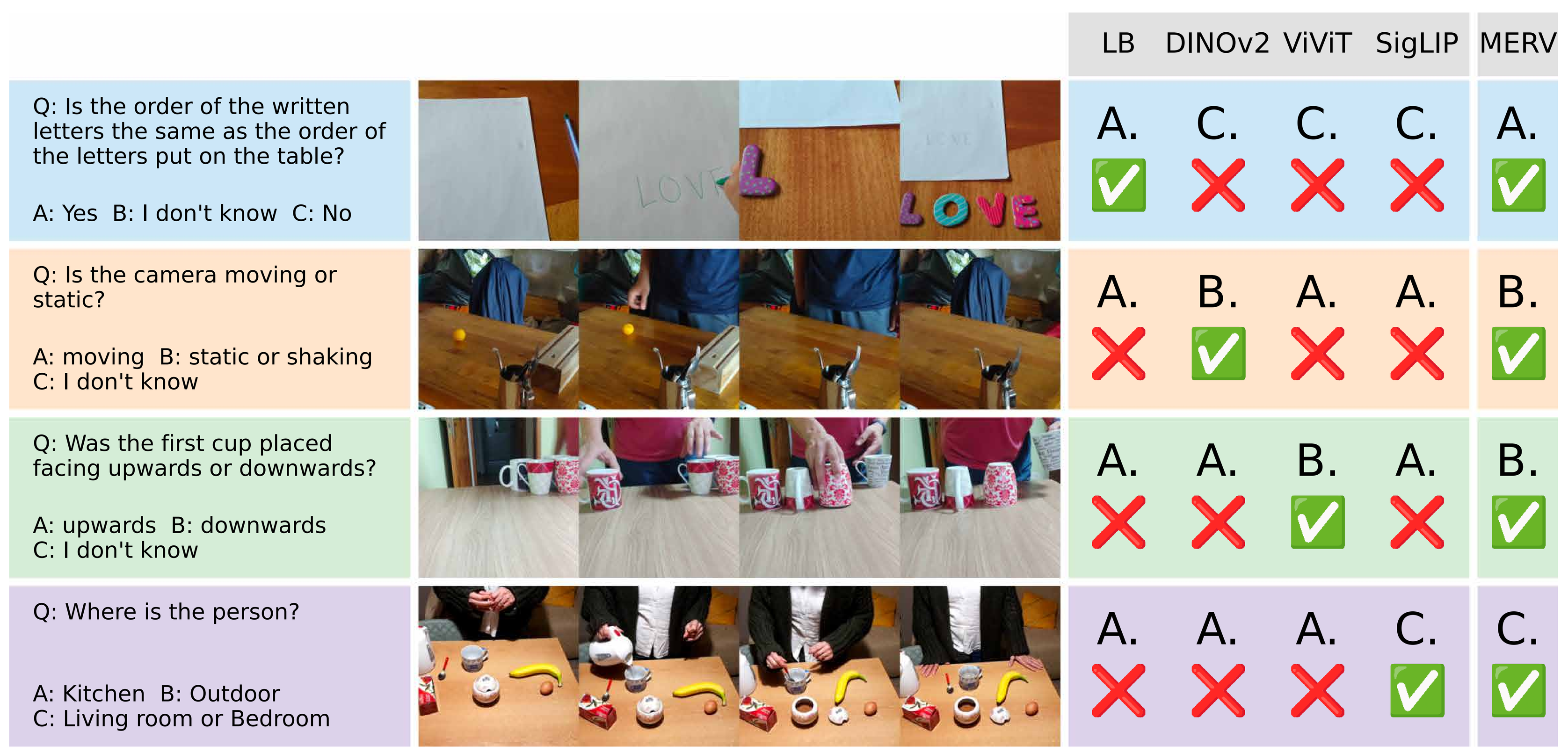
Princeton University[1] Salesforce Research[2]

ICML International Conference On Machine Learning

Princeton Visual AI Lab

Motivation: A General Vision Model

Prior works use a single visual encoder, which limits the type of visual information your VideoLLM can process [1].



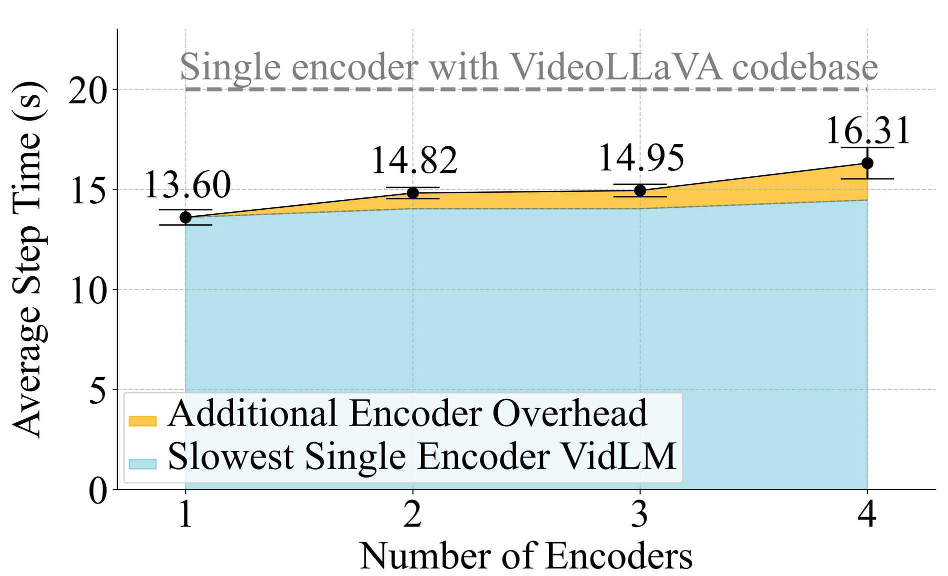
Can we create a generally capable video model by combining multiple pretrained visual models? **Yes!**

Video Benchmark Results

Methods	MSVD-QA		MSRVTT-QA		TGIF-QA		Perception	ActivityNet-QA	
	Acc	Score	Acc	Score	Acc	Score	Acc	Acc	Score
Alternative data mixes									
Video-Chat (Li et al., 2023c)	56.3	2.8	45.0	2.5	-	-	-	26.5	2.2
LLaMA-Adapter (Zhang et al., 2024b)	54.9	3.1	43.8	2.7	-	-	-	34.2	2.7
Video-ChatGPT (Maaz et al., 2024)	64.9	3.3	49.3	2.8	-	-	-	35.2	2.7
LLaMA-VID-7B (Li et al., 2024b)	69.30	3.74	57.84	3.24	51.31	3.26	41.64	46.45	3.22
LLaMA-VID-13B (Li et al., 2024b)	70.25	3.77	58.58	3.26	51.26	3.26	41.54	46.79	3.23
Same data mixes									
Video-LLaVA (Lin et al., 2024)	67.74	3.69	56.90	3.18	47.99	3.17	44.22	47.08	3.27
MERV	**70.97**	3.76	**59.03**	**3.25**	51.1	3.26	46.21	**50.87**	**3.34**
Gains to Video-LLaVA	+3.23	+.07	+2.13	+.07	+3.11	+.09	+1.99	+3.79	+.07

We **outperform** prior works with similar data mixes, especially Video-LLaVA with same data on standard video benchmarks.

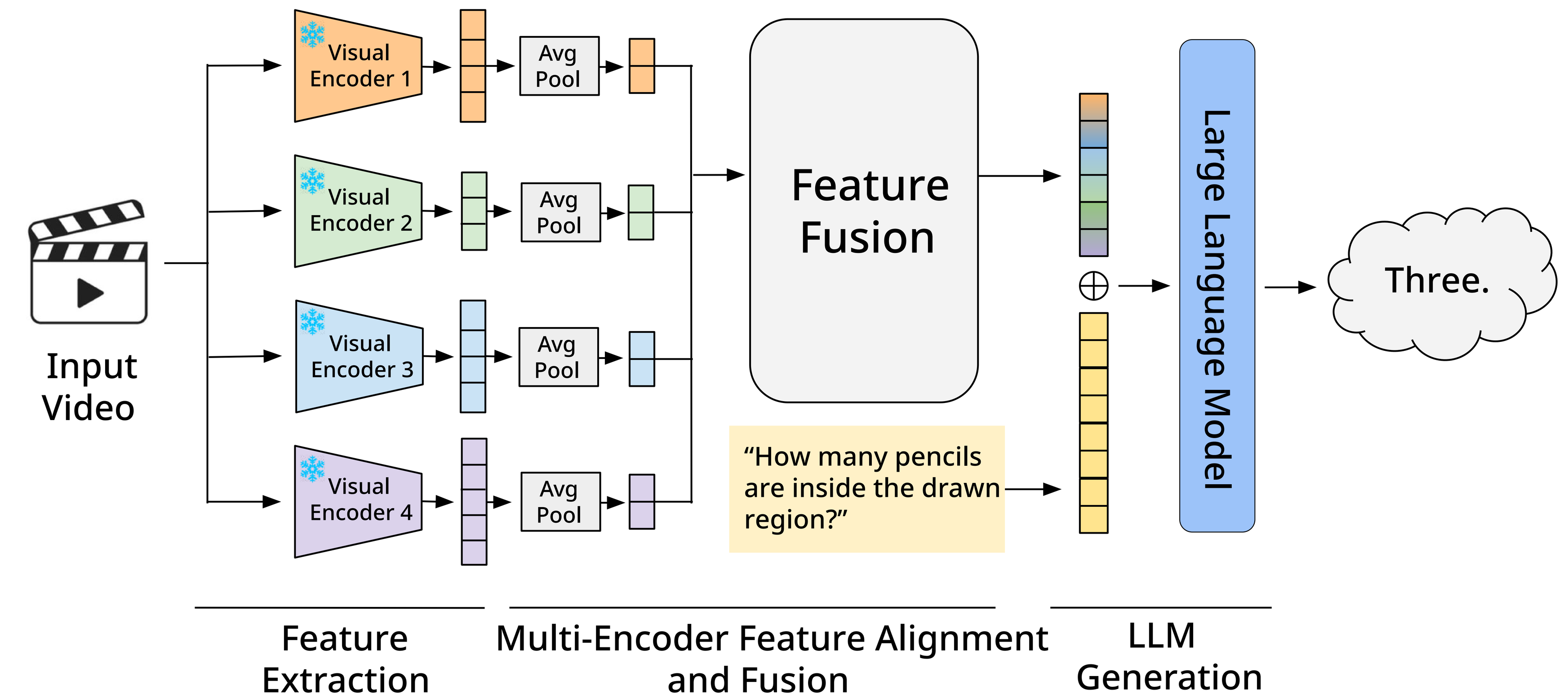
Computational Efficiency



Our method scales well using parallelism over multiple GPUs!

- VLMs are bottlenecked by vision, so || models is OK
- Overhead from multiple is minimal compared to 1 model

Architecture



We use four visual experts varying in visual format (image vs. video) and data (vision vs. vision+language)

- Match visual encoders across space, time, and dim (MLP)
- Space: 2D Avg Pool for spatial alignment was best
- Time: Sample frames for input so that output was aligned

[Feature Fusion] $\mathbf{O} := \mathrm{Softmax}\left(\frac{\mathbf{Q}\overline{\mathbf{X}}^\top}{\sqrt{d}}\right)\mathbf{X} \in \mathbb{R}^{\ell \times d}$

Simple cross attention method using a single learnable query Q over averaged features from each embedding X

Architecture Ablations

Pre-fusion Projector

Projector	Avg Acc	Params	FLOPs
257 tok	54.76	-	-
class tok	52.05	-	-
2D Avg	54.96	0	2.1M
2D Avg*	**55.86**	0	4.2M
2D Attn	52.12	12.7M	9.7G
2D Conv	54.23	237M	241G
3D Avg*	55.09	0	4.2M
3D Conv	55.42	113M	232G

Pre-fusion Projector Tokens

Tkns	MSVD	MSRVTT	TGIF
1	61.94	54.64	41.41
4	64.47	55.72	45.32
16	67.23	56.44	47.75
64	**69.08**	**58.00**	**50.01**
100	68.38	57.47	48.78
144	68.65	57.73	48.81
256	68.46	57.72	48.66

Feature Fusion Strategy

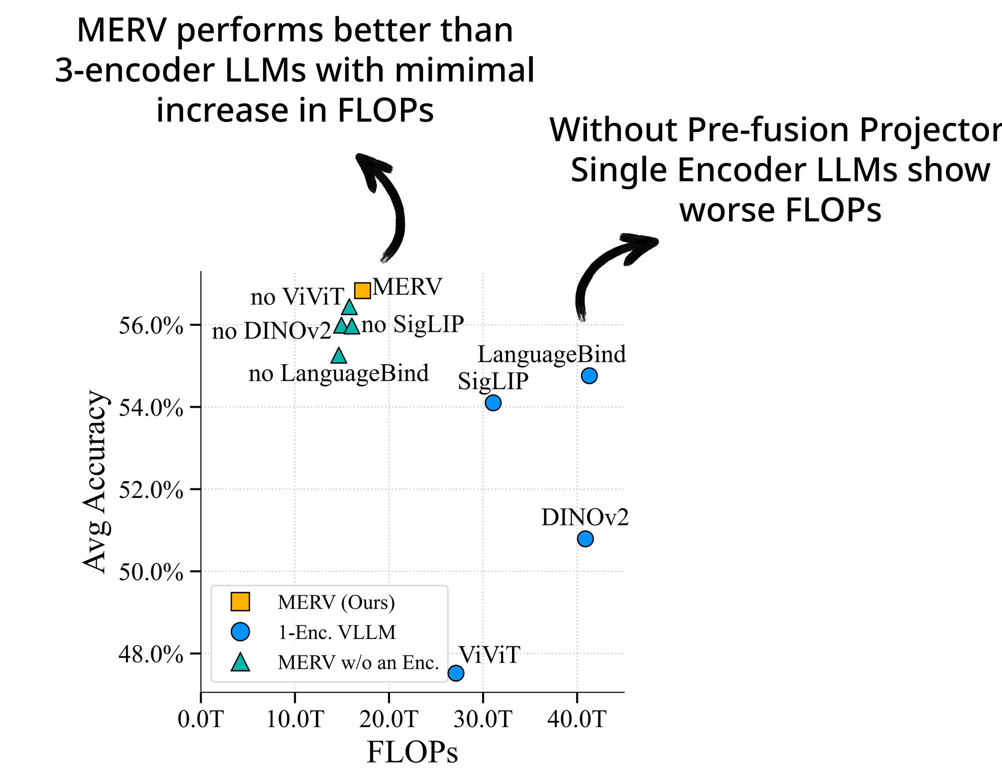
Strategy	Avg Acc	FLOPs
Cross-Attn	**56.83**	17.19 T
Concat (Seq.)	54.45	43.09 T
Concat (Ch.)	56.64	16.29 T
Learnable W	55.01	16.24 T
25% - Mixed	54.19	16.39 T

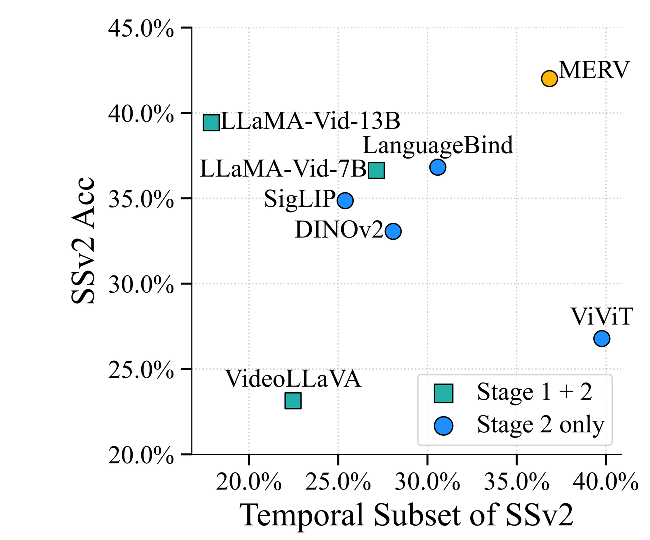
Related Works and Links

[1] Eyes Wide Shut? Exploring the Visual Shortcomings of Multimodal LLMs, Shengbang Tong, et al., CVPR 2024
[2] Video-LLaVA: Learning United Visual Representation by Alignment Before Projection, Bin Lin, et al., EMNLP 2024
[3] LanguageBind: Extending Video-Language Pretraining to N-modality by Language-based Semantic Alignment, Bin Zhu, et al, ICLR 2024
[5] ViViT: A Video Vision Transformer, Anurag Arnab, et al., ICCV 2021
[6] Sigmoid Loss for Language Image Pre-Training, Xiaohua Zhai, et al., ICCV 2023

Sponsored By National Science Foundation Grant No. 2107048

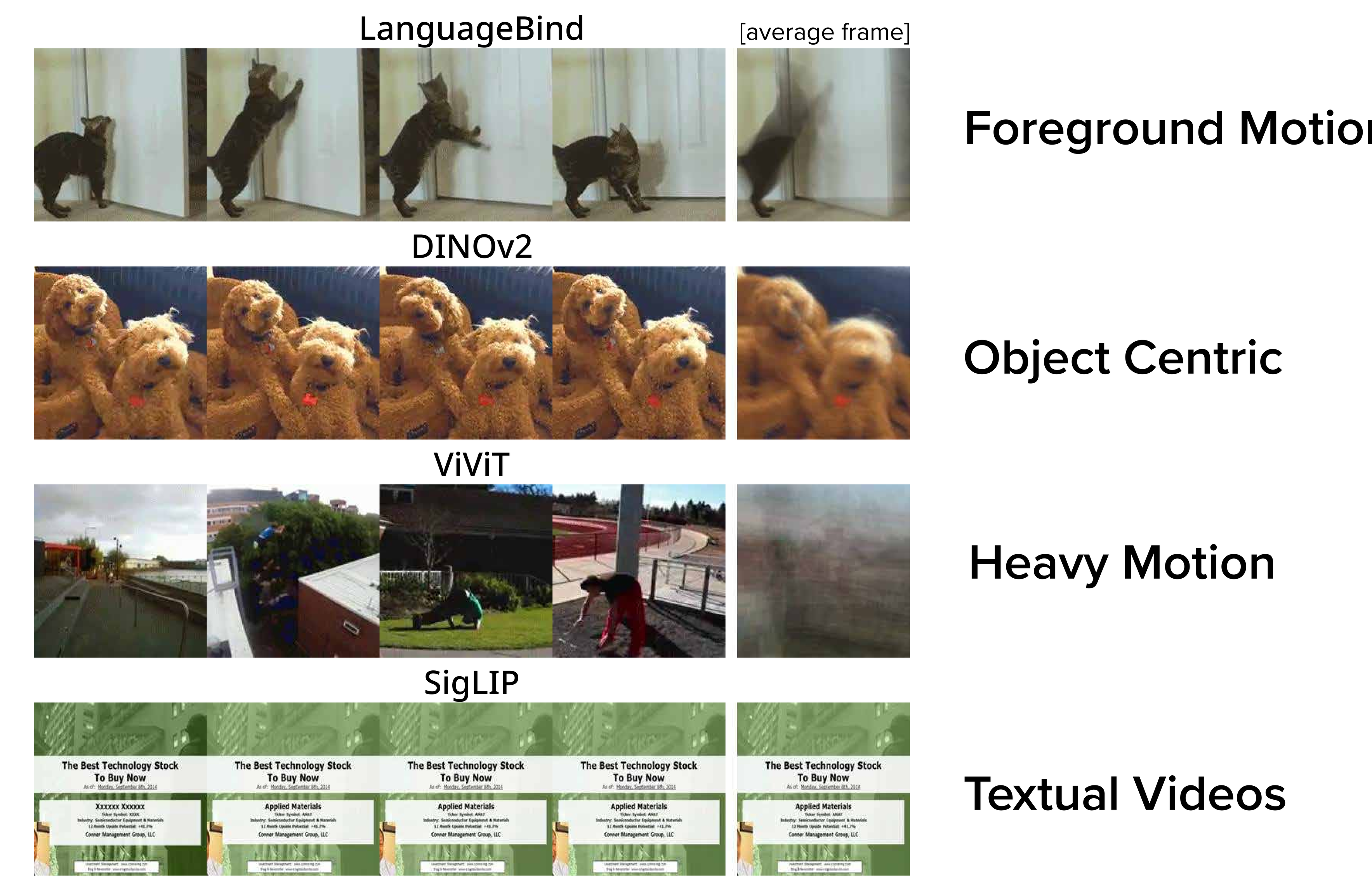
MERV is Capable, Efficient and Video-aware



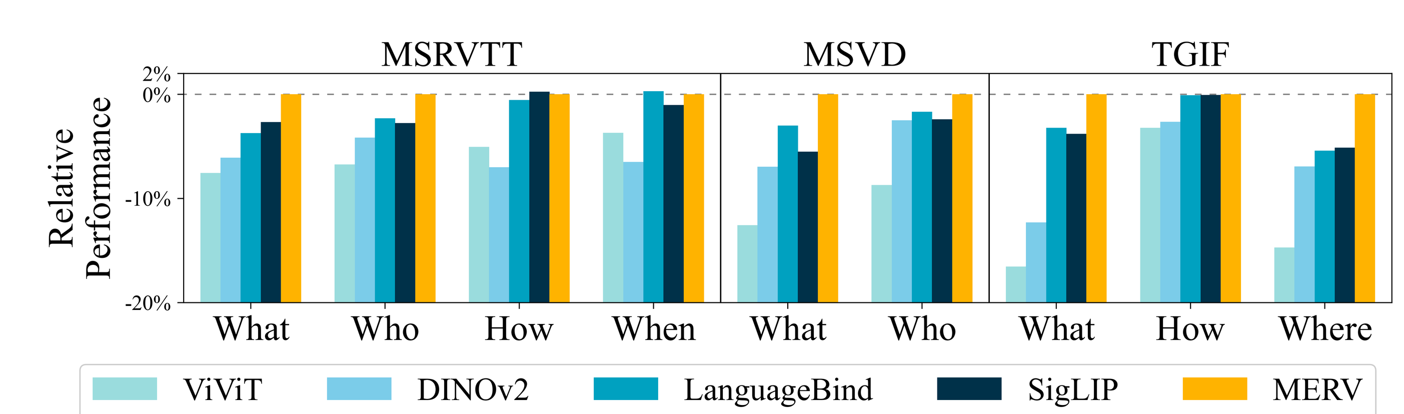


Ours outperforms existing models on entire Something-Something benchmark, with significant increase in classes that require strong temporal understanding ability

Visual Encoders Have Individual Strengths



Maximally activating videos for each encoder from X-attn weights reveal that different encoders are specialized for different types of videos



MERV outperforms single-encoder LLMs on various video tasks.