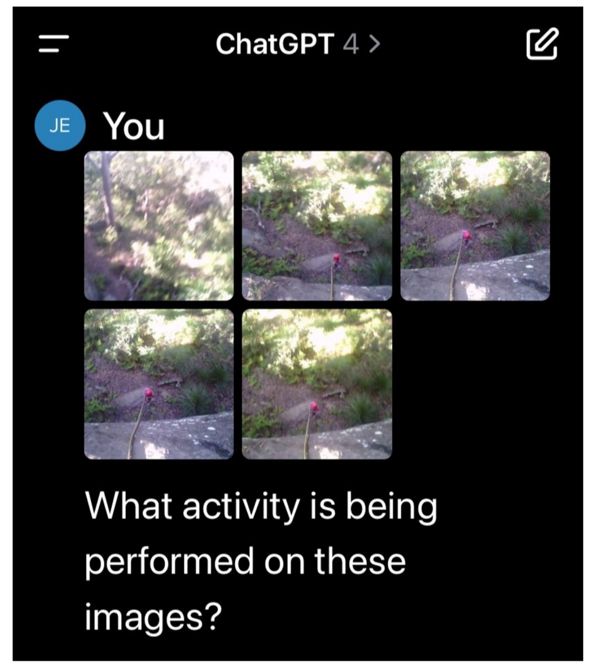
ViC-MAE: Self-Supervised Representation Learning from Images and Video with Contrastive Masked Autoencoders

Jefferson Hernandez[1], Ruben Villegas[2], Vicente Ordonez[1]

[1]Rice University [2]Google DeepMind

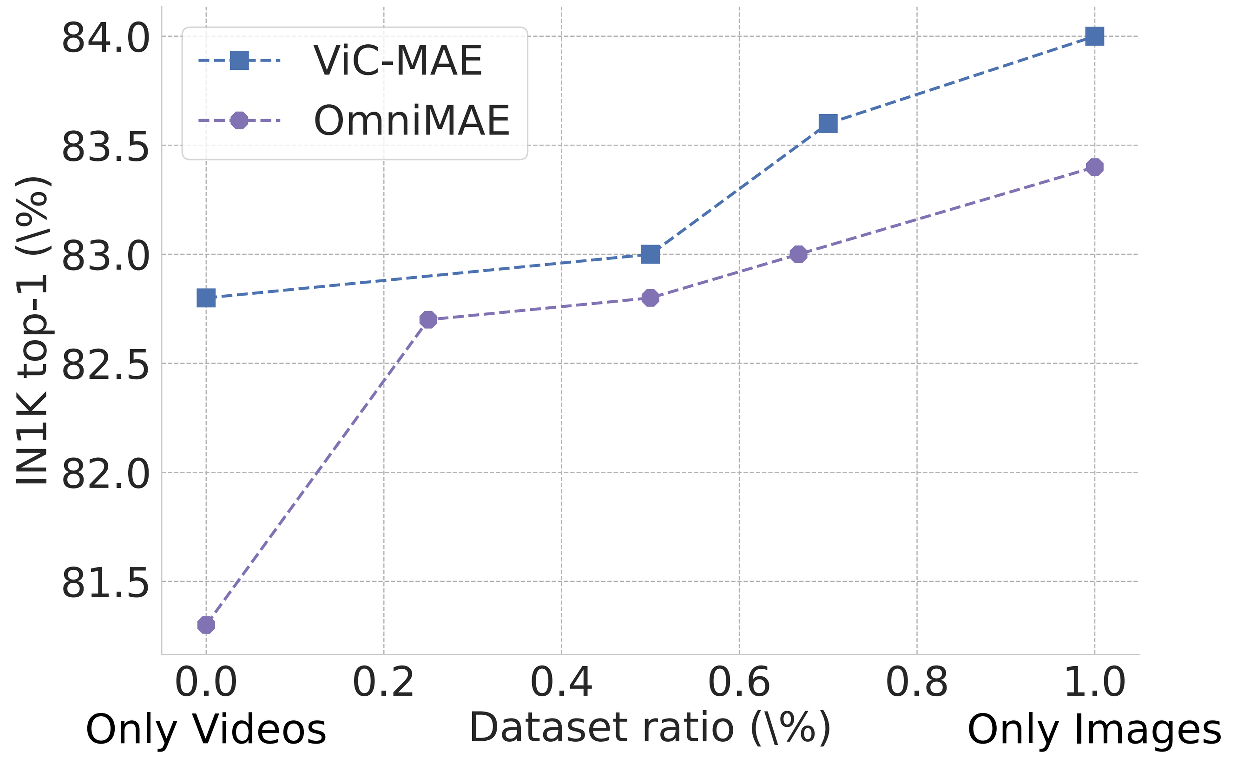
Video is an unrepresented category in foundation models

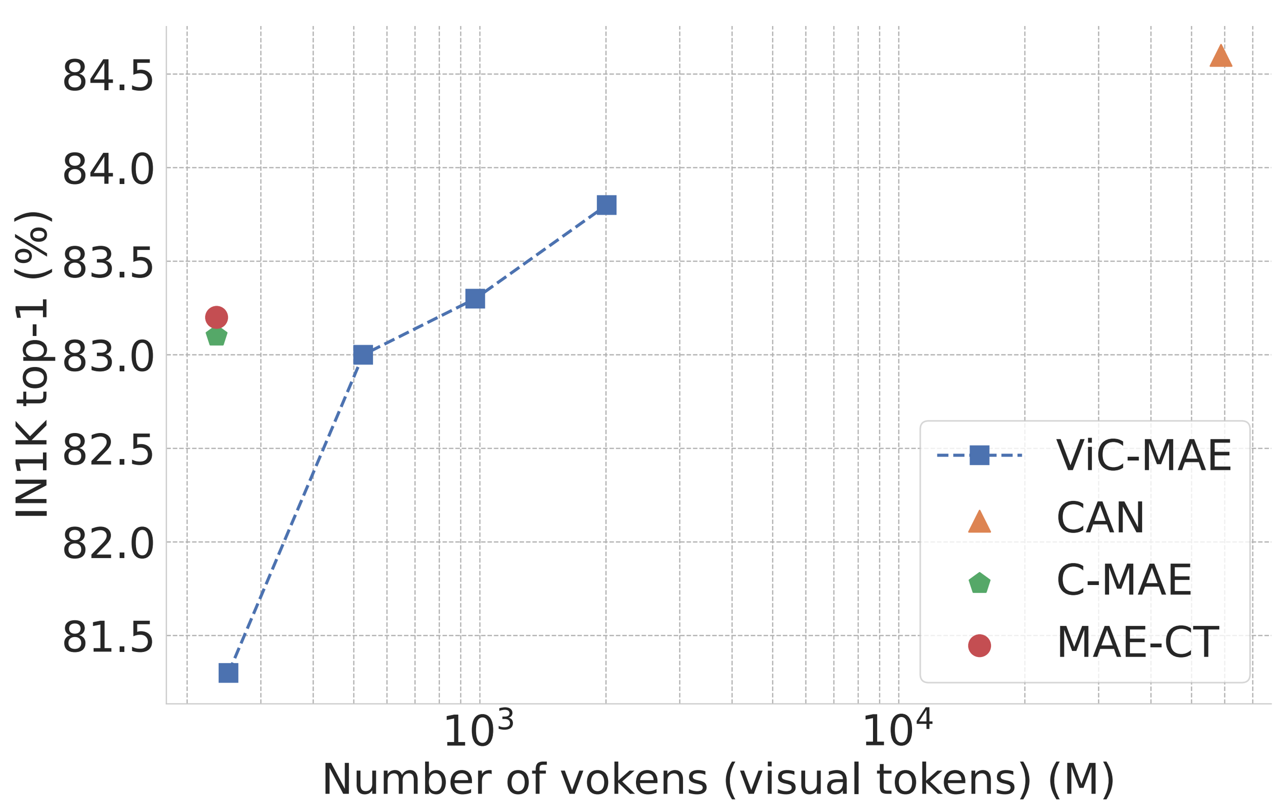


GPT4-v answers **Rock Climbing** when the action is **Abseiling**

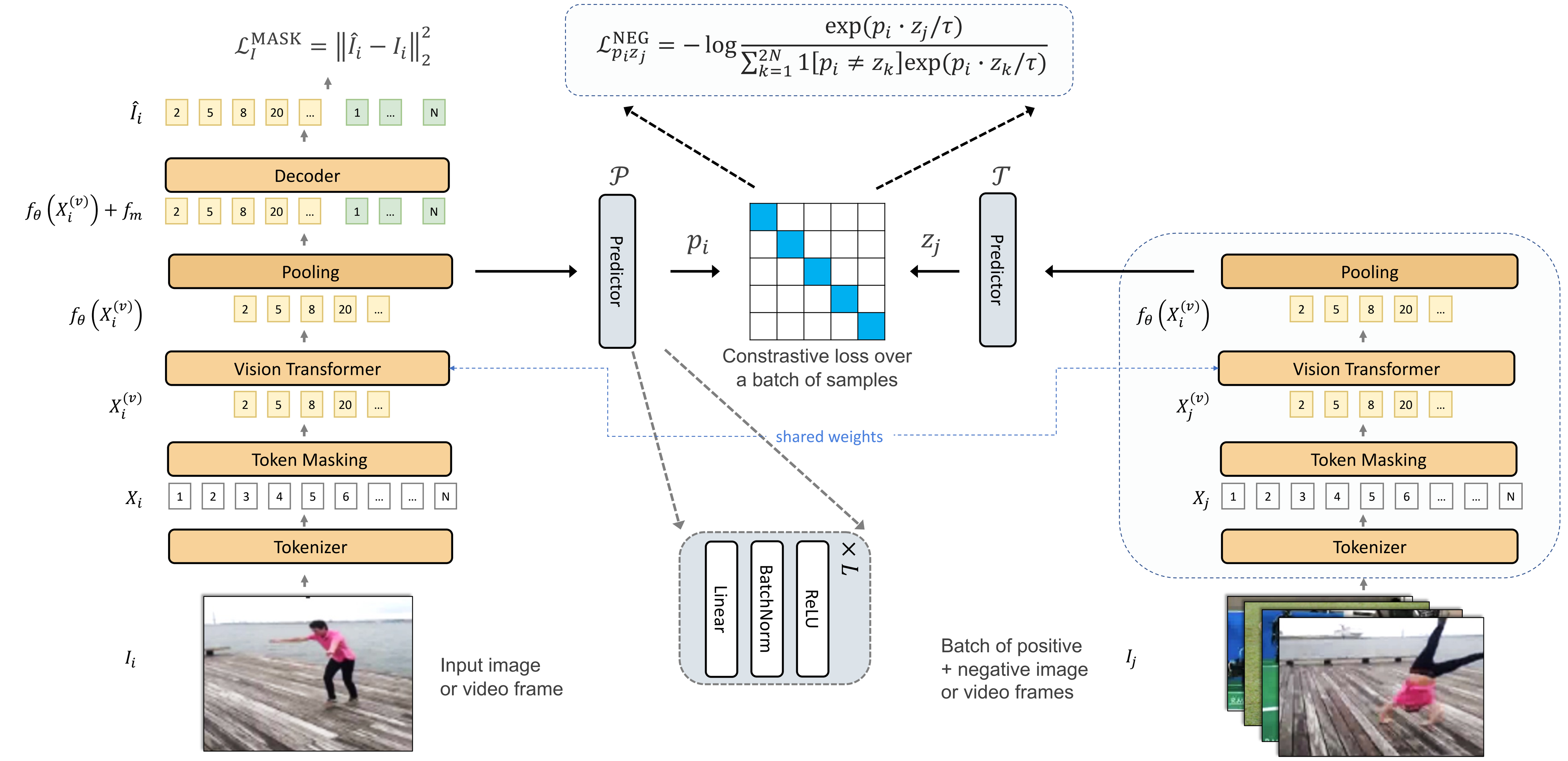
- There is discrepancy in the way we train LLMs (using autoregression, *masking*) vs, how we train vision models (*contrastive learning*)
- There is no unified approach to model video data that has been shown to scale and produce better results than just *averaging over frames*.
- **Image-to-video** transfer learning is very common, But the latter, **video-to-image** transfer learning has not been very successful with models reaching <50% accuracy.

ViC-MAE uses image and video data better, and scales better than other methods





ViC-MAE: Visual Contrastive Masked Auto-Encoders



ViC-MAE is a powerful method to model images and video data

Main result

	Method	Arch.	Pre-training Data	In-Domain		Out-of-Domain	
				IN1K	K400	Places-365	SSv2
Supervised	ViT [23] *ICML'20*	ViT-B	IN1K	82.3	68.5	57.0	61.8
	ViT [23] *ICML'20*	ViT-L	IN1K	82.6	78.6	58.9	66.2
	COVeR [86] *arXiv'21*	TimeSFormer-SR	JFT-3B+ K400+ MiT + IN1K	86.6	87.2	-	70.9
	OMNIVORE [30] *CVPR'22*	ViT-B	IN1K + K400 + SUN RGB-D	84.0	83.3	59.2	68.3
	OMNIVORE [30] *CVPR'22*	ViT-L	IN1K + K400 + SUN RGB-D	86.0	84.1	–	–
	TubeViT [63] *CVPR'23*	ViT-B	K400 + IN1K	81.4	88.6	–	–
	TubeViT [63] *CVPR'23*	ViT-L	K400 + IN1K	–	90.2	–	76.1
Self-Supervised	MAE [34] *CVPR'22*	ViT-B	IN1K	83.4	–	57.9	59.6
	MAE [34] *CVPR'22*	ViT-L	IN1K	85.5	82.3	59.4	57.7
	ST-MAE [26] *NeurIPS'22*	ViT-B	K400	81.3	81.3	57.4	69.3
	ST-MAE [26] *NeurIPS'22*	ViT-L	K400	81.7	84.8	58.1	73.2
	VideoMAE [72] *NeurIPS'22*	ViT-B	K400	81.1	80.0	–	69.6
	VideoMAE [72] *NeurIPS'22*	ViT-L	K400	–	85.2	–	74.3
	OmniMAE [29] *CVPR'23*	ViT-B	K400 + IN1K	82.8	80.8	58.5	69.0
	OmniMAE [29] *CVPR'23*	ViT-L	K400 + IN1K	84.7	84.0	59.4	73.4
	ViC-MAE	ViT-L	K400	85.0	85.1	59.5	73.7
	ViC-MAE	ViT-L	MiT	85.3	84.9	59.7	73.8
	ViC-MAE	ViT-B	K400 + IN1K	83.0	80.8	58.6	69.5
	ViC-MAE	ViT-L	K400 + IN1K	86.0	86.8	60.0	75.0
	ViC-MAE	ViT-B	K710+ MiT + IN1K	83.8	80.9	59.1	69.8
	ViC-MAE	ViT-L	K710 + MiT + IN1K	**87.1**	**87.8**	**60.7**	**75.9**

Video-to-image transfer

	Model	Pre-train.	Food	CIFAR10	CIFAR100	Birdsnap	SUN397	VOC2007	DTD	Caltech101
ViT/B-16	MAE [34] ‡	K400	74.54	94.86	79.49	46.51	64.33	83.07	78.01	93.28
	MAE [34] ‡	MiT	76.23	94.47	79.50	47.98	65.32	83.46	78.21	93.08
	ViC-MAE (ours)	K400	76.56	93.64	78.80	47.56	64.75	83.74	78.53	92.27
	ViC-MAE (ours)	MiT	**77.39**	**94.92**	**79.88**	**48.21**	**65.64**	**84.77**	**79.27**	**93.53**
ViT/L-16	MAE [34]	IN1K	77.5	95.0	82.9	49.8	63.2	83.3	74.5	94.8
	OmniMAE [29]	SSv2+IN1K	76.2	94.2	82.2	50.1	62.6	82.7	73.9	94.4
	ViC-MAE (ours)	IN1K+K400	81.9	95.6	85.4	52.8	67.3	84.2	76.8	94.9
	ViC-MAE (ours)	K710+MiT+IN1K	**82.9**	**96.8**	**86.5**	**53.5**	**68.1**	**85.3**	**77.8**	**96.1**

Ablations

(a) **Ablation on frame separation**. 0: sample same frame, D: distant sampling, and > 0 continuous sampling.

Frame separation	ImageNet-1K	
	Top-1	Top-5
0	63.25	83.34
2	64.47	84.31
4	65.25	84.64
8	65.89	84.91
D	**67.66**	**86.22**

(b) **Ablation on pooling type**. The hyperparameter λ is set to 0.025 and introduced using a schedule.

Pooling type	Top-1	Top-5
GeM	66.92	85.50
max	67.01	85.59
mean	**67.66**	**86.22**

(c) **Ablation on different augmentations**. We use a combination of different color and spatial augs.

Color Augm.	Spatial Augm.	ImageNet-1K	
		Top-1	Top-5
✓		65.40	84.03
	✓	66.03	85.01
✓	✓	**67.66**	**86.22**