Selective Structured State-Spaces for Long-Form Video Understanding

Jue Wang, Wentao Zhu, Pichao Wang, Xiang Yu, Linda Liu, Mohamed Omar, Raffay Hamid

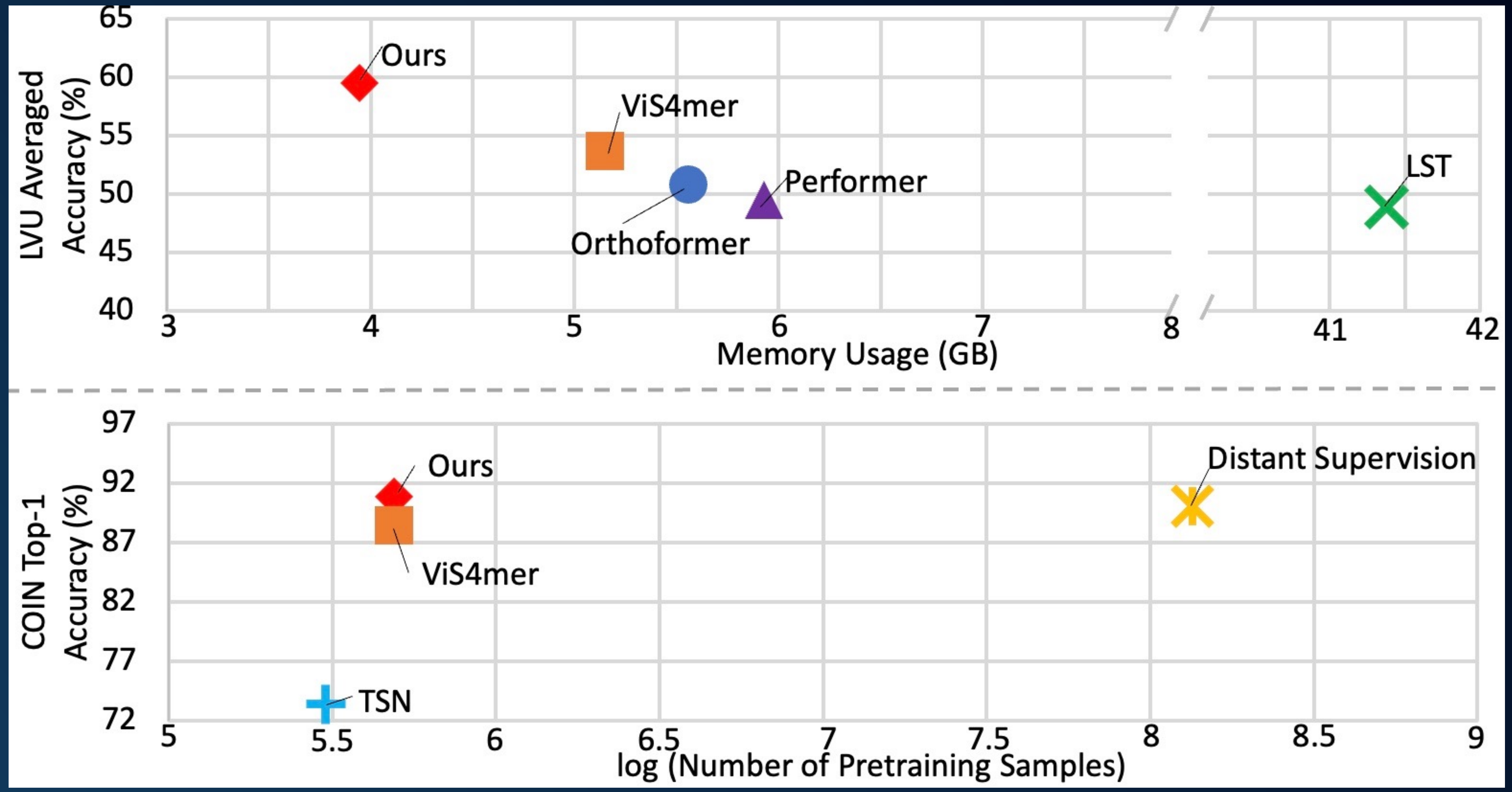
Paper Tag: TUE-PM-216

prime video

Full Paper

OVERVIEW

- We propose a **Selective S4 (S5)** model that leverages the global sequence-context information from S4 features to adaptively **choose informative tokens in a task- specific way**.
- We introduce a novel **long-short masked contrastive learning approach (LSMCL)** that enables our model to be **tolerant to the mis-predicted tokens** and exploit longer duration spatiotemporal context by using shorter duration input videos, leading to improved robustness in the S5 model.
- We demonstrate that two proposed novel techniques (S5 model and LSMCL) are seamlessly suitable and effective for long-form video understanding, achieving the **state-of- the-art performance** on three challenging benchmarks.
- Compared to vanilla video transformer, our work offers **10%** memory and **2.6x** throughput improvements when dealing with the same input.

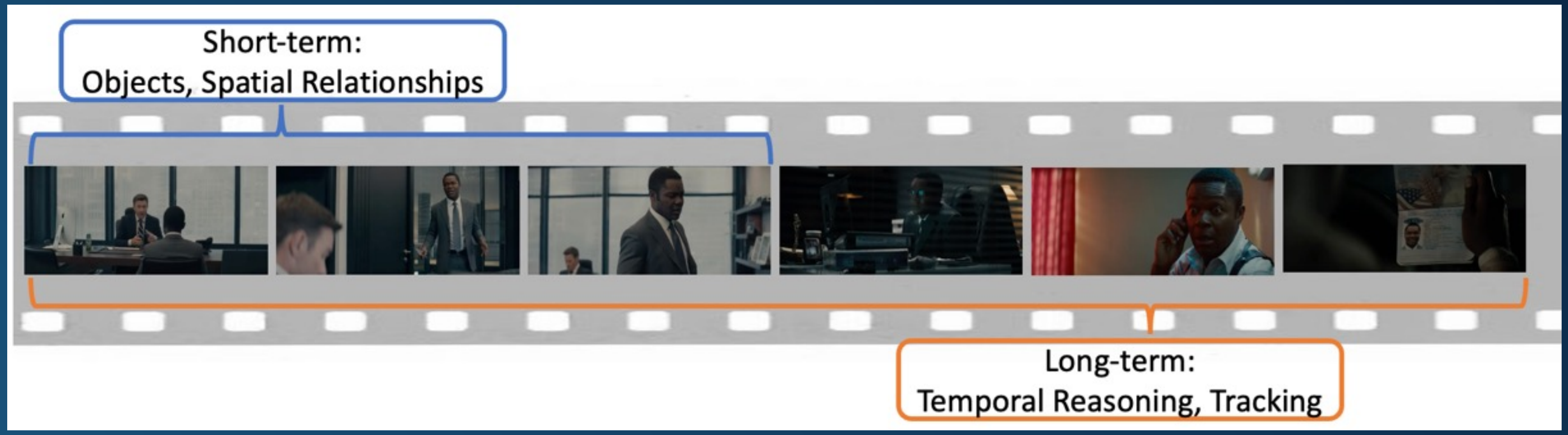


Reference:

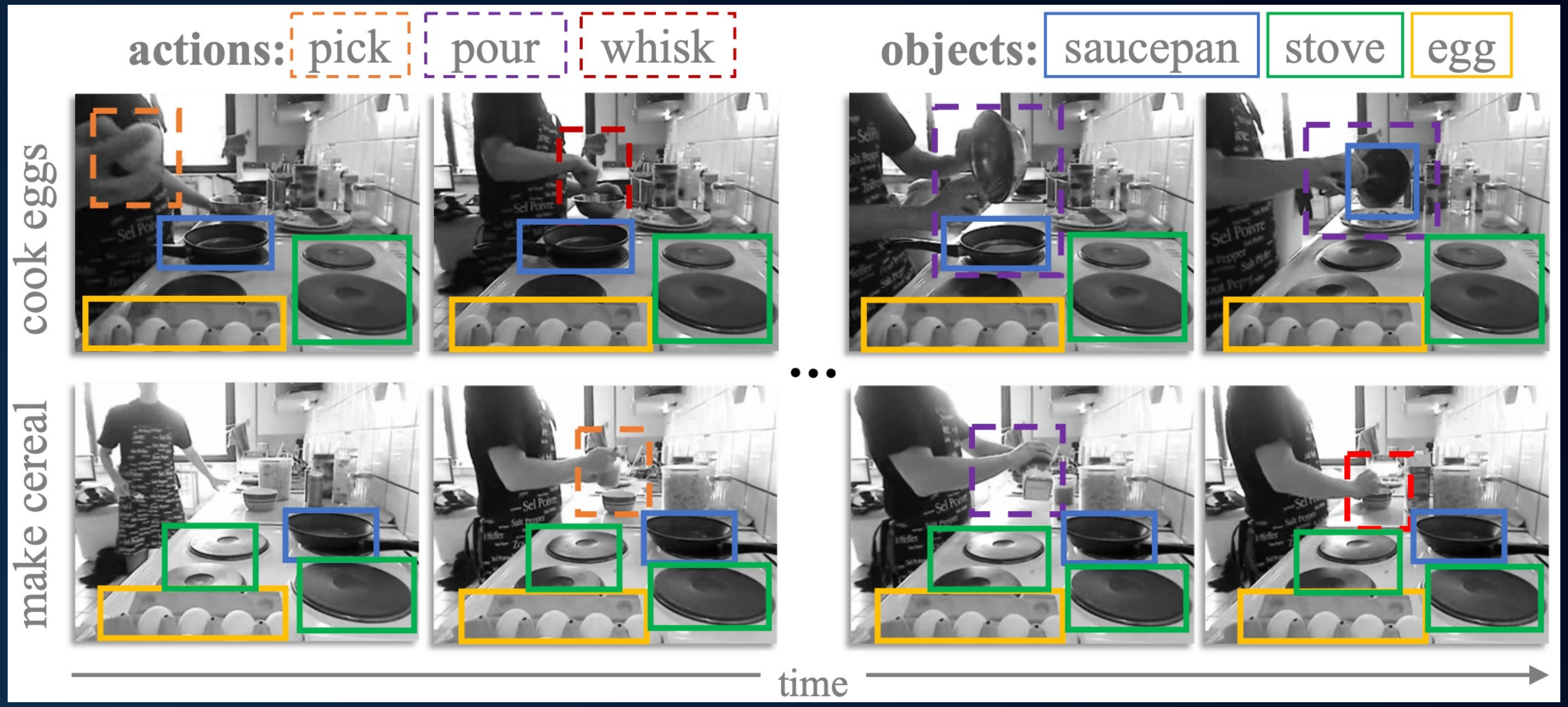
1. ViS4mer, LST: Long movie clip classification with state-space video models. ECCV 2022
2. Orthoformer: Long movie clip classification with state-space video models. NIPS 2021
3. Performer: Rethinking attention with performers. ICLR 2021
4. TSN: Comprehensive in structional video analysis: The coin dataset and performance evaluation. PAMI 2020
5. Distant Supervision: Learning to recognize procedural activities with distant supervision. CVPR 2022

prime video

BACKGROUND



A snapshot of GRINGO from Amazon Studios, showing the complex content of long-form videos.



These two videos heavily overlap in terms of objects (e.g., eggs, saucepan and stove), and actions (e.g., picking, whisking and pouring).

1. **Effectiveness**: Modeling long-term spatiotemporal dependencies for richer representations in various tasks.
2. **Efficiency**: To achieve the high effectiveness, the memory and computational burden become more severe due to the large volume of input.

BACKGROUND

Structured State-Spaces Sequence (S4) Model

$$x'(t) = Ax(t) + Bu(t)$$
$$y(t) = Cx(t) + Du(t)$$

$$\Rightarrow \quad x_k = \bar{A}x_{k-1} + \bar{B}u_k, \quad y_k = \bar{C}x_k \quad \Rightarrow \quad y = \bar{K} \circledast u$$

$$HiPPO: A_{n,k} = -\begin{cases} (2n+1)^{0.5}(2k+1)^{0.5}, & if\ n > k \\ n+1, & if\ n = k \\ 0, & if\ n < k \end{cases}$$

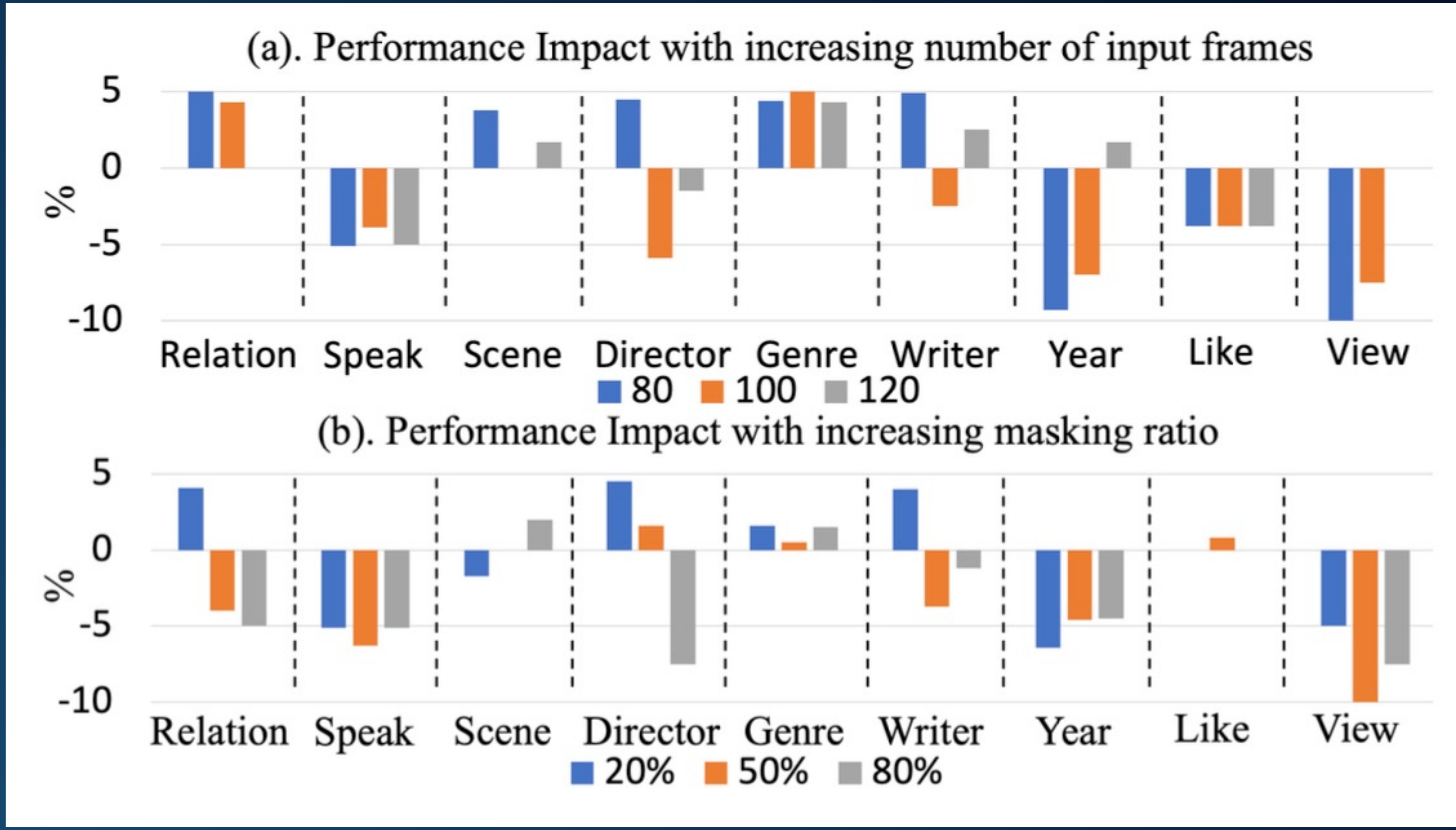
Sequence length (L), batch size (B), and hidden dimension (H). Tildes denote log factors.

	Self-attention	State-space
Parameters	H^2	H^2
Memory	$B(L^2 + HL)$	BLH
Training	$B(L^2H + LH^2)$	$BH(\tilde{H} + \tilde{L}) + B\tilde{L}H$
Inference	$L^2H + LH^2$	H^2

Long Movie Clip Classification with State-Space Video Models, ECCV 2022

Reference:
1. Long Movie Clip Classification with State-Space Video Models, ECCV 2022
2. Efficiently Modeling Long Sequences with Structured State Spaces, ICLR 2022
3. Hippo: Recurrent memory with optimal polynomial projections , NIPS 2020

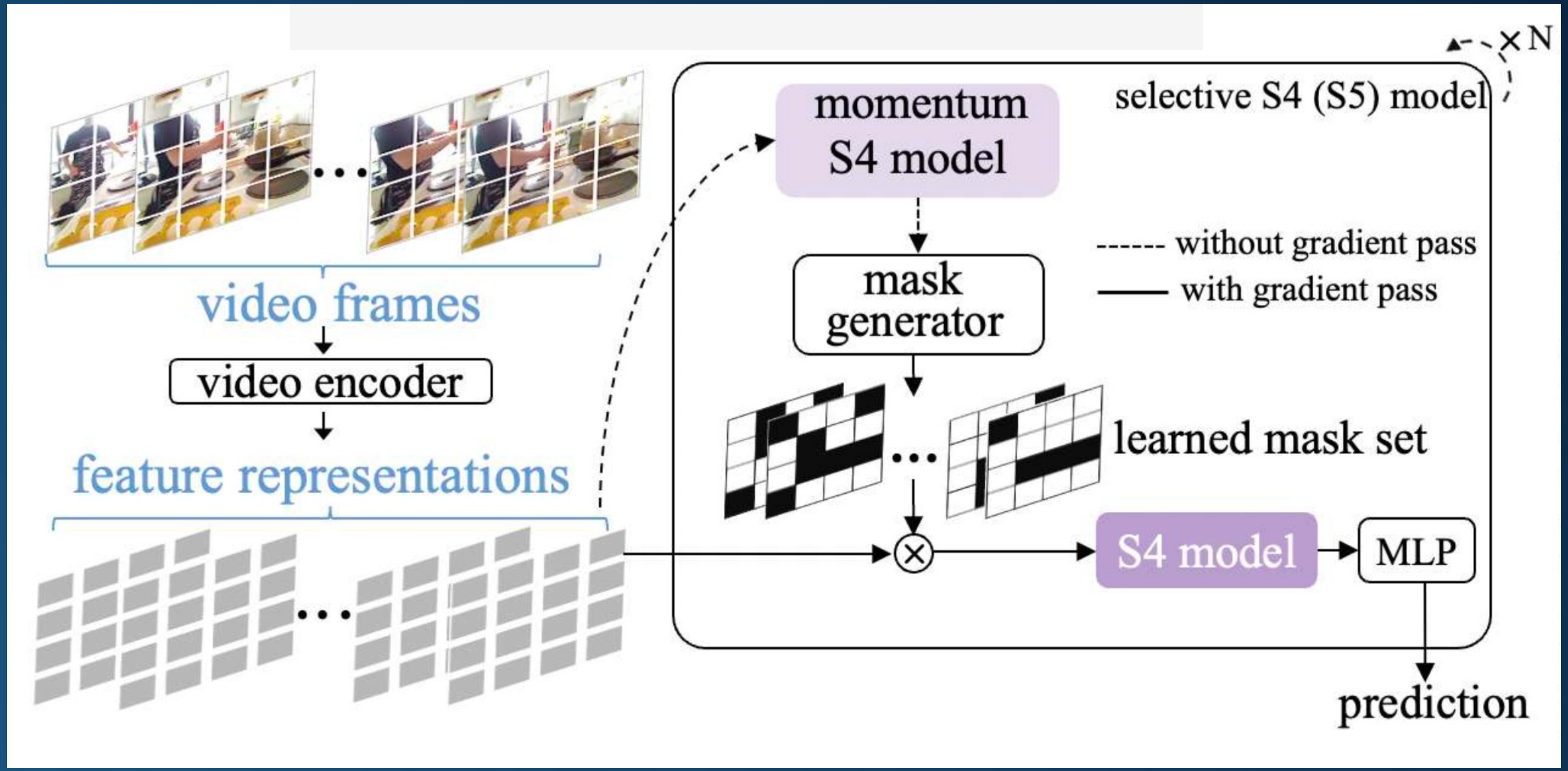
MOTIVATION



Performance gain/loss of ViS4mer on LVU dataset with different settings of input frames and random masking ratio, where we conclude: (a). The performance is not substantially improved with increasing number of input frames. (b) Random masking strategy cannot effectively reduce redundant tokens.

METHOD

Selective Structured State-Spaces Sequence (S5) Model



MG *is trained for a classifcation task on* $\mathbb{C} = \{C_1, C_2, \cdots C_{ST}\}$, *where ST is the total number of tokens.*

$$MG(x_{S_4}) = p(c|x_{S_4}) \in [0,1], \text{so that} \sum_{c=C_1}^{c=C_{ST}} p(c|x_{S_4}) = 1$$

We apply Gumbel SoftMax with Straight-Through tricks in the Mask Generator, and the gradient for each selected token can be written as:

$$G \approx \nabla_{MG} \frac{\exp(logp(c|x_{S_4}) + g(c)/\rho)}{\sum_{c'=C_1}^{c'=C_{ST}} \exp(logp(c'|x_{S_4}) + g(c')/\rho)}$$

$$x_{S_4} = \widehat{S_4}\,(x_{input})$$

$$Mask = MG\,(x_{S_4})$$

$$\hat{x}_{input} = x_{input} \otimes Mask$$

$$x_{output} = MLP(S_4(\hat{x}_{input}))$$

Where $\widehat{S_4} = m\widehat{S_4} + (1-m)S_4$,

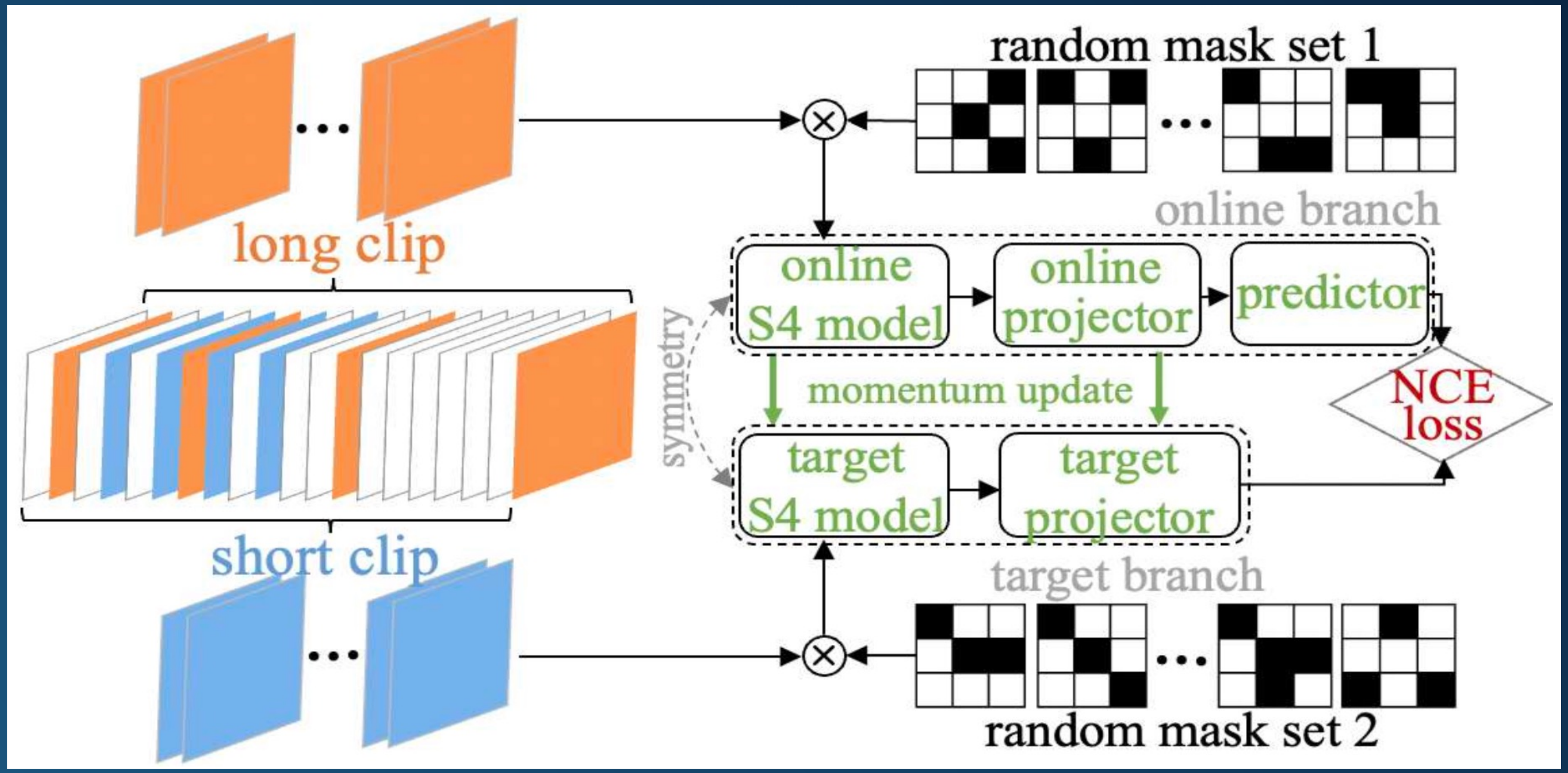
m is the momentum,

MG indicates Mask Generator,

Layer Norm is incorporated in Each Projection

METHOD

Long-Short Masked Contrastive Learning



$In\ Batch\ B:$

f_q: *query encoder* f_k: *key encoder*

short clips: $X_S = \{x_S^1, x_S^2, \cdots x_S^B\}$

long clips: $X_L = \{x_L^1, x_L^2, \cdots x_L^B\}$

Long and short clips can alternatively become queries and keys

$$f_q = mf_q + (1-m)f_k$$

$$Given: q = f_q(\mathcal{R}_{mask}(x_S, \eta)), k = f_k(\mathcal{R}_{mask}(x_L, \eta))$$

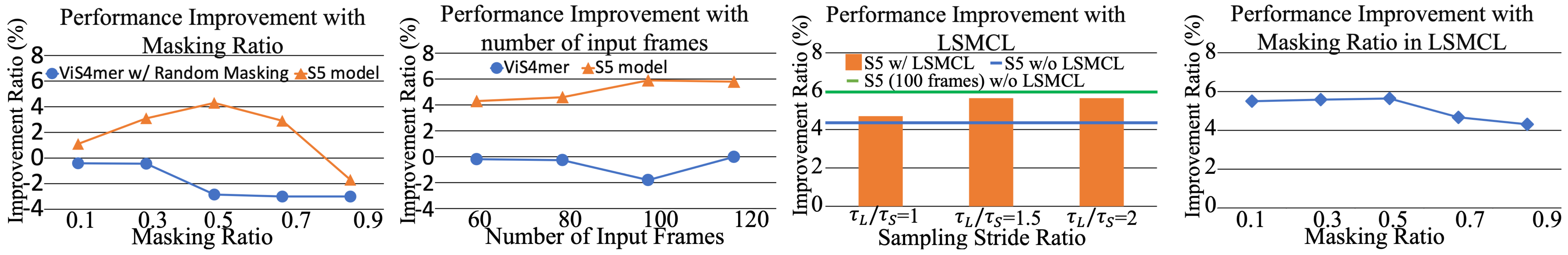
$$\mathcal{L}_{LSMCL} = \sum_i -log \frac{\exp(q^{i^T} k^i/\rho)}{\exp(q^{i^T} k^i/\rho) + \sum_{j\neq i} \exp(q^{i^T} k^j/\rho)}$$

m is the momentum,
ρ is the temperature hyperparameter

prime video

RESULTS

Mask Generator	Content (↑)			Metadata (↑)				User (↓)		
	Relation	Speak	Scene	Director	Genre	Writer	Year	Like	View	
No Mask (ViS4mer [29])	57.14	40.79	67.44	62.61	54.71	48.80	44.75	0.26	3.63	
Random	54.81	38.22	67.44	63.60	54.97	47.00	42.70	0.25	4.00	
Single TX	57.85	40.79	68.66	63.98	55.12	48.85	43.46	0.26	3.82	
Single TX_{S4}	**60.54**	**41.21**	**69.83**	**66.43**	**57.55**	**49.47**	**44.15**	**0.25**	**3.51**	+3.4
Stacked TXs	59.51	41.21	69.83	64.91	55.12	51.83	47.55	0.25	3.42	
Stacked TXs_{S4}	**61.98**	**41.75**	**70.94**	**67.34**	**59.16**	**51.83**	**47.55**	**0.24**	**3.42**	+2.5
Linear	54.81	40.28	67.44	63.90	54.97	48.17	42.77	0.26	3.95	
Linear_{S4}	**61.98**	**41.75**	**69.88**	**66.40**	**58.80**	**50.60**	**47.70**	**0.25**	**3.51**	+6.7



(a) (b) (c) (d)

prime video

RESULTS

Model	Content (↑)			Metadata (↑)				User (↓)		GPU Usage (GB) (↓)
	Relation	Speak	Scene	Director	Genre	Writer	Year	Like	View	
Obj. T4mer [67]	54.76	33.17	52.94	47.66	52.74	36.30	37.76	0.30	3.68	N/A
Performer [11]	50.00	38.80	60.46	58.87	49.45	48.21	41.25	0.31	3.93	5.93
Orthoformer [49]	50.00	38.30	66.27	55.14	55.79	47.02	43.35	0.29	3.86	5.56
VideoBERT [53]	52.80	37.90	54.90	47.30	51.90	38.50	36.10	0.32	4.46	N/A
LST [29]	52.38	37.31	62.79	56.07	52.70	42.26	39.16	0.31	3.83	41.38
ViS4mer [29]	57.14	40.79	67.44	62.61	54.71	48.80	44.75	0.26	3.63	5.15
$\text{Ours}_{\text{60 frames}}$	**61.98**	**41.75**	**69.88**	**66.40**	**58.80**	**50.60**	**47.70**	**0.25**	**3.51**	**3.85**
$\text{Ours}_{\text{60 frames+LSMCL}}$	**61.98**	**41.75**	**72.53**	**66.40**	**61.34**	**50.60**	**47.70**	**0.24**	**3.51**	**3.85**
$\text{Ours}_{\text{100 frames}}$	**66.71**	**41.78**	**73.28**	**66.64**	**63.65**	**50.60**	**47.85**	**0.25**	**3.51**	**3.95**
$\text{Ours}_{\text{100 frames+LSMCL}}$	**67.11**	**42.12**	**73.49**	**67.32**	**65.41**	**51.27**	**47.95**	**0.24**	**3.51**	**3.95**

Method	P.T. Dataset	P.T. Samples	Accuracy
TSN [57]	Kinetics-400	306K	73.40
D-Sprv. [39]	HowTo100M	136M	90.00
ViS4mer [29]	Kinetics-600	495K	88.41
Ours	Kinetics-600	495K	**90.42**
$\text{Ours}_{\text{+LSMCL}}$	Kinetics-600	495K	**90.81**

Method	P.T. Dataset	P.T. Samples	Accuracy
VideoGraph [28]	Kinetics-400	306K	69.50
Timeception [27]	Kinetics-400	306K	71.30
GHRM [73]	Kinetics-400	306K	75.50
D-Sprv. [39]	HowTo100M	136M	89.90
ViS4mer [29]	Kinetics-600	495K	85.10*
Ours	Kinetics-600	495K	**90.14**
$\text{Ours}_{\text{+LSMCL}}$	Kinetics-600	495K	**90.70**

prime video

THANK YOU!