LookWhen? Fast Video Recognition by Learning When, Where, and What to Compute



Ali Salamatian

UBC



Anthony Fuller

Carleton University



Pritam Sarkar

UBC



James R. Green

Carleton University



Leonid Sigal

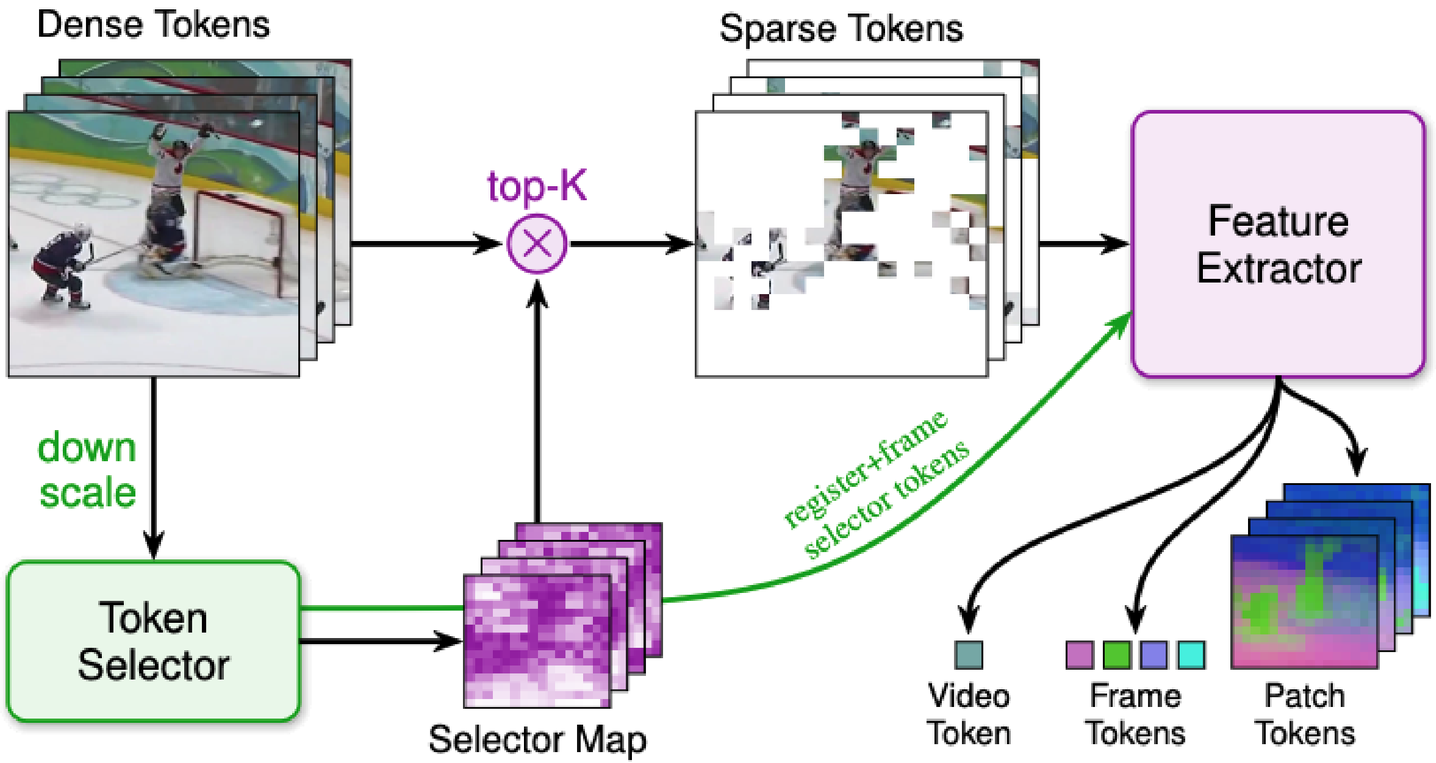
UBC



Evan Shelhamer

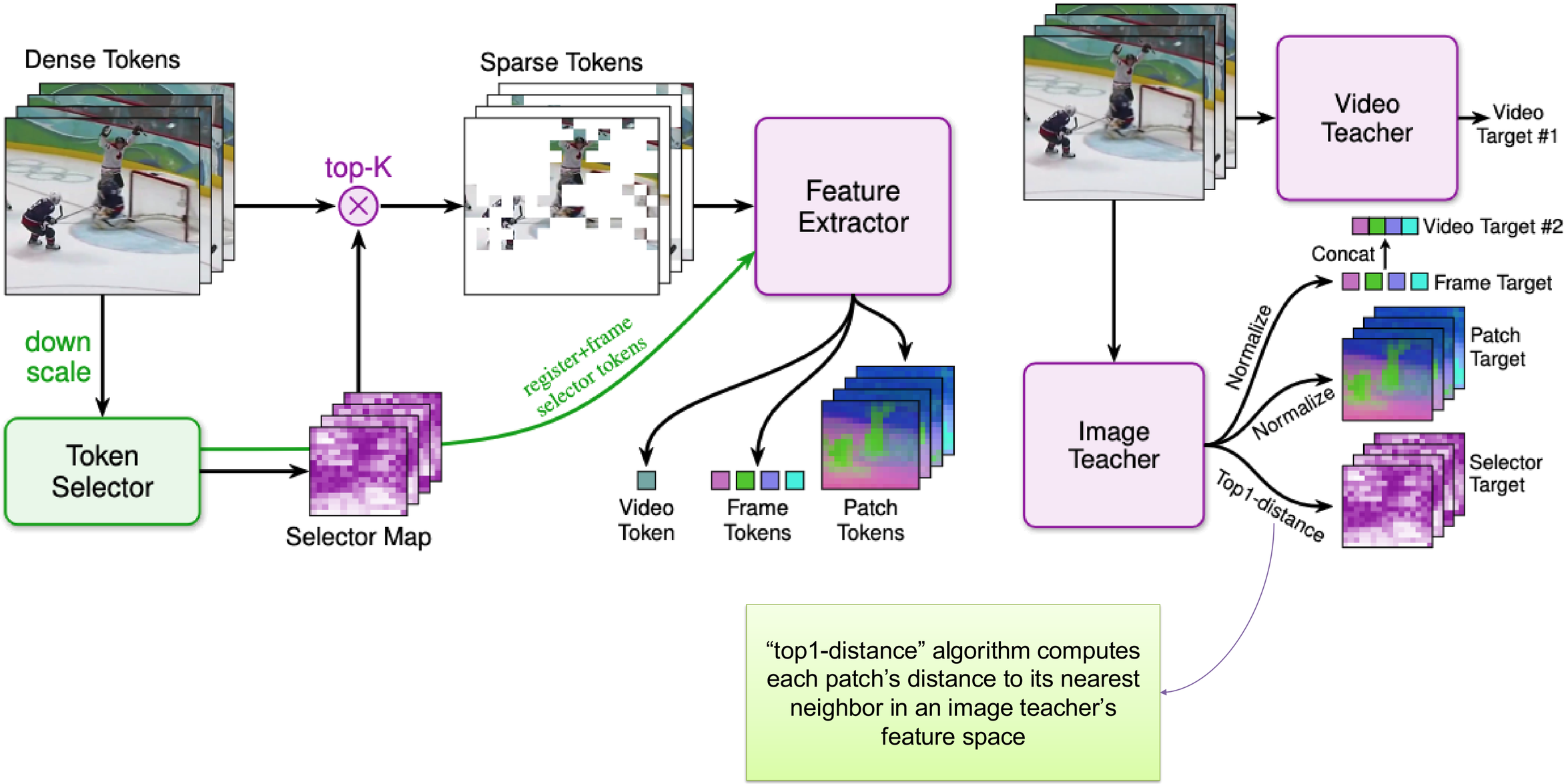
UBC

Learning which video tokens to process



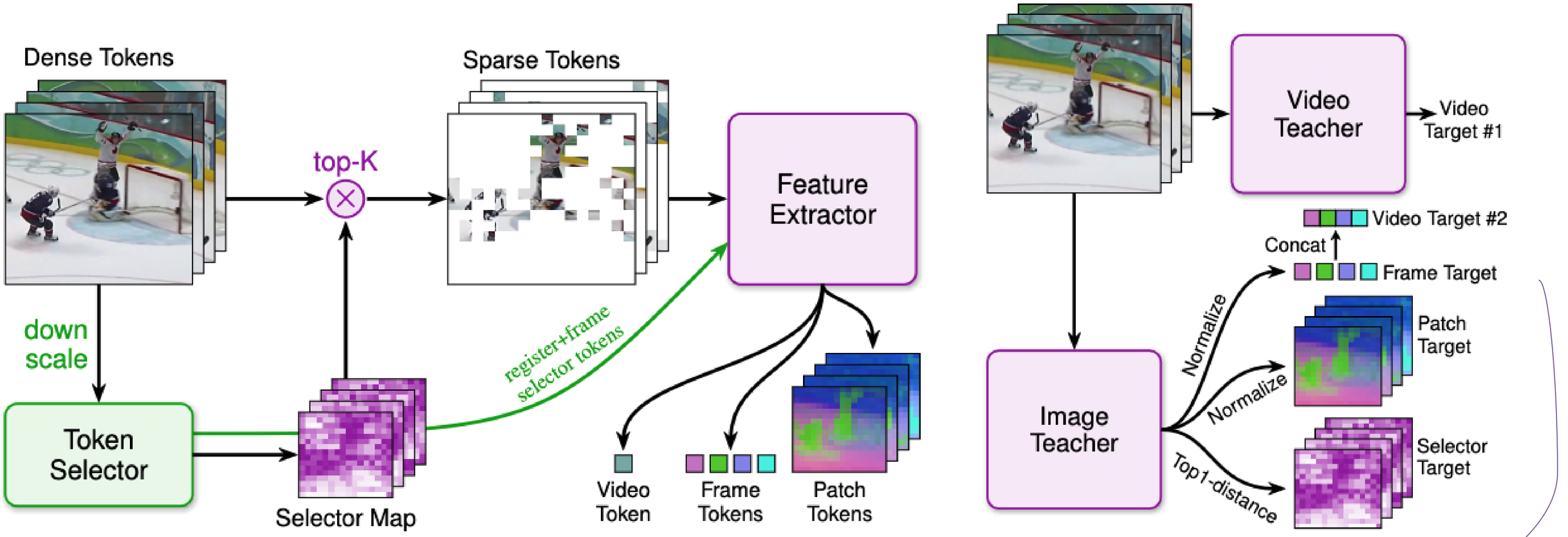
selector-extractor framework: selecting the top-k most unique tokens and extracting representations using only those

Learning which video tokens to process



"top1-distance" algorithm computes each patch's distance to its nearest neighbor in an image teacher's feature space

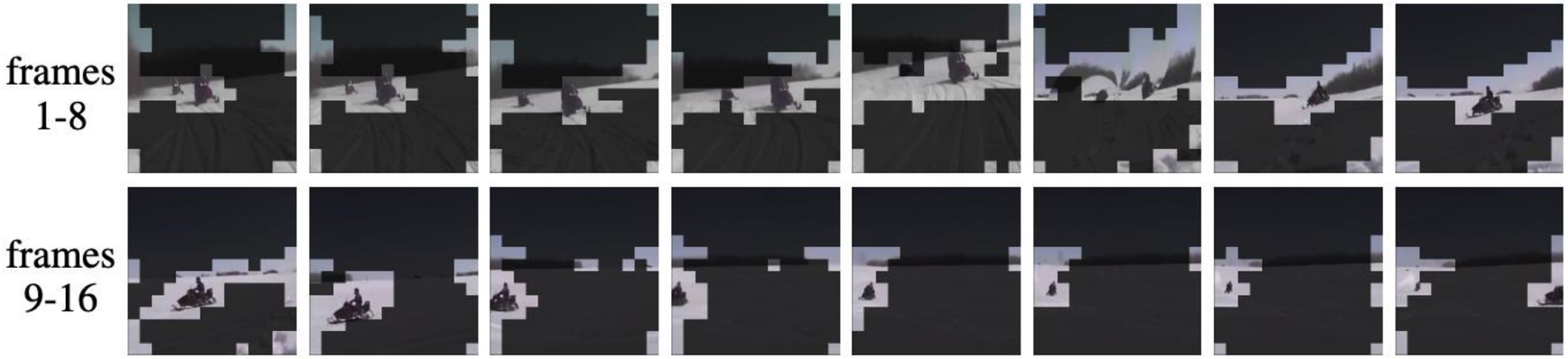
Learning which video tokens to process



Target features are constructed from a video teacher (InternVideo2), supervising the global video token, and an image teacher (DINOv3), supervising frame, patch, and the video token.

Qualitative Results

Kinetics-400
In-distribution

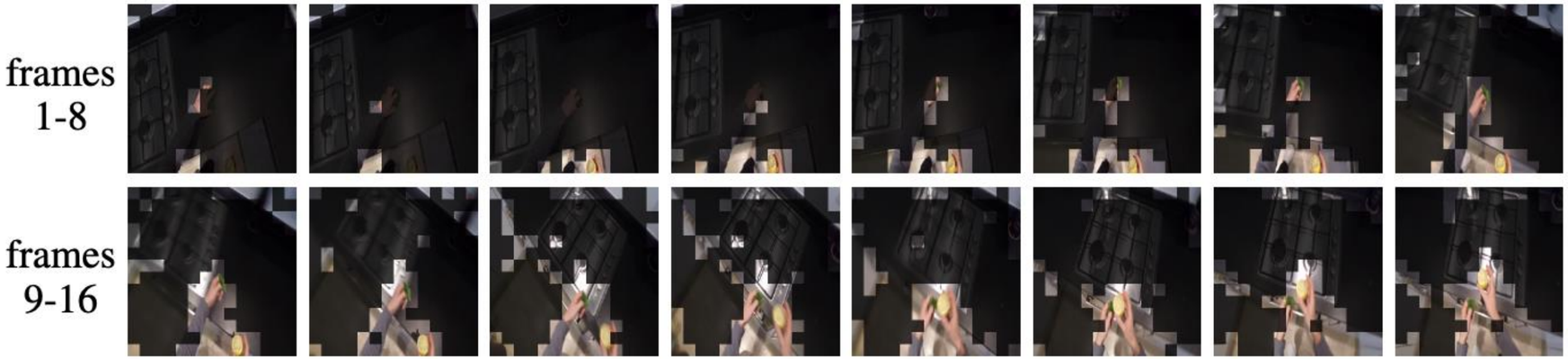


Qualitative Results

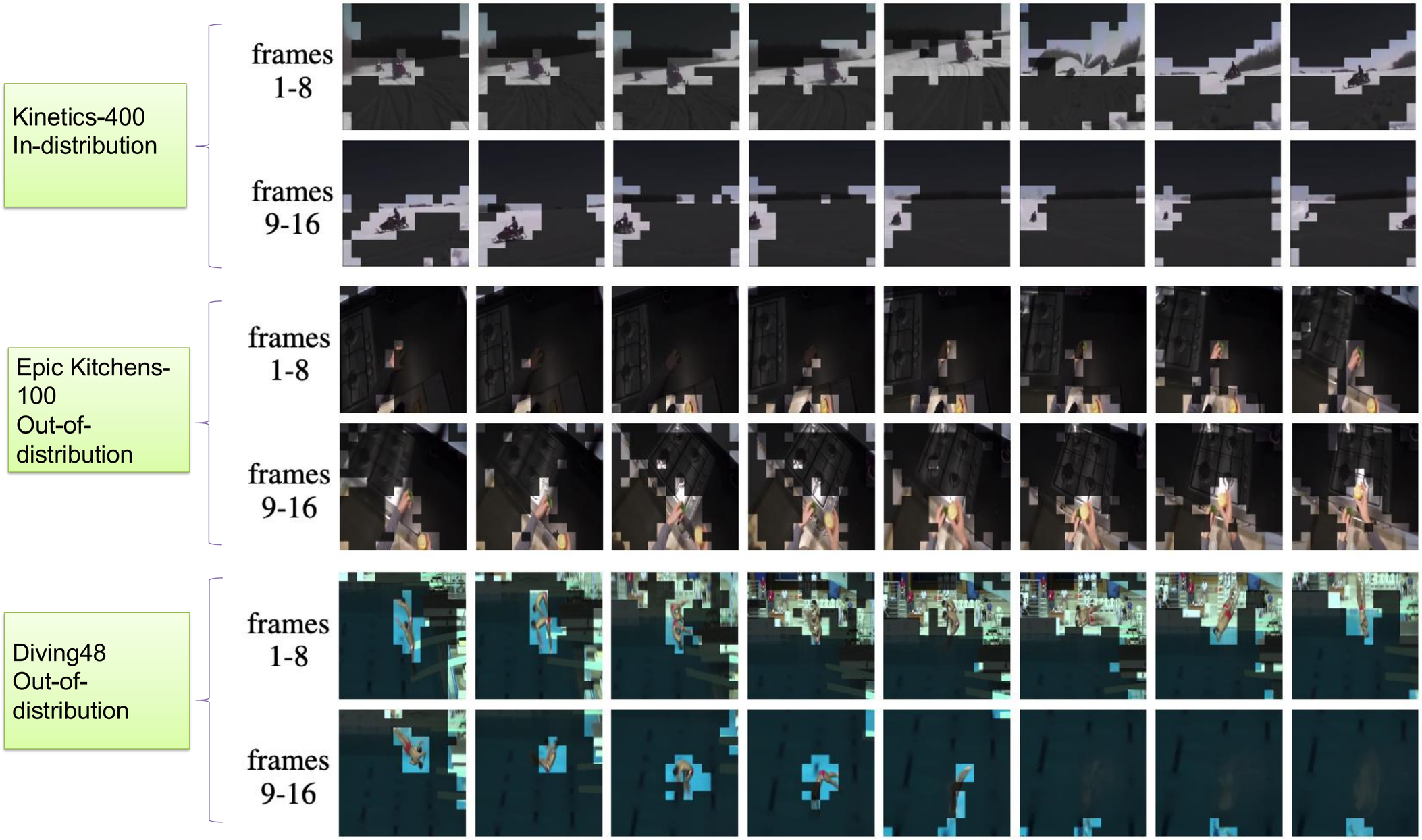
Kinetics-400
In-distribution



Epic Kitchens-100
Out-of-distribution



Qualitative Results



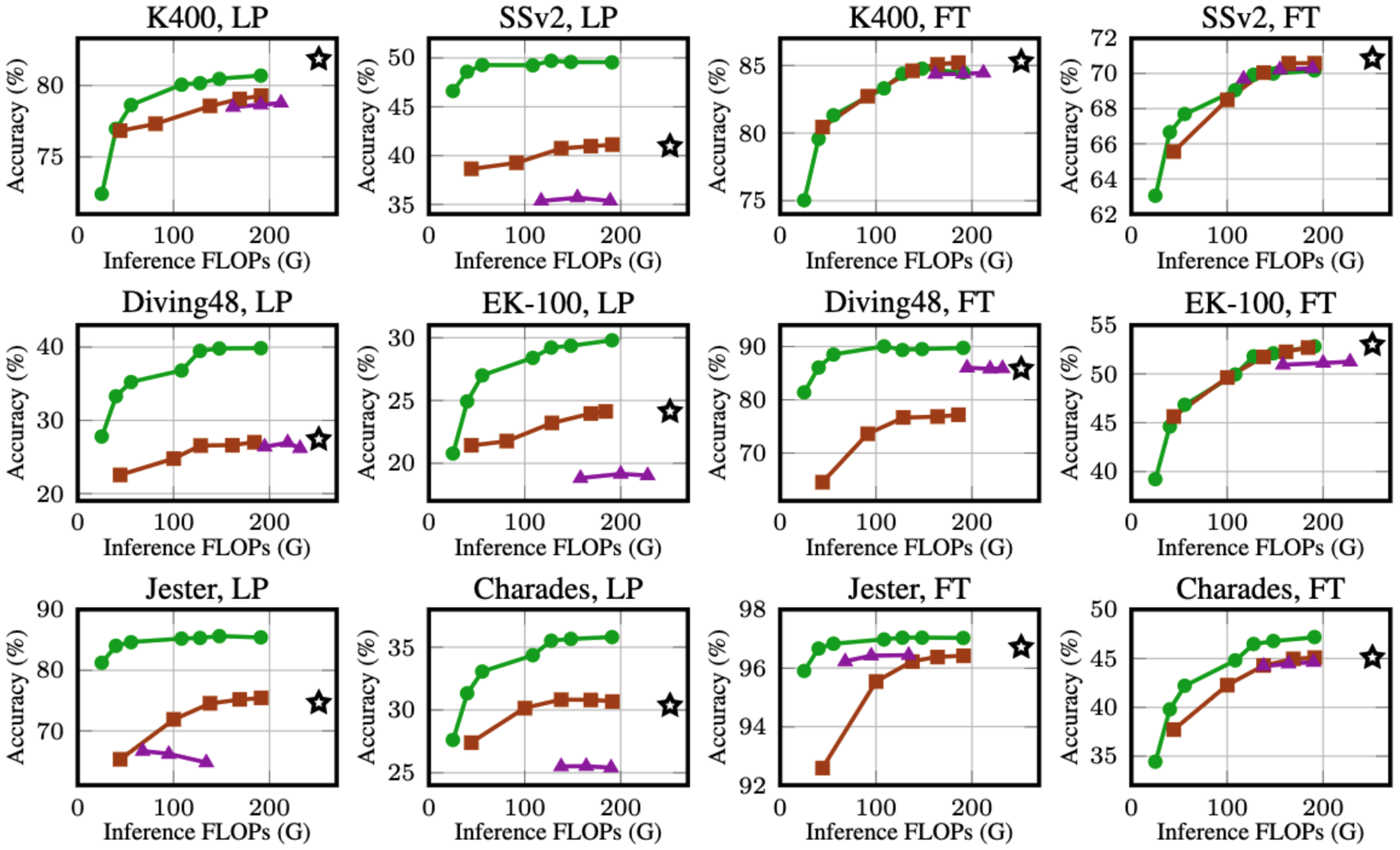
Quantitative Results

Model	Params M	Kinetics-400 FLOPs G×T×S	Kinetics-400 Top-1 % acc.	SSv2 FLOPs G×T×S	SSv2 Top-1 % acc.
Larger models for reference only, not for direct comparisons					
V-JEPA-2 [25]	355	935×8×3	85.1	935×2×3	73.7
InternVideo2 [18]	1020	2500×4×3	89.4	2500×2×3	69.7
ViT-B, Swin-B, or Mamba-M models					
UMT-B800e [29]	87	180×4×3	85.7	180×2×3	70.8
VideoMAE [3]	87	180×5×3	81.5	180×2×3	70.8
VideoMAEv2 [5]	87	180×5×3	81.5	180×2×3	71.2
VideoMamba-M800e [30]	74	202×4×3	83.4	202×2×3	71.0
VideoMambaPro [31]	72	392×4×3	84.0	183×4×3	69.4
VideoSwin + STTS ($T_0^{0.6}$) [32]	89	181×4×3	81.4	190×1×3	68.1
VideoMAE + LITE (K=0.3) [26]	87	46×5×3	78.4	46×2×3	68.3
VideoMAE + ToMe (r=64) [10, 8]	87	131×4×3	80.0	131×4×3	69.7
VideoMAE + RLT (τ=0.1) [8]	87	120×4×3	80.1	120×4×3	70.2
VideoMAE + vid-TLDR [11]	87	—	—	57×unk	69.6
LookWhen (90% sparse)	106	40×4×3	82.6	40×2×3	69.3
LookWhen (70% sparse)	106	108×4×3	84.6	108×2×3	72.0

LookWhen achieves the best inference accuracy-computation trade-off

Quantitative Results

Linear probing (LP) and fine-tuning (FT) accuracy vs. FLOPs



LookWhen (•) mostly outperforms the baselines (vid-TLDR (■), RLT (▲)) and in some cases the video teacher (InternVideo2 (⋆)) in a controlled settings

Ablations

Teacher	Method	K400-20K		SSv2-20K		Diving48		EK100-20K		Jester-20K		Charades	
		LP	FT	LP	FT	LP	FT	LP	FT	LP	FT	LP	FT
DINOv3	top1-dist	72.9	73.6	31.5	45.2	**29.7**	**83.2**	**11.6**	**22.2**	**63.4**	**94.7**	**29.3**	**36.7**
DINOv3	attn	72.1	72.8	30.1	41.1	28.0	77.7	9.1	17.6	53.8	90.1	26.6	33.2
DINOv3	Δattn	**73.5**	**74.2**	**31.7**	**45.6**	27.6	76.0	11.0	21.3	**63.4**	94.4	28.6	35.7
InternVideo2	top1-dist	64.7	64.7	27.0	31.2	15.8	44.6	7.3	13.4	56.8	89.2	21.9	26.0
InternVideo2	attn	68.6	68.9	28.6	35.4	26.2	72.9	9.3	15.2	58.2	89.9	24.4	30.4
none	random	72.5	72.7	30.4	42.8	18.3	57.2	10.2	20.3	57.2	91.8	25.9	30.8

Message 1: Training to select unique tokens beats highly-attended tokens, if we have suitable teacher features (e.g. DINOv3)

Ablations

Teacher	Method	K400-20K LP	K400-20K FT	SSv2-20K LP	SSv2-20K FT	Diving48 LP	Diving48 FT	EK100-20K LP	EK100-20K FT	Jester-20K LP	Jester-20K FT	Charades LP	Charades FT
DINOv3	top1-dist	72.9	73.6	31.5	45.2	**29.7**	**83.2**	**11.6**	**22.2**	**63.4**	**94.7**	**29.3**	**36.7**
DINOv3	attn	72.1	72.8	30.1	41.1	28.0	77.7	9.1	17.6	53.8	90.1	26.6	33.2
DINOv3	Δattn	**73.5**	**74.2**	**31.7**	**45.6**	27.6	76.0	11.0	21.3	**63.4**	94.4	28.6	35.7
InternVideo2	top1-dist	64.7	64.7	27.0	31.2	15.8	44.6	7.3	13.4	56.8	89.2	21.9	26.0
InternVideo2	attn	68.6	68.9	28.6	35.4	26.2	72.9	9.3	15.2	58.2	89.9	24.4	30.4
none	random	72.5	72.7	30.4	42.8	18.3	57.2	10.2	20.3	57.2	91.8	25.9	30.8

Message 1: Training to select unique tokens beats highly-attended tokens, if we have suitable teacher features (e.g. DINOv3)

IntVid2	DINOv3 Vid	DINOv3 Frame	DINOv3 Patch	DINOv3 Norm	K400-20K LP	K400-20K FT	SSv2-20K LP	SSv2-20K FT	Diving48 LP	Diving48 FT	EK100-20K LP	EK100-20K FT	Jester-20K LP	Jester-20K FT	Charades LP	Charades FT
✔	✔	✘	✘	✔	**73.3**	**74.0**	**37.9**	48.0	29.9	81.9	**12.8**	**23.3**	69.0	94.6	**30.3**	**37.4**
✔	✔	✘	✘	✘	72.1	73.3	31.5	45.1	**32.1**	84.2	11.3	22.4	63.9	94.9	29.5	36.9
✔	✘	✘	✘	✘	**73.3**	73.9	32.1	45.6	27.9	83.1	11.0	22.0	61.6	94.4	29.0	36.3
✘	✔	✘	✘	✔	17.9	61.1	26.7	**48.8**	26.7	**85.7**	2.6	18.7	**70.4**	**95.2**	14.0	30.5

Message 2: Predicting **both** InternVideo2's video token and a video token we make from time-normalized DINOv3's frame-wise class tokens improves representations

Ablations

Teacher	Method	K400-20K LP	K400-20K FT	SSv2-20K LP	SSv2-20K FT	Diving48 LP	Diving48 FT	EK100-20K LP	EK100-20K FT	Jester-20K LP	Jester-20K FT	Charades LP	Charades FT
DINOv3	top1-dist	72.9	73.6	31.5	45.2	**29.7**	**83.2**	**11.6**	**22.2**	**63.4**	**94.7**	**29.3**	**36.7**
DINOv3	attn	72.1	72.8	30.1	41.1	28.0	77.7	9.1	17.6	53.8	90.1	26.6	33.2
DINOv3	Δattn	**73.5**	**74.2**	**31.7**	**45.6**	27.6	76.0	11.0	21.3	**63.4**	94.4	28.6	35.7
InternVideo2	top1-dist	64.7	64.7	27.0	31.2	15.8	44.6	7.3	13.4	56.8	89.2	21.9	26.0
InternVideo2	attn	68.6	68.9	28.6	35.4	26.2	72.9	9.3	15.2	58.2	89.9	24.4	30.4
none	random	72.5	72.7	30.4	42.8	18.3	57.2	10.2	20.3	57.2	91.8	25.9	30.8

Message 1: Training to select unique tokens beats highly-attended tokens, if we have suitable teacher features (e.g. DINOv3)

IntVid2	DINOv3 Vid	DINOv3 Frame	DINOv3 Patch	DINOv3 Norm	K400-20K LP	K400-20K FT	SSv2-20K LP	SSv2-20K FT	Diving48 LP	Diving48 FT	EK100-20K LP	EK100-20K FT	Jester-20K LP	Jester-20K FT	Charades LP	Charades FT
✔	✔	✘	✘	✔	**73.3**	**74.0**	**37.9**	48.0	29.9	81.9	**12.8**	**23.3**	69.0	94.6	**30.3**	**37.4**
✔	✔	✘	✘	✘	72.1	73.3	31.5	45.1	**32.1**	84.2	11.3	22.4	63.9	94.9	29.5	36.9
✔	✘	✘	✘	✘	**73.3**	73.9	32.1	45.6	27.9	83.1	11.0	22.0	61.6	94.4	29.0	36.3
✘	✔	✘	✘	✔	17.9	61.1	26.7	**48.8**	26.7	**85.7**	2.6	18.7	**70.4**	**95.2**	14.0	30.5

Message 2: Predicting **both** InternVideo2's video token and a video token we make from time-normalized DINOv3's frame-wise class tokens improves representations

Model	Sparsity	FLOPs ratio	K400-20K LP	K400-20K FT	SSv2-20K LP	SSv2-20K FT	Diving48 LP	Diving48 FT	EK100-20K LP	EK100-20K FT	Jester-20K LP	Jester-20K FT	Charades LP	Charades FT
LookWhen (224px)	90%	1.0×	72.5	73.8	**38.1**	**50.1**	**37.6**	84.8	13.5	23.8	**79.0**	95.2	31.3	39.3
Dense extractor (224px)	0%	11.3×	**73.7**	**75.8**	34.5	47.7	37.2	**87.4**	**13.9**	**27.8**	78.1	**95.5**	**34.5**	**44.7**
Dense extractor (80px)	0%	0.9×	68.4	70.8	31.9	45.2	29.4	82.5	8.9	18.7	77.6	95.2	26.3	34.4

Message 3: Selection can improve accuracy while reducing compute. With just 10% of the patches, LookWhen beats the full-resolution dense extractor on SSv2 in both linear probing and fine-tuning.

Takeaways

- **Select cheaply, extract sparsely.** Scan a downscaled video and process only the top-K patches at full resolution.
- **Select unique patches across space and time.** Measure uniqueness by nearest-neighbor distance in DINOv3 feature space.
- **Learn from image and video teachers.** Distill InternVideo2 features and time-normalized DINOv3 features.
- **LookWhen achieves better throughput**. 6.7× the inference throughput of InternVideo2-B at matched average accuracy.

📅 Wed, Dec 9, 2026 • 10:00 AM – 1:00 PM AEDT

📍 Hall 1-4

Paper

Project

