Paper & Poster

Paxion: Patching Action Knowledge in Video-Language Foundation Models

NeurIPS2023 Spotlight

Zhenhailong Wang, Ansel Blume, Sha Li, Genglin Liu, Jaemin Cho, Zineng Tang, Mohit Bansal, Heng Ji

ILLINOIS

THE UNIVERSITY of NORTH CAROLINA at CHAPEL HILL

NEURAL INFORMATION PROCESSING SYSTEMS

Background: Current VLMs struggle to understand concepts beyond nouns

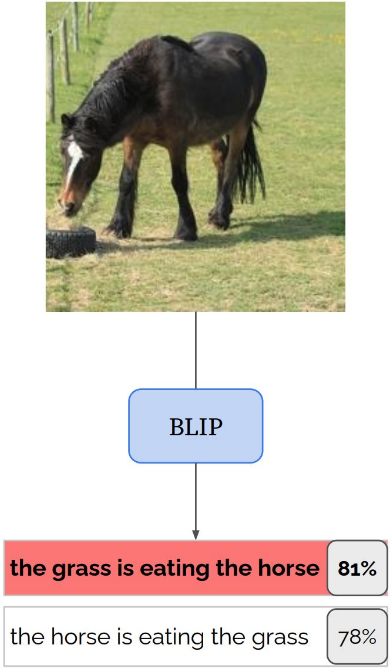
Visual Genome Relation
Assessing relational understanding (23,937 test cases)

✔ *the horse is eating the grass*
X *the grass is eating the horse*

Visual Genome Attribution
Assessing attributive understanding (28,748 test cases)

✔ *the paved road and the white house*
X *the white road and the paved house*

https://arxiv.org/abs/2210.01936



Recent VLMs face challenges in understanding visual language concepts beyond object nouns (e.g., recognizing attributes, relations, states)

Background: How about actions?

?

?

"Book **falling** like a rock" ✓

Original Action Text

"Book **rising** like a rock" ✗

Action Antonym Text

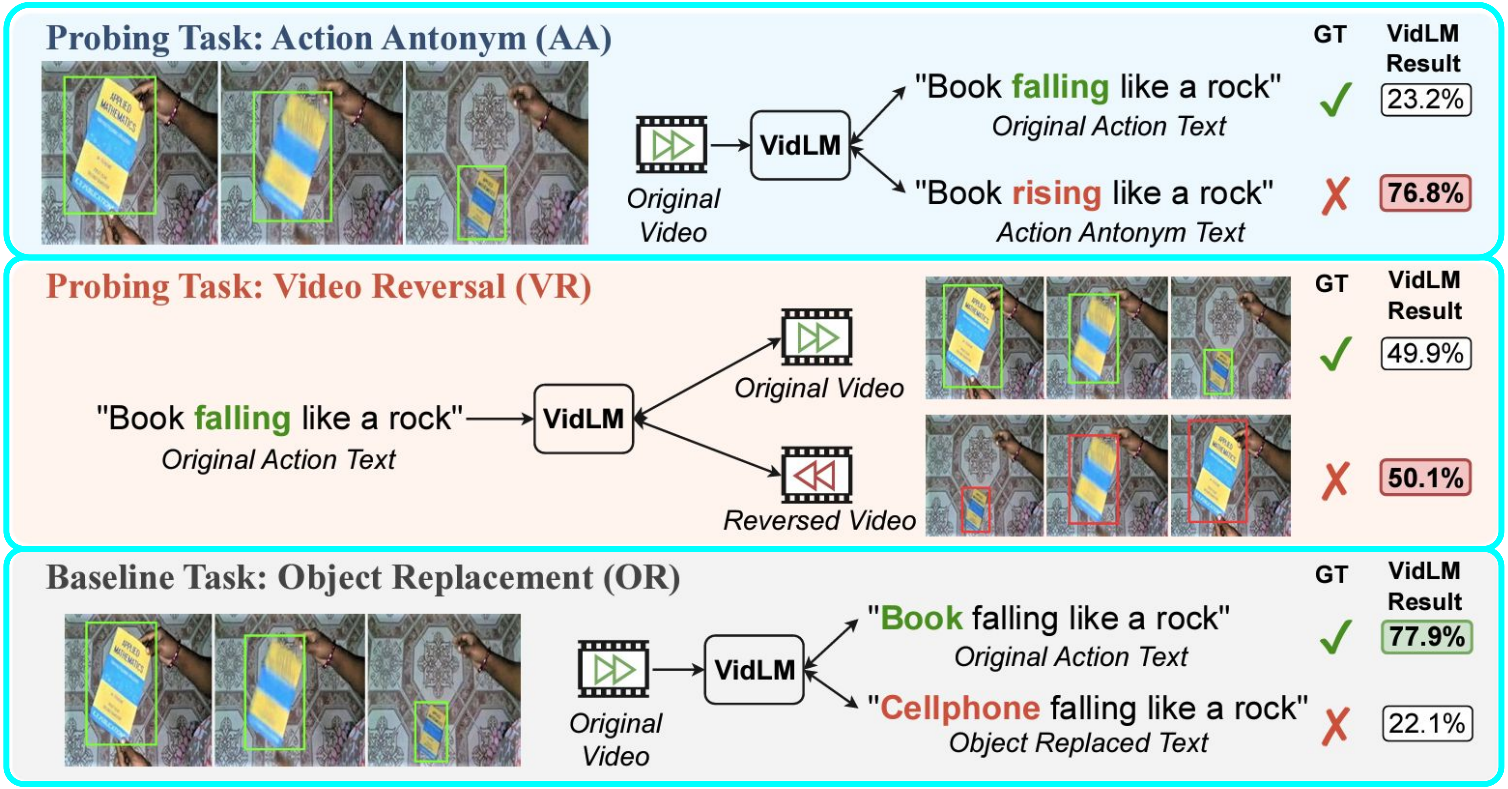
The understanding of the cause and effect of actions in textual, visual, and temporal dimensions

Action Knowledge

ActionBench: Do SOTA VidLM really understand actions?

➤ **Action Dynamics Benchmark (ActionBench)** based on two VL datasets: SSv2, Ego4d

- **Probing tasks: Action Antonym (AA), Video Reversal (VR)**
- **Baseline task: Object Replacement (OR)**

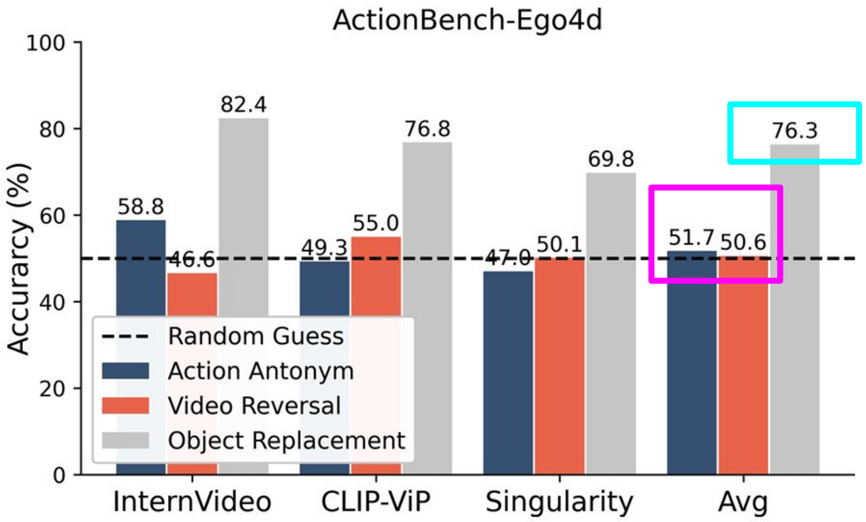


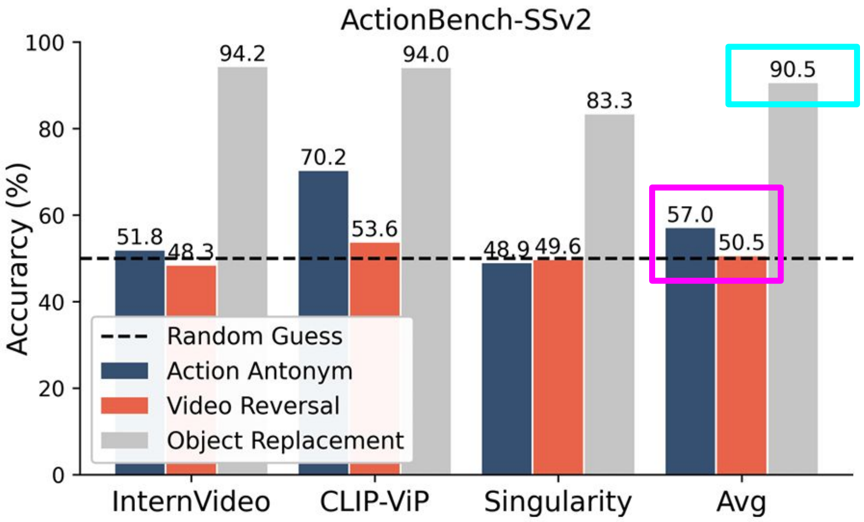
ActionBench: Do SOTA VidLM really understand actions?

➤ Evaluating SOTA VidLMs on ActionBench

Near random performance on **Action Antonym (AA)** and **Video Reversal (VR)**

Clear **biases towards object nouns** compared to actions

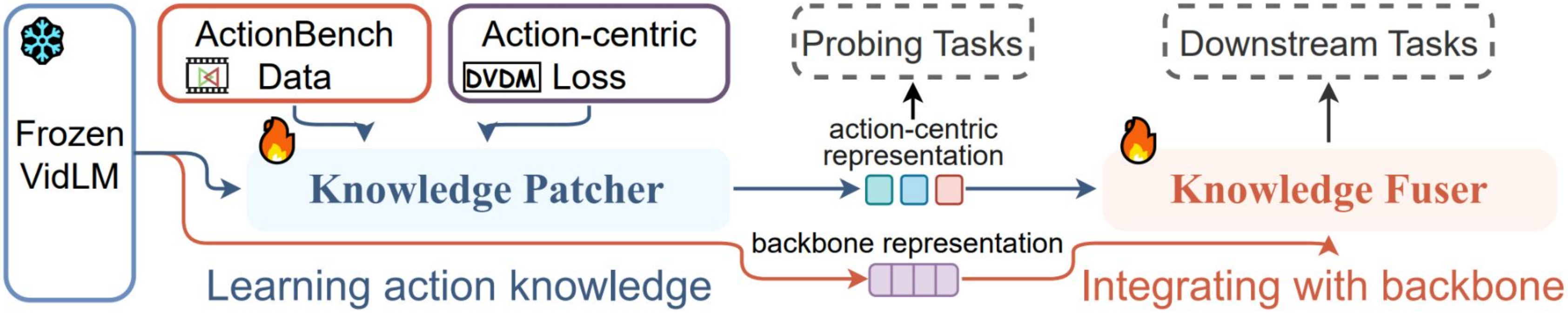




PAXION Framework Overview

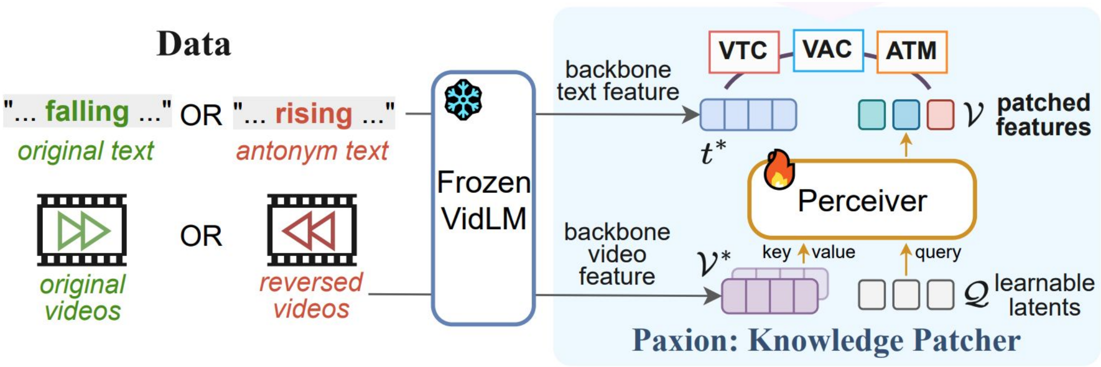
➤ **Patch ⟶ Fuse**

How can we **patch action knowledge** into existing VidLMs **without compromising their general VL capabilities?**



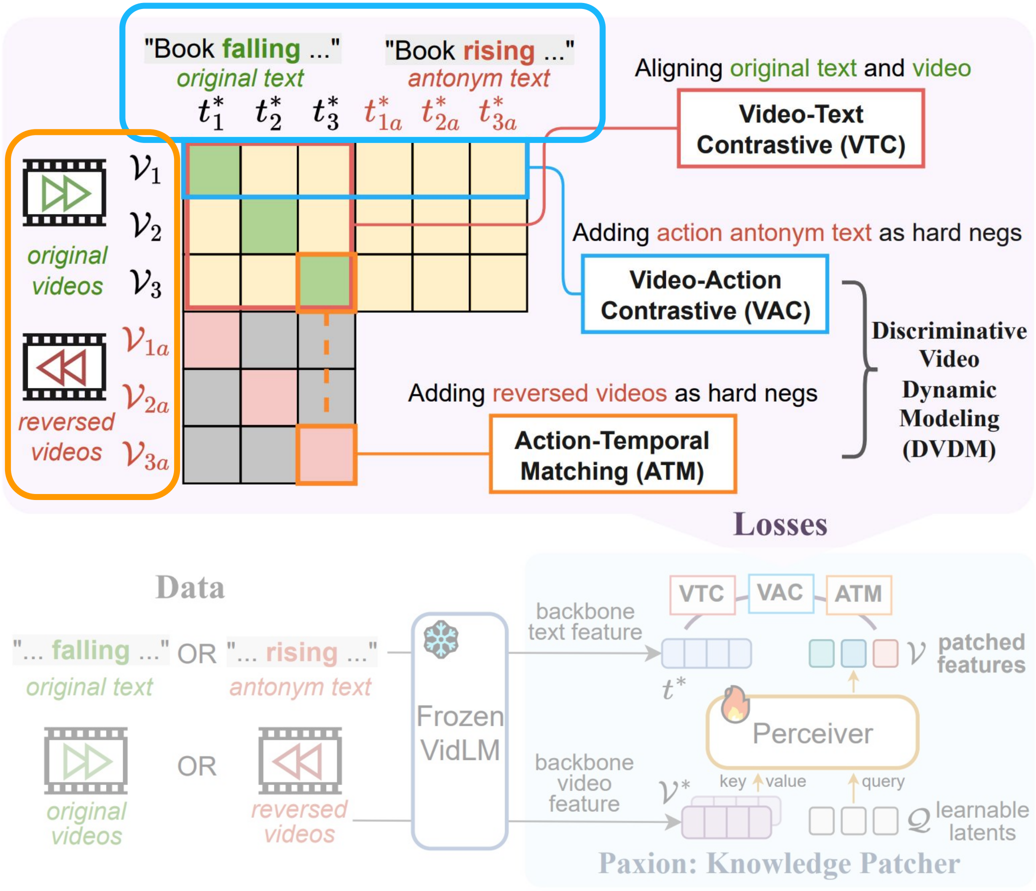
PAXION

Knowledge Patcher: Patching frozen VLMs with Action Knowledge



A light-weight **Perceiver-based module** attached to a frozen VidLM

Knowledge Patcher: Patching frozen VLMs with Action Knowledge



New training objects DVDM (VAC, ATM losses) to force the model to encode action dynamics

Video-Action Contrastive (VAC): encourages learning the **alignment** between the **video** and the **action verbs**

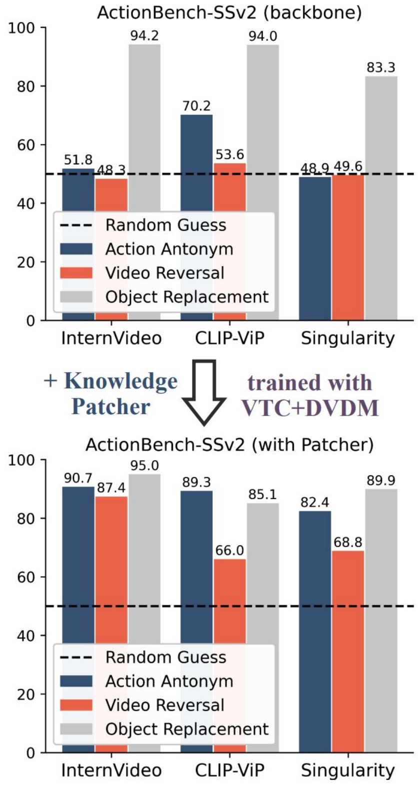
Action-Temporal Matching: encourages learning the correct **temporal ordering** implied by the **action text**

Knowledge Patcher: Patching frozen VLMs with Action Knowledge

DVDM objectives **significantly improves action understanding (near-random ⇒ ~80%)**

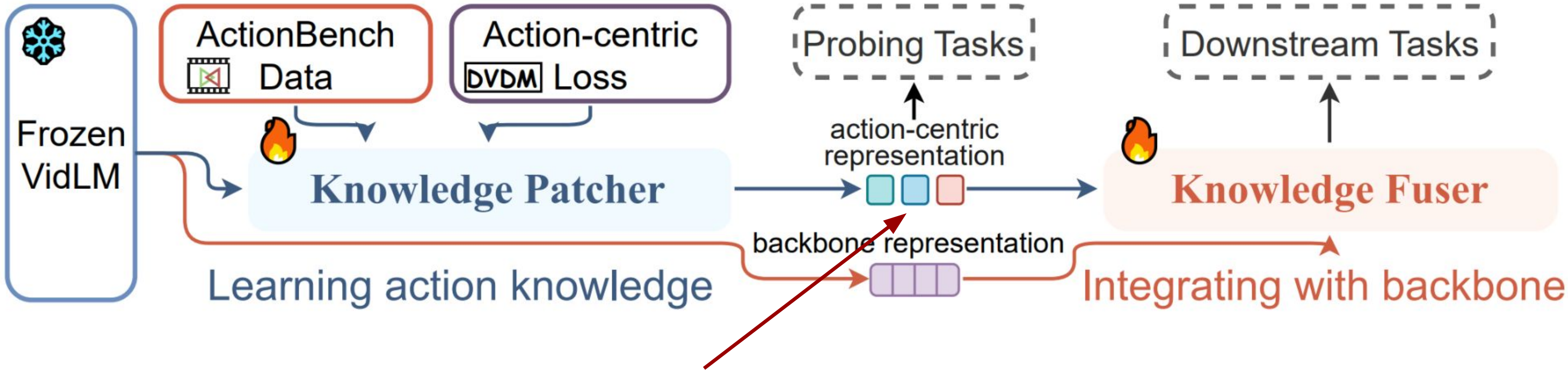
Action Dynamics Benchmark (ActionBench) Results

Backbone	**Method** [**Patcher Training Loss**]	**Trainable Param#**	**AA** (Ego4d)	**VR** (Ego4d)	**AA** (SSv2)	**VR** (SSv2)	**Avg**
InternVideo	Backbone	-	58.8	46.2	51.8	48.3	51.3
	KP-Transformer [VTC]	8.4M (1.8%)	68.2	62.8	65.5	60.6	64.3
	KP-Perceiver [VTC]	4.2M (0.9%)	66.5	63.6	69.8	71.0	67.7
	KP-Perceiver [VTC+**DVDM**]	4.2M (0.9%)	**90.1**	**75.5**	**90.7**	**87.4**	**85.9**
Clip-ViP	Backbone	-	49.3	55.0	70.2	53.6	57.0
	KP-Transformer [VTC]	3.9M (2.6%)	61.9	53.4	72.2	54.3	60.5
	KP-Perceiver [VTC]	2.4M (1.6%)	61.9	54.6	71.5	48.8	59.2
	KP-Perceiver [VTC+**DVDM**]	2.4M (1.6%)	**89.3**	**56.9**	**89.3**	**66.0**	**75.4**
Singularity	Backbone	-	47.0	50.1	48.9	49.6	48.9
	KP-Transformer [VTC]	3.9M (1.8%)	61.9	48.2	63.8	49.5	55.9
	KP-Perceiver [VTC]	1.3M (0.6%)	60.3	46.1	63.3	51.5	55.3
	KP-Perceiver [VTC+**DVDM**]	1.3M (0.6%)	**83.8**	**58.9**	**82.4**	**68.8**	**73.5**
Human			92.0	78.0	96.0	90.0	89.0



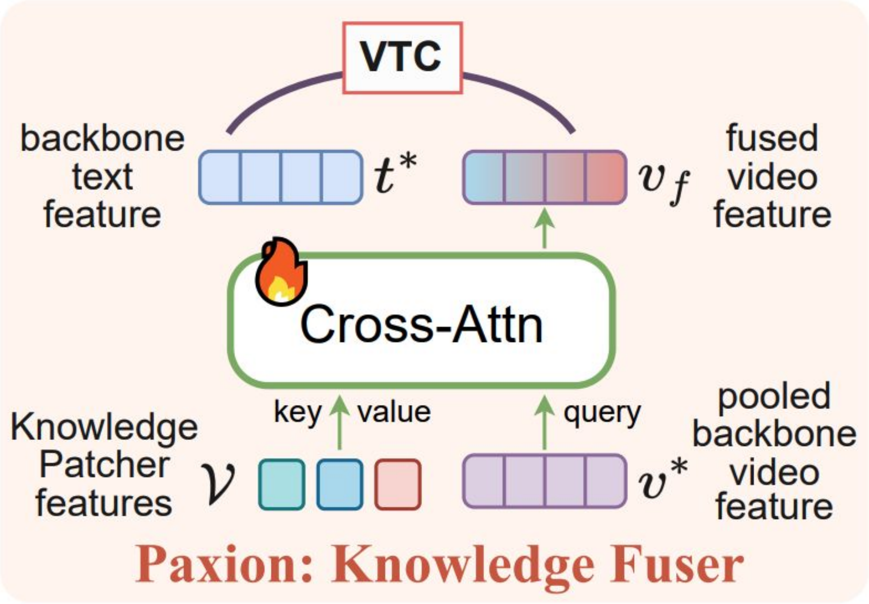
Knowledge Fuser: Retaining VL capabilities while leveraging the patched action knowledge

How can we patch action knowledge into existing VidLMs **without compromising their general VL capabilities?**



The KP representation is highly specialized in action understanding

Knowledge Fuser: Retaining VL capabilities while leveraging the patched action knowledge



A light-weight cross-attention module which **fuses the learned Knowledge Patcher features** with the **frozen backbone features**

Knowledge Fuser: Retaining VL capabilities while leveraging the patched action knowledge

➢ Downstream Tasks

Video-Text Retrieval	**Causal-Temporal VQA**	**Video-to-Action Retrieval**	
SSv2-Label	NExT-QA	SSv2-Template	Temporal

⟵⟶

More object-centric — Require causal and temporal understanding of objects and actions — More action-centric

Video-Text Retrieval Example

SSv2-label

"pushing scissors so that it falls off the table"

Video-to-Action Retrieval Example

SSv2-template (where the main object is obfuscated)

"pushing something so that it falls off the table"

Knowledge Fuser: Retaining VL capabilities while leveraging the patched action knowledge

➤ **Downstream Task Results**

PAXION with Knowledge Fuser outperforms/performs competitively with VTC-only baselines on both object-centric and action-centric tasks

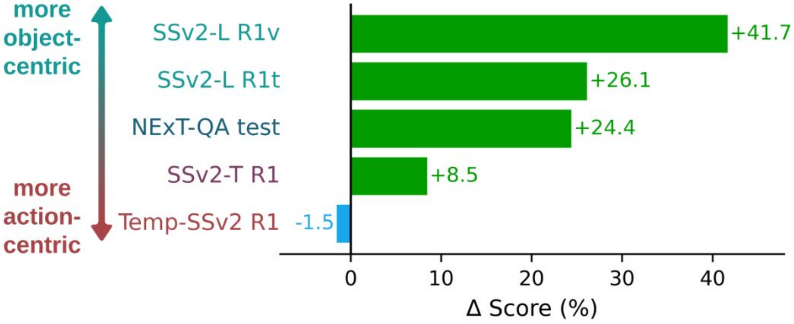
Method [Patcher Training Loss]	Video-Text Retrieval SSv2-label				Video-to-Action Retrieval SSv2-template		Temporal-SSv2	
	$R1_{v2t}$	$R5_{v2t}$	$R1_{t2v}$	$R5_{t2v}$	$R1$	$R5$	$R1$	$R5$
InternVideo Backbone	18.8	39.9	19.9	40.0	5.6	15.9	11.2	35.8
KP-Transformer FT [VTC]	24.1	50.0	21.7	46.0	21.1	55.9	41.1	88.9
KP-Perceiver FT [VTC]	27.0	57.4	27.1	**56.8**	24.8	59.7	42.5	91.3
Side-Tuning [61] [VTC+DVDM]	30.9	59.2	26.6	53.1	22.2	55.1	50.2	90.9
PAXION [VTC+DVDM]	**32.3**	**61.2**	**28.0**	54.3	**26.9**	**61.5**	**51.2**	**91.9**

Method [Patcher Training Loss]	NExT-QA Original				ATP-hard [7]		
	C	T	D	all	C	T	all
InternVideo Backbone	43.3	38.6	52.5	43.2	27.0	27.3	27.1
KP-Transformer FT [VTC]	46.1	45.0	61.3	48.1	32.5	33.6	33.0
KP-Perceiver FT [VTC]	46.0	46.0	58.9	48.0	30.1	31.6	30.7
Side-Tuning [60] [VTC+DVDM]	54.9	52.0	**69.8**	56.3	37.4	36.0	36.8
PAXION [VTC+DVDM]	**56.0**	**53.0**	68.5	**57.0**	**38.8**	**38.1**	**38.5**

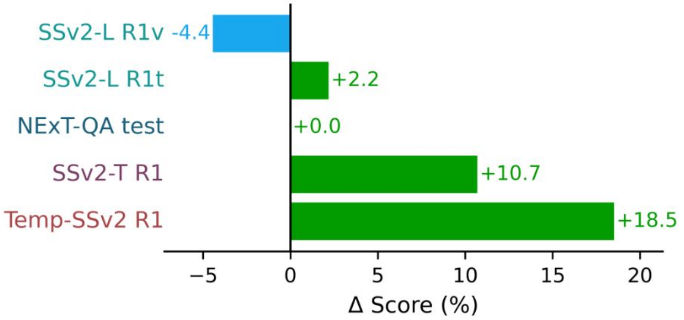
Paxion **helps more on T and C questions**, and on **ATP-hard** where the temporal and action knowledge is emphasized

Knowledge Fuser: Retaining VL capabilities while leveraging the patched action knowledge

➤ Analysis



Finetune v.s. **Fuse**



VTC v.s. **VTC+DVDM**

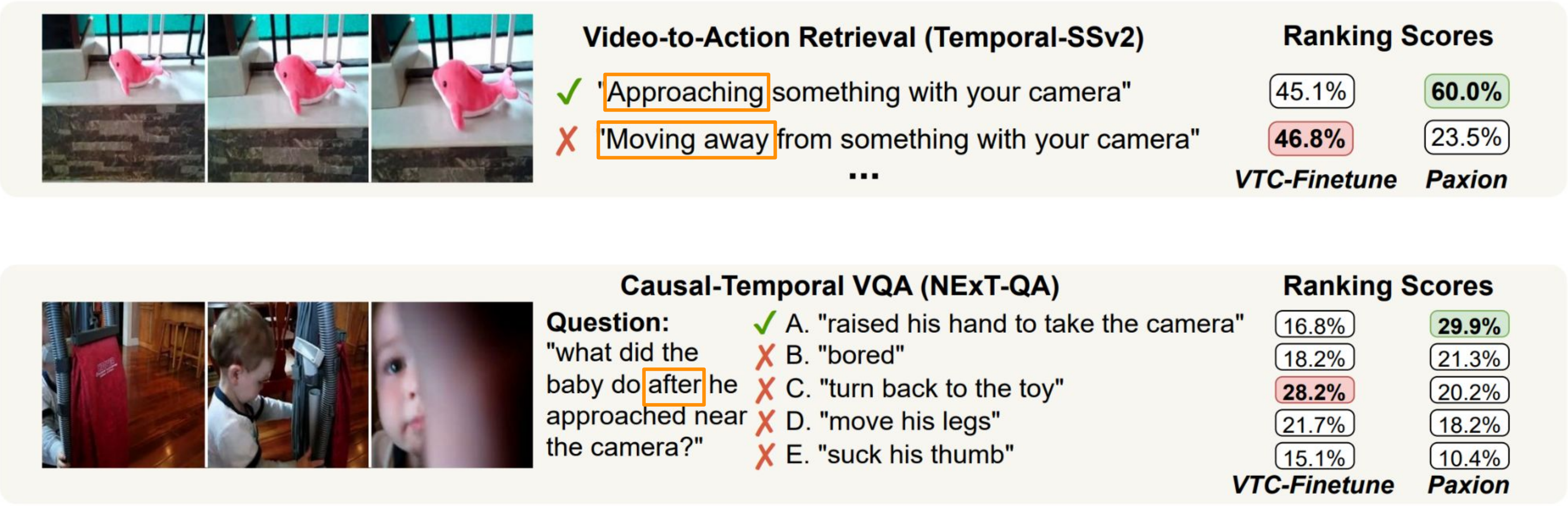
Compared setting: Finetune Knowledge Patcher[VTC+DVDM] **w/o a Knowledge Fuser**

Compared setting: Add Knowledge Fuser to a **Knowledge Patcher trained with only VTC loss**

Knowledge Fuser is **essential for retaining object understanding** capabilities

DVDM patching improves action understanding on downstream tasks

Qualitative Examples of PAXION

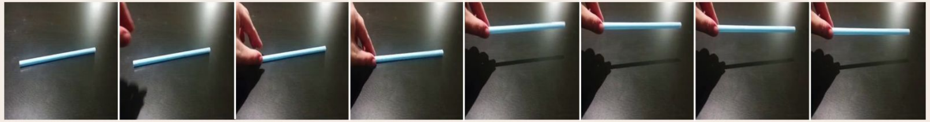


Qualitative Examples of PAXION

Remaining challenges

Video-Text Retrieval & Video-to-Action Retrieval Failure Examples

Negation



Dataset	GT	Text Candidates	VTC-Finetune Score	VTC-Finetune Rank	Paxion Score	Paxion Rank
Temporal-SSv2	✓	"Lifting something up completely *without* letting it drop down"	25.4%	2	28.4%	2
	✗	"Lifting up one end of something, *then* letting it drop down"	**47.2%**	**1**	**39.0%**	**1**
		...				

Object identification

Dataset	GT	Text Candidates	VTC-Finetune Score	VTC-Finetune Rank	Paxion Score	Paxion Rank
SSv2-label	✓	"bending tube so that it deforms"	0.02%	286	0.05%	213
	✗	"holding soaps over tooth paste"	**13.9%**	**1**	**15.4%**	**1**
		...				

NExT-QA Failure Example

Counting

Question	GT	Answer Candidates	VTC-Finetune Score	VTC-Finetune Rank	Paxion Score	Paxion Rank
"how many goats can be spotted?"	✓	A. "eight"	8.5%	5	17.3%	5
	✗	B. "one"	23.3%	3	17.5%	4
	✗	C. "two"	**27.2%**	**1**	20.2%	3
	✗	D. "three"	24.8%	2	**24.3%**	**1**
	✗	E. "four"	16.2%	4	20.6%	2

Paper & Poster

Paxion: Patching Action Knowledge in Video-Language Foundation Models

NeurIPS2023 Spotlight

Zhenhailong Wang, Ansel Blume, Sha Li, Genglin Liu, Jaemin Cho, Zineng Tang, Mohit Bansal, Heng Ji

ILLINOIS

THE UNIVERSITY of NORTH CAROLINA at CHAPEL HILL

NEURAL INFORMATION PROCESSING SYSTEMS