Google

VideoPrism: A Foundational Visual Encoder for Video Understanding

Long Zhao[*] Nitesh B. Gundavarapu[*] Liangzhe Yuan[*] Hao Zhou[*] Shen Yan[†] Jennifer J. Sun[†] Luke Friedman[†] Rui Qian[†]
Tobias Weyand Yue Zhao Rachel Hornung Florian Schroff Ming-Hsuan Yang David A. Ross Huisheng Wang Hartwig Adam
Mikhail Sirotenko[‡] Ting Liu[‡] Boqing Gong[‡]

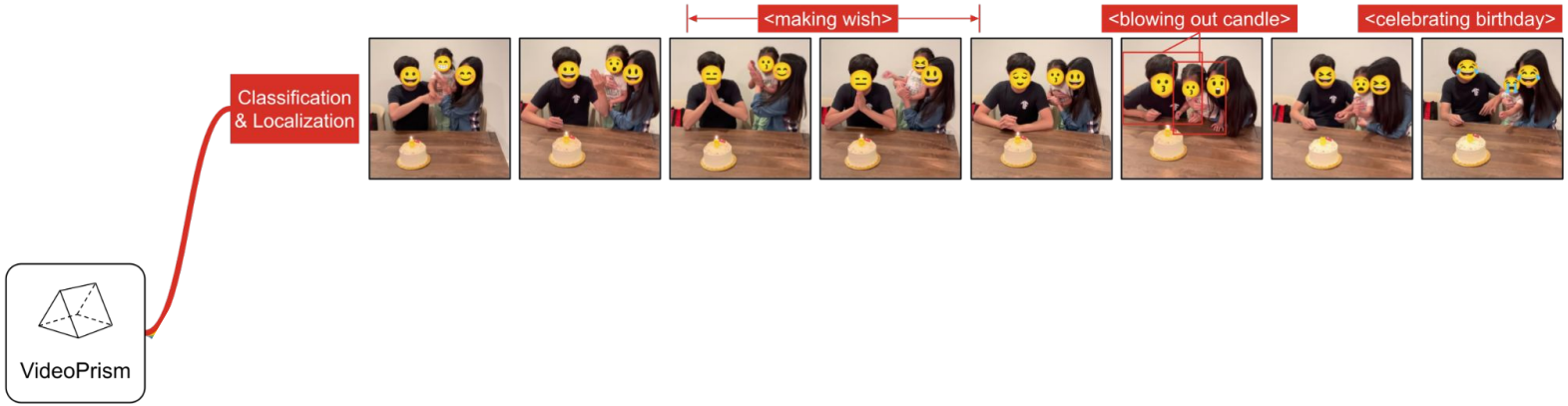
[*]Equal primary contribution [†]Equal core technical contribution [‡]Equal senior contribution, project leads

What is VideoPrism?

A foundational **video encoder** that enables **state-of-the-art** performance for **video understanding**

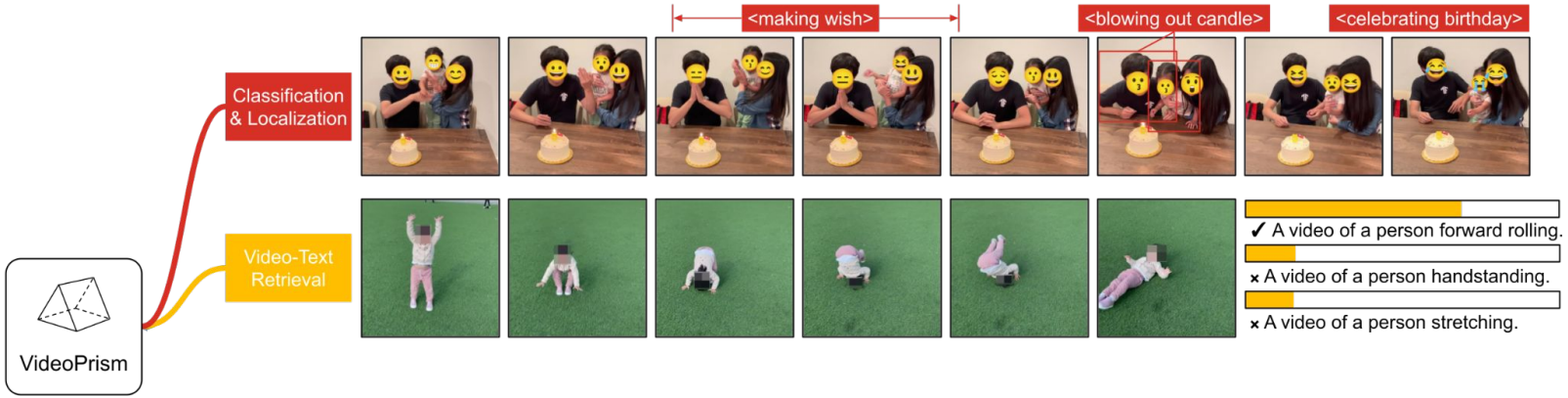
What is VideoPrism?

A foundational **video encoder** that enables **state-of-the-art** performance for **video understanding**



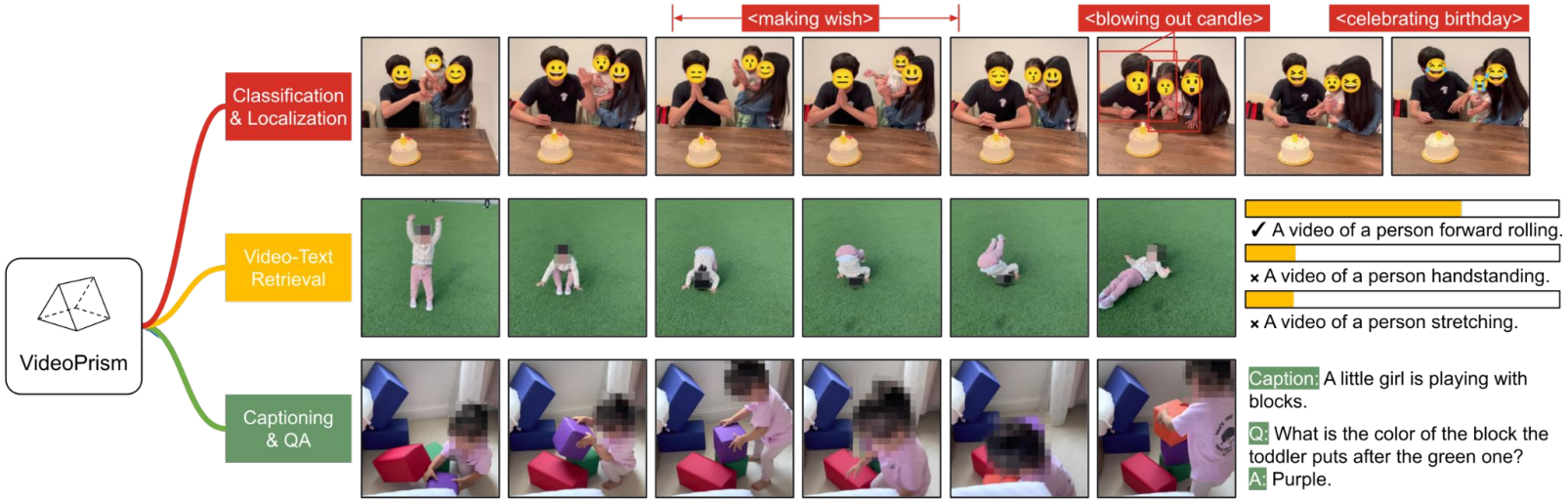
What is VideoPrism?

A foundational **video encoder** that enables **state-of-the-art** performance for **video understanding**



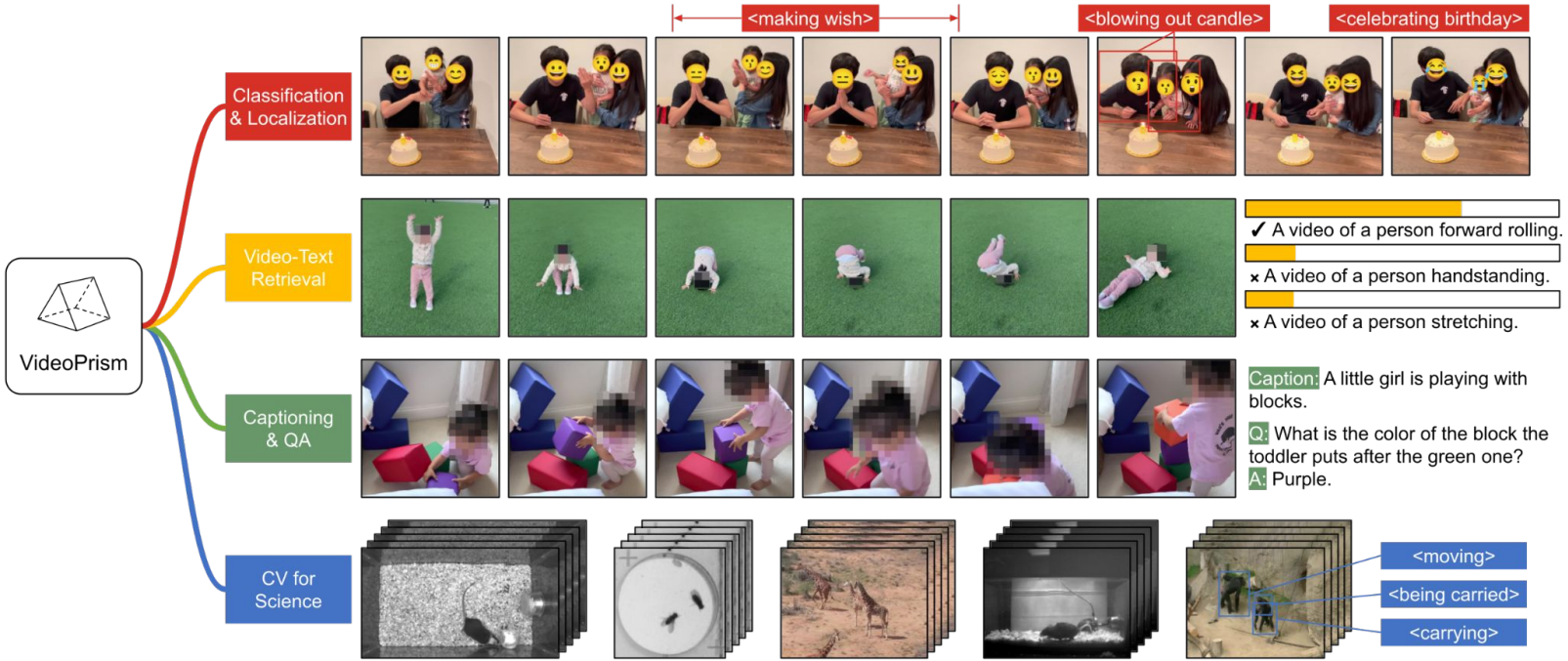
What is VideoPrism?

A foundational **video encoder** that enables **state-of-the-art** performance for **video understanding**



What is VideoPrism?

A foundational **video encoder** that enables **state-of-the-art** performance for **video understanding**



VideoPrism: Overview

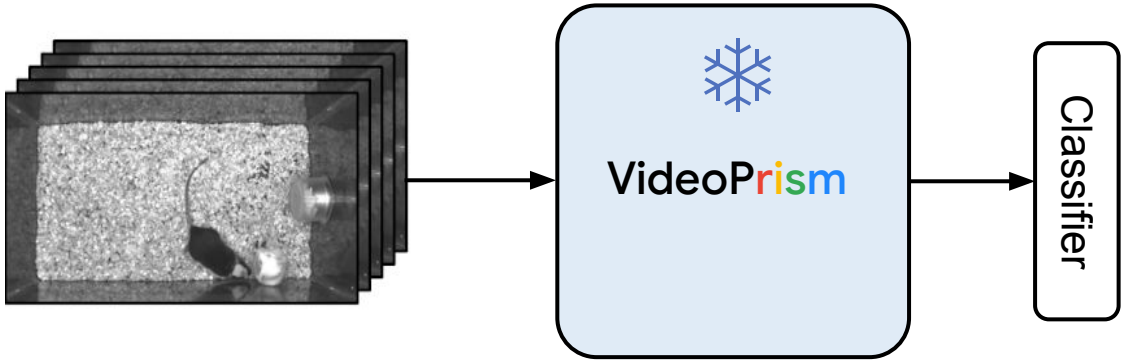
SOTA results on both **appearance** and **motion** understanding with a **single, frozen** backbone.

Models (frozen backbone, all models at Base scale)	**Kinetics-400** (appearance)	**Something-Something v2** (motion)	**AVA** (spatiotemporal)
CoCa	73.1	41.5	23.3
InternVideo	69.3	58.2	13.4
UMT	77.1	47.7	20.7
VideoPrism	**84.2** (+7.1)	**63.6** (+5.4)	**30.6** (+9.5)

VideoPrism: Overview

SOTA results on both **appearance** and **motion** understanding with a **single, frozen** backbone.

Generalizes well to **unseen domain**, outperforming **domain expert models** in AI4Science.

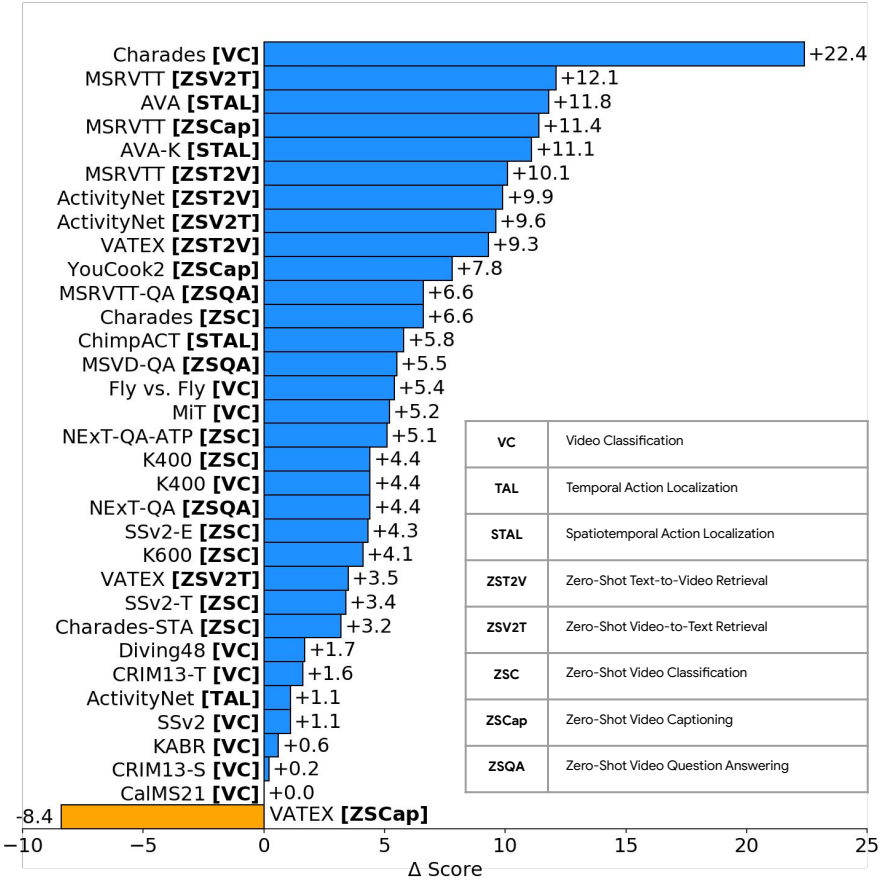


VideoPrism: Overview

SOTA results on both **appearance** and **motion** understanding with a **single, frozen** backbone.

Generalizes well to **unseen domain**, outperforming **domain expert models** in AI4Science.

VideoPrism outperforms prior-art foundation models on **31 out of 33** video understanding benchmarks.

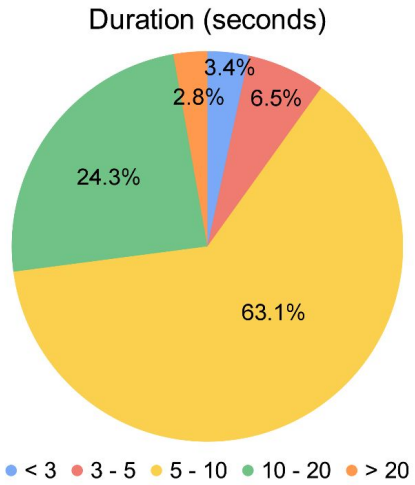


How is VideoPrism trained?

Large scale training data: **619M** video-text pairs:
(36M with high-quality captions + 583M with noisy parallel text).

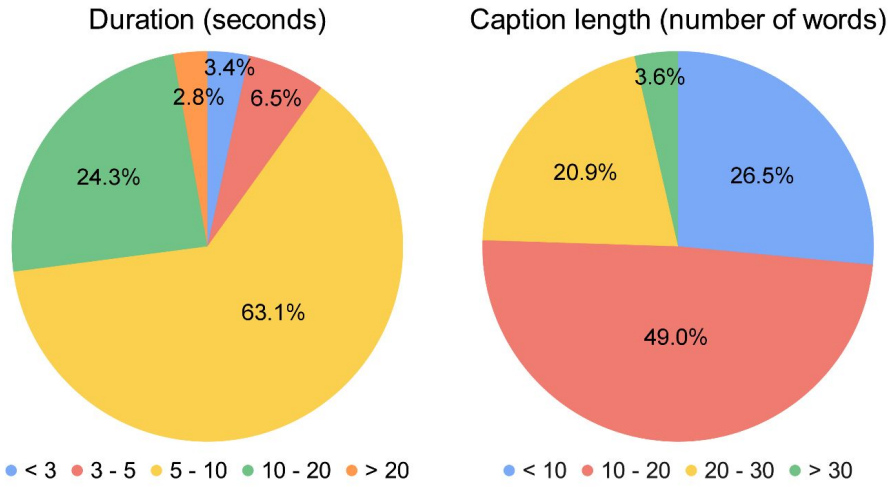
How is VideoPrism trained?

Large scale training data: **619M** video-text pairs:
(36M with high-quality captions + 583M with noisy parallel text).



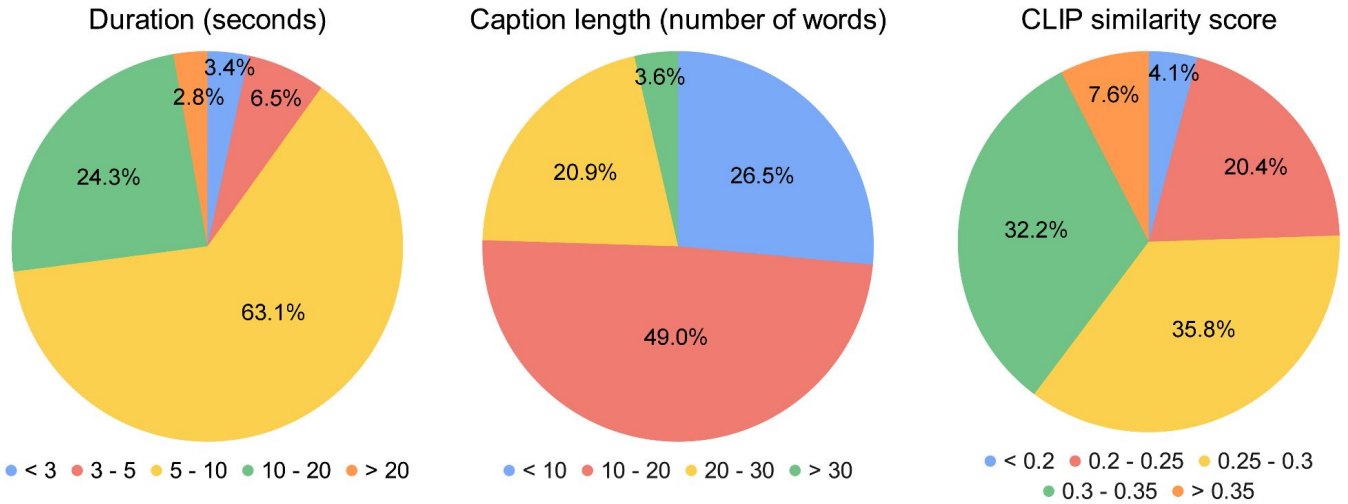
How is VideoPrism trained?

Large scale training data: **619M** video-text pairs:
(36M with high-quality captions + 583M with noisy parallel text).



How is VideoPrism trained?

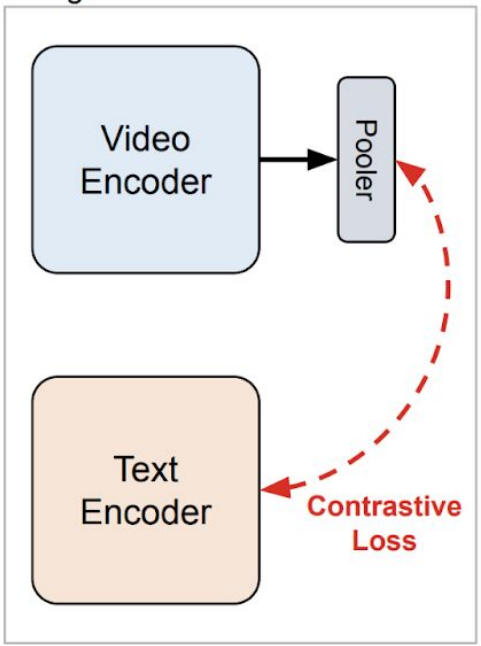
Large scale training data: **619M** video-text pairs:
(36M with high-quality captions + 583M with noisy parallel text).



How is VideoPrism trained?

Two stage training:

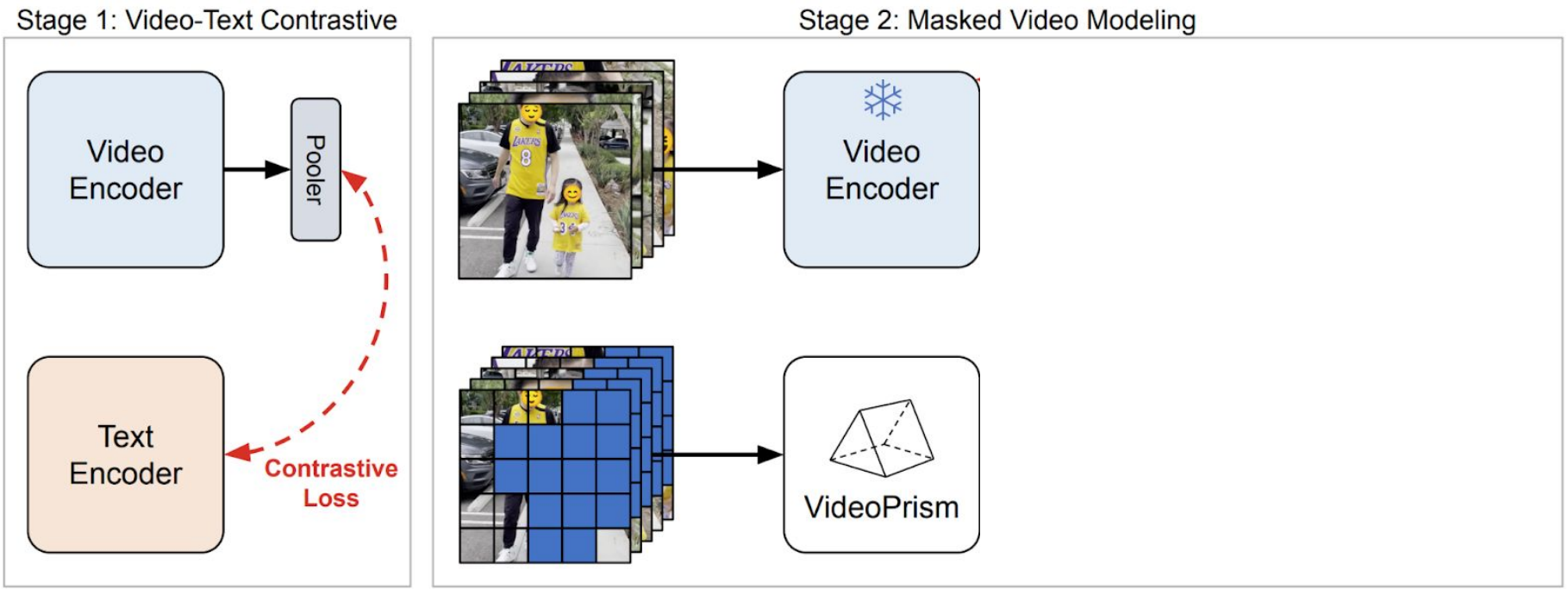
Stage 1: Video-Text Contrastive



Stage 2: Masked Video Modeling

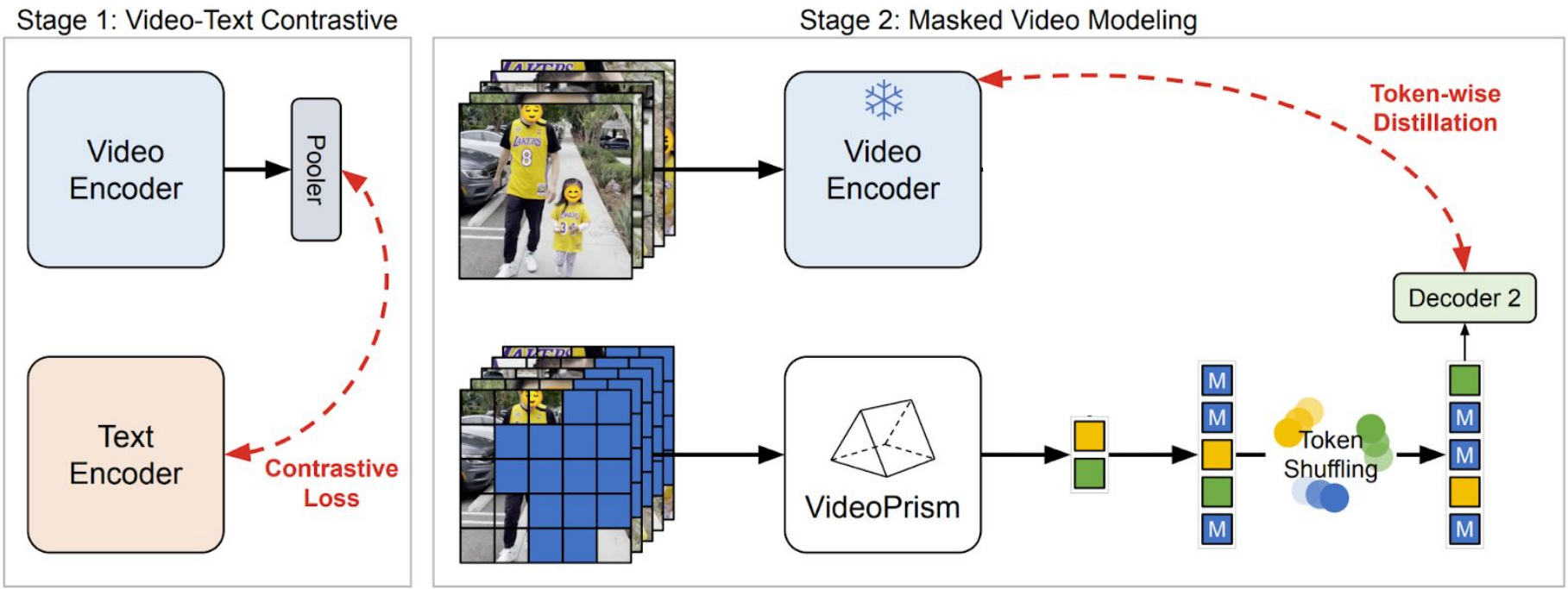
How is VideoPrism trained?

Two stage training:



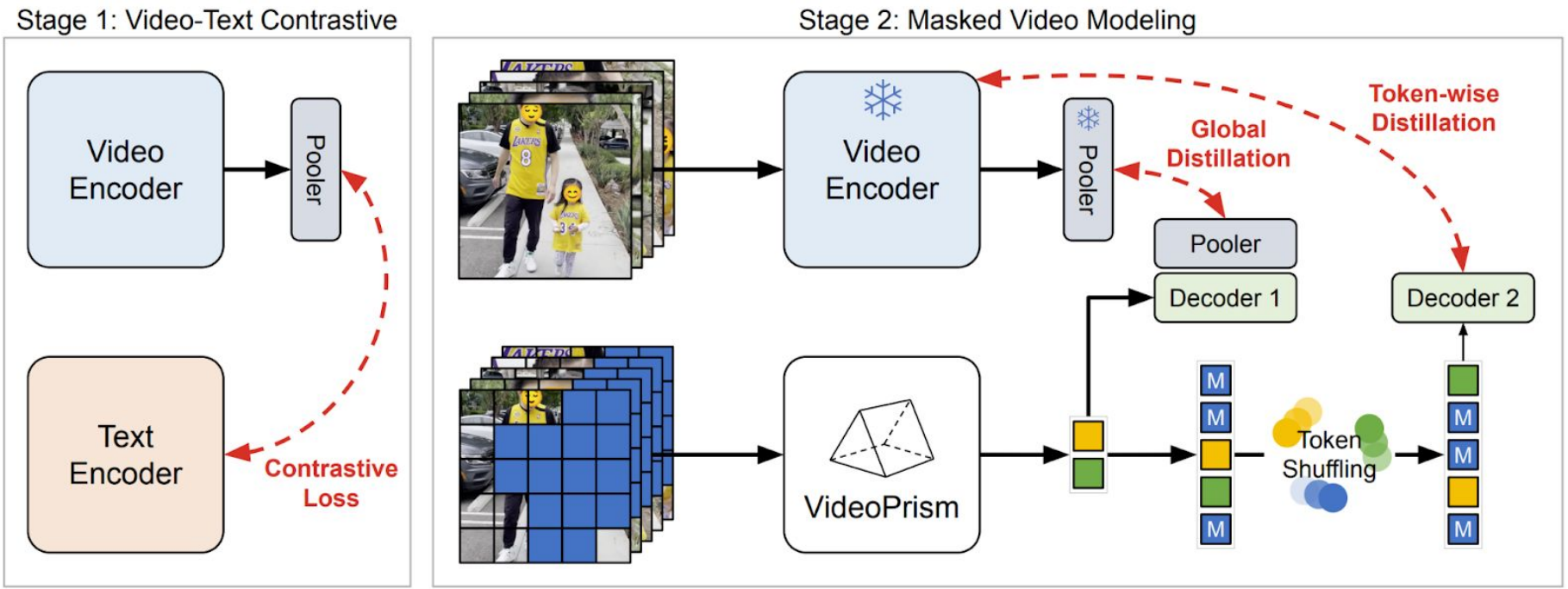
How is VideoPrism trained?

Two stage training:

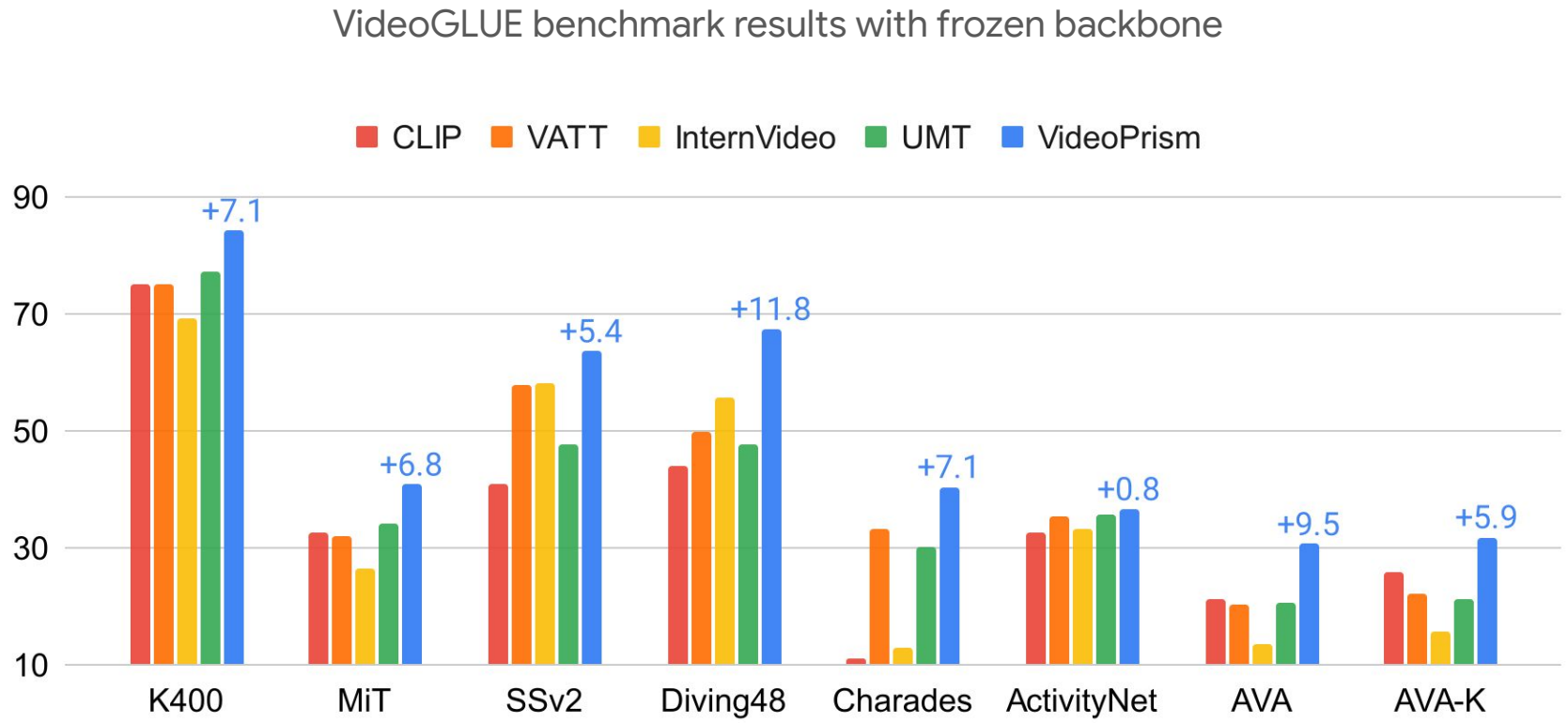


How is VideoPrism trained?

Two stage training:

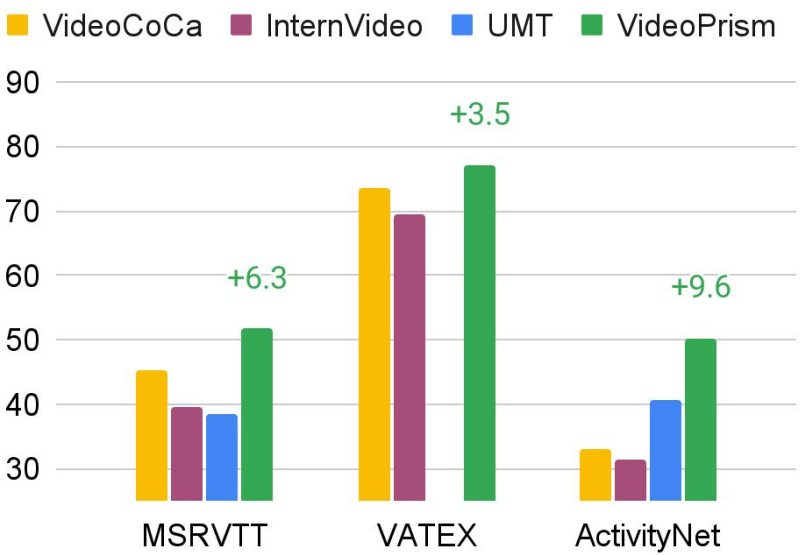


How does VideoPrism perform?

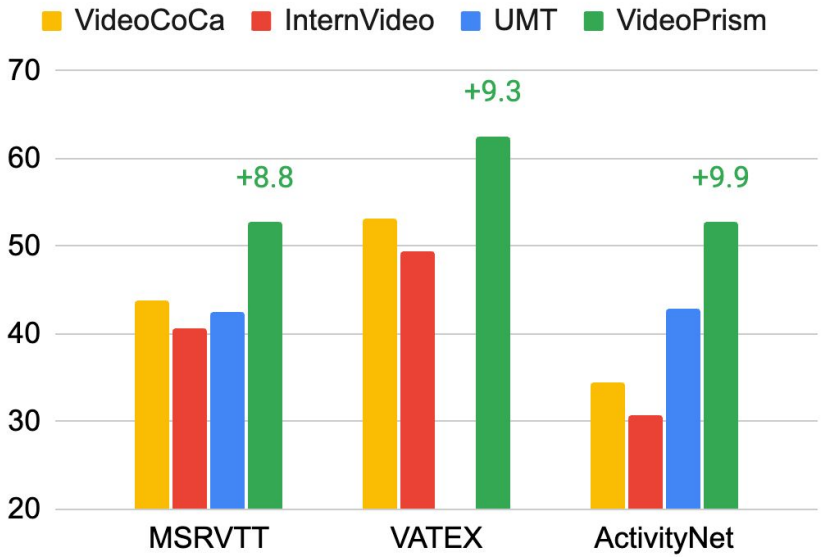


How does VideoPrism perform?

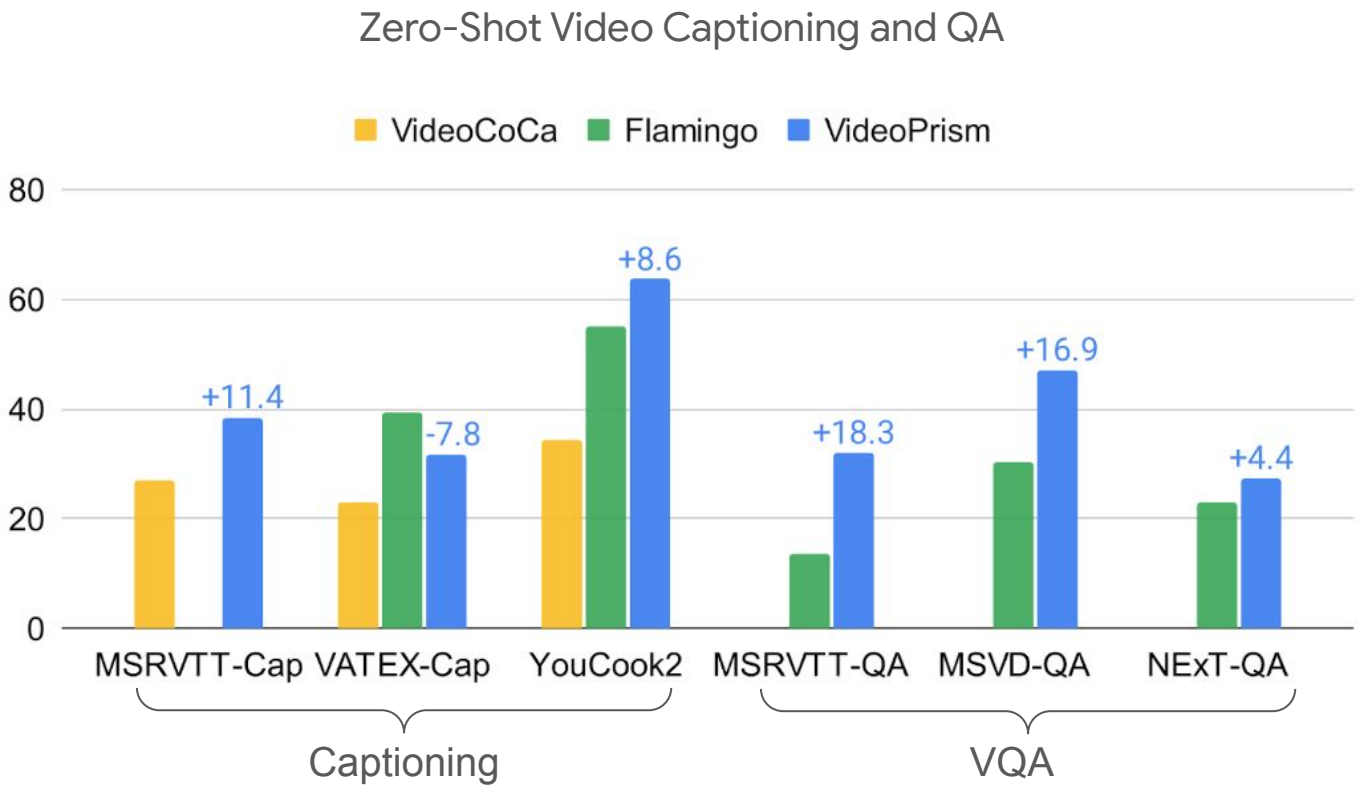
Zero-Shot Video-Text Retrieval



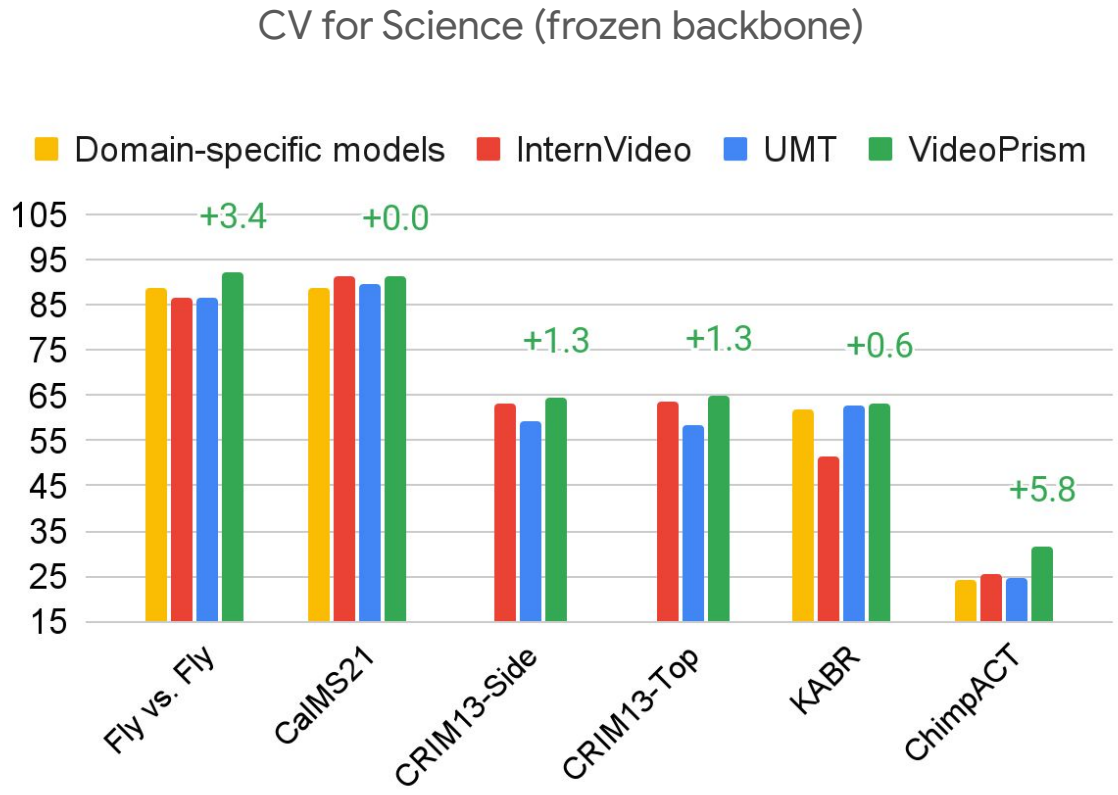
Zero-Shot Text-Video Retrieval



How does VideoPrism perform?



How does VideoPrism perform?



See more at our poster session:

Location: Hall C 4-9 #205

Time: Thu 25 Jul 11:30 a.m. CEST — 1 p.m. CEST