Learning Streaming Video Representation via Multitask Training

Yibin Yan, Jilan Xu, Shangzhe Di, Yikun Liu, Yudi Shi, Qirui Chen, Zeqian Li, Yifei Huang, Weidi Xie

School of Artificial Intelligence, Shanghai Jiao Tong University

Oct 2025

Paradigm Shifting of Video Backbone

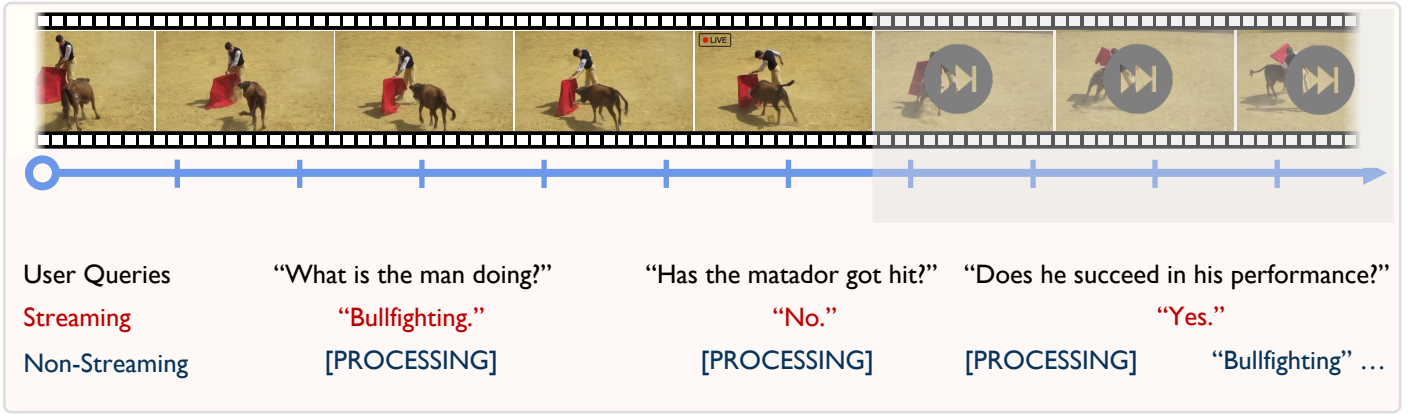
- Why Streaming Video?

Embodied AI

Autonomous Driving

Online Action Detection[1]

- Key Attributes
  - frame-by-frame processing
  - low-latency decision making
  - long-term context preservation

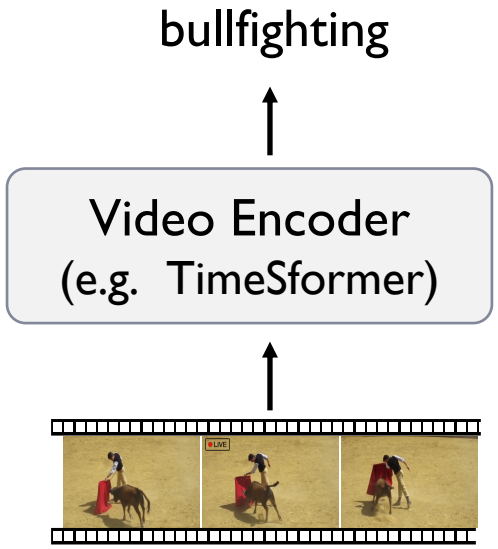


[1] Roeland De Geest, et al. "Online Action Detection." *ECCV 2016*

Paradigm Shifting of Video Backbone

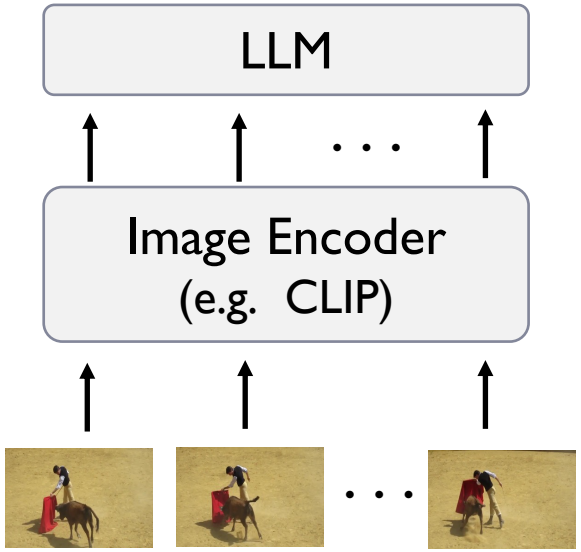
Offline Video Backbone

- Responding after seeing the entire video
- Video-text contrastive pairs
- OOM when processing hour-long videos
- Fine-grained video representation



VideoLLM

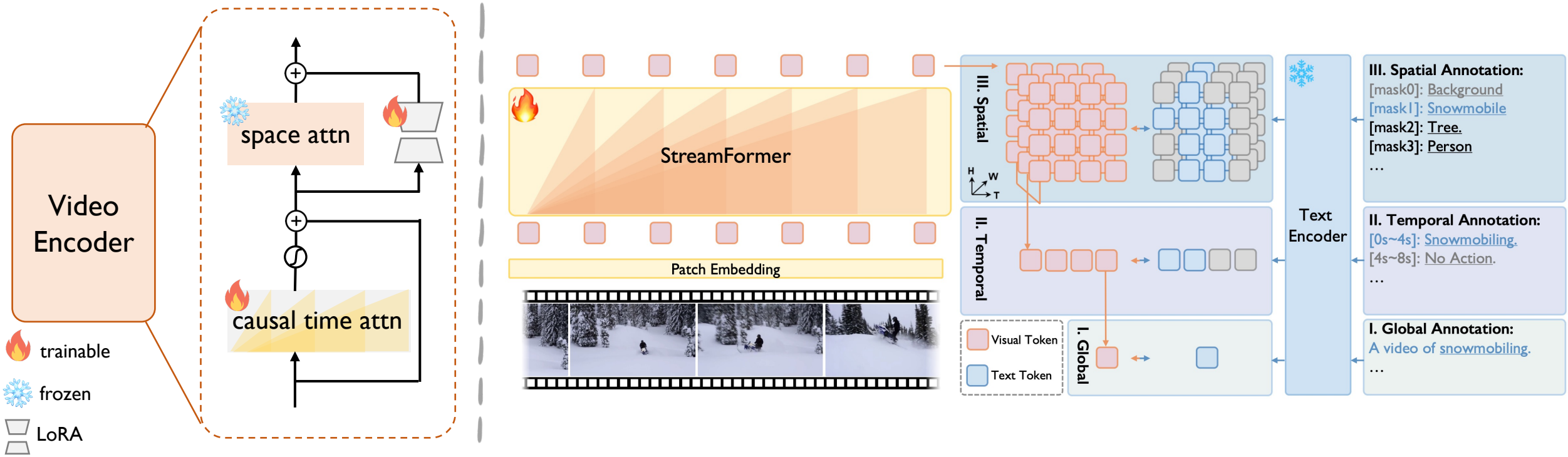
- Requires large amount of data
- Unable to offer fine-grained video representation
- Only supports text output
- Accepts streaming input



Main Contribution

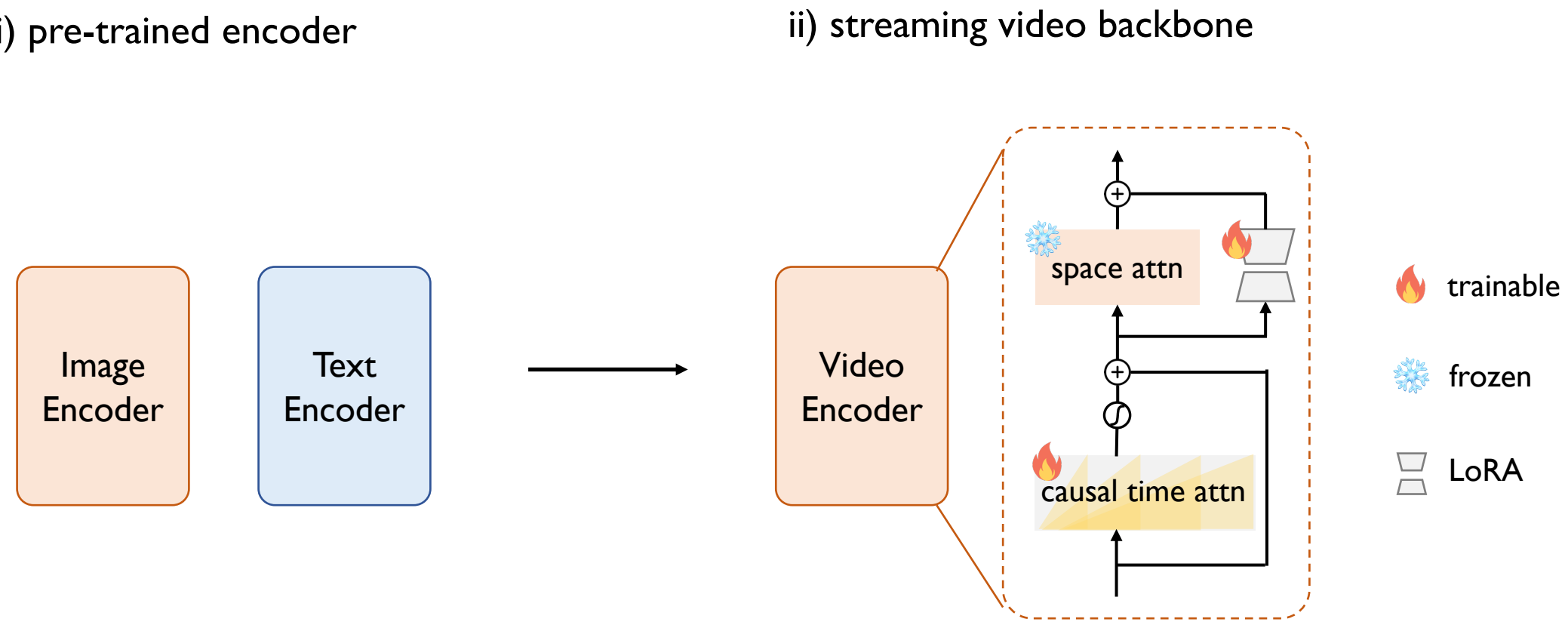
StreamFormer

- **Architecture**: Divided space-time attention by combing (1) causal temporal attention and (2) SigLIP w/ LoRA.
- **Method**: Unifying multiple spatiotemporal video tasks into a visual-language alignment framework.
- **Data**: Instead of web-scale video-text pairs, human-annotated video datasets of various granularities.



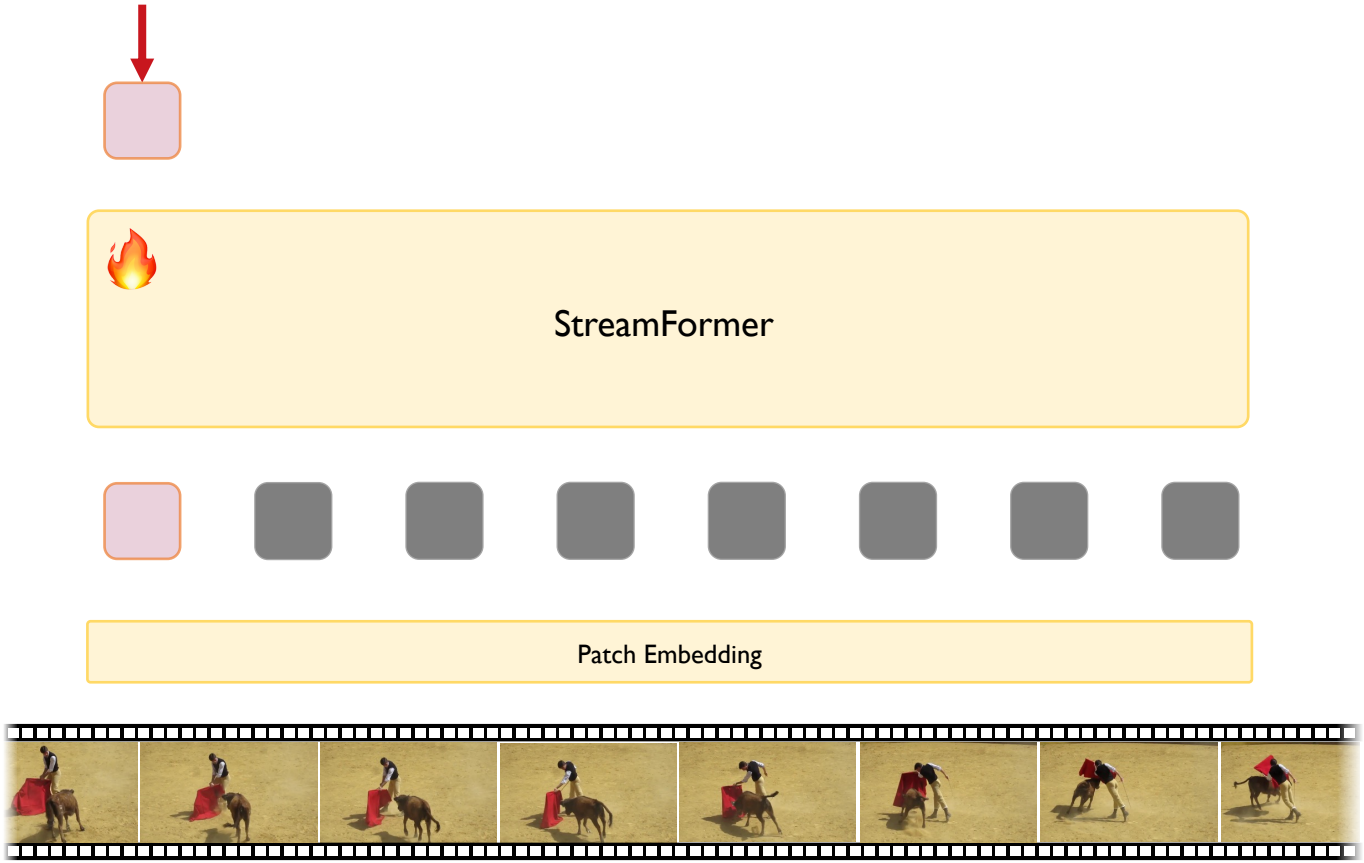
Architecture

- Lifting a pre-trained image encoder (e.g. SigLIP) to a streaming video backbone.



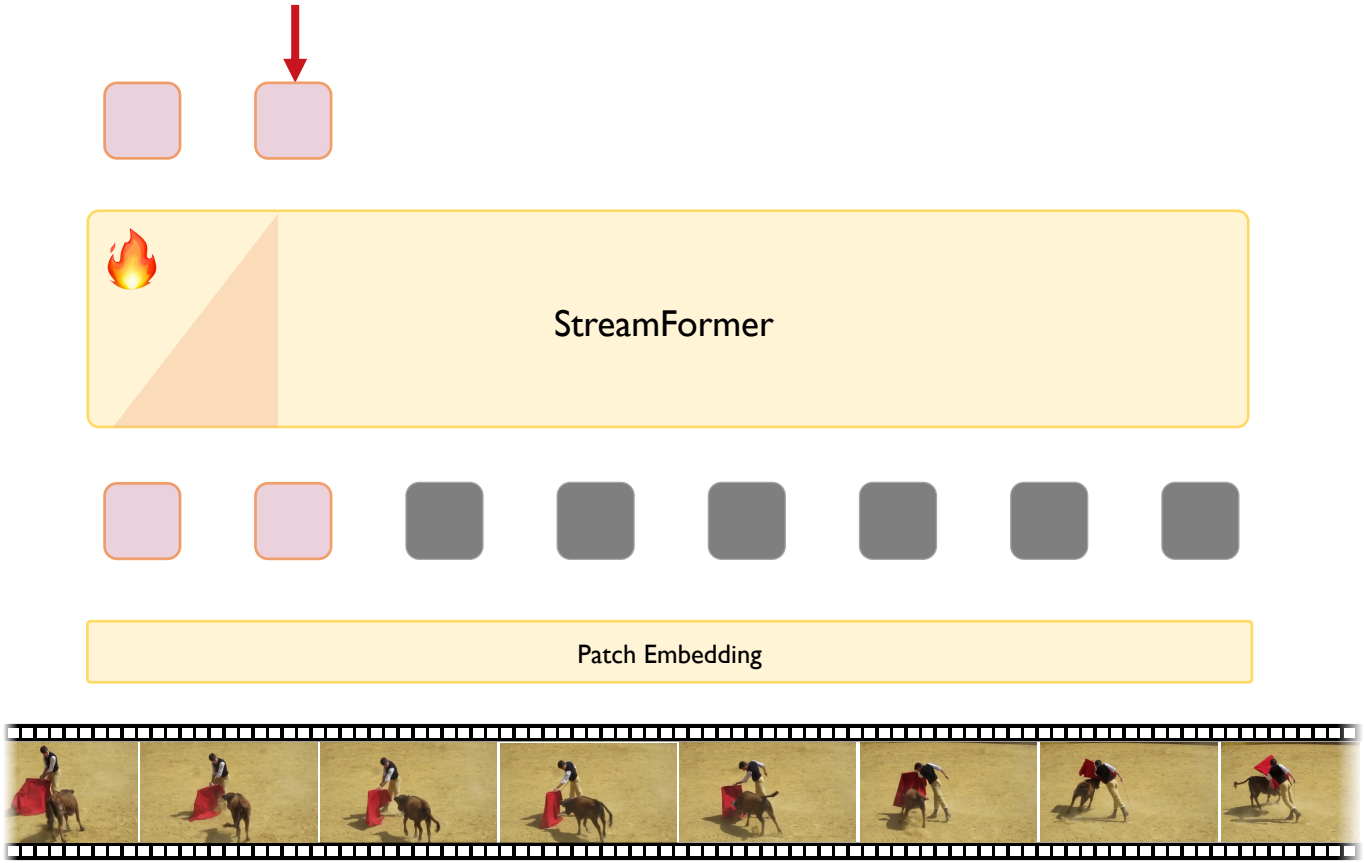
Architecture

- Lifting a pre-trained image encoder (e.g. SigLIP) to a streaming video backbone.



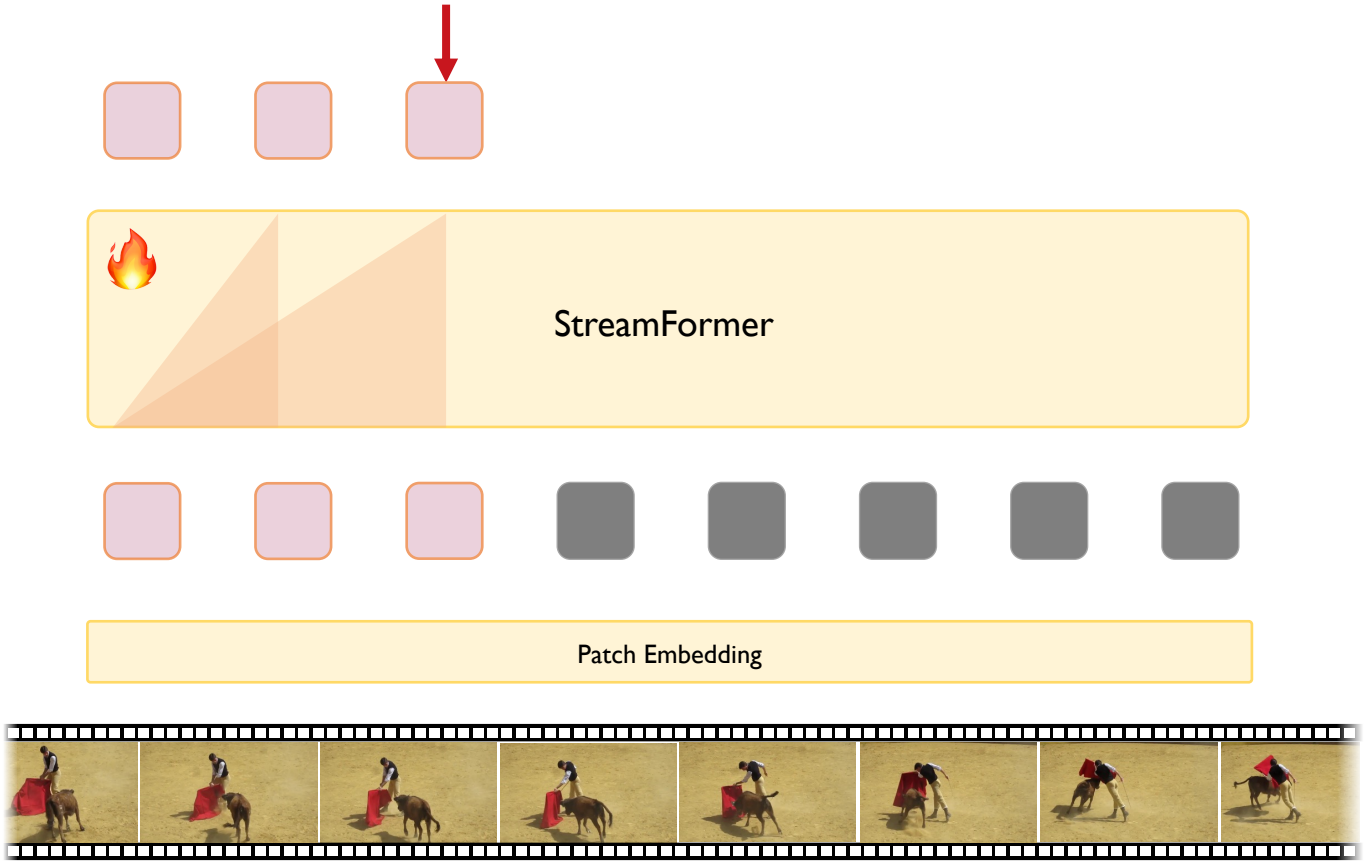
Architecture

- Lifting a pre-trained image encoder (e.g. SigLIP) to a streaming video backbone.



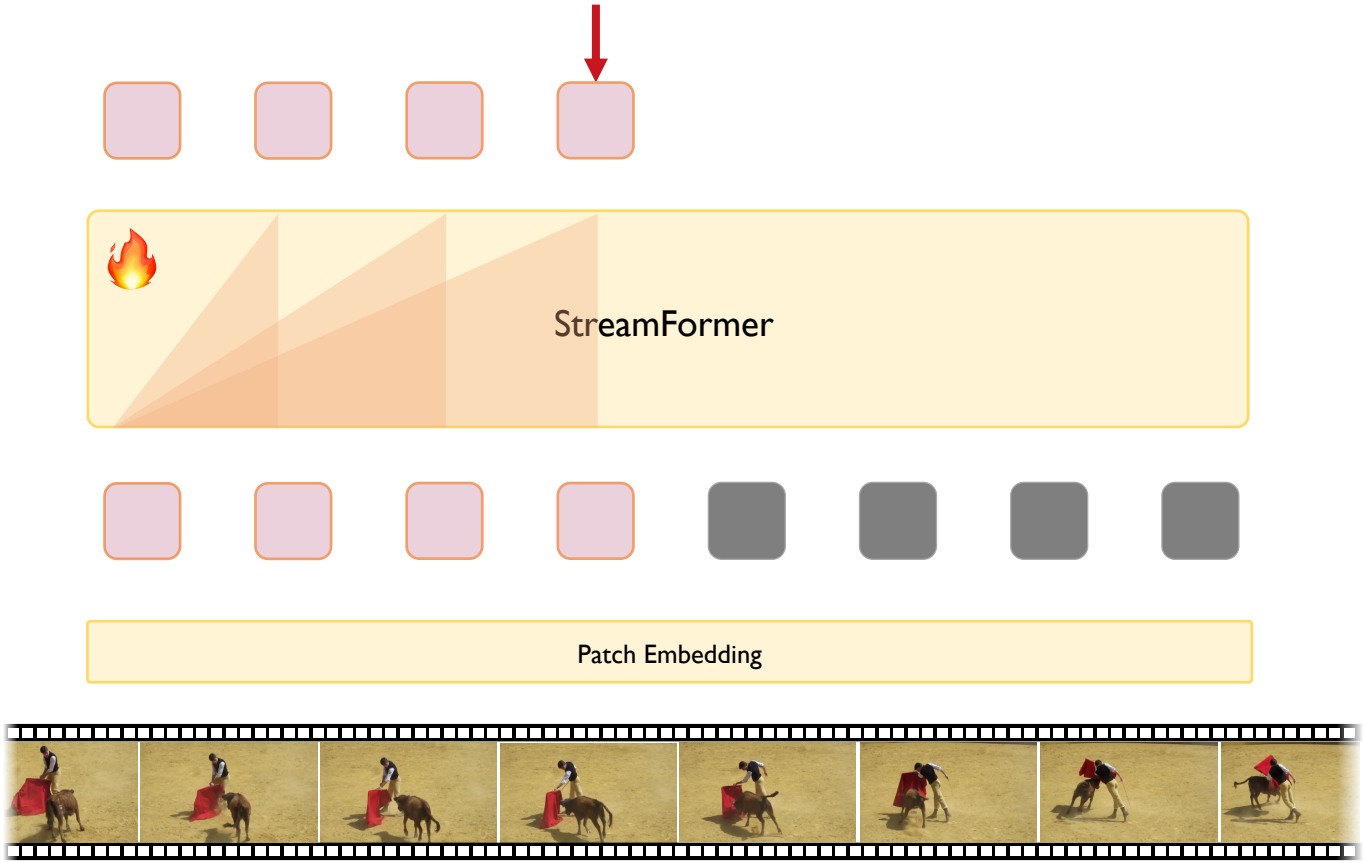
Architecture

- Lifting a pre-trained image encoder (e.g. SigLIP) to a streaming video backbone.



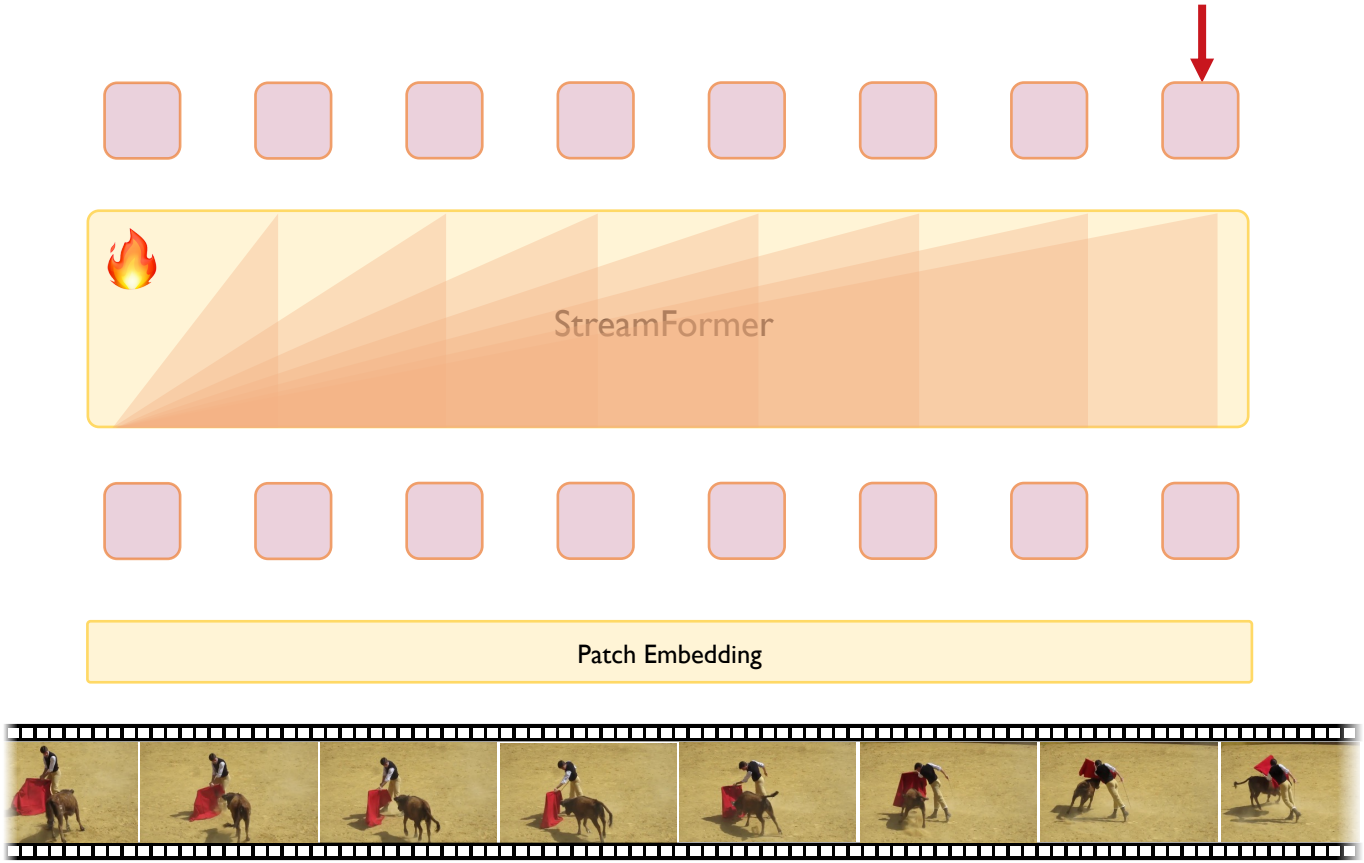
Architecture

- Lifting a pre-trained image encoder (e.g. SigLIP) to a streaming video backbone.



Architecture

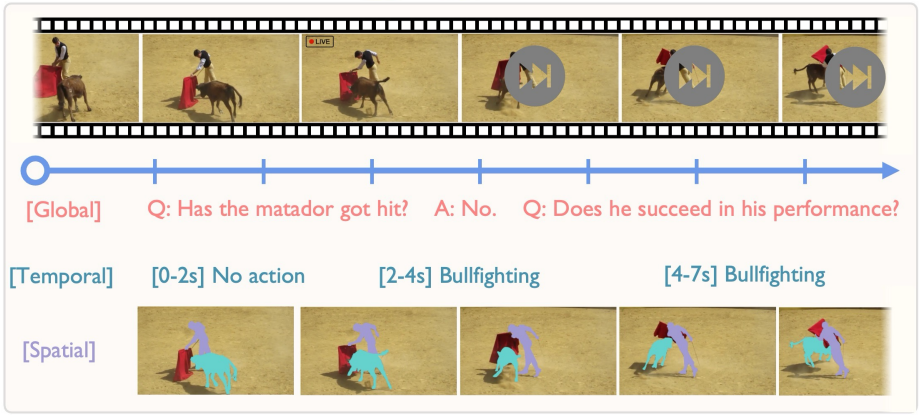
- Lifting a pre-trained image encoder (e.g. SigLIP) to a streaming video backbone.



Video-Language Pre-training

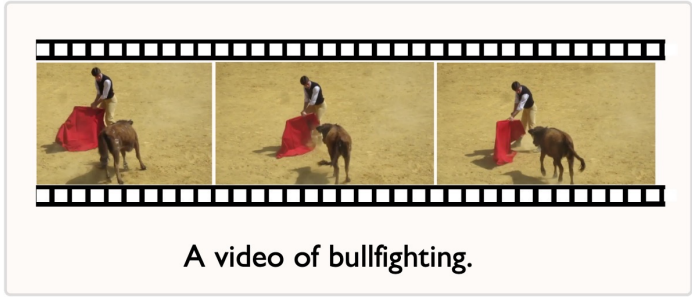
- Video-language alignment is more than a simple video-text pair:

Learning from
comprehensive annotations



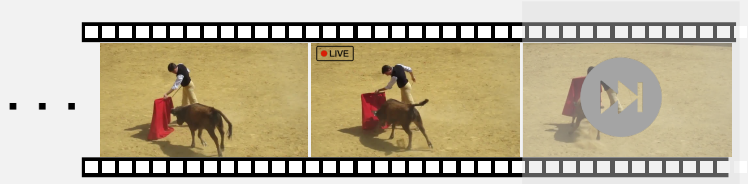
VS

Learning from
naïve video-text pairs



Video-Language Pre-training

- Multitask formulation: unify various video understanding tasks in a visual-text alignment framework.
  - Global-level tasks
    - ➢ one label/narration per video clip (e.g. action recognition, video-text retrieval)
  - Temporal-level tasks
    - ➢ one label/narration per frame (e.g. temporal action localisation, temporal video grounding)
  - Spatial-level tasks
    - ➢ one label/narration per pixel (e.g. video object segmentation, referring video object segmentation)



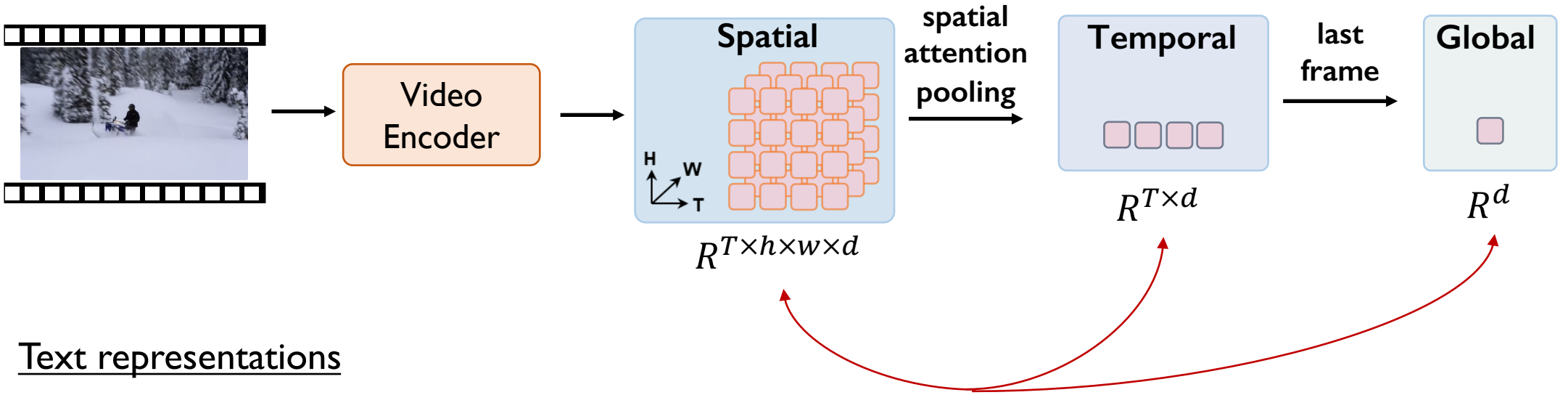
$\mathcal{A} = g(V, T)$

The bullfighter waves the muleta at the bull.

We train StreamFormer by optimising the visual-text alignment score $\mathcal{A}$

Video-Language Pre-training

- Video representations

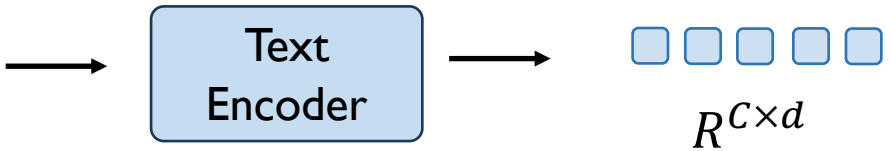


- Text representations

A video clip of *snowmobiling.*

A video clip of *playing football.*

⋮



visual-text alignment

Video-Language Pre-training

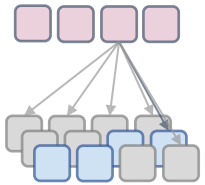
- Multitask formulation: unify various video understanding tasks in a visual-text alignment framework.
  - Global-level tasks (e.g. recognition)

$v_i \in R^{1\times d}$

$t_i \in R^{C\times d}$

$\mathcal{A}_{\text{global}} = v_i t_i^{\mathrm{T}} \in R^{1\times C}$ $\overset{L_{CE}}{\longleftrightarrow}$ action label

  - Temporal-level tasks (e.g. localisation)

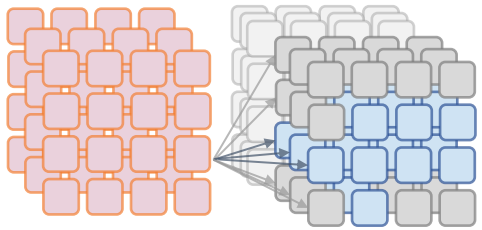


$f_i \in R^{T\times d}$

$t_i \in R^{T\times C\times d}$

$\mathcal{A}_{\text{temporal}} = f_i t_i^{\mathrm{T}} \in R^{T\times C}$ $\overset{L_{CE}}{\longleftrightarrow}$ frame label

  - Spatial-level tasks (e.g. segmentation)



$F_i \in R^{T\times h\times w\times d}$

$t_i \in R^{T\times h\times w\times C\times d}$

$\mathcal{A}_{\text{spatial}} = F_i t_i^{\mathrm{T}} \in R^{T\times w\times h\times C}$ $\overset{L_{CE}}{\longleftrightarrow}$ pixel label

Video-Language Pre-training

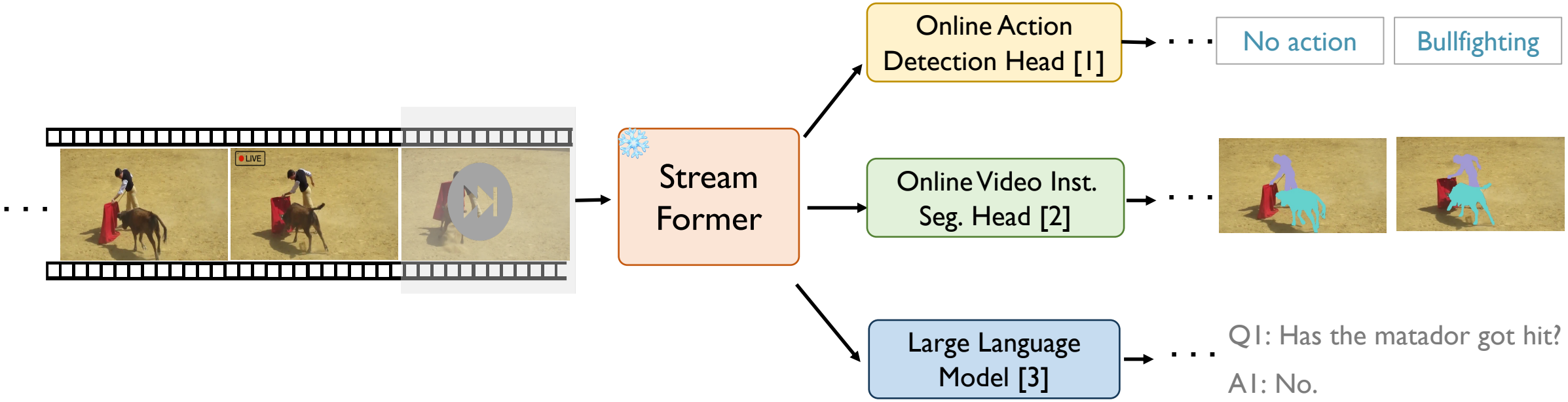
- Pre-training datasets

Task	Pre-training Dataset	Scale
Global-level		
Action Recognition	K400, SSv2	400K
Video Text Retrieval	MSRVTT, MSVD, ActivityNet, DiDeMo, LSMDC, VATEX	94K
Temporal-level		
Temporal Action Localisation	ActivityNet-1.3, FineAction, HACS	180K
Temporal Video Grounding	CharadesSTA, QVHighlights, TaCoS, ANet-Captions, DiDeMo, QuerYD	120K
Spatial-level		
Video Instance Segmentation	YouTubeVIS-19, LVVIS, COCO	120K
Referring Video Object Segmentation	MEVIS, Refer-YouTube-VOS	36K
Total	-	∼1M

† We sample data from the same task at each mini-batch, and use gradient accumulation to perform backpropagation and parameter update collectively after iterating through all tasks.

Downstream Application

- Freeze the pre-trained video backbone and append task-specific head



[1] Jiahao Wang, et al. "Memory-and-anticipation transformer for online action understanding." *ICCV 2023*.
[2] Kaining Ying, et al. "Ctvis: Consistent training for online video instance segmentation." *ICCV 2023*.
[3] Haotian Liu, et al. "Visual instruction tuning." *NeurIPS 2023*.

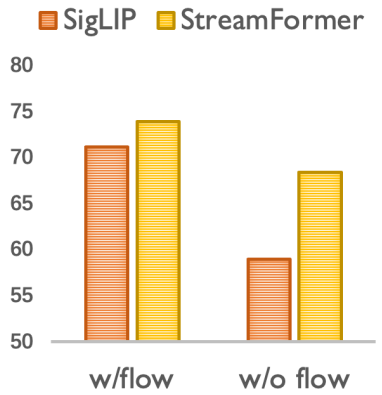
Downstream Application

Experimental Setting

- Keep all settings the same except for visual encoder (StreamFormer vs. SigLIP)
- Add specific task heads for each downstream
  - OAD: training MAT[1] with extracted video features.
  - OVIS: ViT-Adapter with CTVIS[2] training.
  - VideoQA: LLaVA-Next[3] pipeline added with video samples.

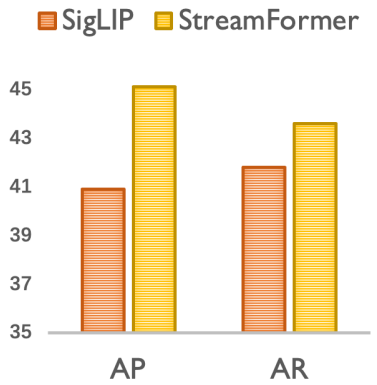
i) Online Action Detection

THUMOS (mAP)



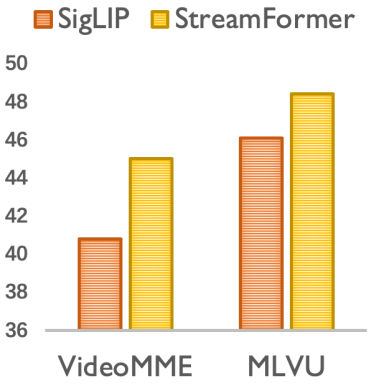
ii) Online Video Instance Segmentation

YouTube-VIS-19 (AP, AR)



iii) Video Question Answering

VideoMME, MLVU (Acc)

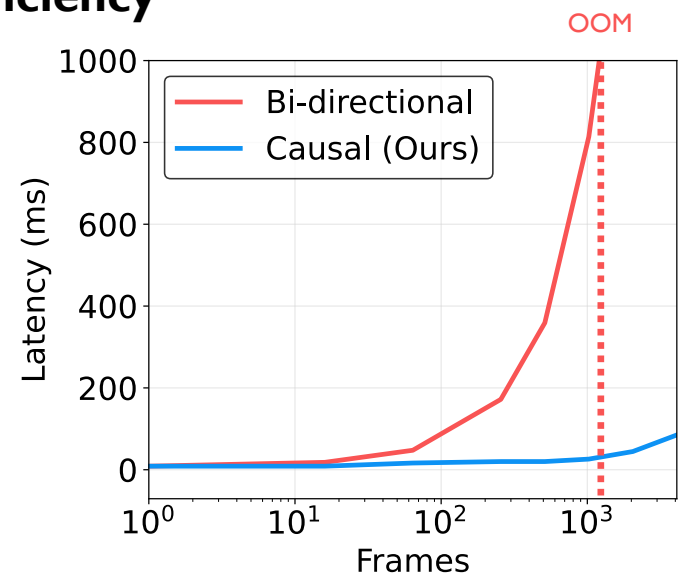


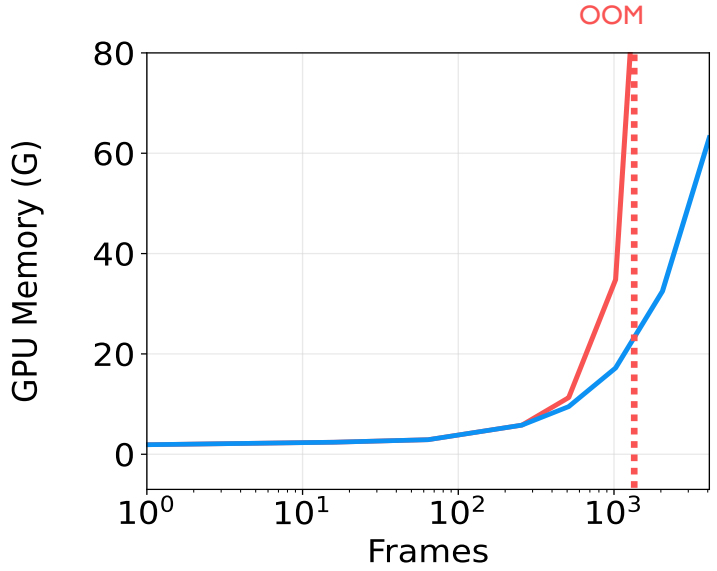
[1] Jiahao Wang, et al. "Memory-and-anticipation transformer for online action understanding." *ICCV 2023*.
[2] Kaining Ying, et al. “Ctvis: Consistent training for online video instance segmentation.” *ICCV 2023*.
[3] Haotian Liu, et al. “Visual instruction tuning.” *NeurIPS 2023*.

Downstream Application

- **Inference Efficiency**

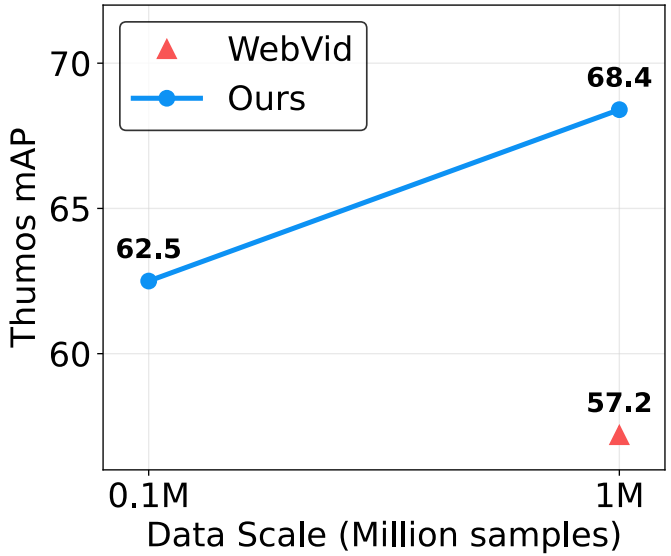
Bi-directional vs Causal time attention w/ kv-cache (ours)

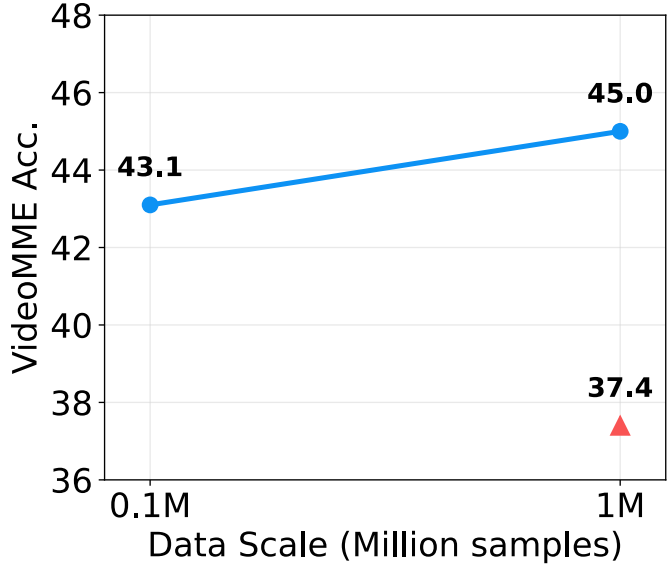




- **Data Efficiency**

Contrastive Learning vs Multitask Learning (ours)





Conclusion

- Streaming video representation learning made possible by multi-granularity task training.
- Video representation learning needs to be revolutionized!

Feel free to chat!

- Homepage: go2heart.github.io
- Email: cain.y.yan@gmail.com

Web

Github

Twitter/X

Wechat