Revisiting the "Video" in Video-Language Understanding

Shyamal Buch[1], Cristóbal Eyzaguirre[1], Adrien Gaidon[2], Jiajun Wu[1], Li Fei-Fei[1], Juan Carlos Niebles[1]

1 Stanford Vision & Learning
2 Toyota Research Institute

CVPR June 19-24 2022, New Orleans • Louisiana

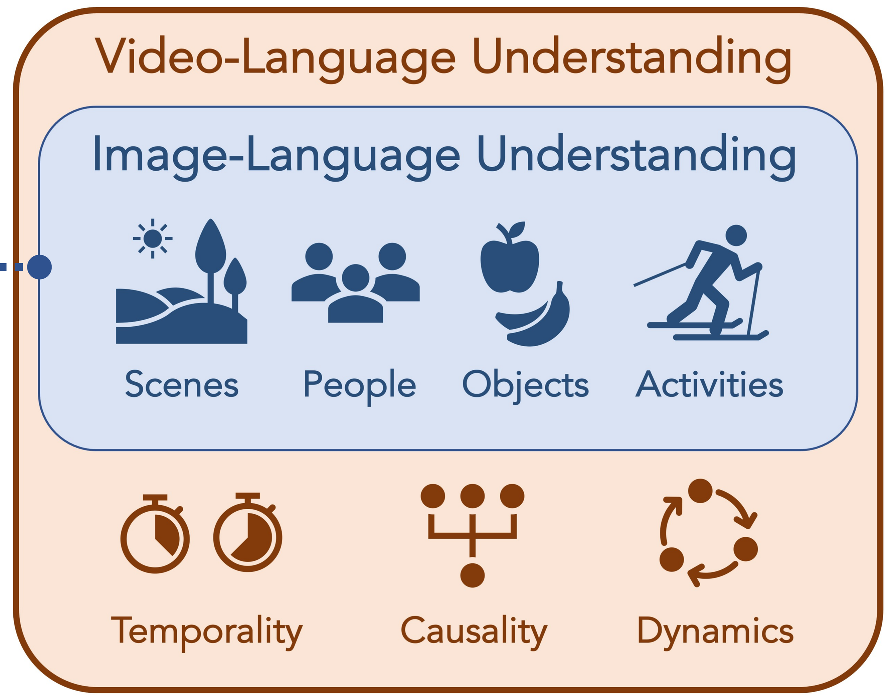
Project website: atp-video-language.stanford.edu

Motivation

Video understanding offers the potential to go *beyond* image-level semantics (scenes, objects) towards event temporality + causality.

Our work re-examines a foundational question [5,6] in video research:

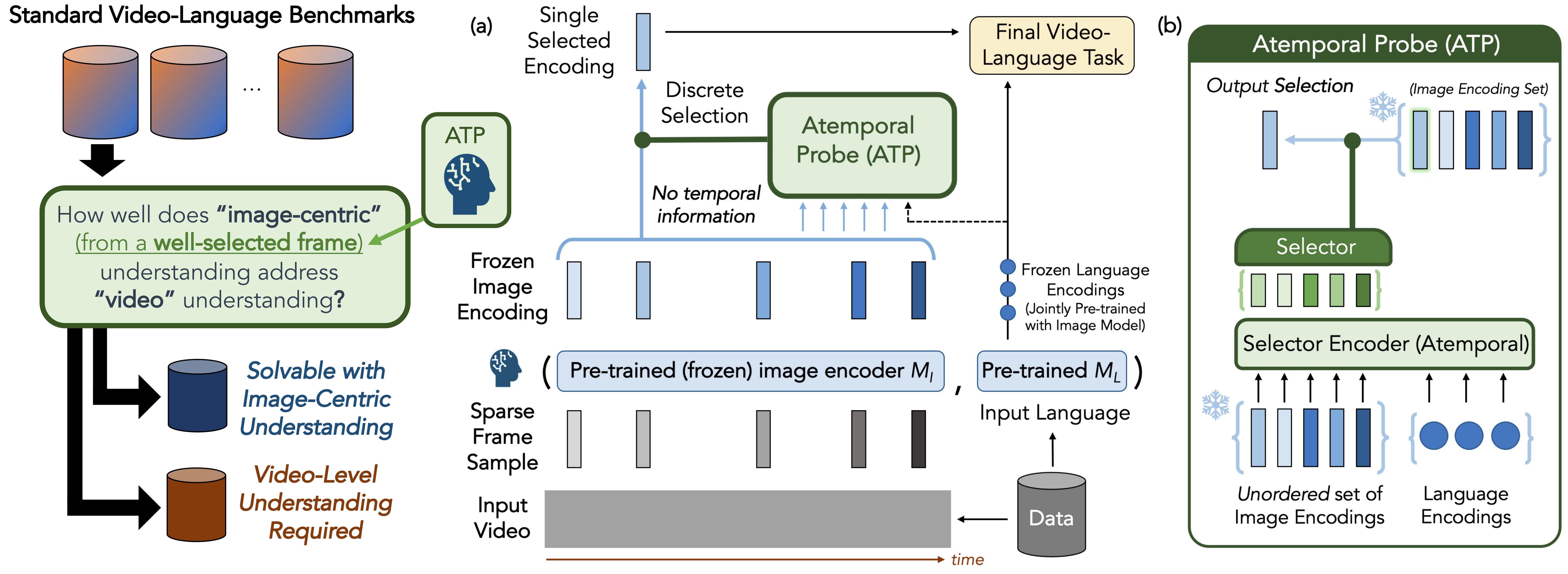
★ What makes a video task uniquely suited for **videos**, beyond what can be understood from a **single image**?



Our focus is on **video + language**, where *language* has the potential to describe richer event properties and relationships in *videos*.

Challenge: Standard approaches may **under-represent "image-centric" bound** for video-language understanding!

Atemporal Probe (ATP) for Video-Language Analysis

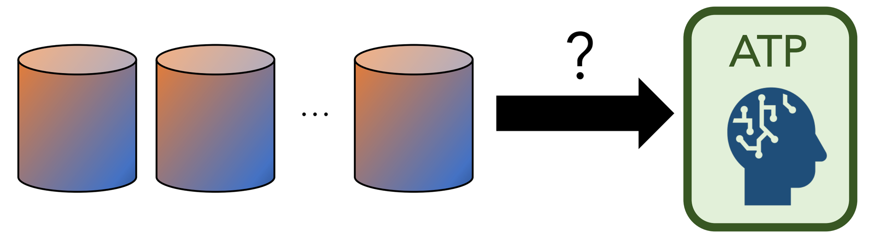


Our **Atemporal Probe (ATP)** model builds on progress in self-supervised image-language understanding [1], and learns to **discretely select a** ***frozen*** **image-level encoding** (*without* using any temporal information). This encoding of a single frame is sent downstream – unmodified – to the **final video-language understanding task** (video question answering, text-to-video retrieval; e.g. [2,7]).

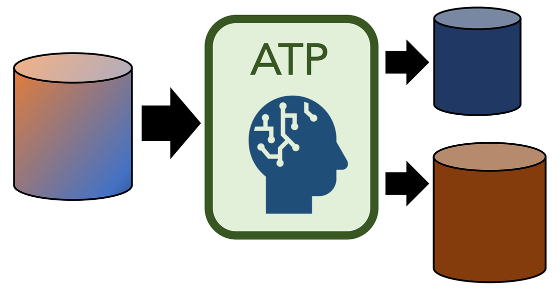
★ Overview

We propose the **Atemporal Probe (ATP)**:

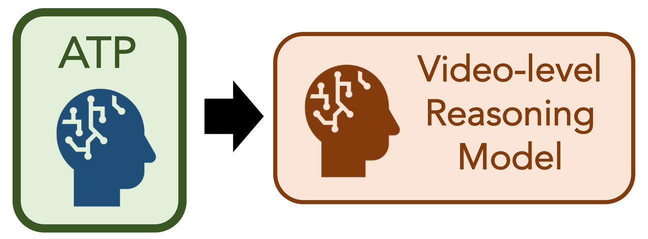
✓ To analyze **current standard video-language** benchmarks (stronger bound on image-centric understanding)



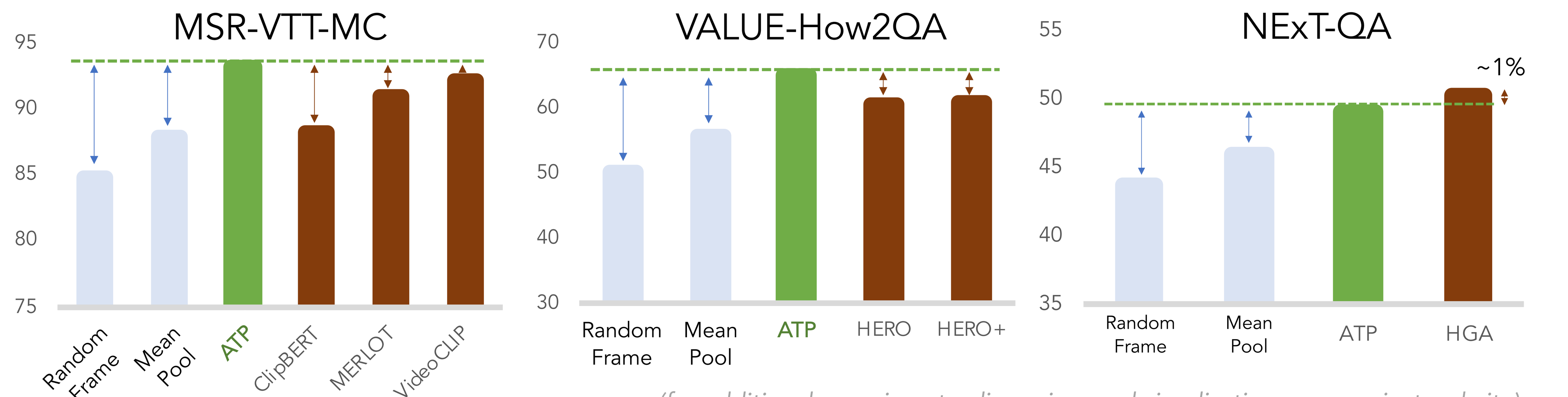
✓ **Improve dataset design** (disentangling unintended biases)



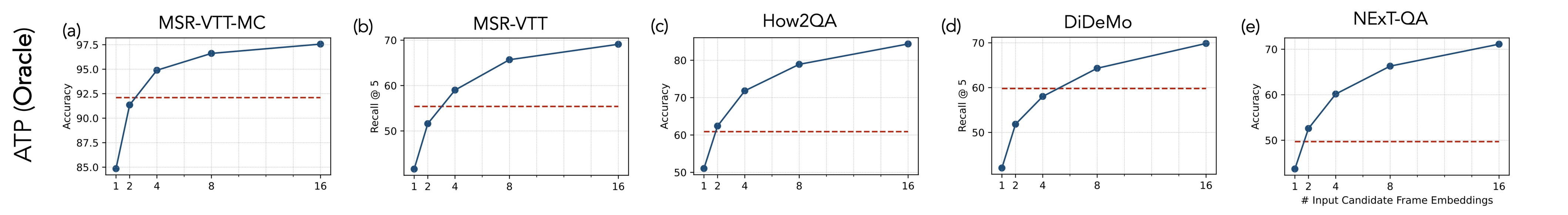
✓ **Improve model design** (better efficiency and accuracy)



Experiments and Analysis



(for additional experiments, discussion, and visualizations, see project website)

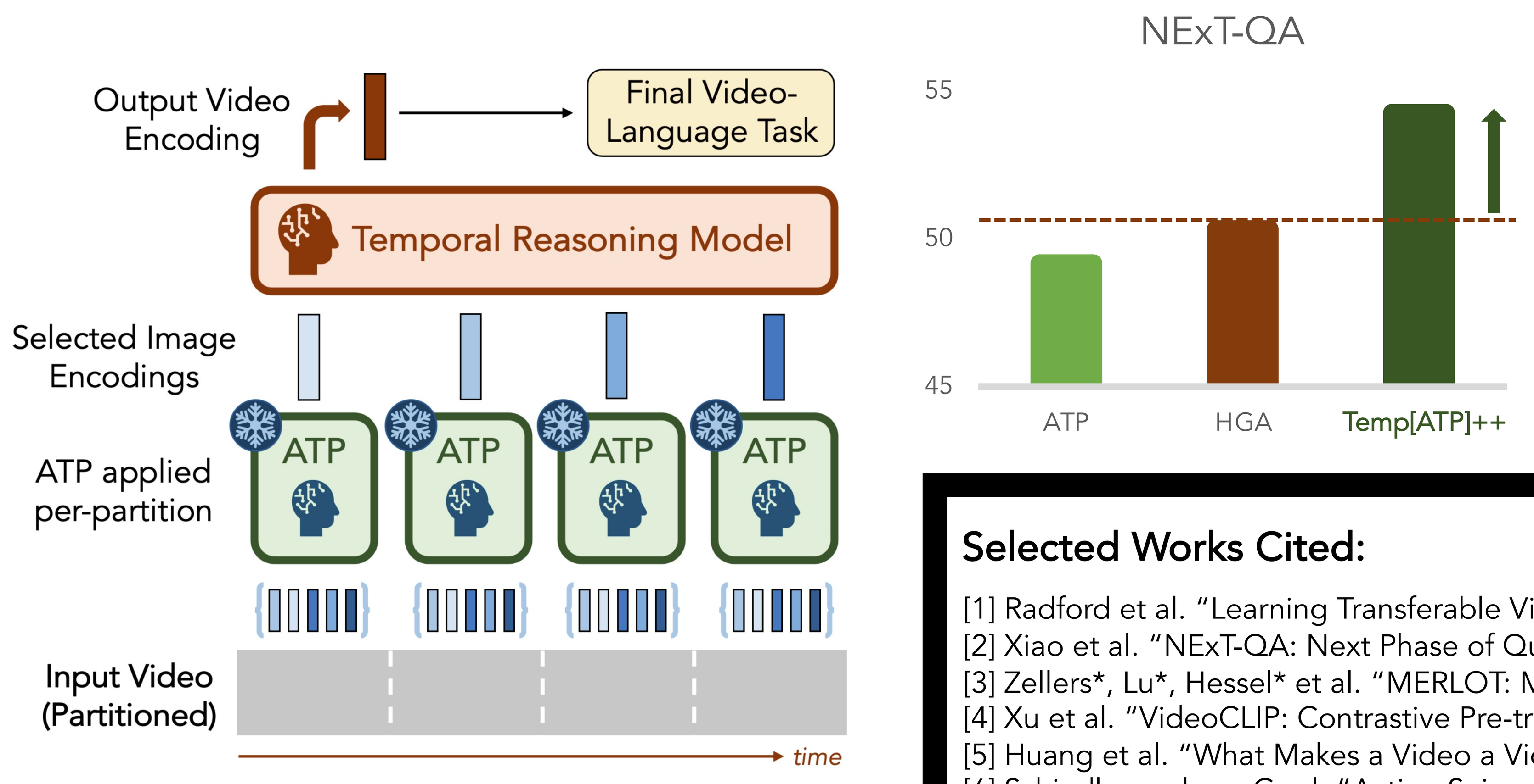


Takeaways: (hold even when dataset *explicitly designed* for temporal + causal video-language)

1. **Datasets** can be (surprisingly) well-addressed by image-centric understanding
2. Video-level **models** may be significantly impacted by processing noisy frames

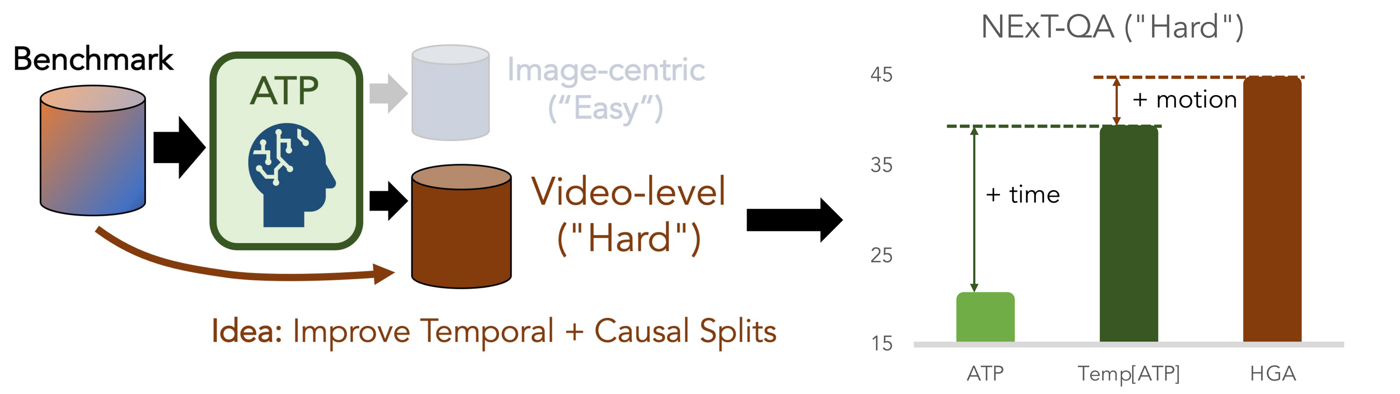
Improving Model Design with ATP

ATP can **improve temporal (multi-frame) model design** by forwarding semantically useful candidates to reason over (reducing noise). A combined model with ATP achieves higher accuracy with fewer frames needed.



Improving Dataset Design with ATP

ATP can help identify (multi-frame) temporally challenging data, and is a **promising tool** for in-the-loop **dataset design.**



Selected Works Cited: (full list of references + acknowledgements in paper)

[1] Radford et al. "Learning Transferable Visual Models from Natural Language Supervision." ICML 2021.
[2] Xiao et al. "NExT-QA: Next Phase of Question-Answering to Explaining Temporal Actions." CVPR 2021.
[3] Zellers*, Lu*, Hessel* et al. "MERLOT: Multimodal Neural Script Knowledge Models." NeurIPS 2021.
[4] Xu et al. "VideoCLIP: Contrastive Pre-training for Zero-Shot Video-Text Understanding." EMNLP 2021.
[5] Huang et al. "What Makes a Video a Video: Analyzing Temporal Information in Videos." CVPR 2018.
[6] Schindler and van Gool. "Action Snippets: How many frames does human action recognition require?" CVPR 2008
[7] Li*, Lei* et al. "VALUE: A Multi-task Benchmark for Video-and-Language Understanding Evaluation." NeurIPS 2021 (D&B)

Acknowledgements:
HAI Stanford University Human-Centered Artificial Intelligence
salesforce
SAMSUNG