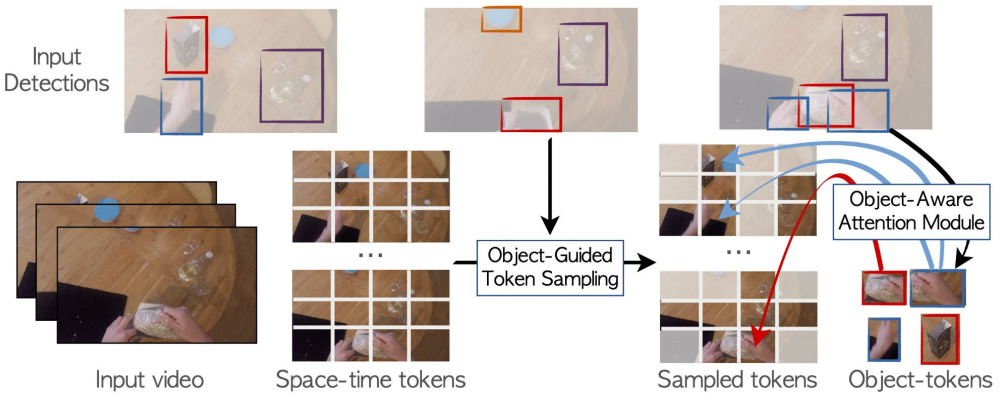
# How can objects help action recognition?

Xingyi Zhou, Anurag Arnab, Chen Sun, Cordelia Schmid
Google Research
Poster @ **TUE-AM-225**

# Overview

- Goal: using object detection results from an detector to **help** action recognition.
- Help **efficiency**: we drop non-object pixels in the input.
- Help **accuracy**: we design novel object-aware attention module.
- Overall: allow us to process fewer tokens with better accuracy.




Google Research

# Motivation

- Most of the video regions are redundant.
- The action is defined by a few key objects.
- Tracked objects connects video pixels across space and time.

Google Research

# Framework

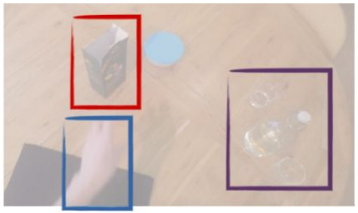
Input video

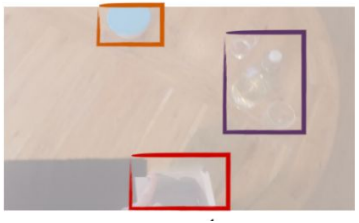
Put something
inside something

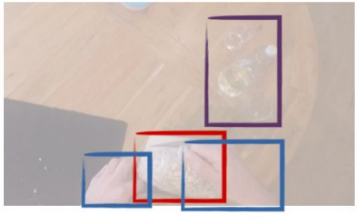
Google Research

# Framework

Input Detections





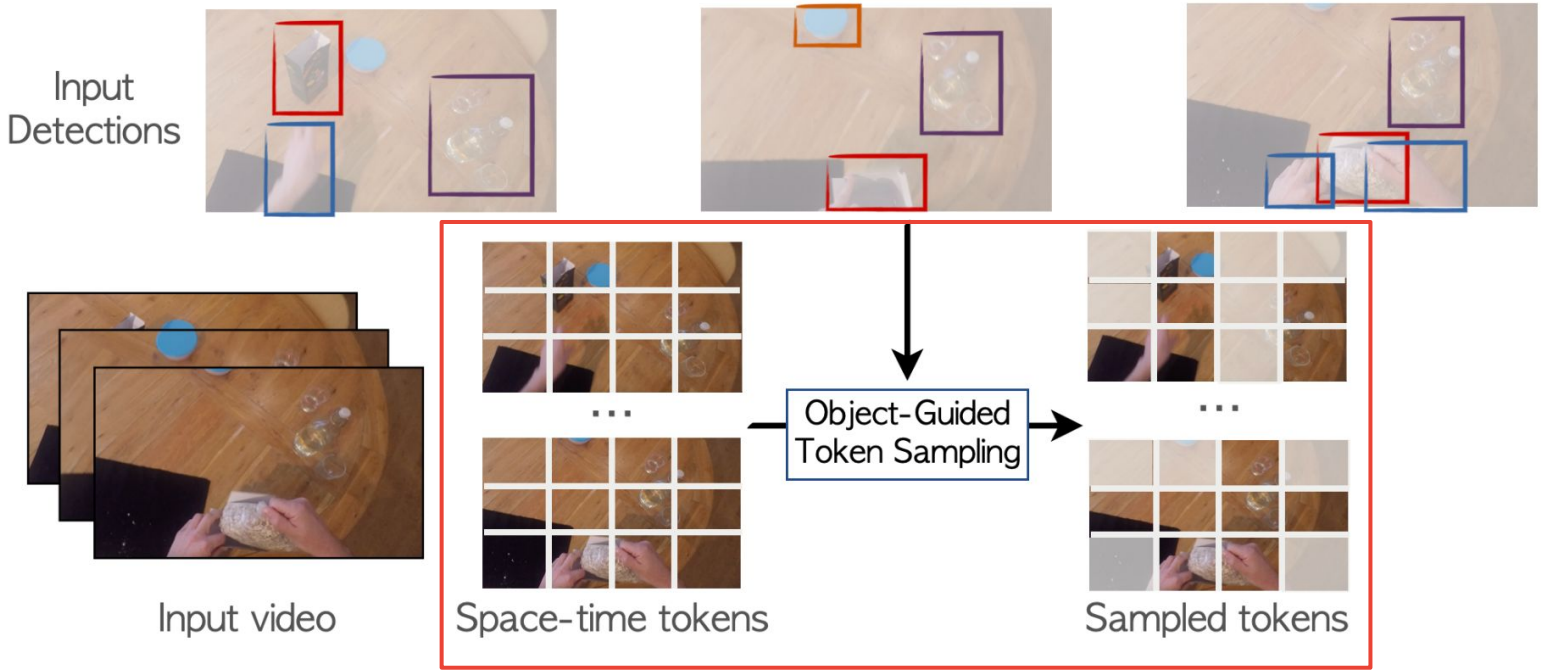


Input video

→ Put something inside something

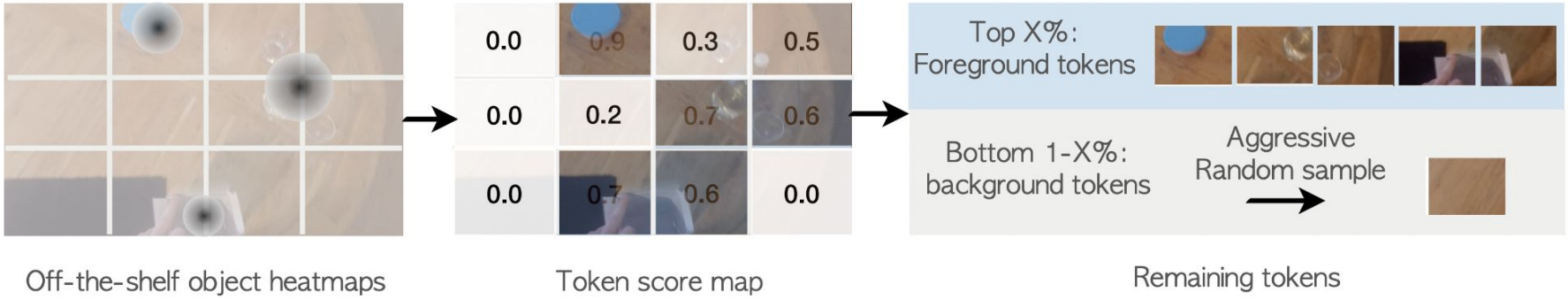
Google Research

# Framework



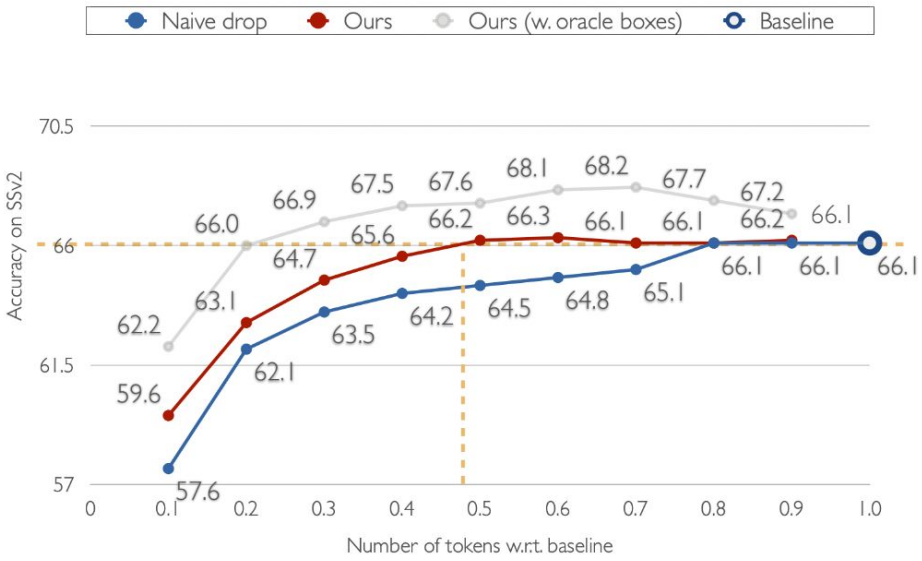

Google Research

# Object-guided token sampling

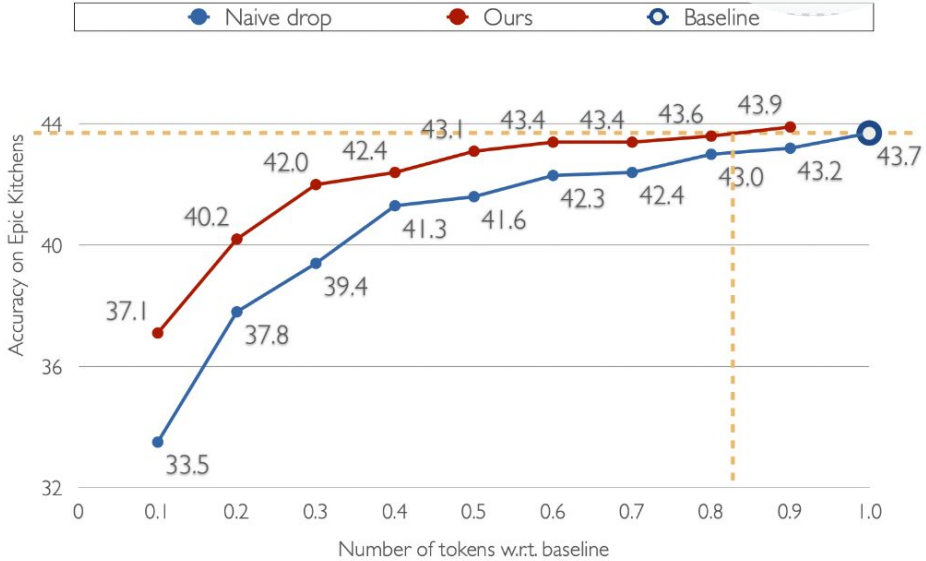



- Configurable number of remaining tokens.
- Always keep background information.

# Object-guided token sampling



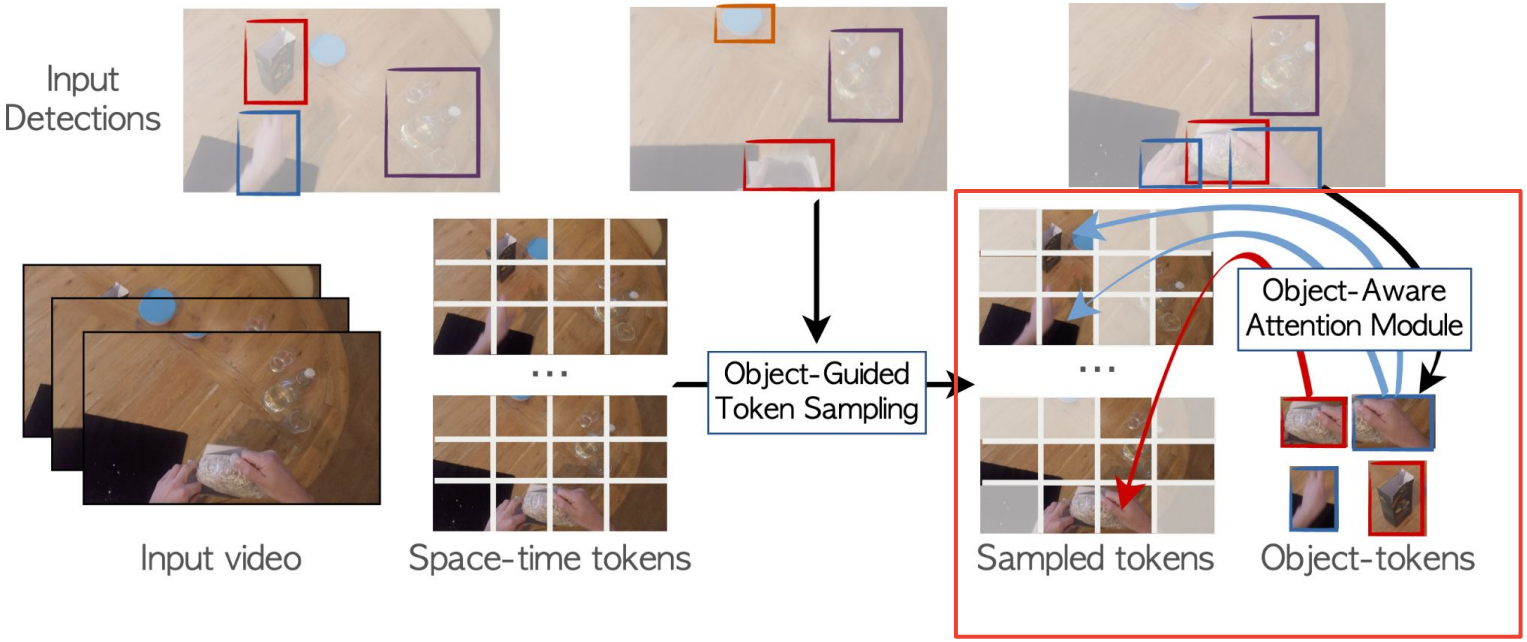

(b) Something-Something v2




(c) Epic-Kitchens

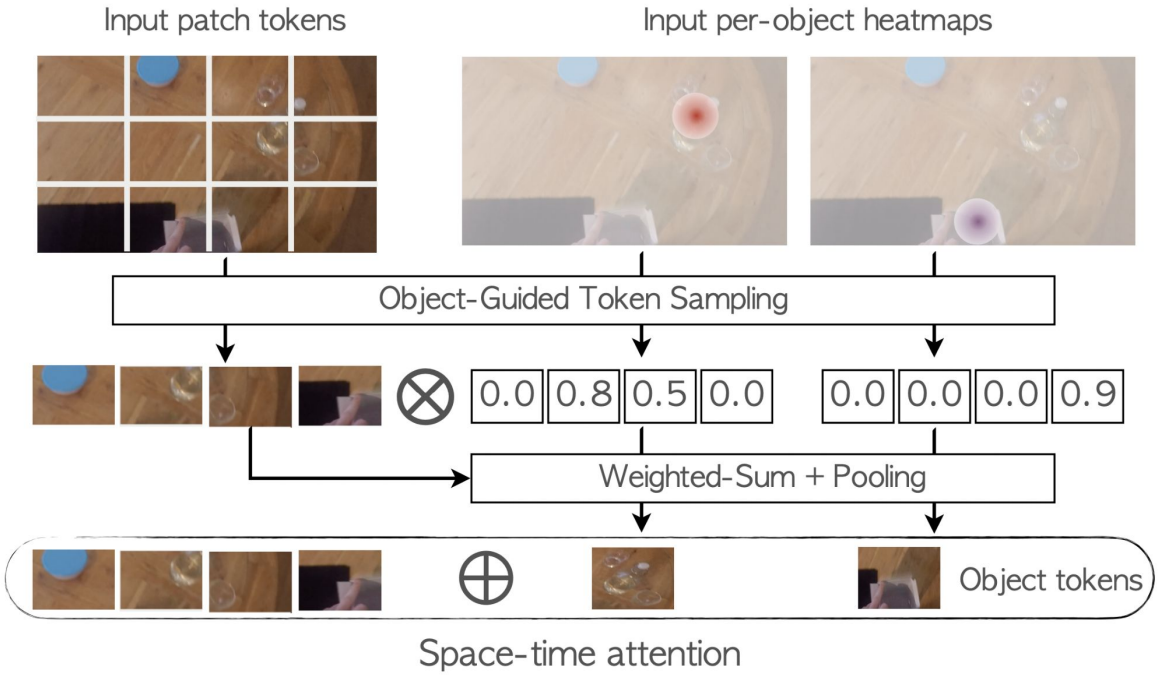
Google Research

# Framework

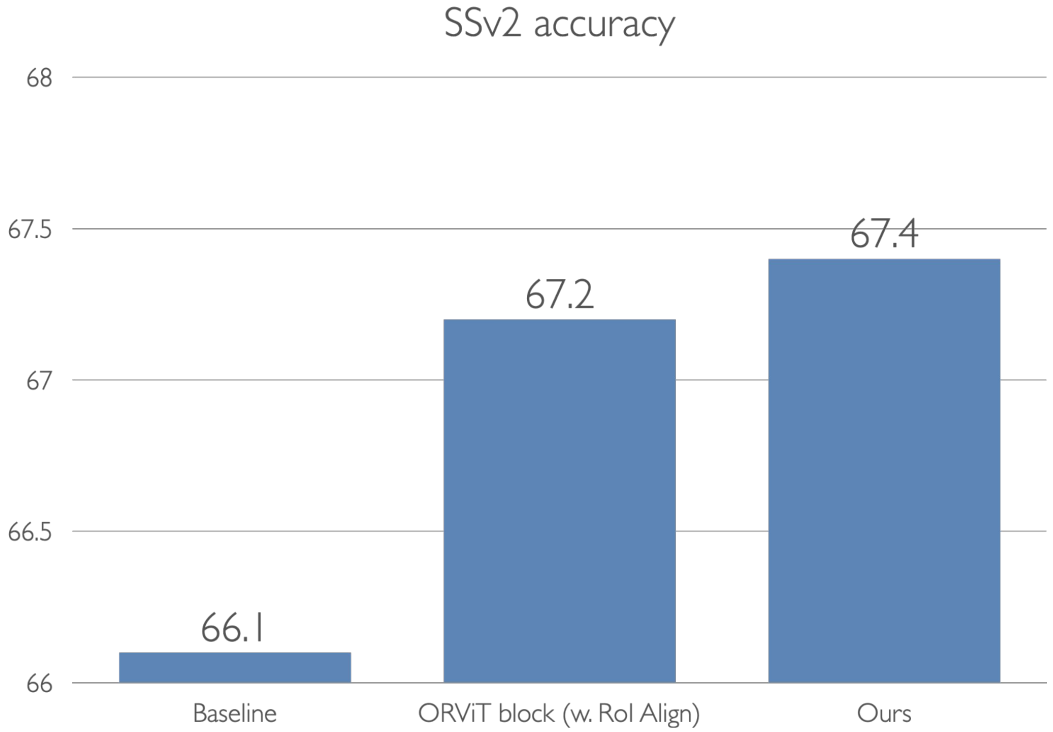



Google Research

# Object-aware attention module



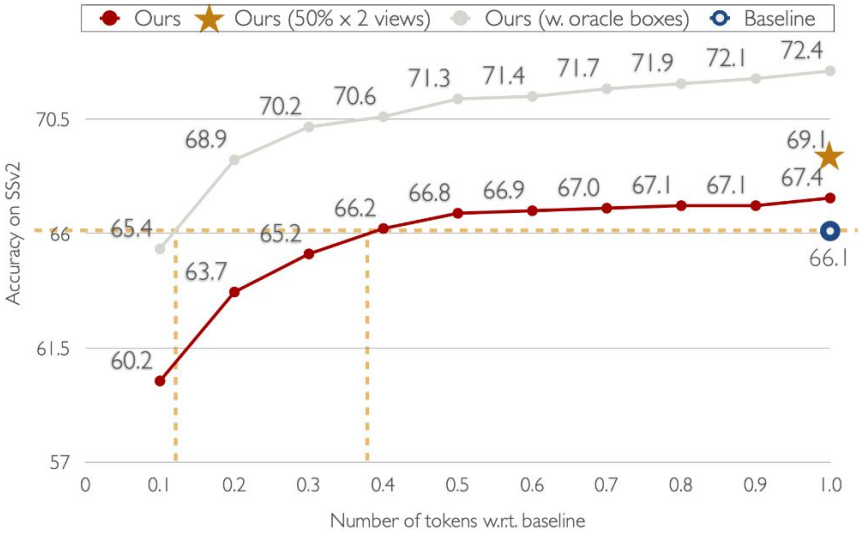
# Object-aware attention ablation



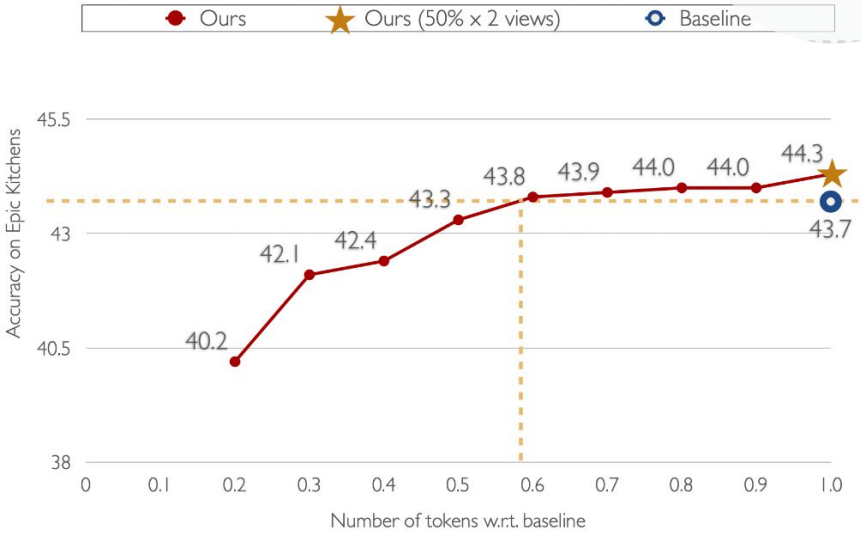

Object-Region Video Transformers, Herzig et al. CVPR 22

Google Research

# Apply both sampling & attention




(b) Something-Something v2

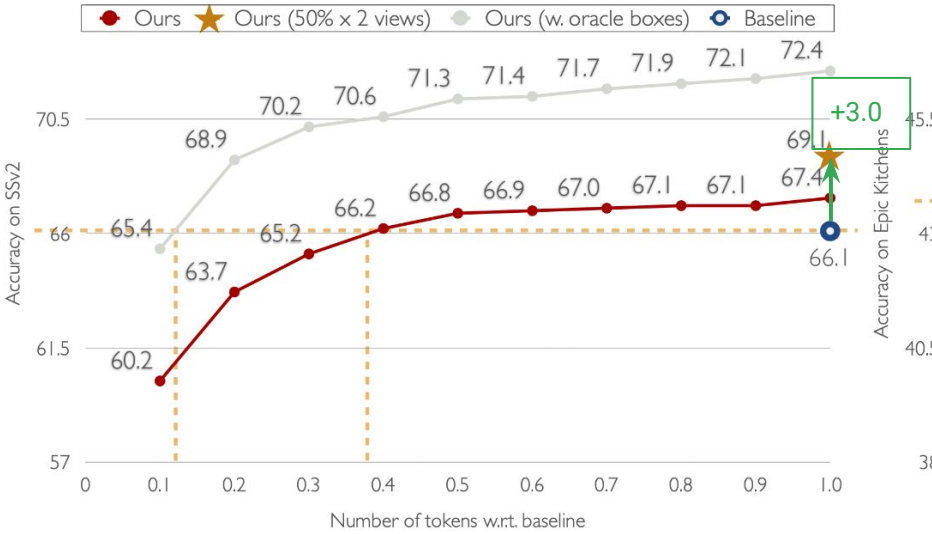



(c) Epic-Kitchens

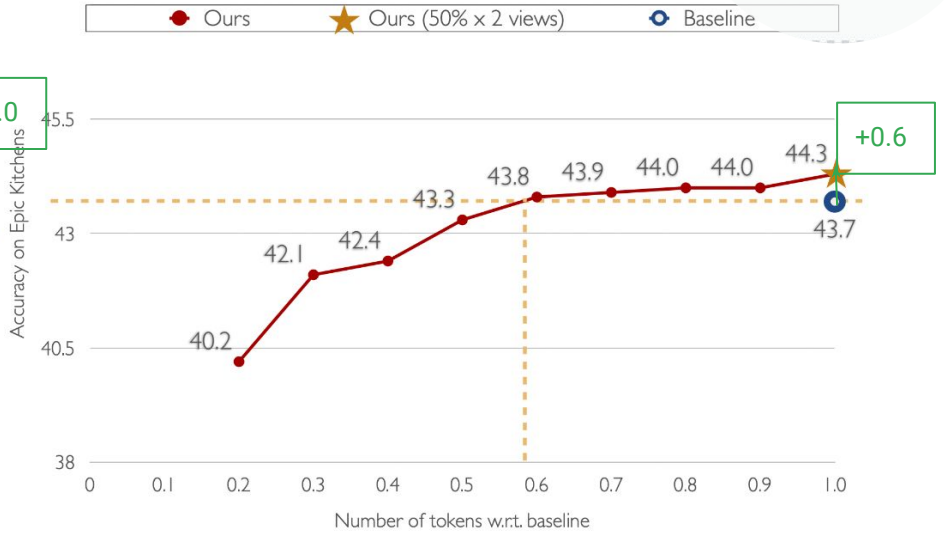
- Retaining baseline performance using 40-60% tokens.
- Less tokens → more testing views under the same total tokens.

Google Research

# Apply both sampling & attention




(b) Something-Something v2

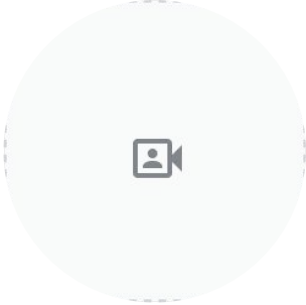



(c) Epic-Kitchens

- Retaining baseline performance using 40-60% tokens.
- Less tokens → more testing views under the same total tokens.

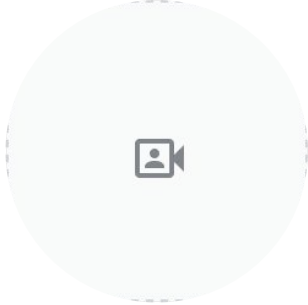
Google Research

# Takeaways



- Objects help action recognition in two ways:
  - Improve token-**efficiency** by downsampling tokens.
  - Improve **accuracy** by gathering feature in attention.

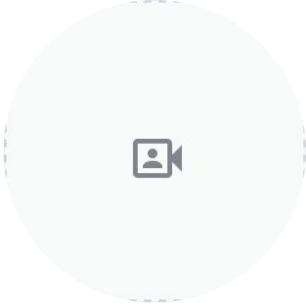
# Limitations & Discussions



- Need external detection inputs, so NOT actually **speed up** if counting detectors.

- The performance relies on detection quality.
  - Currently, domain-specific detectors performs the best.
  - General detectors with many background objects and did not improve as much.

# Learn more



- Poster: **TUE-AM-225**
- Email: zhouxy@google.com, aarnab@google.com
- Code: https://github.com/google-research/scenic

Google Research