FIBER: Fill-in-the-Blanks as a Challenging Video Understanding Evaluation Framework

Santiago Castro, Ruoyao Wang, Pingxuan Huang, Ian Stewart, Oana Ignat, Nan Liu, Jonathan C. Stroud, and Rada Mihalcea

sacastro@umich.edu

What's FIBER?

FIBER is a Video Understanding benchmark using a fill-in-the-blanks strategy applied on VaTeX.

- 28,000 10-second videos
- High human agreement
- Challenging

Two children throw _____ at each other as a video is captured in slow motion.

Correct answers: balloons, balloons filled with water, balloons of water, pink balloon, pink water balloon, things, water, water balloons, water-filled balloons

_____ sits at a drum set and practices playing the drums.

Correct answers: child, drummer, future drummer, girl, kid, little girl, little kid, musician, small child, young girl

A boy is trying to comb his hair while _____ dries it.

Correct answers: another person, friend, girl, his sister, his sister with hairdryer, person, young woman

Motivation

Existing benchmarks have fundamental issues:

- Multiple-choice benchmarks:
  - Unrealistic for production
  - Models learn to rely on distractors
- Free-form benchmarks' automatic evaluation is noisy

FIBER brings balance: both challenging and with a robust evaluation.

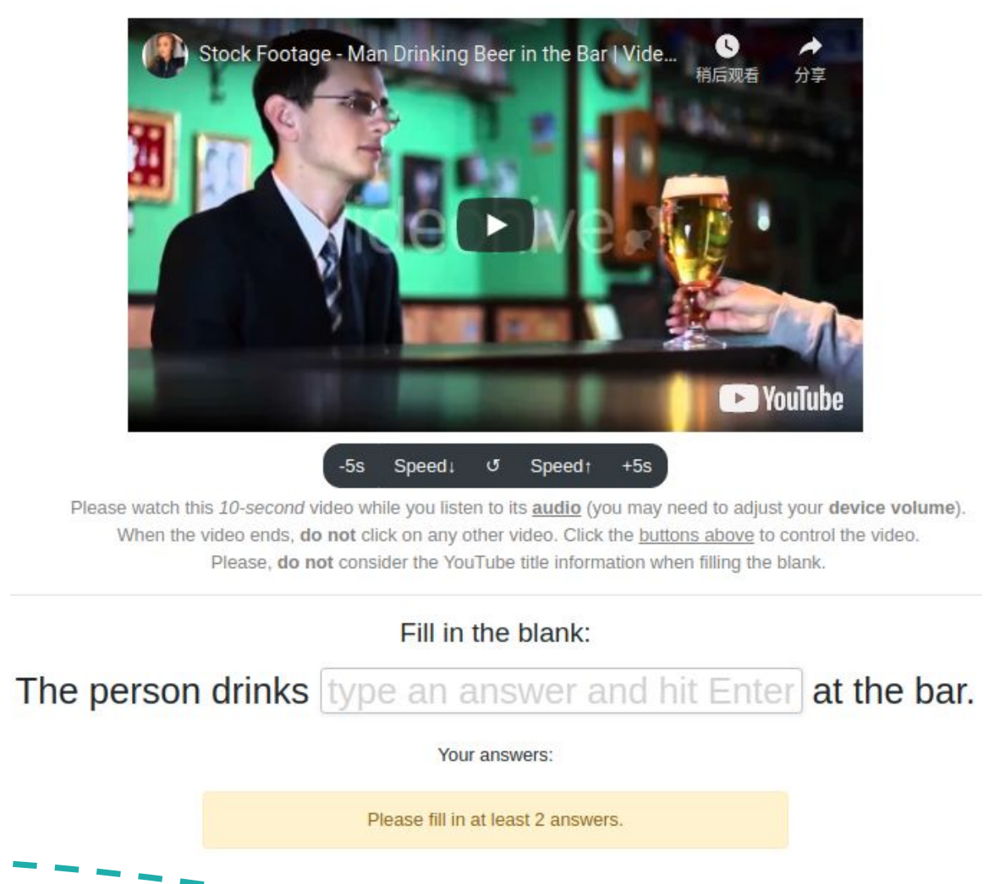
FIBER contains videos along with a sentence description with a noun phrase blank that needs to be fill in.

There are more correct answers than just the originally blanked phrase → we collect additional answers

Data Collection Recipe

1. Take a Video Captioning dataset (VaTeX)
2. For each video caption: extract NPs and blank them
3. Split the data into train, val, and test
4. Collect additional correct answers for val and test (1,000 each in FIBER):
   - Amazon Mechanical Turk
   - Nine annotators per video

Annotation Interface



Please watch this *10-second* video while you listen to its **audio** (you may need to adjust your **device volume**). When the video ends, **do not** click on any other video. Click the buttons above to control the video. Please, **do not** consider the YouTube title information when filling the blank.

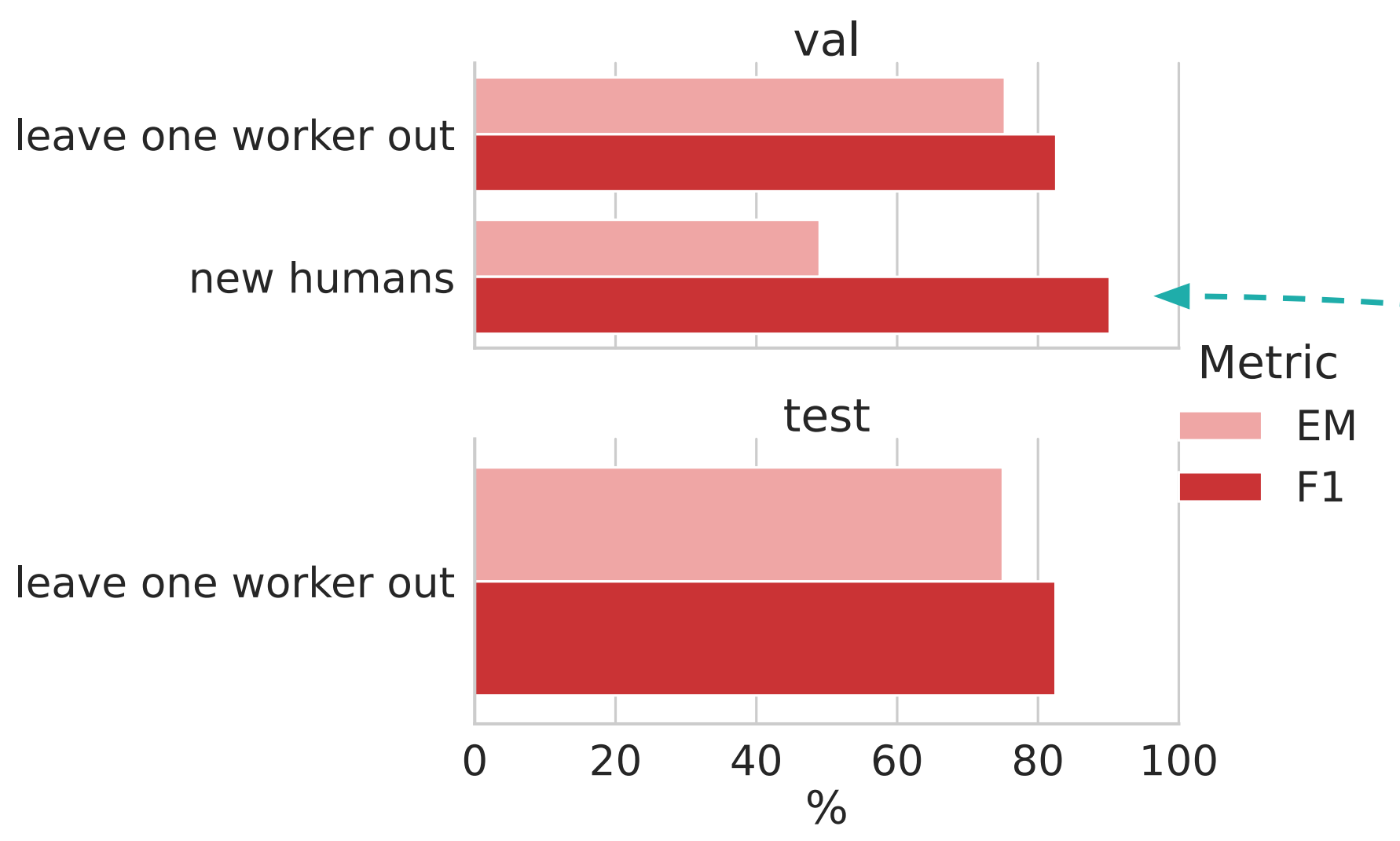
Fill in the blank:

The person drinks [type an answer and hit Enter] at the bar.

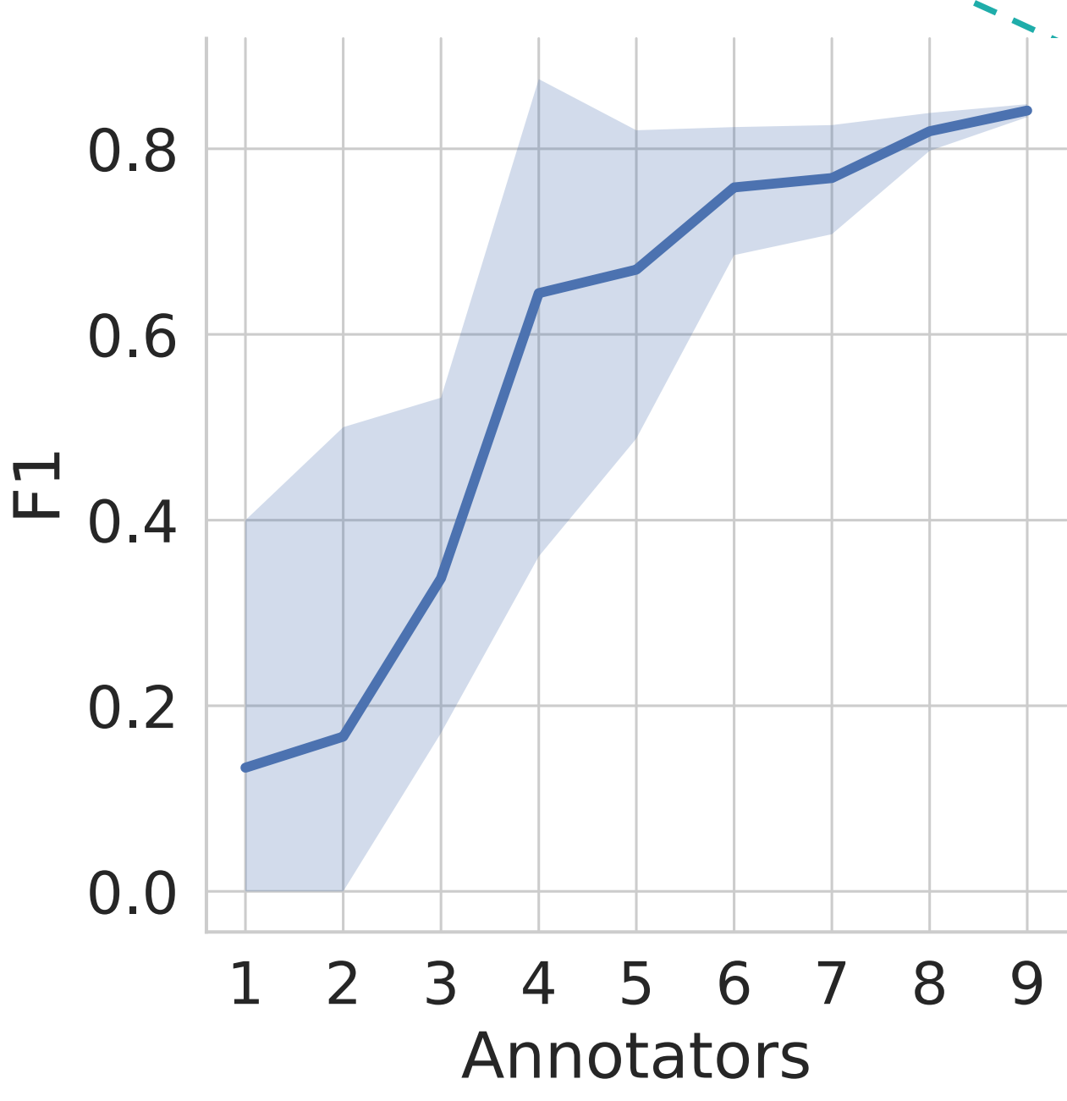
Your answers:

Please fill in at least 2 answers.

Human Metrics

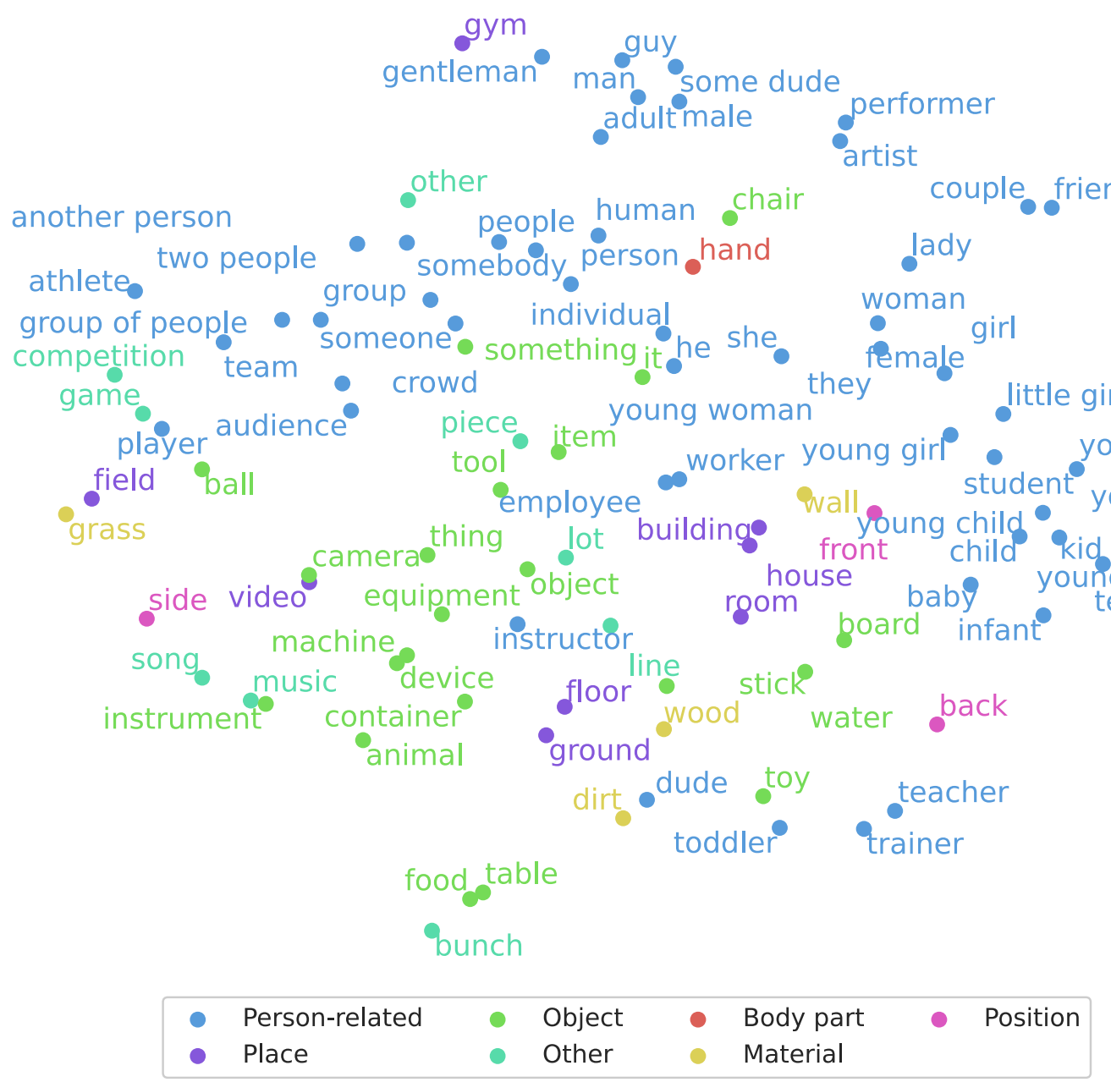


Agreement vs Annotators

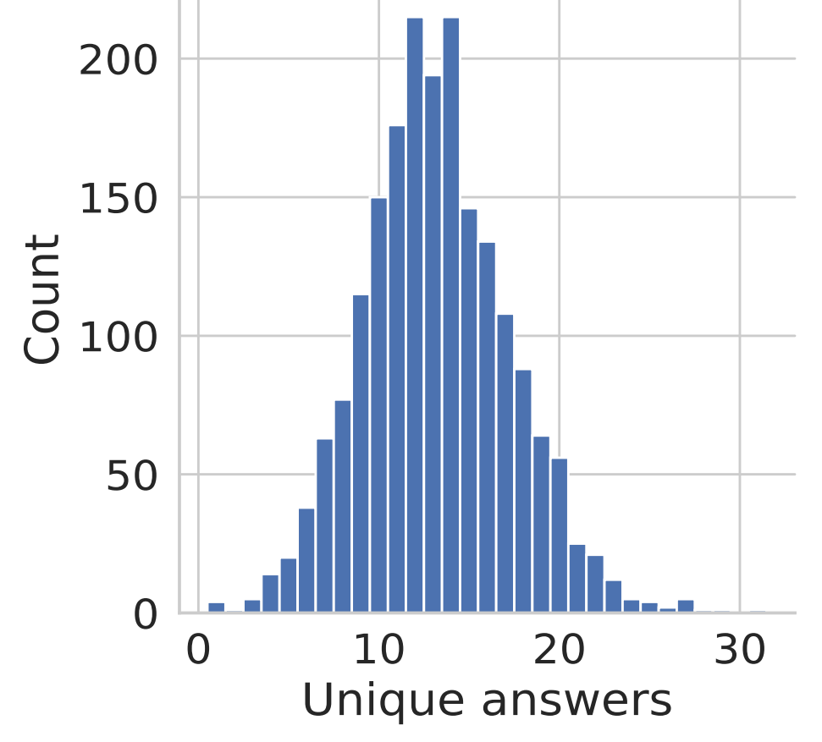


Data Insight

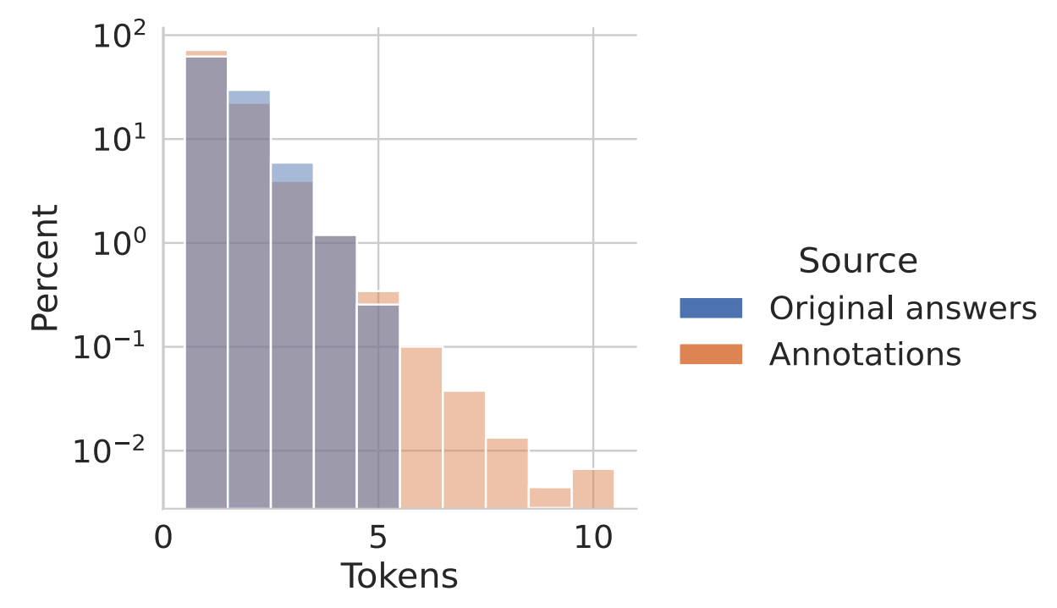
Top 100 Most Freq. Ans.



Answers per Question

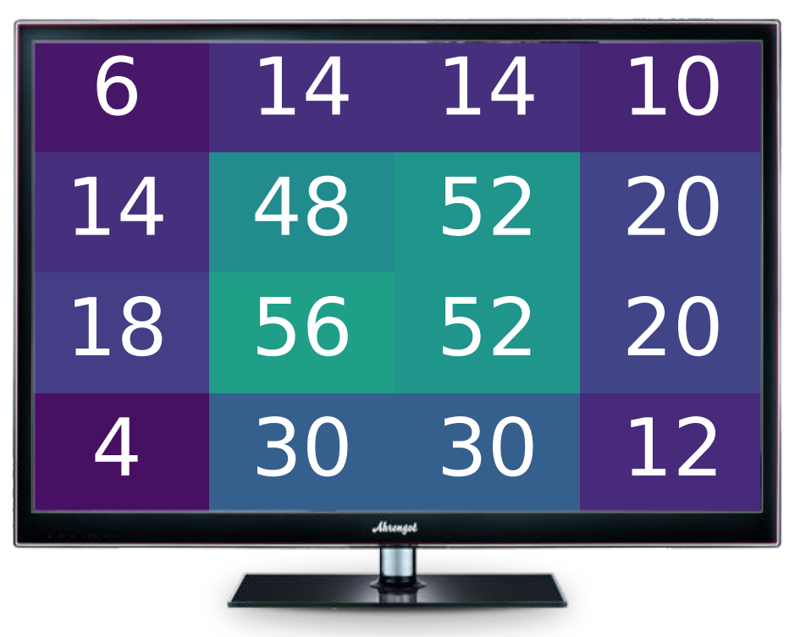


Tokens per Answer



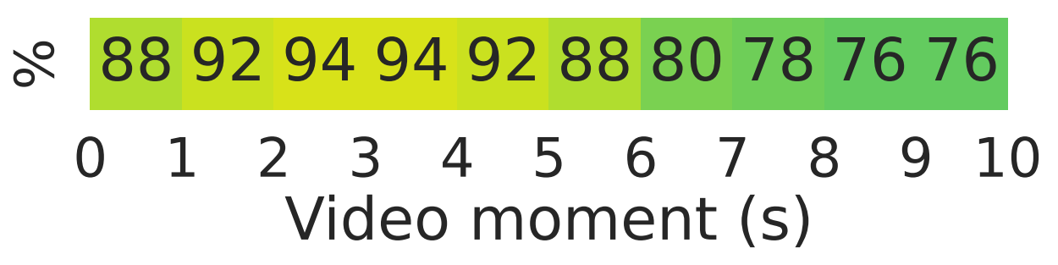
Target Entity

Where?

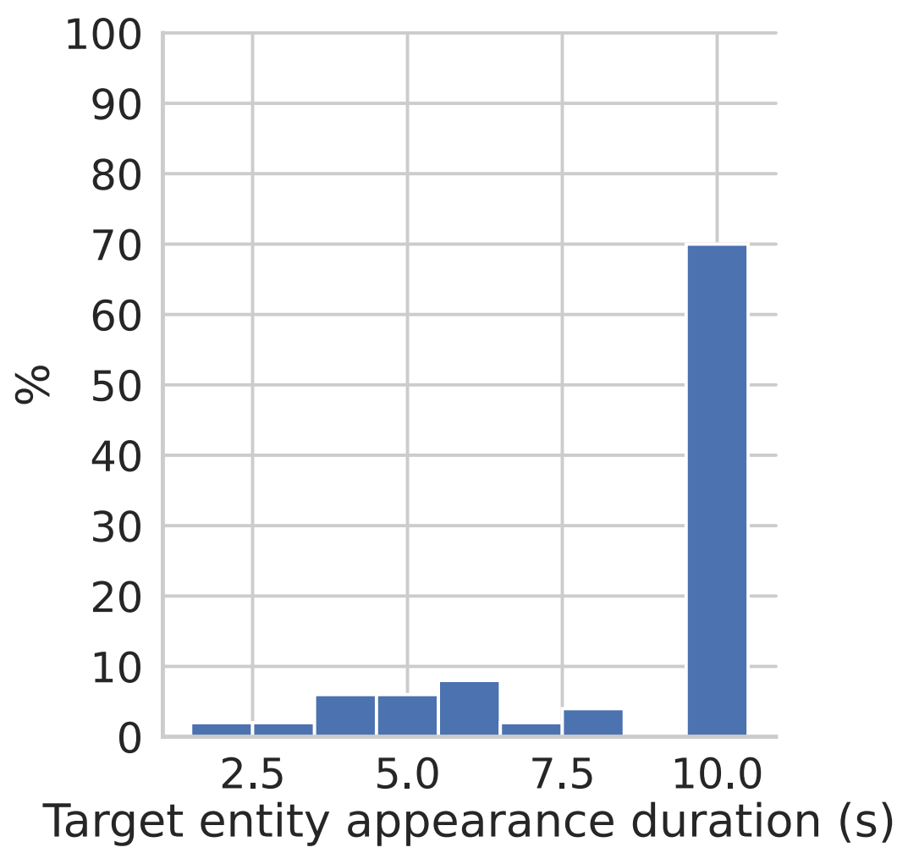


6	14	14	10
14	48	52	20
18	56	52	20
4	30	30	12

When?

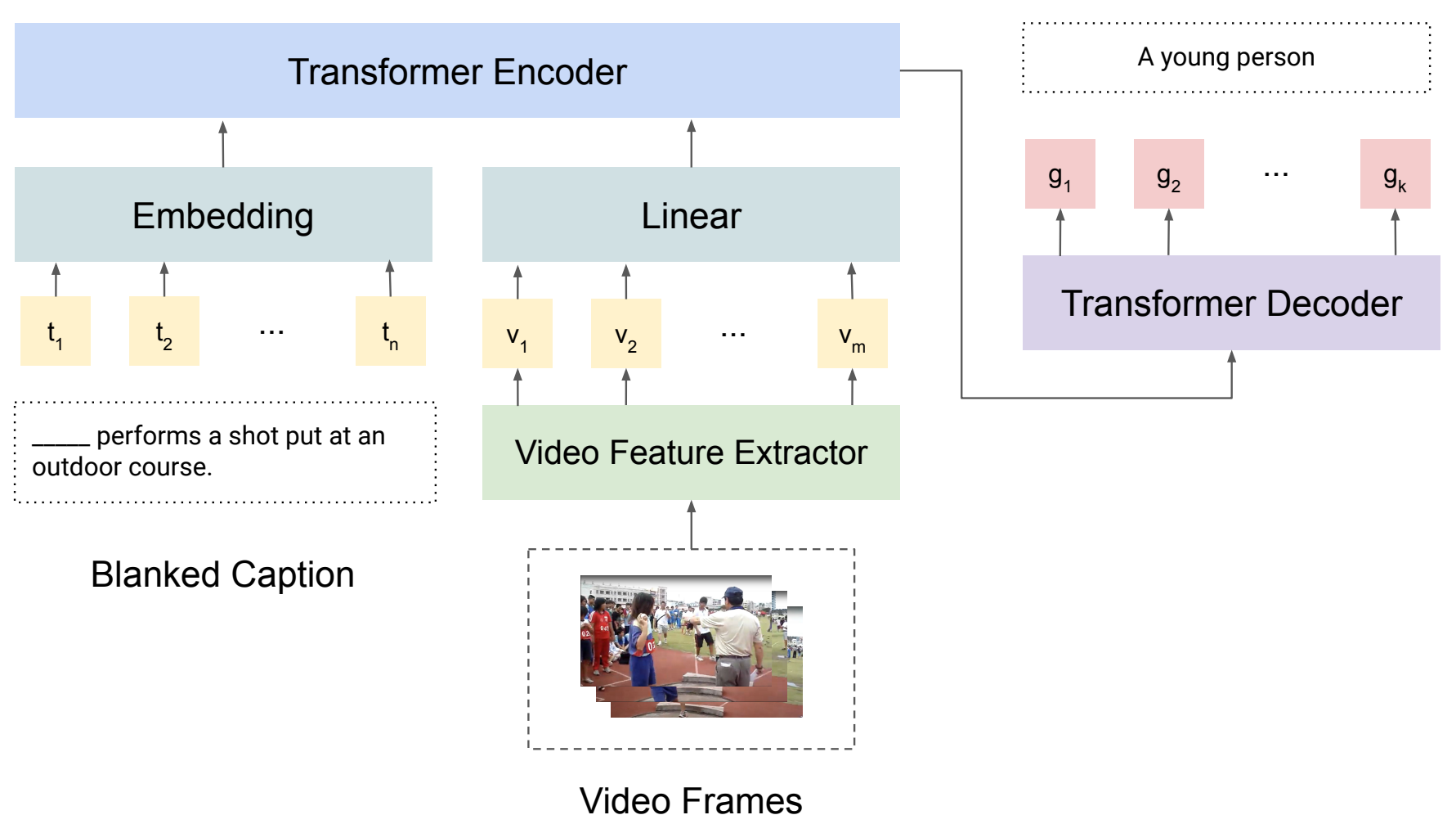


For how long?

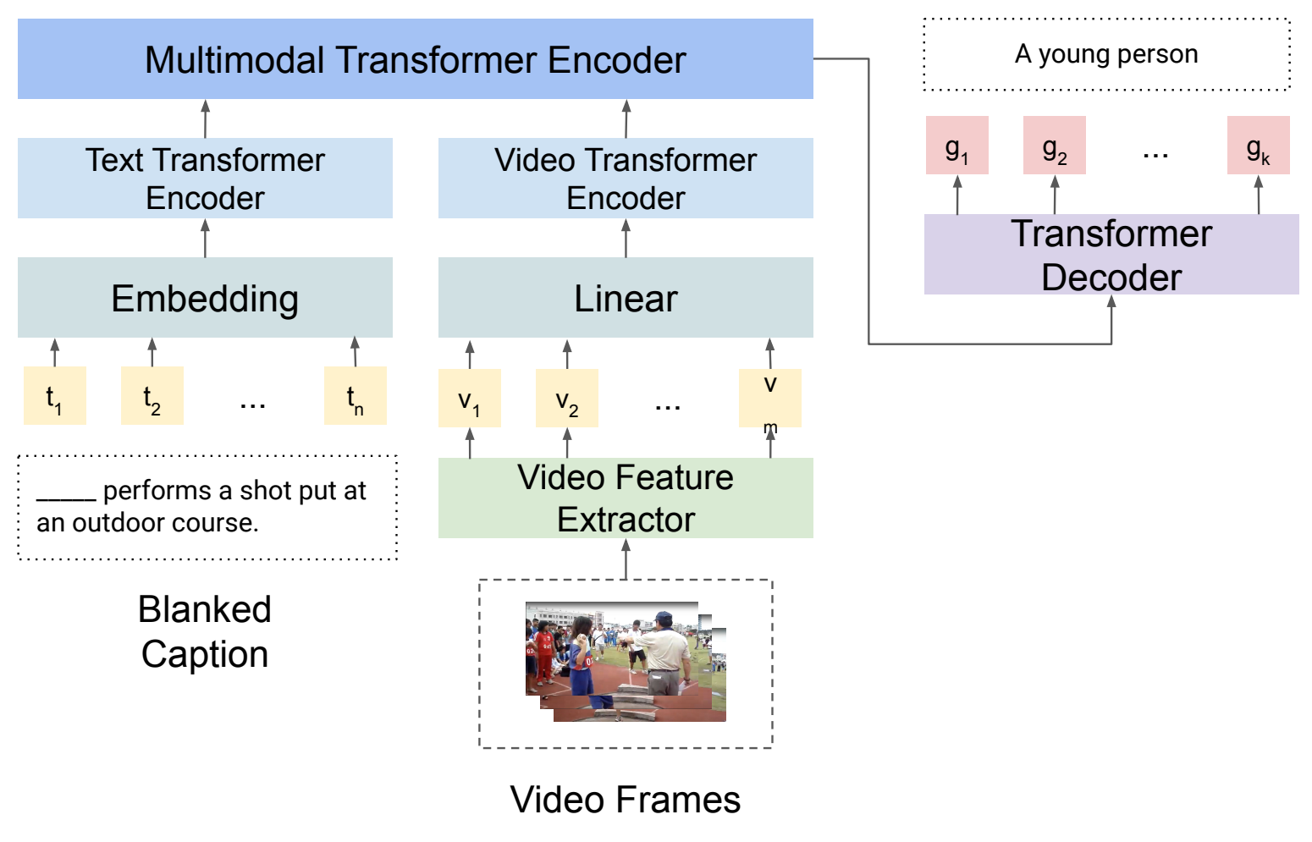


Models

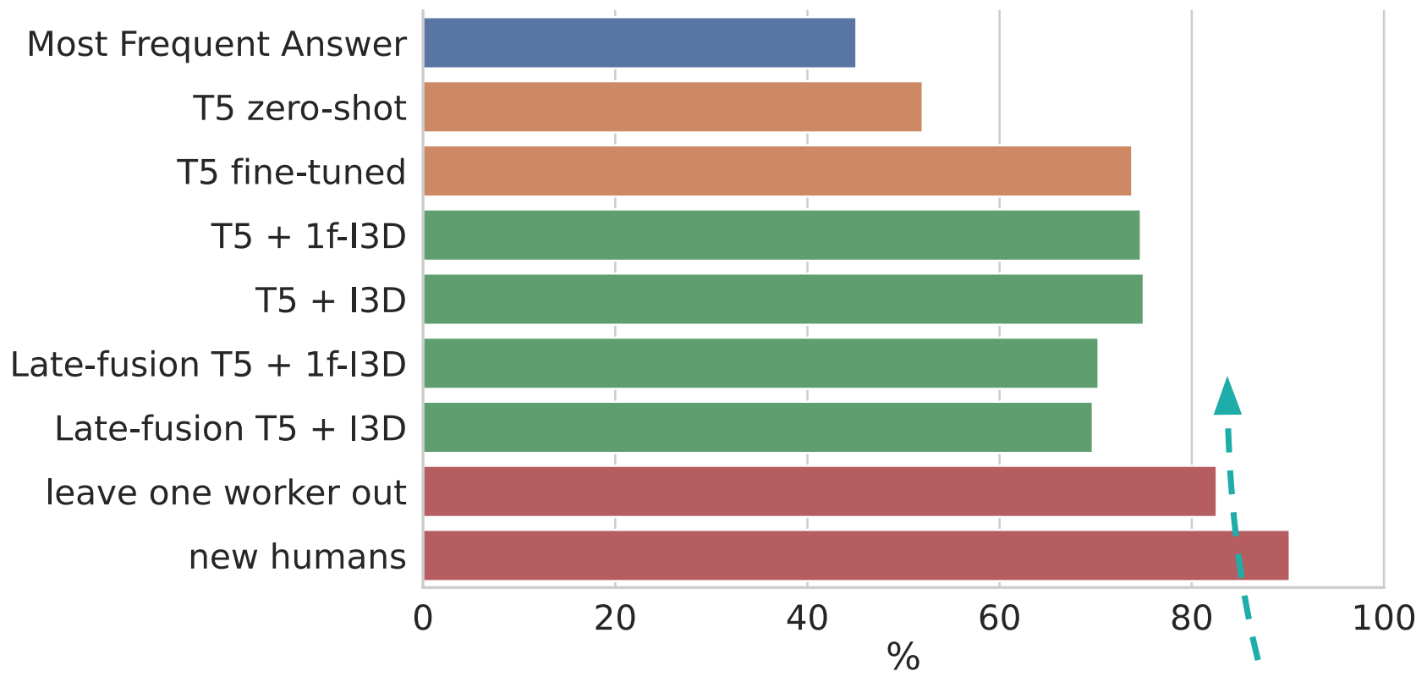
T5 + I3D

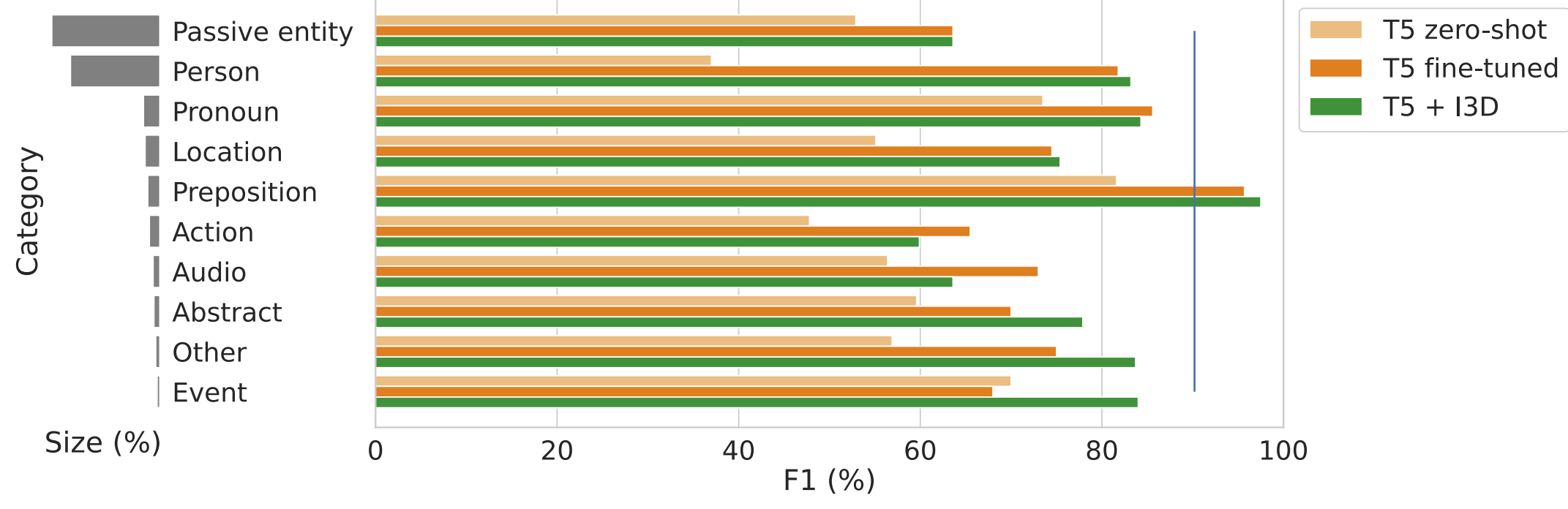


Late-fusion T5 + I3D



Experiments





Takeaways

- FIBER: a new Video Understanding benchmark
  - 28,000 10-second videos and tests based on filling blanks on text descriptions
- FIBER is a robust benchmark for Video Understanding
  - Challenging and unsolved
  - Robust evaluation (high human performance!)
- Our data collection recipe can be replicated to create similar datasets
- We present a T5+I3D transformer model as a strong method

Data + Code: lit.eecs.umich.edu/fiber

SCAN ME

ACL 2022
22ND – 27TH MAY | 60TH MEETING | DUBLIN