COBE: Contextualized Object Embeddings from Narrated Instructional Video

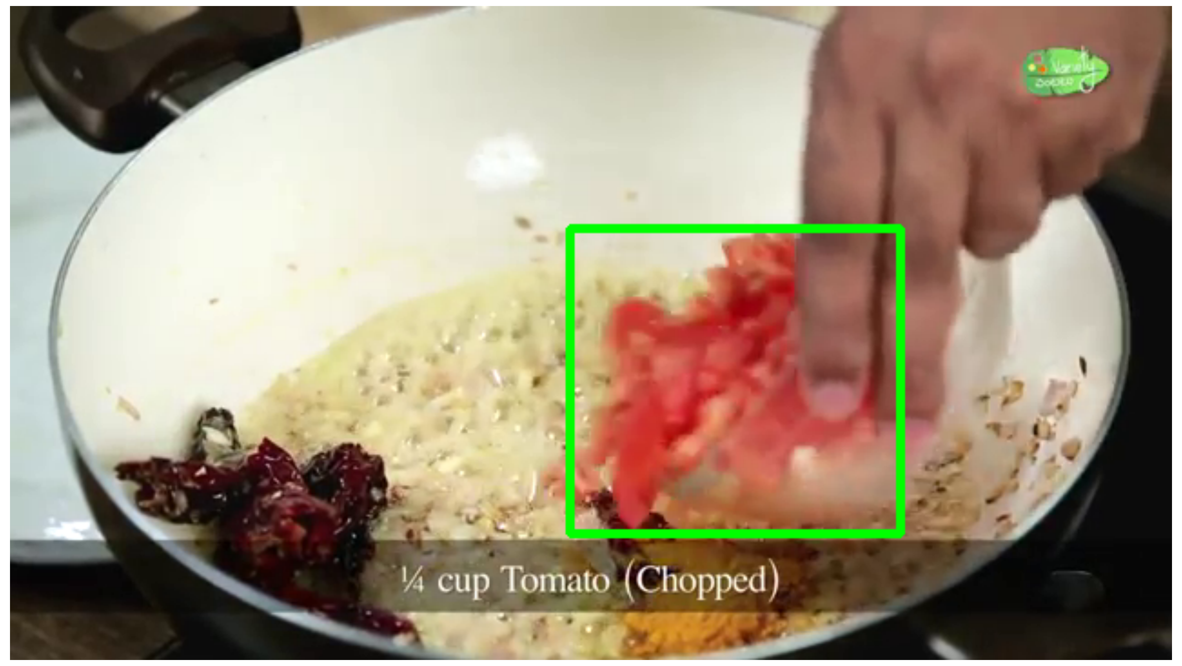
FACEBOOK AI

Gedas Bertasius

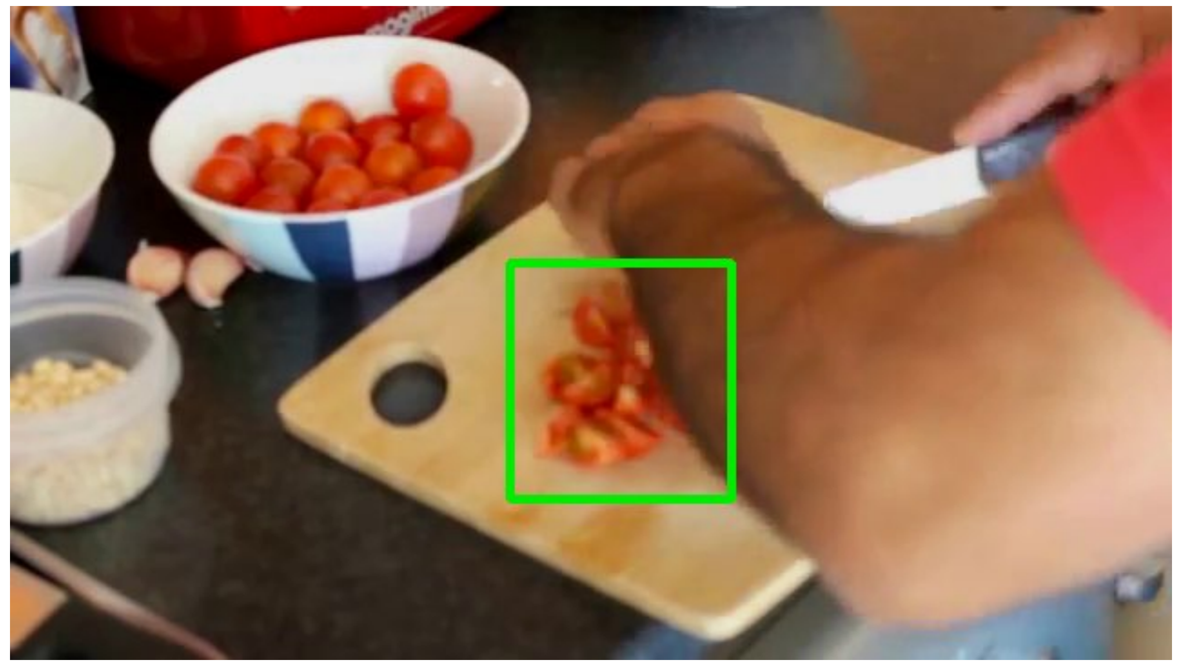
Lorenzo Torresani

Motivation

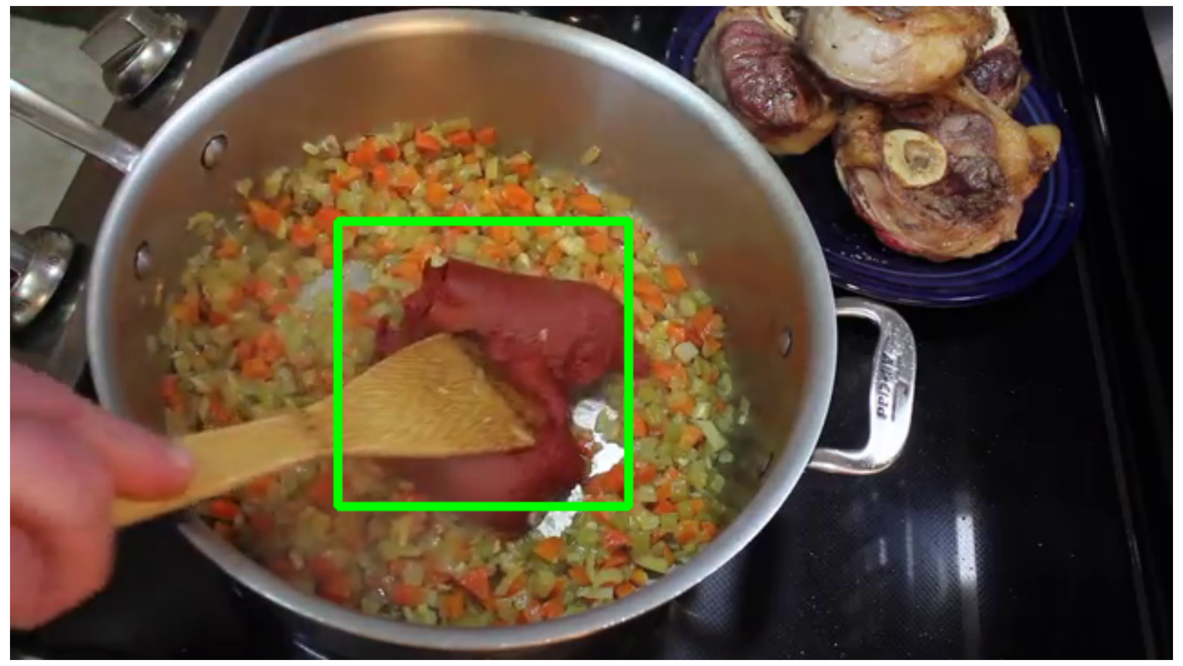
- Many objects in the real-world exhibit dramatic variations in their appearance.



chopped tomatoes

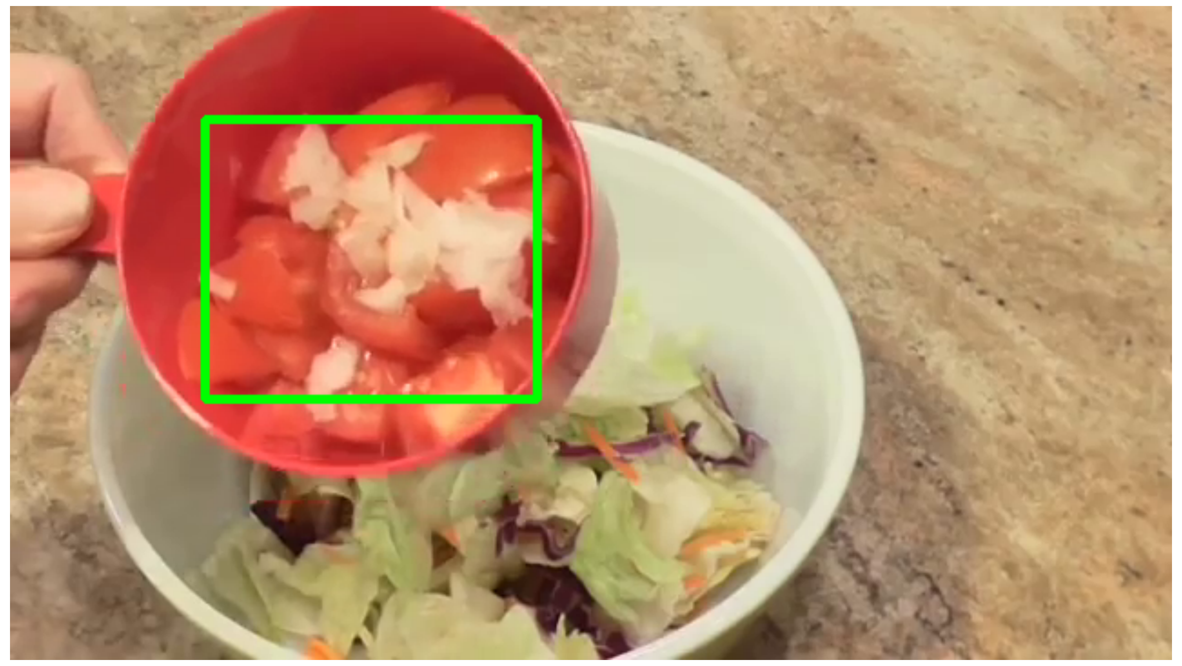


halved tomatoes



tomato paste

cutting a tomato



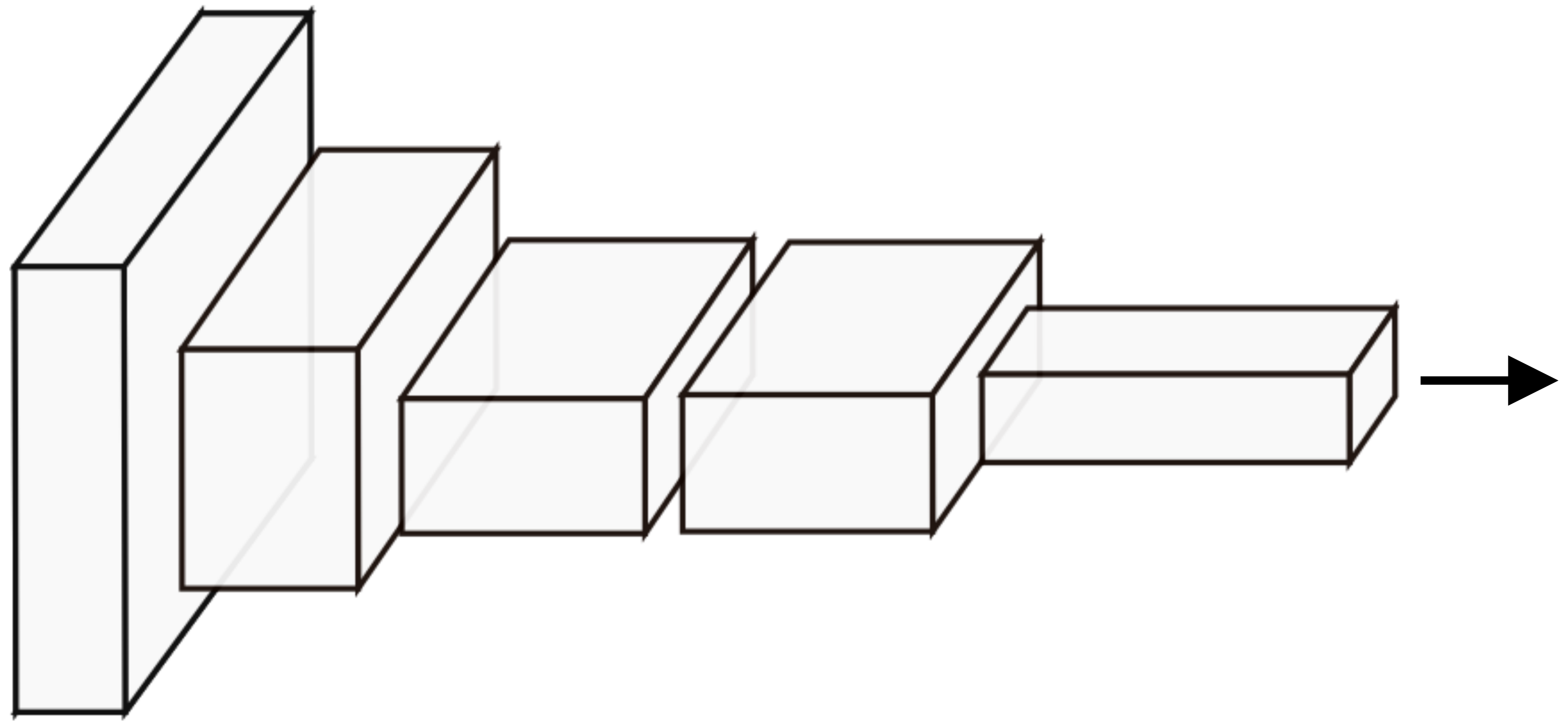
tomatoes & onions

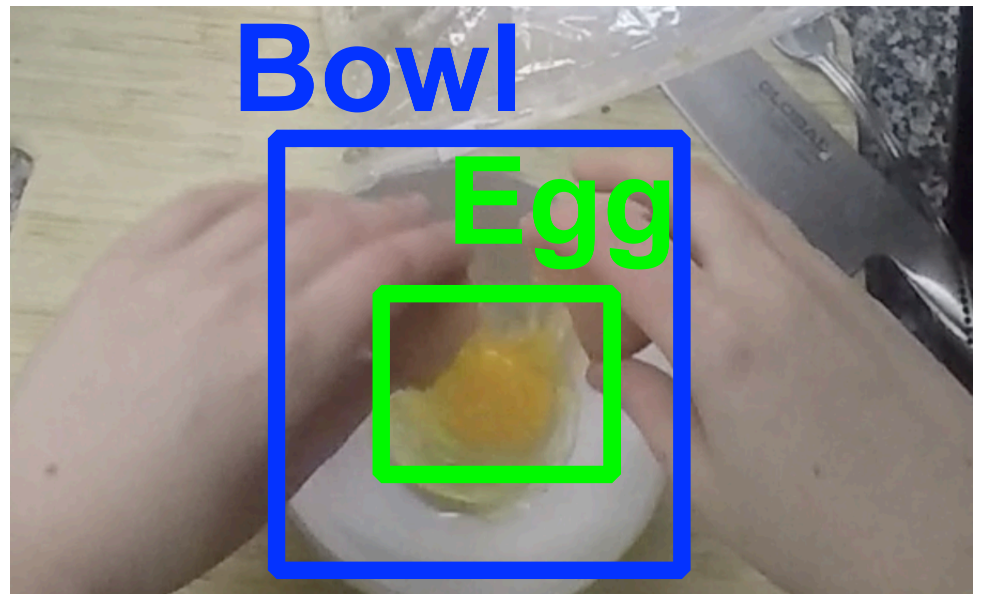
tomato sauce

Motivation

- Most visual models are trained to detect objects at a very coarse level, with label spaces typically expressed in terms of nouns.

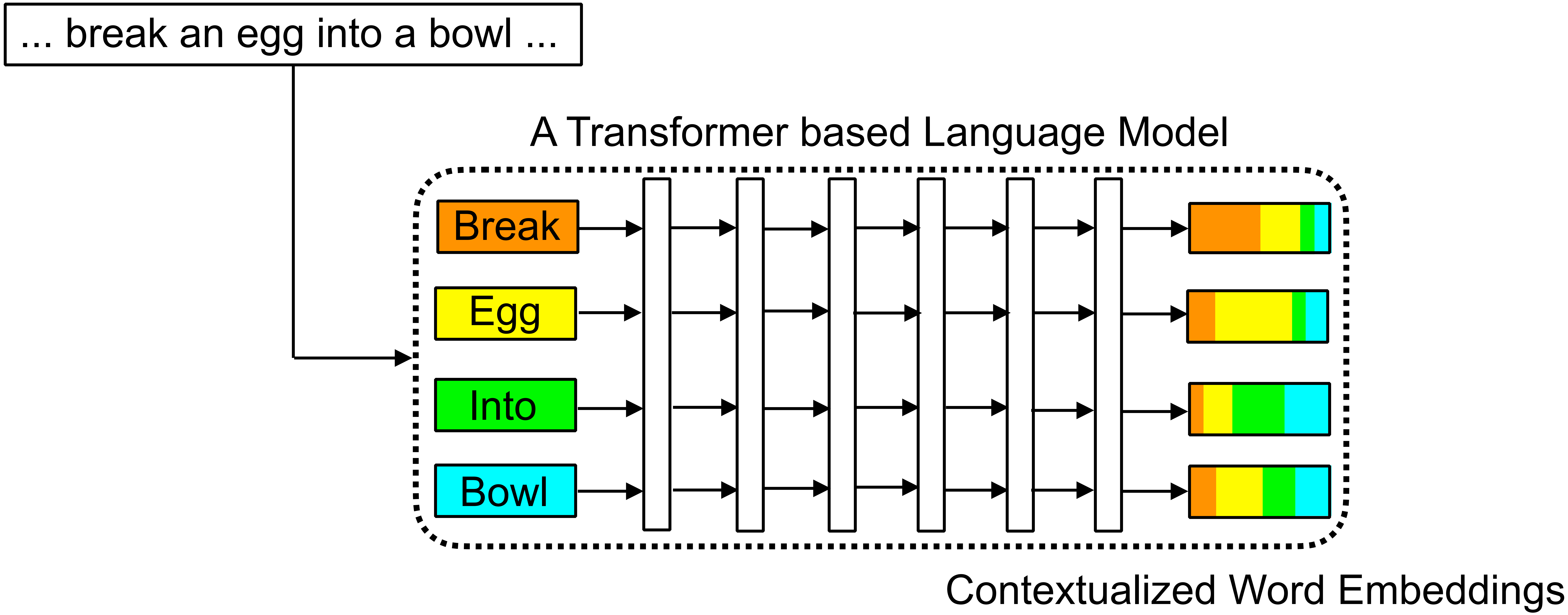
Input Image





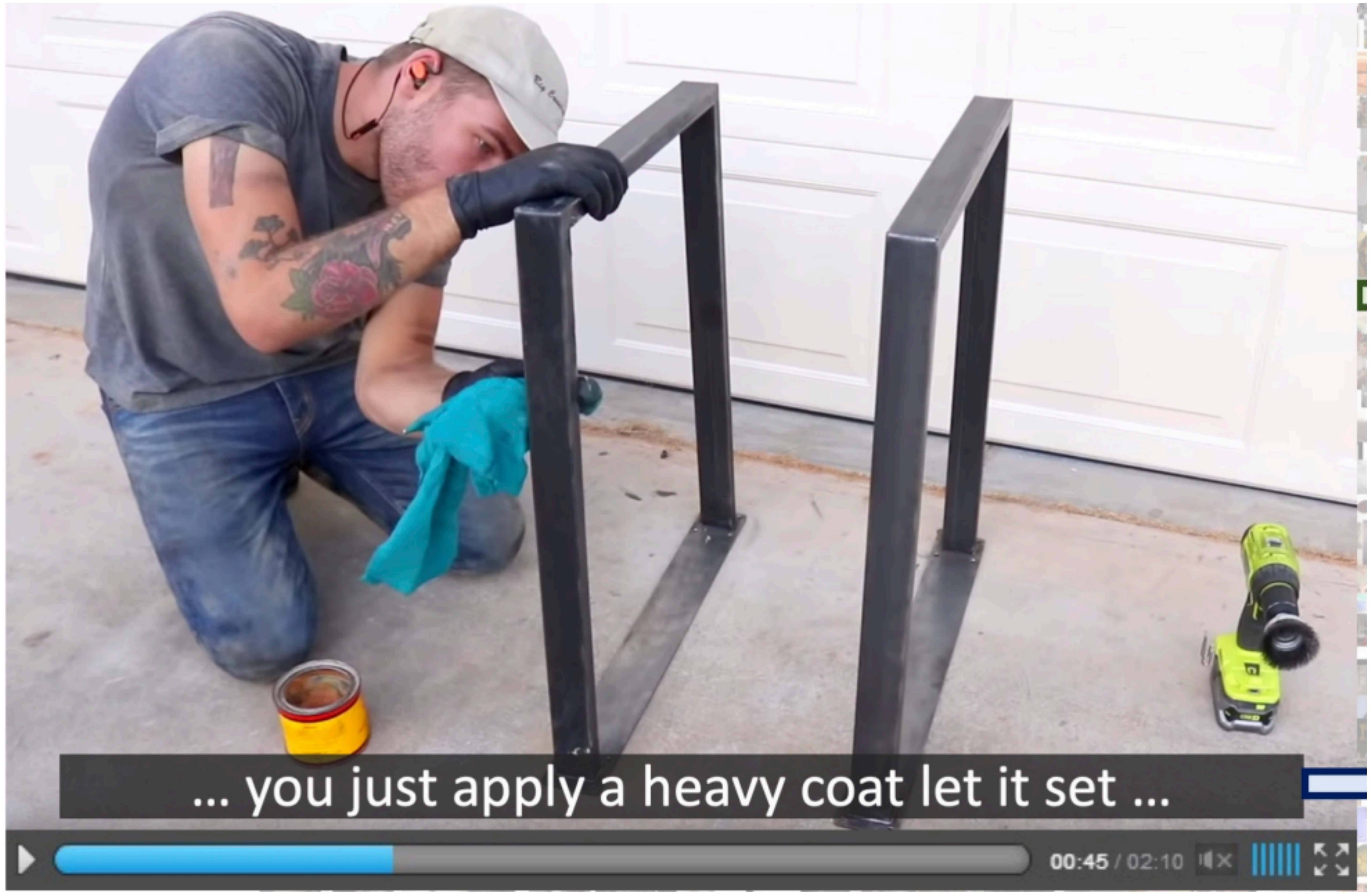
Output Detections

Contextualized Word Embeddings

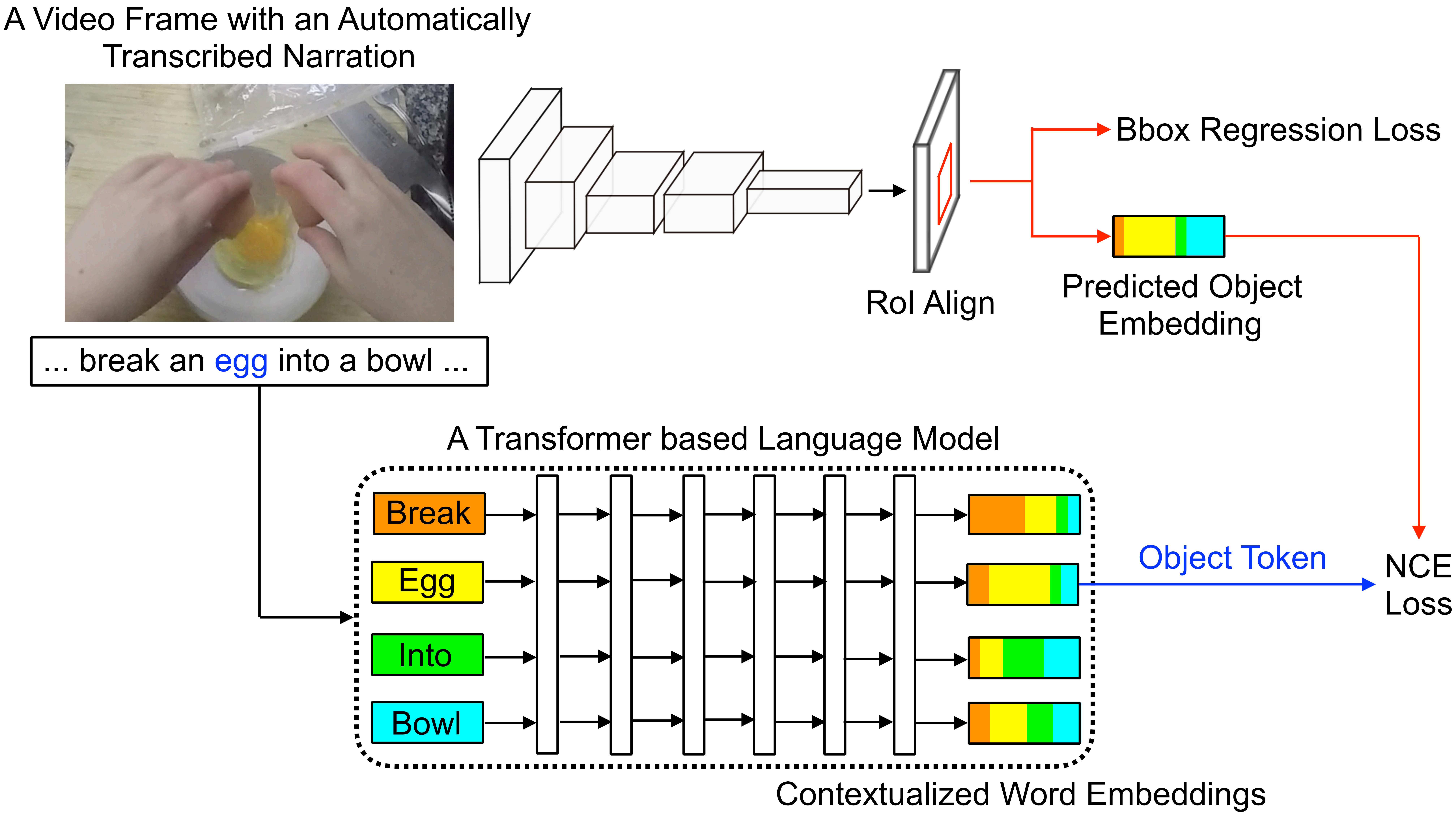

... break an egg into a bowl ...
A Transformer based Language Model
Break
Egg
Into
Bowl
Contextualized Word Embeddings

Dataset

- We leverage the recently introduced HowTo100M dataset which includes over 100M clips sourced from narrated instructional Web videos.



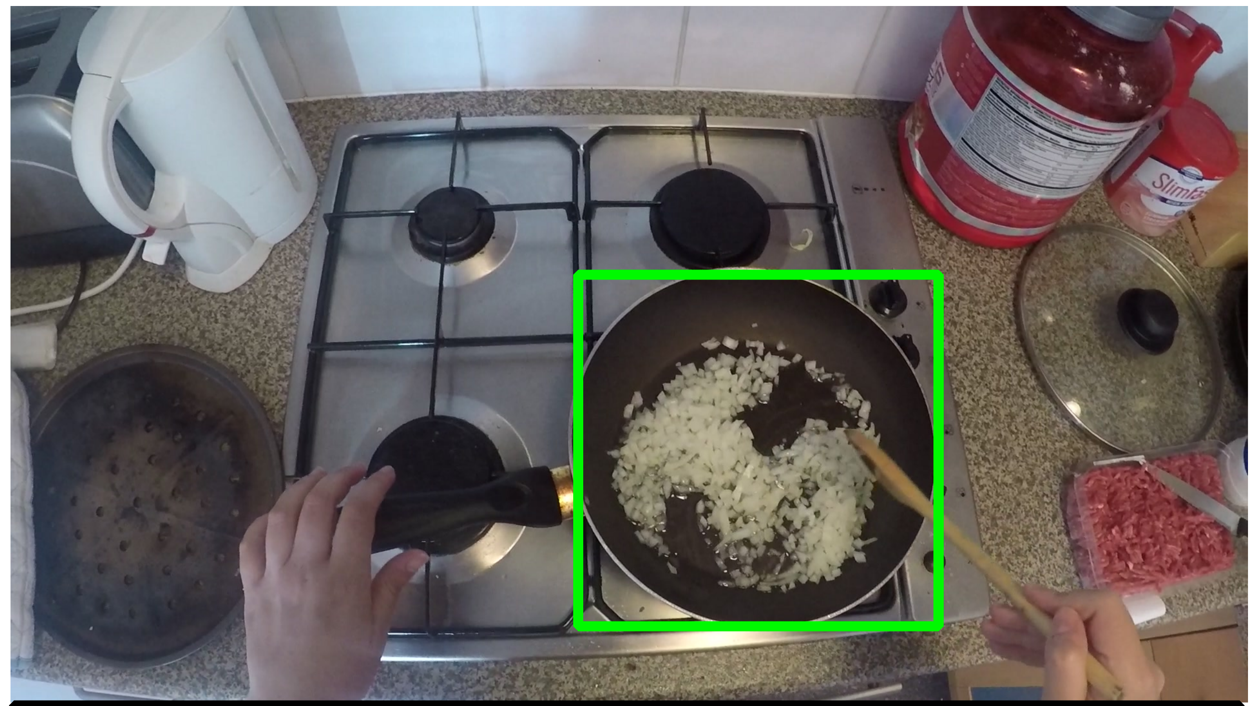
Contextualized Object Embeddings (COBE)


A Video Frame with an Automatically Transcribed Narration
... break an egg into a bowl ...
RoI Align
Bbox Regression Loss
Predicted Object Embedding
A Transformer based Language Model
Break
Egg
Into
Bowl
Object Token
NCE Loss
Contextualized Word Embeddings

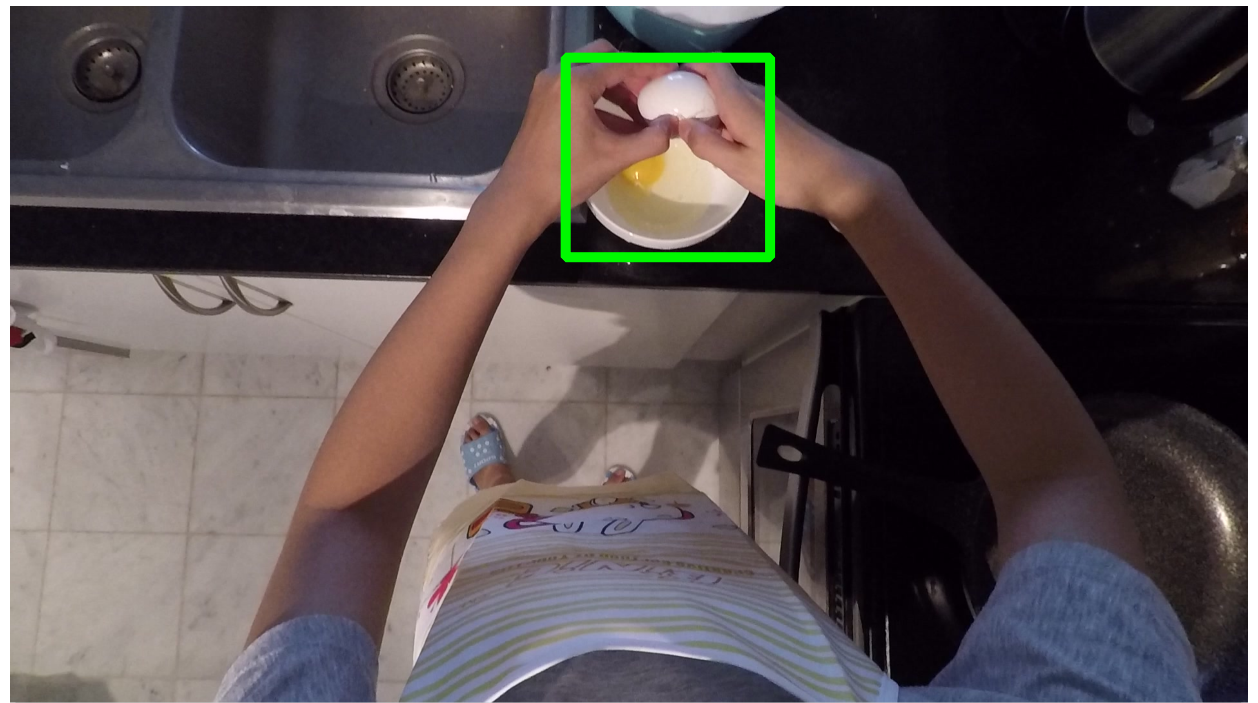
Results

Object-To-Text Retrieval:

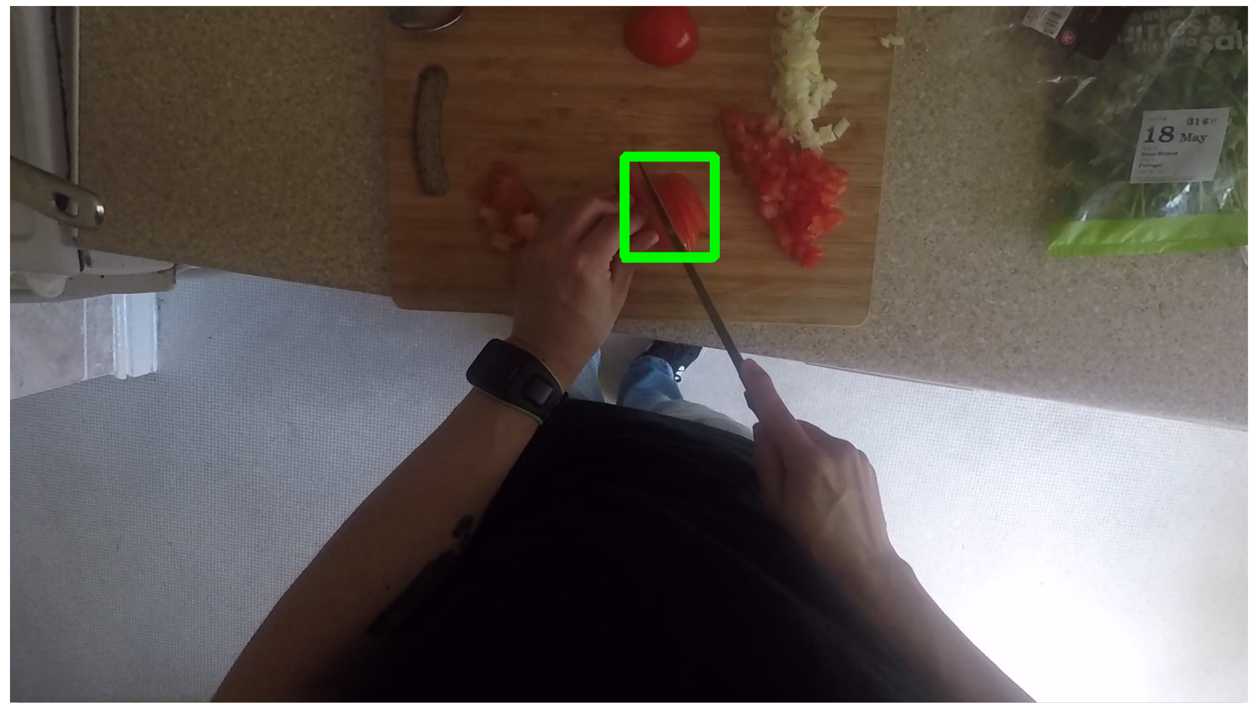
- Given a visual query, we retrieve most similar (**object**, **context**) text pairs in the space of a contextual language model.



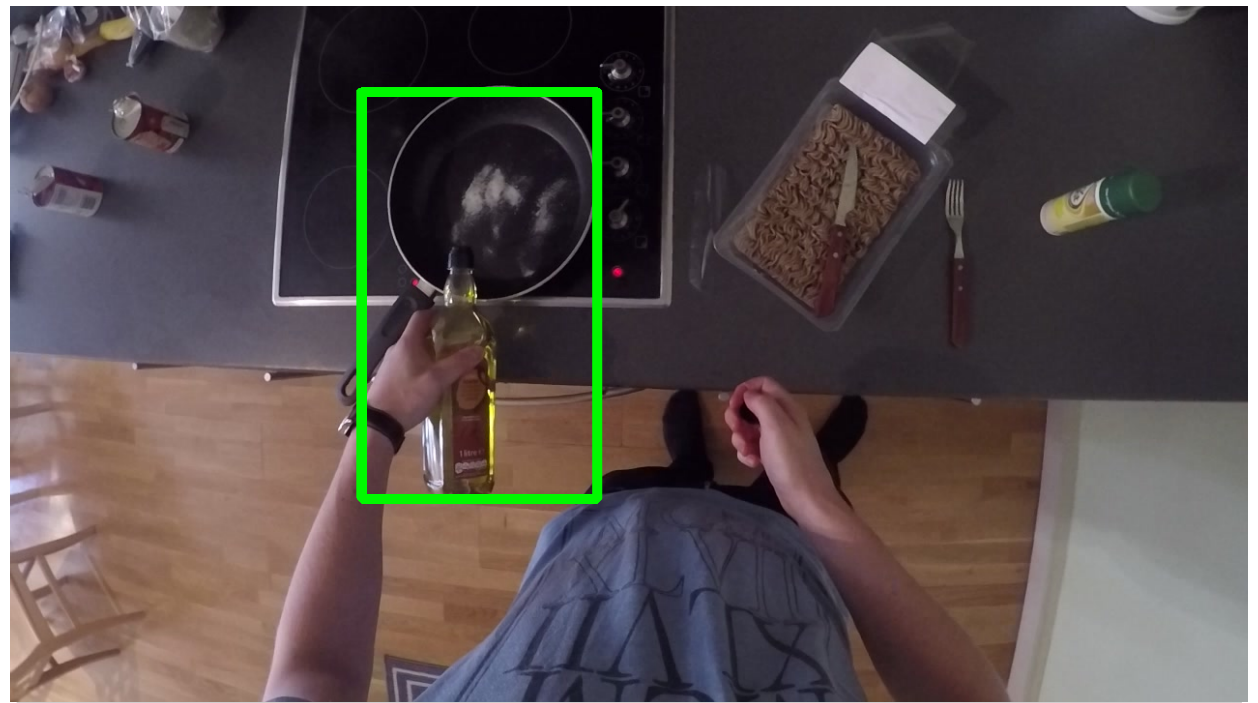
Object:	Context:
onion	chopped
pan	onions
pan	medium
pan	sauté
onion	sauté



Object:	Context:
bowl	egg
egg	whites
egg	crack
egg	bowl
egg	yolk



Object:	Context:
tomato	slice
tomato	slices
tomato	chop
tomato	knife
tomato	cherry

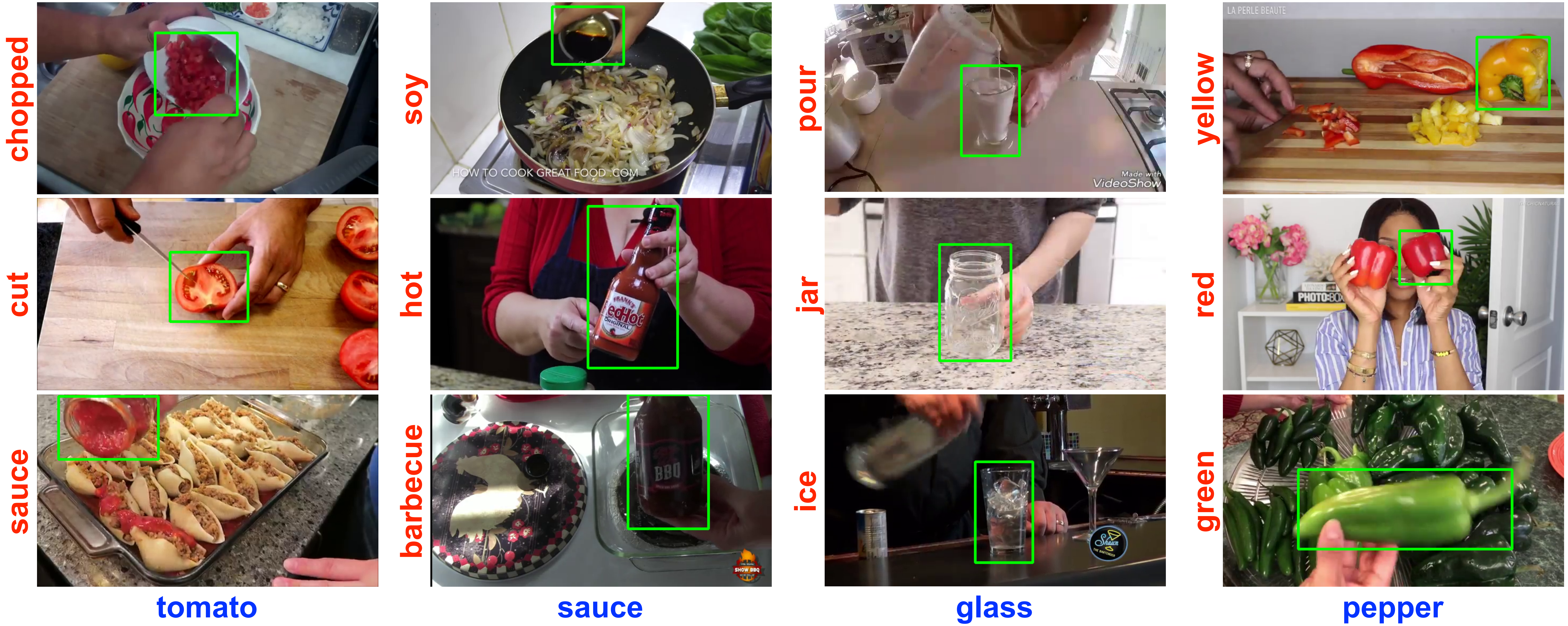


Object:	Context:
pan	oil
oil	pan
pan	stick
oil	olive
pan	pour

Results

Text-To-Object Retrieval:

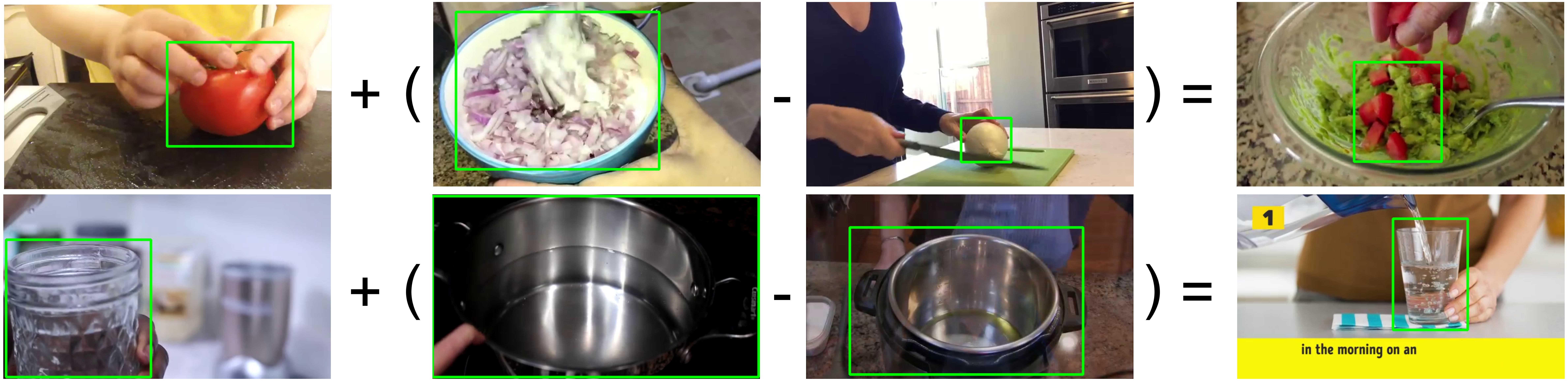
- Given a text query of the form (**object**, **context**), we retrieve most similar object instances in the space defined by the contextual language model.



Results

Visual Object Analogies:

- We can leverage our learned contextualized object embeddings to combine different visual concepts via simple vector arithmetic.



Conclusions

- In contrast to prior work, which focuses on noun-centric object detection, we present a framework for learning object detectors that generalize to novel object states.

Conclusions

- In contrast to prior work, which focuses on noun-centric object detection, we present a framework for learning object detectors that generalize to novel object states.

- Our framework does not require manually labeled text descriptions but instead leverages automatically transcribed narrations of instructional videos.

Conclusions

- In contrast to prior work, which focuses on noun-centric object detection, we present a framework for learning object detectors that generalize to novel object states.

- Our framework does not require manually labeled text descriptions but instead leverages automatically transcribed narrations of instructional videos.

- Our model is effective in the scenarios of zero-shot and few-shot learning.