TempCLR: Temporal Alignment Representation with Contrastive Learning

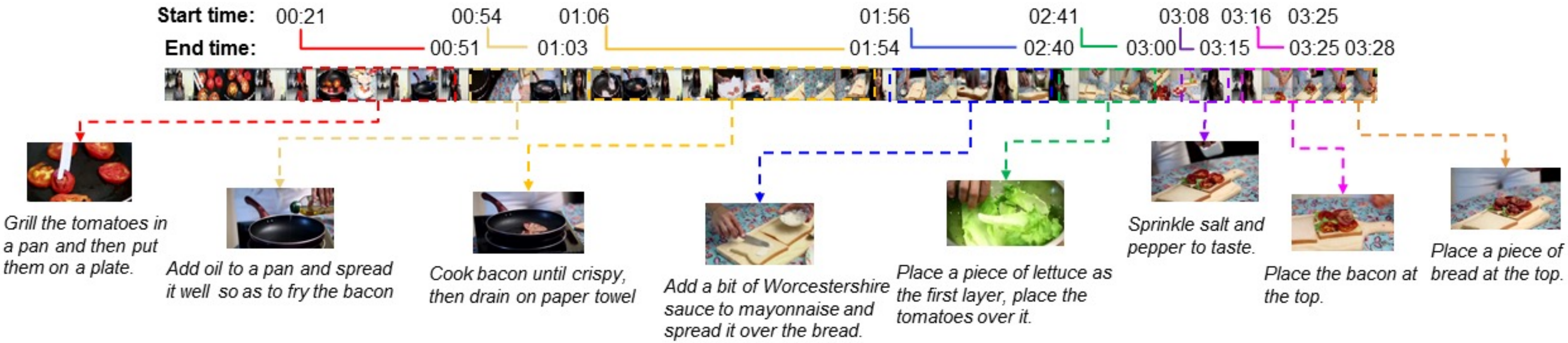
Yuncong Yang, Jiawei Ma*, Shiyuan Huang, Long Chen, Xudong Lin, Guangxing Han, Shih-Fu Chang*

Digital Video | Multimedia Laboratory

Background

Multi Modal Video understanding

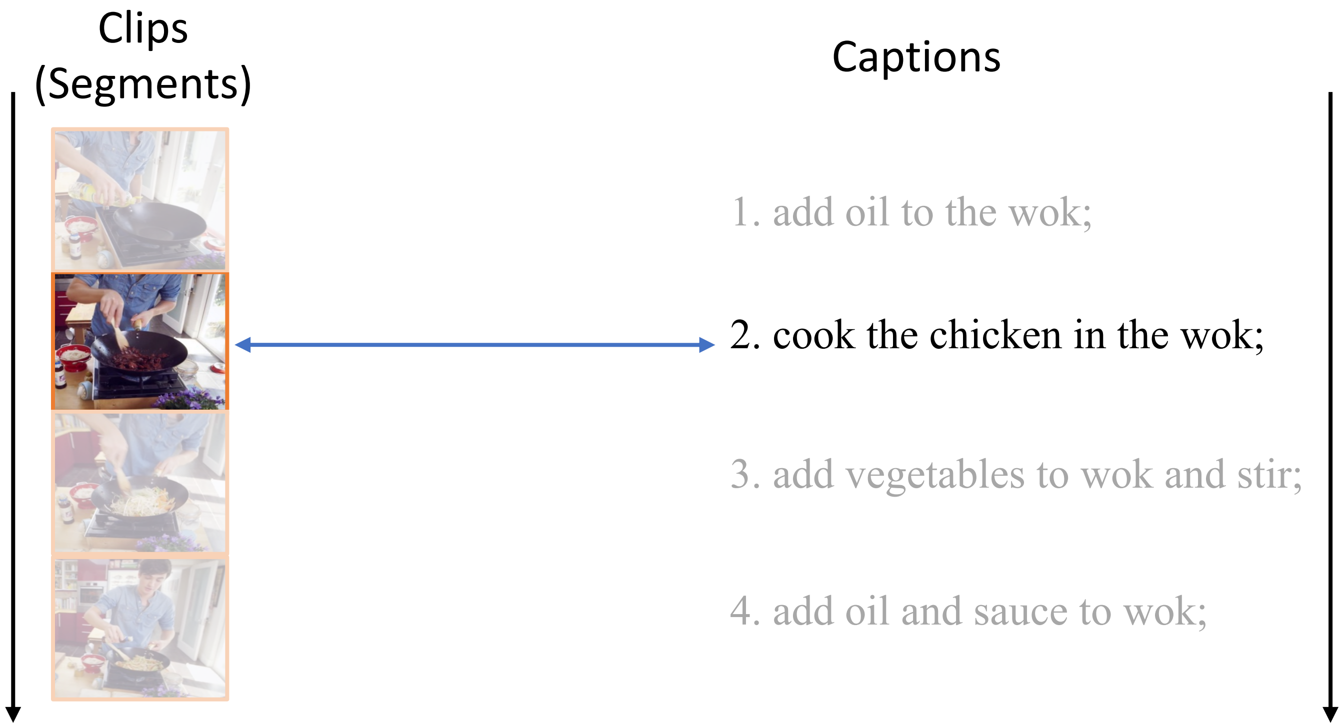
A long video contains multiple steps (i.e., segments) in succession and each step is described by a caption. Aligning all steps with captions is important for multi-modal video understanding.

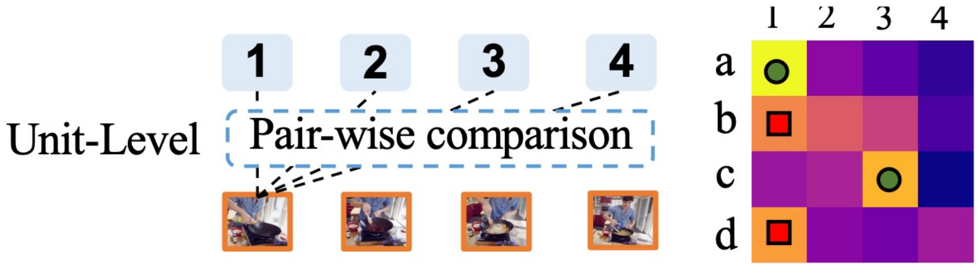


Baseline

Contrastive Clip-Caption Pretraining

To align the video with all captions, one intuitive approach is to compare the segments with captions directly, i.e., unit-level comparison





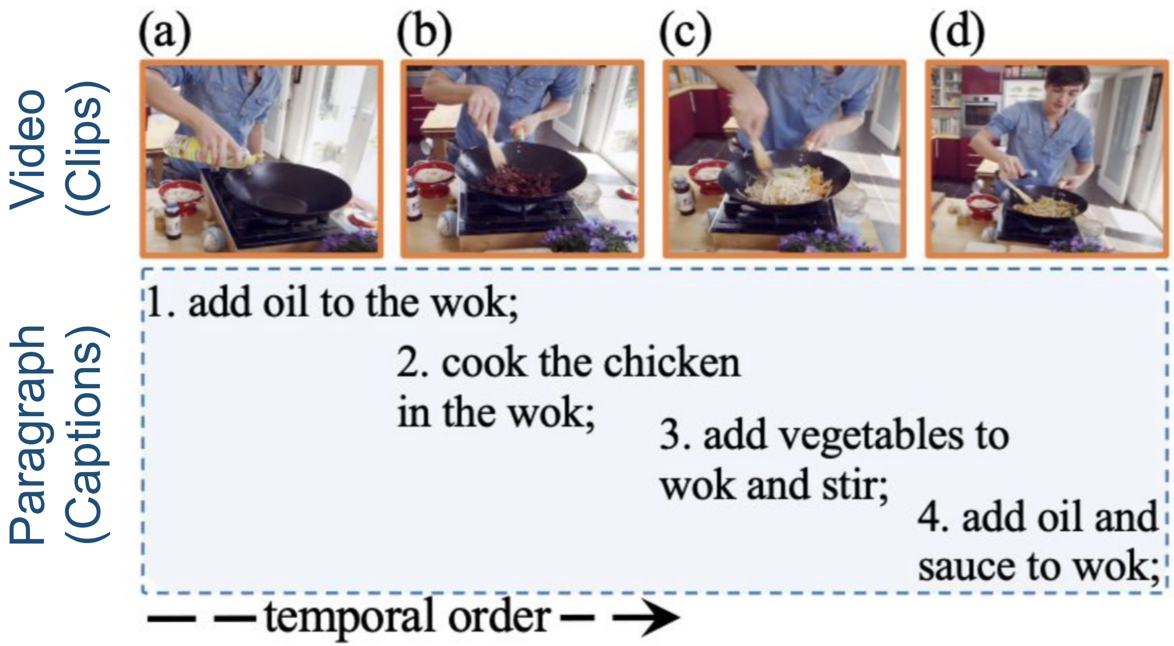
VideoCLIP: Contrastive Pre-training for Zero-shot Video-Text Understanding. [EMNLP 2021]
End-to-End Learning of Visual Representationsfrom Uncurated Instructional Videos. [CVPR 2020]

Observation

Exploring the Prior of Global Temporal Succession

A long video is naturally formulated as a sequence of short video clips.

A paragraph is then formulated as a sequence of sentences.

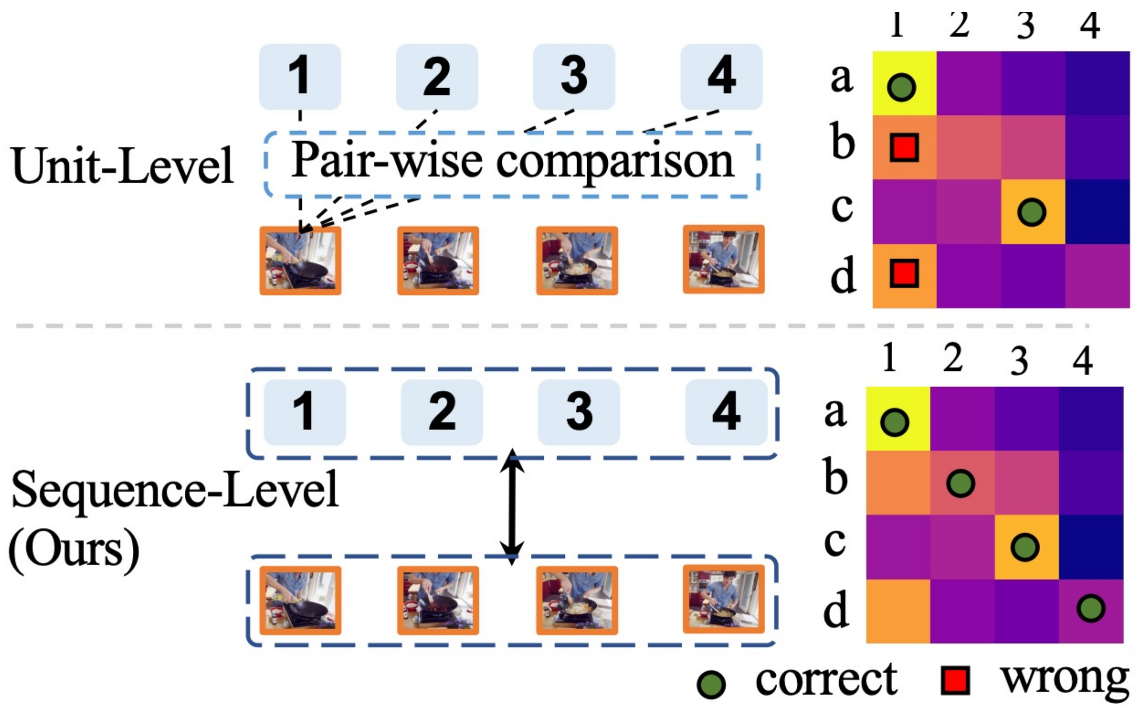


Note: A segment contains multiple consecutive clips.

Motivation

Alignment Video and Paragraph GLOBALLY

We propose to perform the sequence-level comparison between the two sequences, i.e., video and paragraph, and use dynamic time wrapping (or its variant) for distance calculation.

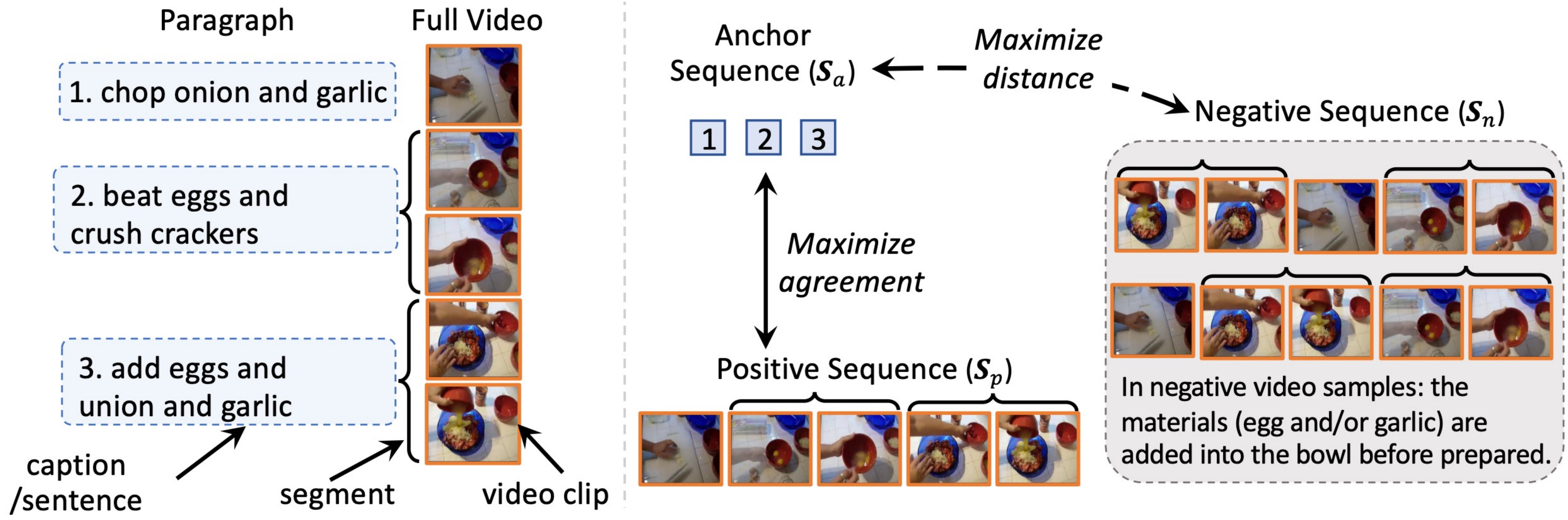


With the temporal order as prior, the confusion caused by visual similarity in unit-level comparison can be avoided

Approach

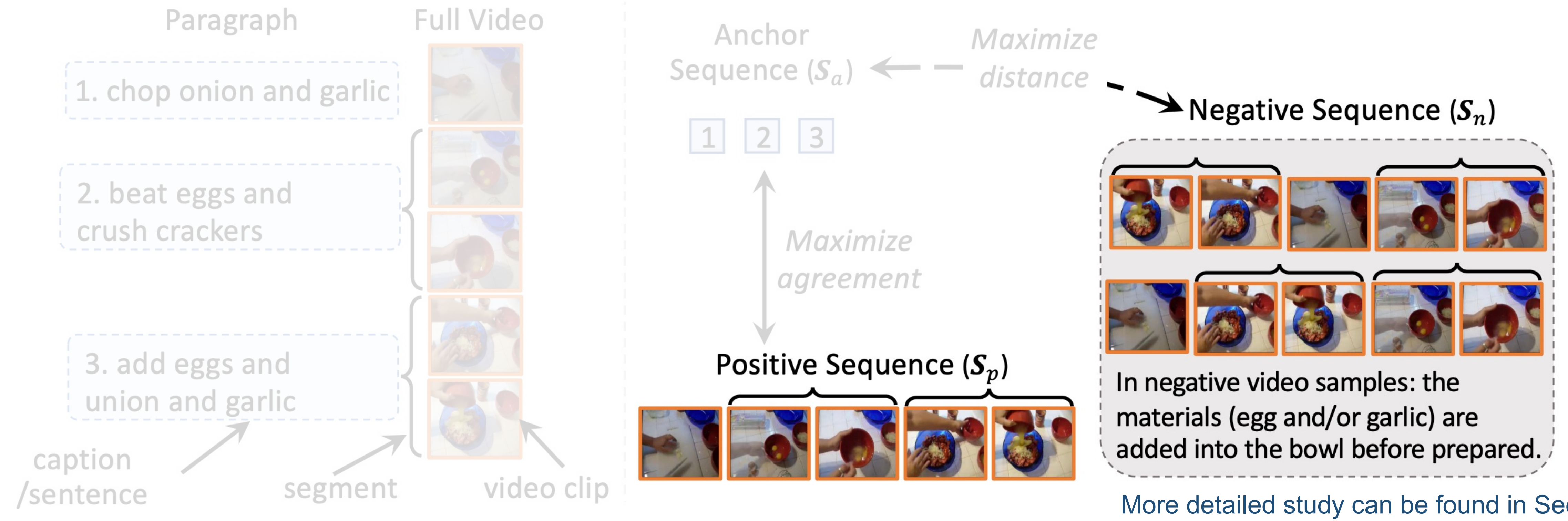
Contrastive Learning on Temporal Succession

Under contrastive learning framework, we use the *consistency of temporal succession* between two sequences to generate positive & negative samples.



Negative Shuffling with Temporal Granularity

To break the temporal consistency, we generate negative sequences by shuffling the clips in the positive sequence with respect to the temporal granularity.



More detailed study can be found in Section 5.2.

Experiments

Our approach achieves consistent performance gain on video retrieval, action step localization, and can be generalized on few-shot action recognition.

Video Retrieval on YouCookII

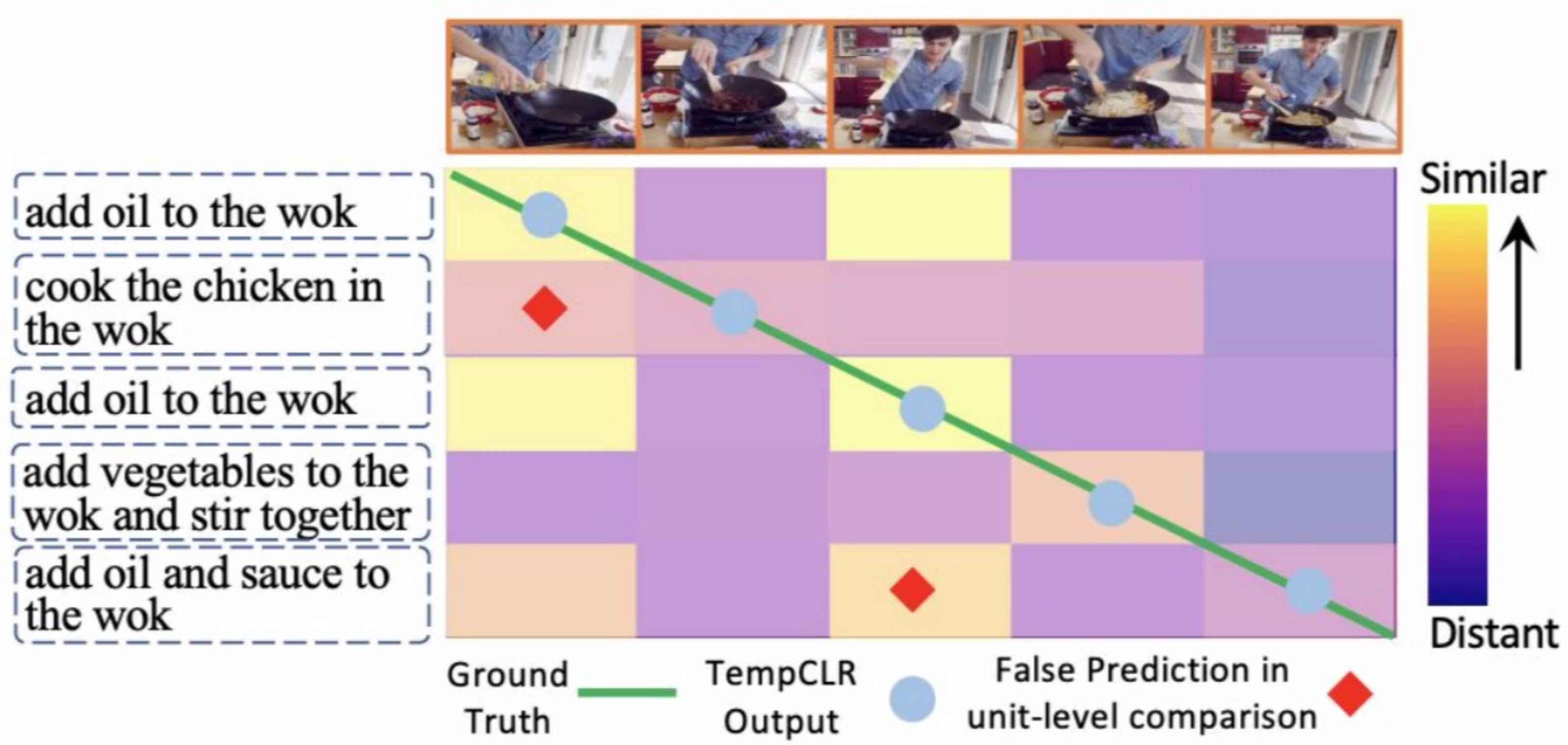
Exp.	*(Background Removed)*	Measure	R@1	R@5	R@10
1	MIL-NCE*	Cap. Avg.	43.1	68.6	79.1
2	HT100M*	Cap. Avg.	46.6	74.3	83.7
3	MCN (Chen et al., 2021)	Cap. Avg.	53.4	75.0	81.4
4	VideoCLIP†	Cap. Avg.	74.5	94.5	97.9
5	VideoCLIP†	DTW	56.0	89.9	96.3
6	TempCLR(Ours)	Cap. Avg.	74.5	94.6	97.0
7	TempCLR(Ours)	DTW	**83.5**	**97.2**	**99.3**
	(Background Kept)	Measure	R@1	R@5	R@10
8	VideoCLIP†	DTW	55.7	93.1	**98.9**
9	TempCLR	DTW	**70.4**	**93.8**	97.9

Action Step Localization on COIN

Approach (*Zero-shot*)	TFS	Recall
HT100M (Miech et al., 2019)	✓	33.6
MIL-NCE (Miech et al., 2020)	✓	40.5
MCN (Chen et al., 2021)	✓	35.1
DWSA (Shen et al., 2021)	✓	35.3
UniVL (Luo et al., 2020)	✓	42.0
VT-TWINS (Ko et al., 2022)	✓	40.7
VideoCLIP (Xu et al., 2021)	✓	33.9
VideoCLIP†		33.5 (↓ 0.4)
TempCLR (Ours)†		**36.9** (↑ **3.0**)

Visualization

With proper sequence-wise alignment, the sequence similarity can be improved. In addition, the aligned units between sequences are also semantically similar.



More details can be found in Section A.6

Analysis on Clip-Caption Matching

By contrasting paragraphs and videos during training, the gradients w.r.t. to each clip-caption pair is re-weighted by the context of the entire sequence.

Positive Pair

$$\frac{\partial \mathcal{L}_{VideoCLIP}}{\partial (m_1 \cdot n_1^T)} \Big/ \frac{\partial \mathcal{L}_{TempCLR}}{\partial (m_1 \cdot n_1^T)} = \frac{(e^{m_1 \cdot n_1^T} + e^{m_1 \cdot n_2^T})^{-1}}{(e^{m_1 \cdot n_1^T} e^{m_2 \cdot n_2^T - m_2 \cdot n_1^T} + e^{m_1 \cdot n_2^T})^{-1}}$$

Negative Pair

$$\frac{\partial \mathcal{L}_{VideoCLIP}}{\partial (m_1 \cdot n_2^T)} \Big/ \frac{\partial \mathcal{L}_{TempCLR}}{\partial (m_1 \cdot n_2^T)} = \frac{(e^{m_1 \cdot n_1^T - m_1 \cdot n_2^T} + 1)^{-1}}{(e^{m_1 \cdot n_1^T - m_1 \cdot n_2^T + m_2 \cdot n_2^T - m_2 \cdot n_1^T} + 1)^{-1}}$$

More details can be found in Section A.5

TempCLR: Temporal Alignment Representation with Contrastive Learning

Contact: {yy3035, jiawei.m}@columbia.edu

GitHub: https://github.com/yyuncong/TempCLR

International Conference On Learning Representations

COLUMBIA | ENGINEERING
The Fu Foundation School of Engineering and Applied Science