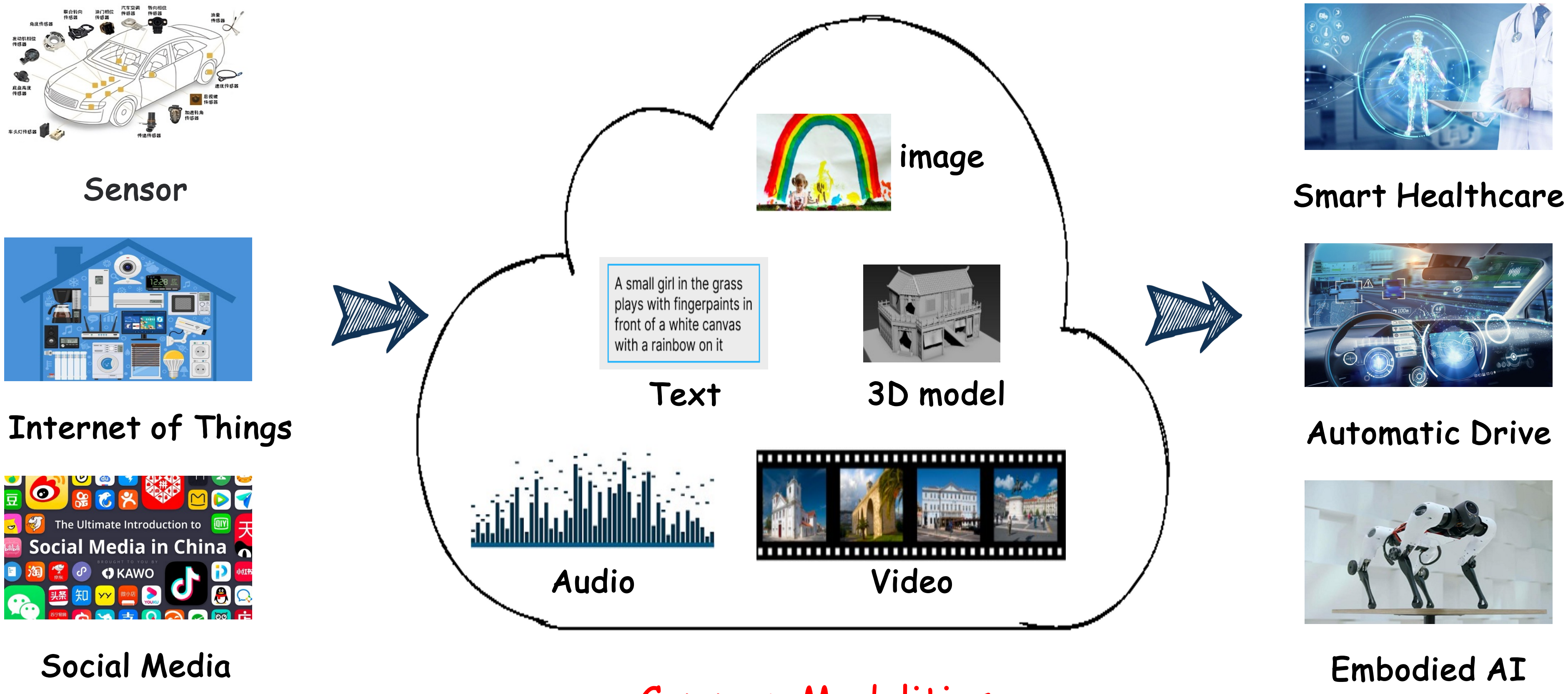
Noisy Correspondence Learning

Multi-granularity Correspondence Learning from Long-term Noisy Videos

Yijie Lin, Jie Zhang, Zhenyu Huang, Jia Liu, Zujie Wen, Xi Peng

Background

With the evolution of sensors, the popularization of smart devices, and the rise of the internet and social media, multi-modal data is showing a rapidly growing trend.

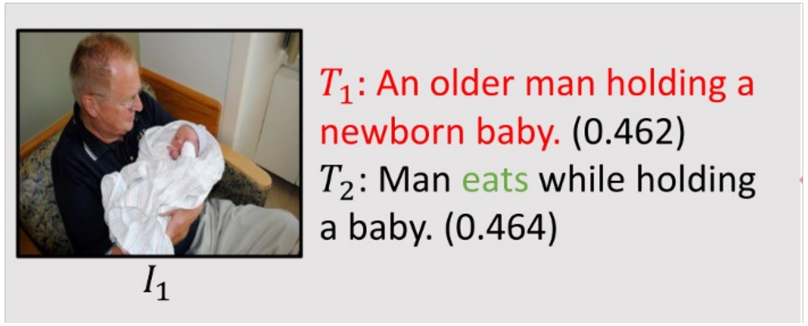


Common Modalities

Background

Traditionally, most machine learning methods aim to build or use the many-to-many or one-to-one correspondence.

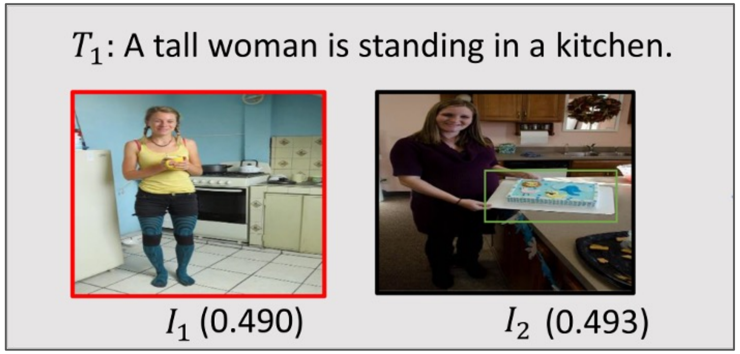
Cross-modal Retrieval



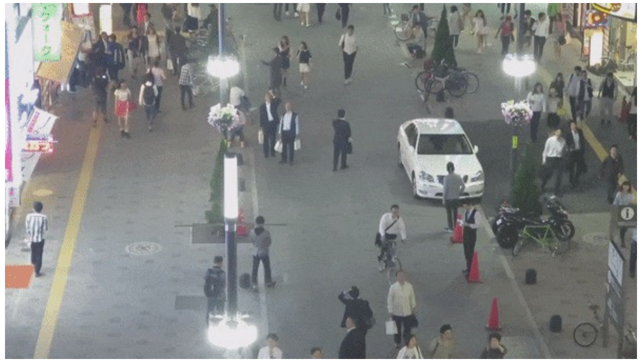
Visual Grounding

- The blue truck in the bottom right corner
- The light blue truck
- The blue truck on the right

Cross-modal Generation



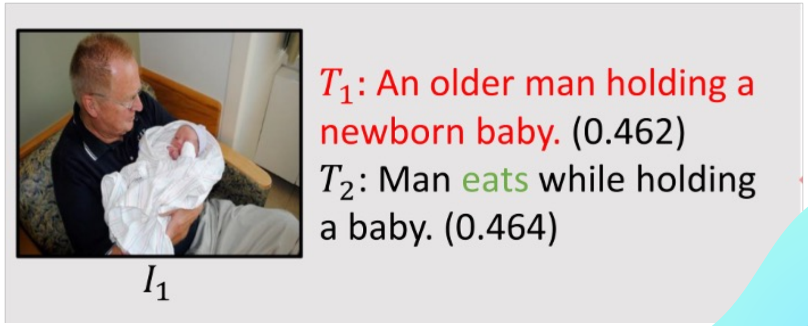
Tracking



Background

These methods heavily rely on the well-established data correspondence!

Cross-modal Retrieval

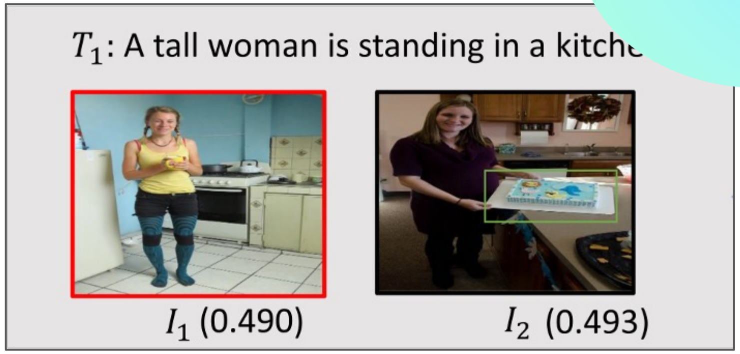


Visual Grounding

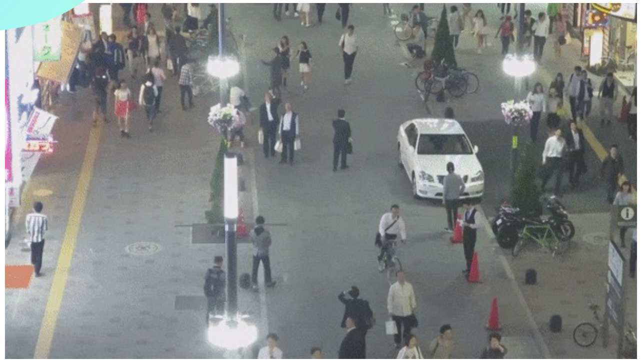
- The blue truck in the bottom right corner
- The light blue truck
- The blue truck on the right

Object/ Sample/ Modality correspondence

Cross-modal Generati



Tracking



Motivation

However, it is impractical to assume that the correspondence is well-established. Instead, noisy pairs are common in the real world.

Read between the lines , and your dream about person will be clear

A large crowd turned out for show .

There is no need to be sad.

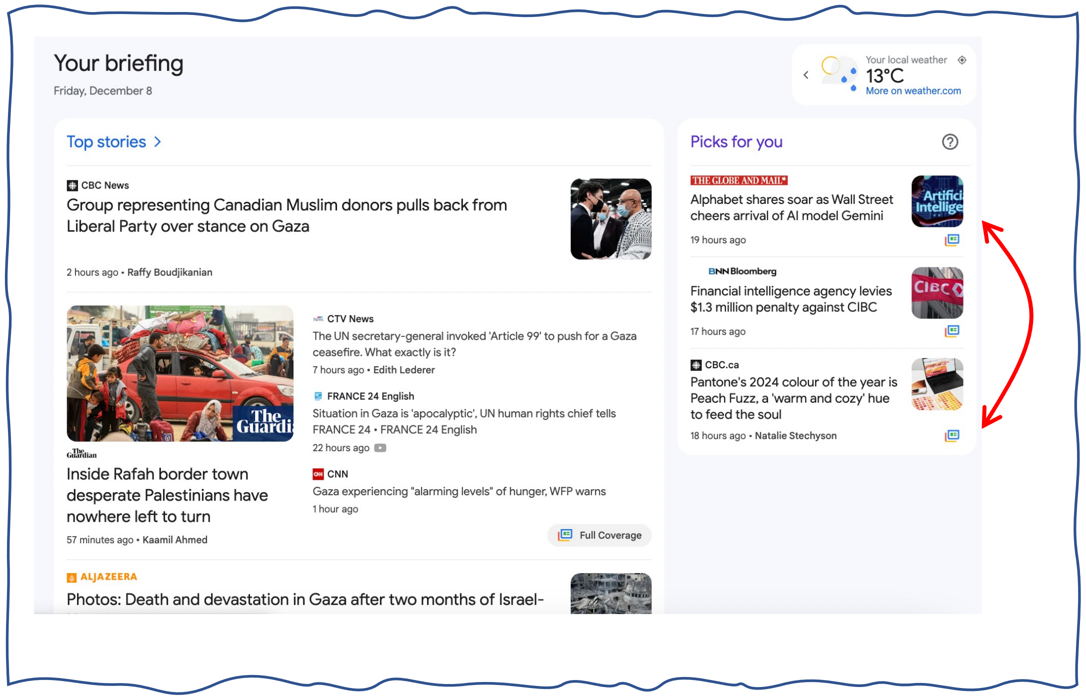
See pictures of first home .

WRONGLY matched image-text pairs from Conceptual Captions dataset[1]

Ref:
1. Conceptual Captions: A Cleaned, Hypernymed, Image Alt-text Dataset For Automatic Image Captioning, ACL 2018

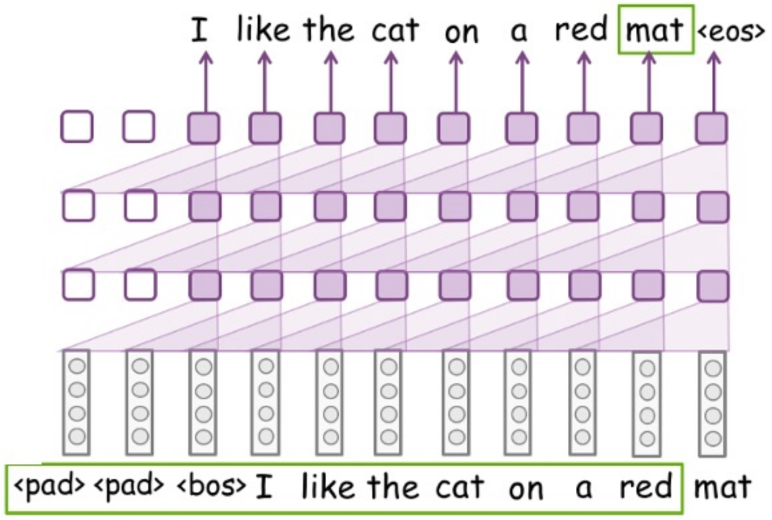
Motivation

Noisy pairs also emerge within language corpus, impacting the next token prediction (*i.e.*, learning the context) in training large language models.



Uncorrelated context crawling from Google news

Predict next token

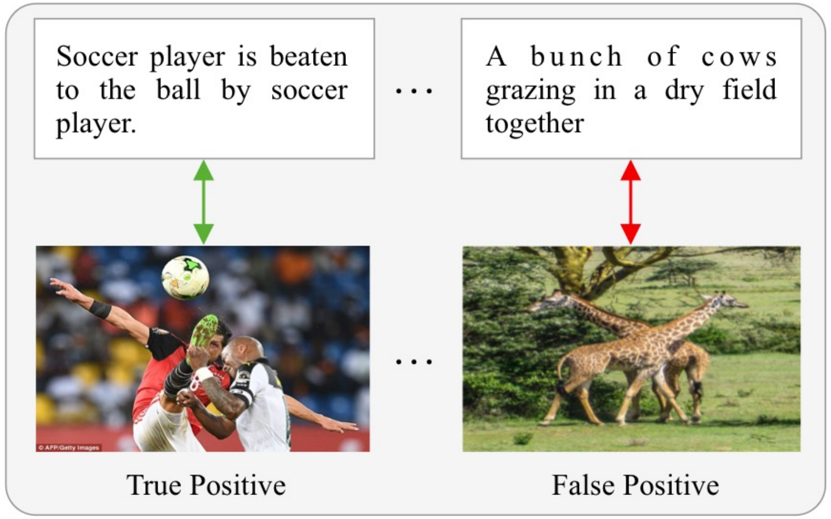


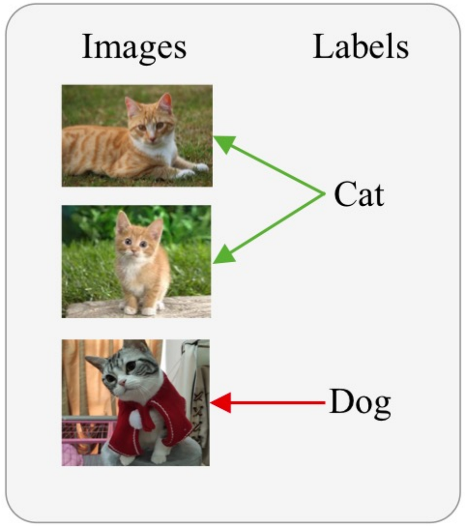
condition on the previous tokens

Motivation

For the first time[1], we reveal the existence and influence of Noisy Correspondence (NC) in a number of applications.

Noisy Correspondence **vs** Noisy Label





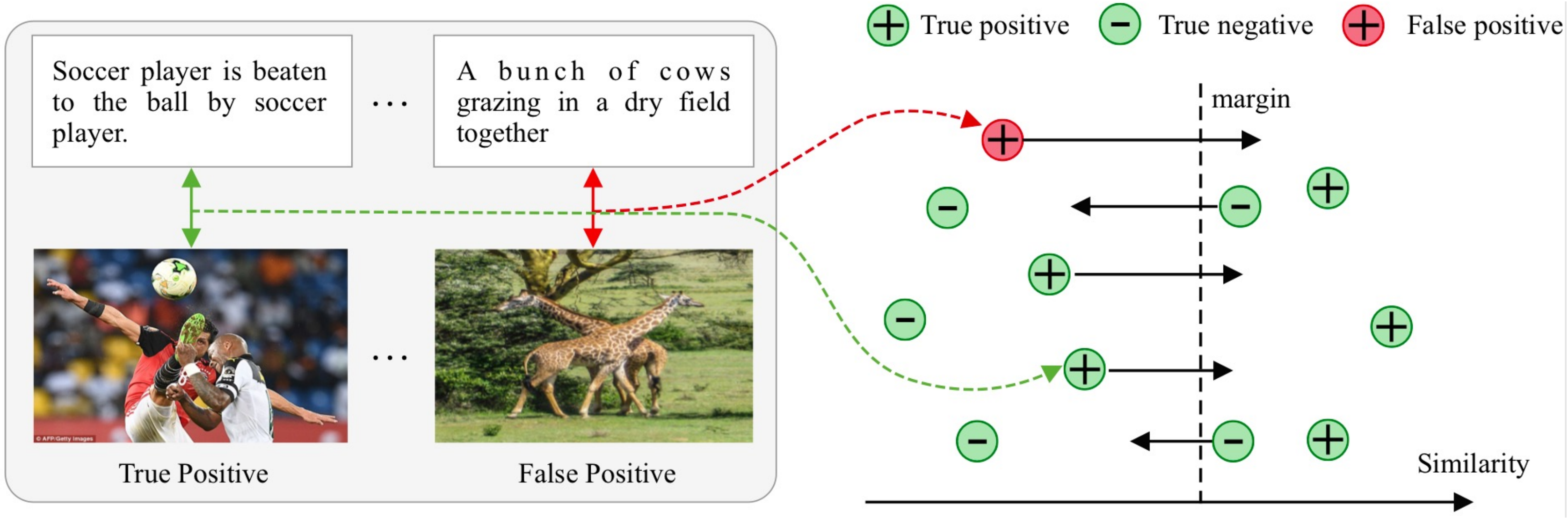
NC refers to the alignment errors in paired data rather than the errors in category annotations

Ref:

1. Learning with Noisy Correspondence for Cross-modal Matching, NeurIPS 2021. (Oral)

Motivation

We show that, Noisy Correspondence will degrade the performance of various tasks including but not limited to Cross-modal Matching, Object ReID, Question Answering, Machine Reading Comprehension, etc.



An example: noisy correspondence in cross-modal matching task

Taxonomy

Multi-modal/view tasks

Different type of Noisy correspondence

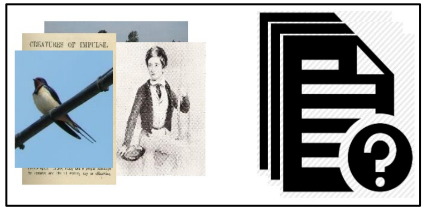
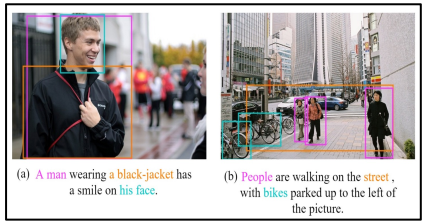
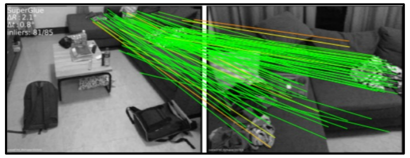


Image-text retrieval

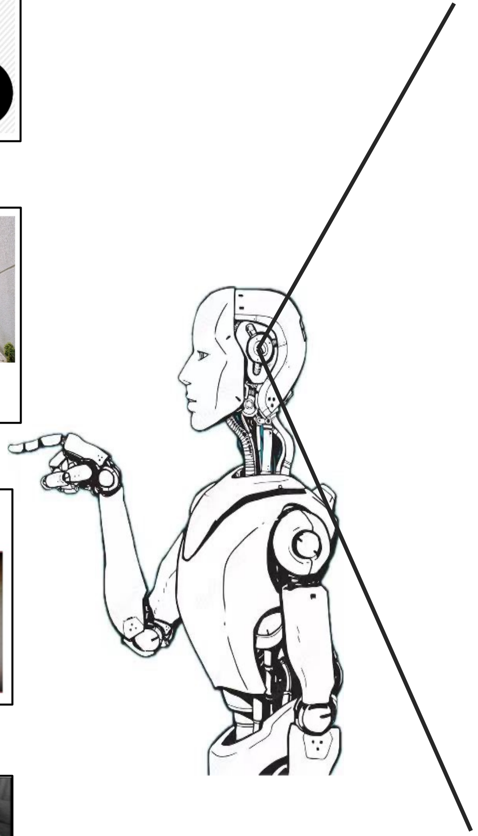


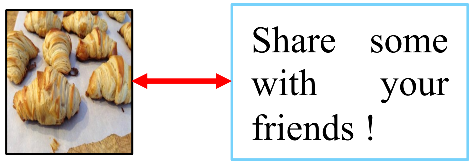
Visual Grounding

VQA

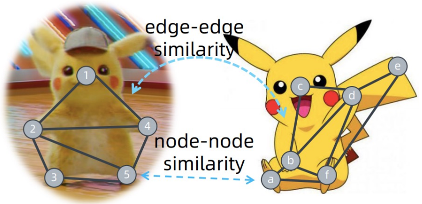


Dense Correspondence

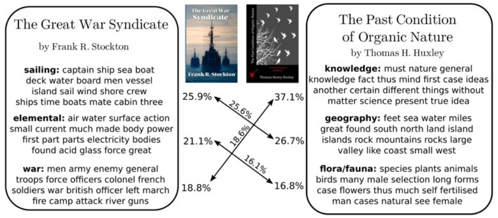




Pair-wise
Noisy Correspondence



Multi-level
Noisy Correspondence



Multi-granularity
Noisy Correspondence

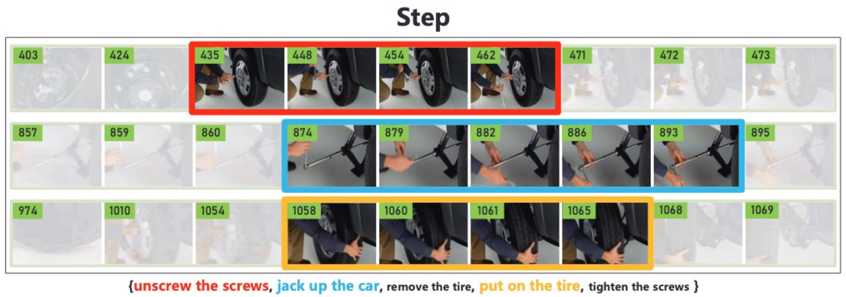
Multi-granularity Correspondence Learning from Long-term Noisy Videos

Background

Video-Language Pre-training (VLP) has emerged as a popular foundation for video understanding.

Video QA

Video Retrieval



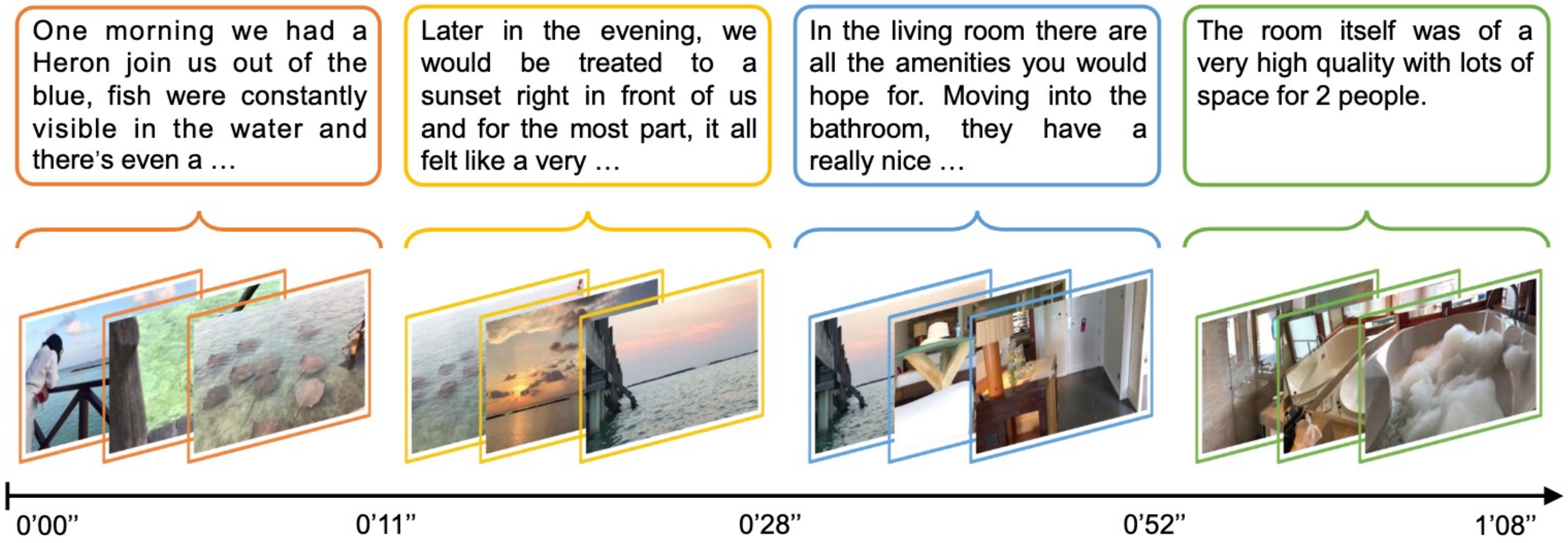
Action Segmentation

Video Classification

Background

Long-term temporal dependency in video plays an indispensable role in understanding the relationships and transitions over time.

However, the modeling of long videos entails an over-high computational cost, constraining this challenging problem rarely explored.



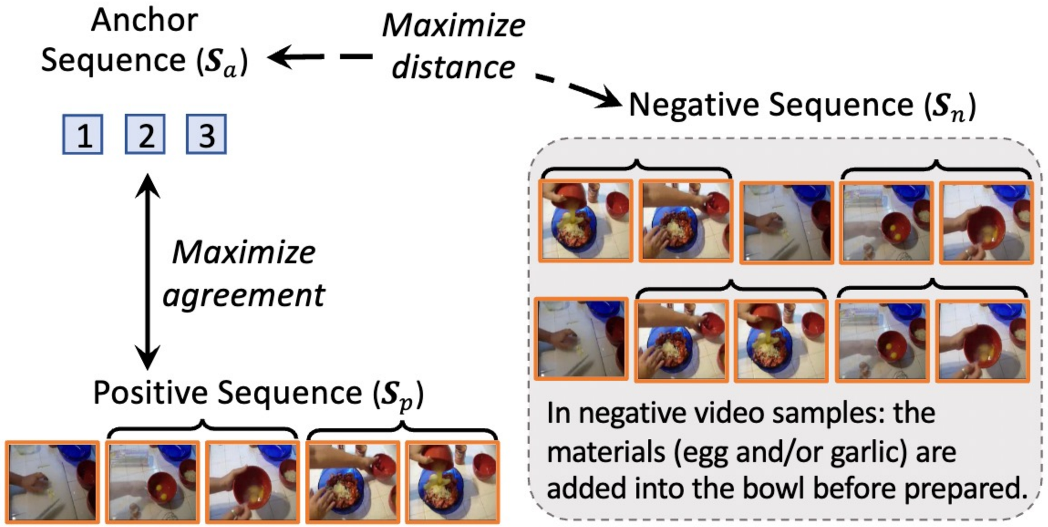
Long-term dependency in Video Learning[1]

Ref:
1. Long-Form Video-Language Pre-Training with Multimodal Temporal Contrastive Learning, NeurIPS 2022

Background

As long videos are typically composed of a sequence of short video clips according to ASR timestamps, an alternative approach is to explore the temporal correlation among video clips and captions.

TempCLR[1] uses Dynamic Time Warping to measure the sequential distance in a late fusion manner.



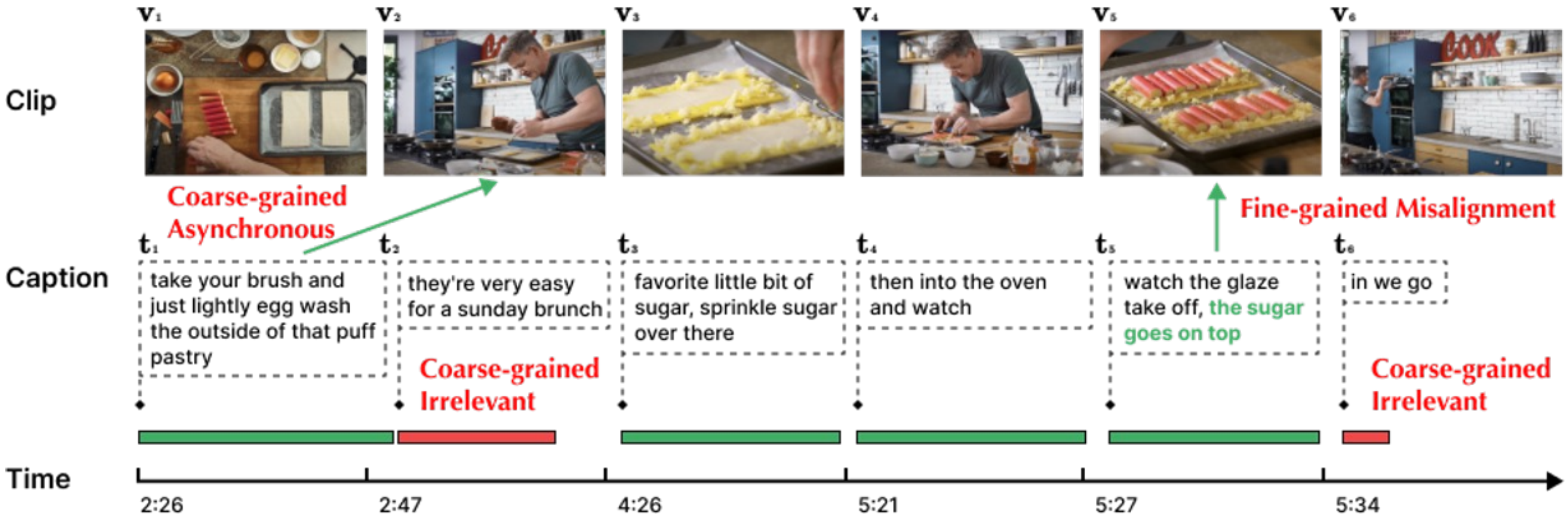
TempCLR

Ref:
1. TempCLR: Temporal Alignment Representation with Contrastive Learning, ICLR 2023

Observation & Motivations

Dividing long videos into short clips would introduce multi-granularity noisy correspondence (MNC) challenge.

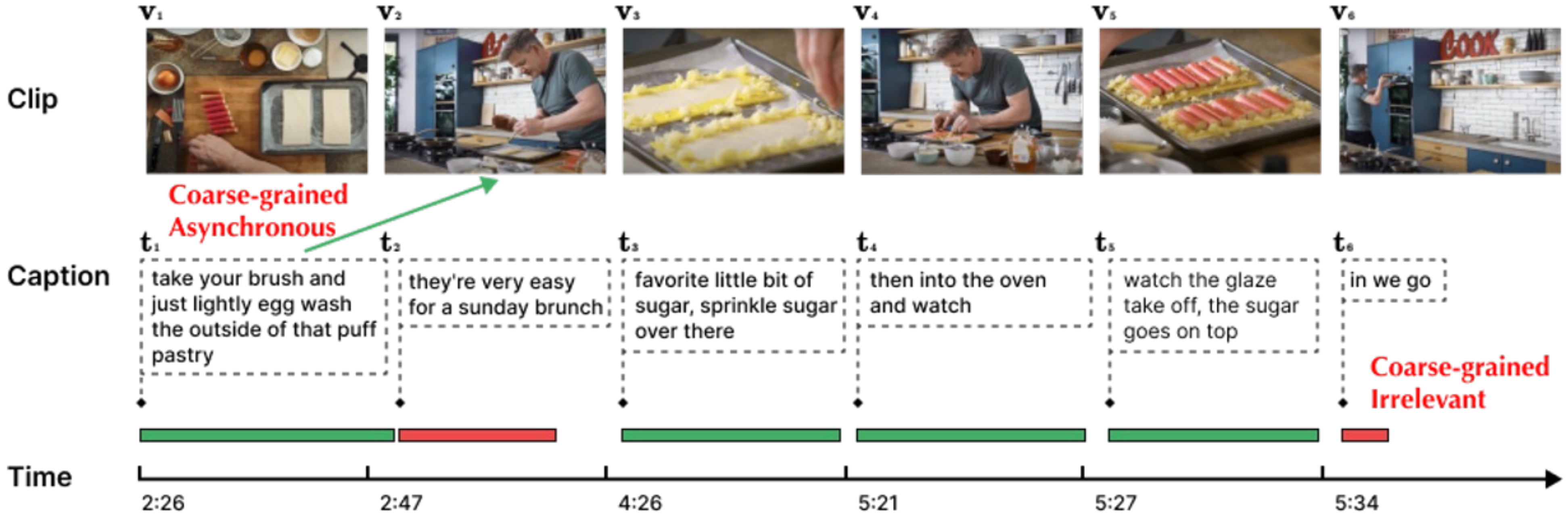
- Coarse-grained misalignment (Clip-caption).
- Fine-grained misalignment (Frame-word)



Observation & Motivations

Coarse-grained Noisy Correspondence (Clip-caption)

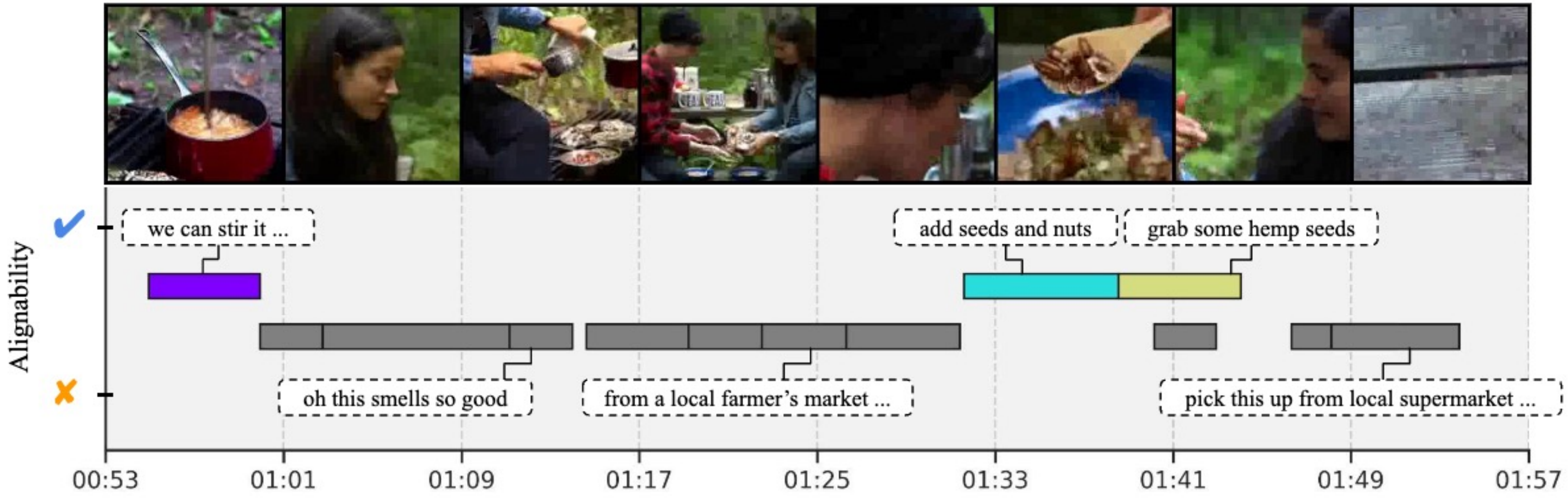
- Asynchronous misalignment refers to temporal misalignment between subtitles and visual clips. It often occurs when people explain their actions before or after actually performing them.
- Irrelevant misalignment refers to irrelevant or meaningless captions that cannot be aligned with any available video, and vice versa for video clips.



Observation & Motivations

Coarse-grained Noisy Correspondence (Clip-caption)

- According to Han et al. (2022)[1], only 30% of clip-caption pairs are visually aligned in HowTo100M, with even fewer 15% being naturally well-aligned;



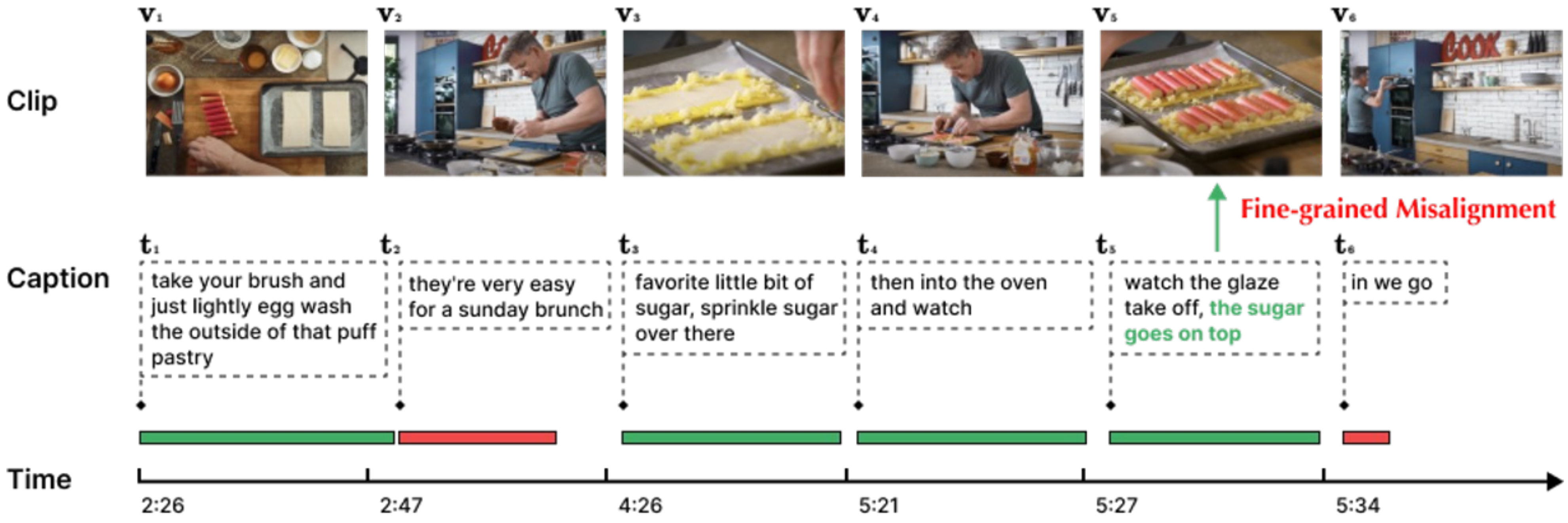
Ref:

1. Temporal Alignment Networks for Long-term Video, CVPR 2022 (Oral)

Observation & Motivations

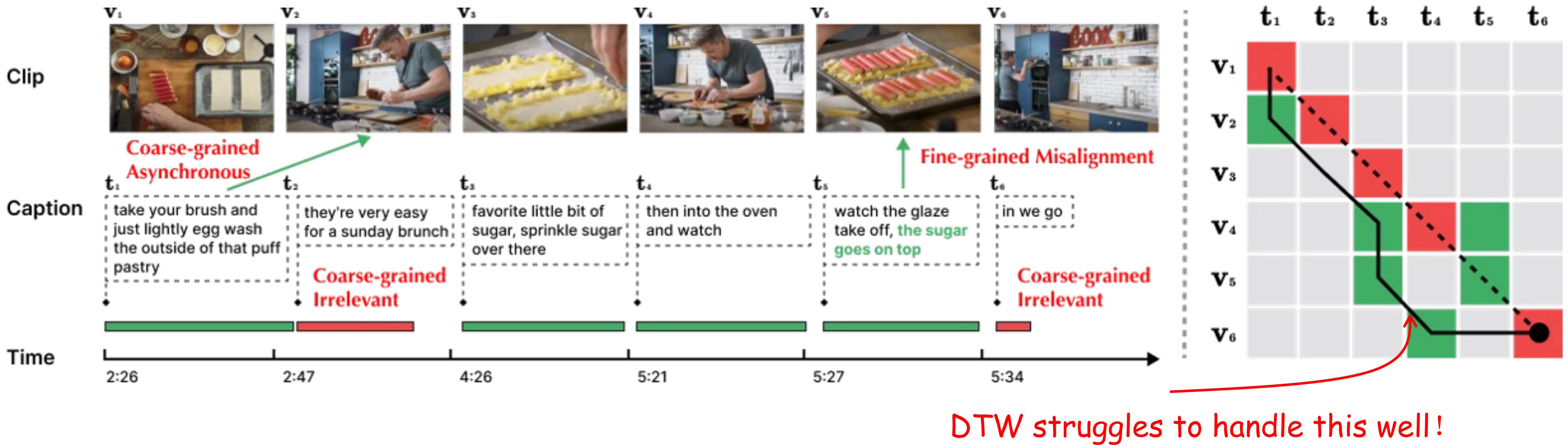
Fine-grained Noisy Correspondence (Frame-word)

- Within each video clip, the narration sentences may only partially correlate with the visual frames.
- Irrelevant words or frames can distort the identification of crucial ones and result in inaccurate similarity measurements, further contaminating the clip-caption alignment.



Observation & Motivations

Dividing long videos into short clips would introduce multi-granularity noisy correspondence (MNC) challenge.



Observation & Motivations

Challenge 1: Directly modeling long videos entails heavy computation demands

Align between short clips and captions

Challenge 2: Multi-granularity noisy correspondence (~~DTW-based method~~)

Unified Optimal Transport Solution

Method

Unified Optimal Transport Solution

$$\mathcal{L} = \mathcal{L}_{\text{clip}} + \lambda \mathcal{L}_{\text{video}},$$

- Video-paragraph contrastive loss (video-level)

  unifies the multi-granularity learning in a fine to coarse perspective through a noise-robust temporal optimal transport distance.

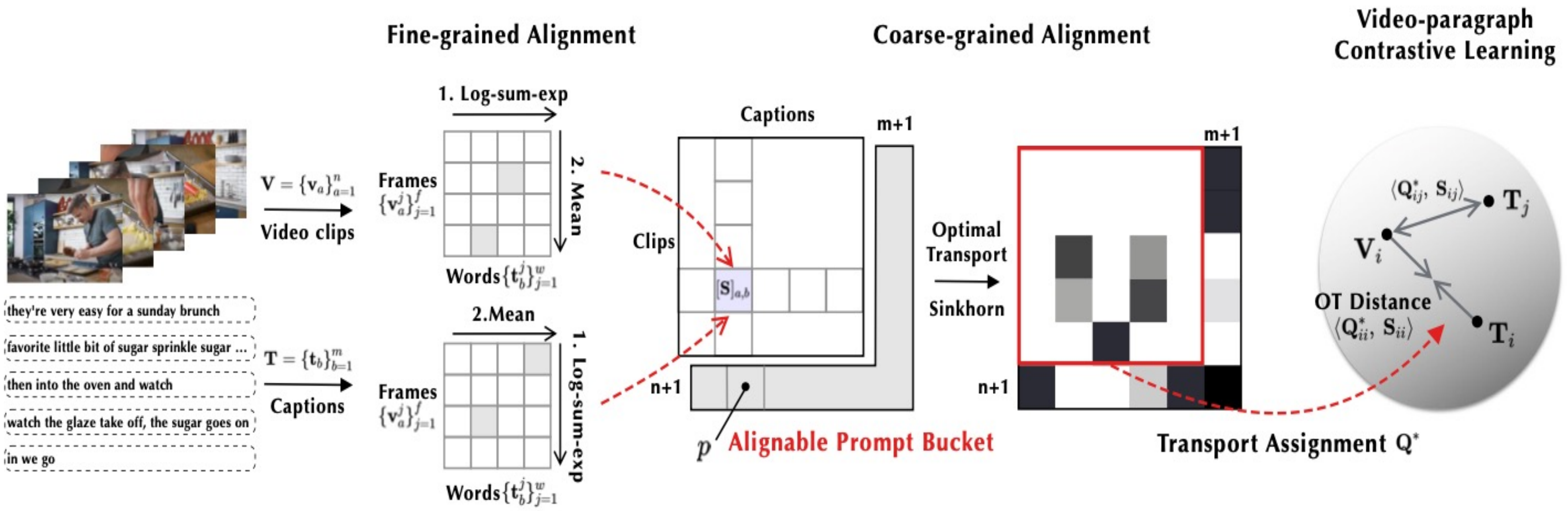
- Clip-caption contrastive loss (clip-level)

  exploits potential false negative pairs (pair-wise NC) to improve clip representation and ensure accurate temporal modeling.

Method

Multi-granularity correspondence learning (Video level)

From fine-to-coarse

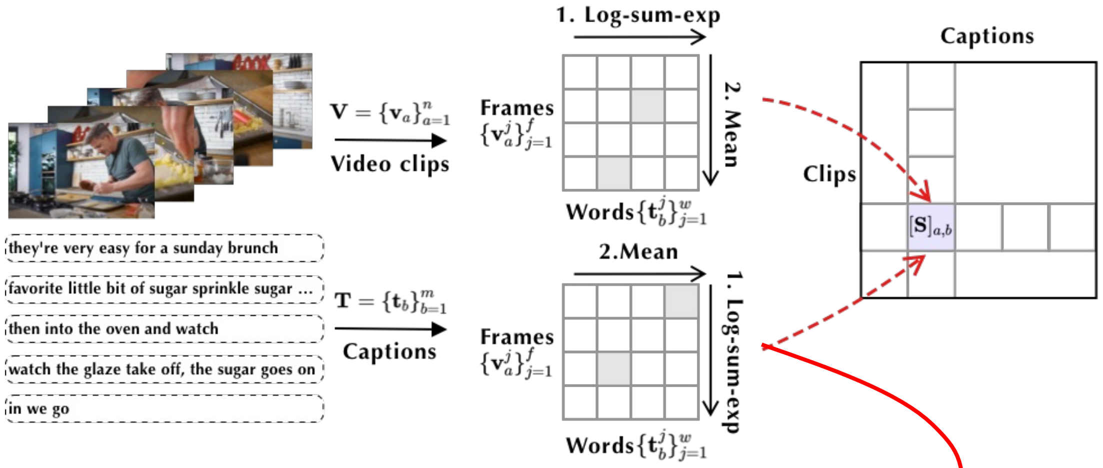


Video-paragraph contrastive learning captures long-term temporal correlations from a fine-to-coarse perspective.

Method

Fine-grained Alignment – Soft-maximum Operation

- Identify the most important word/frame by log-sum-exp approximation in a late interactive manner
- Average soft-maximum similarities of all frames/words as clip-caption level similarity
- α controls the importance



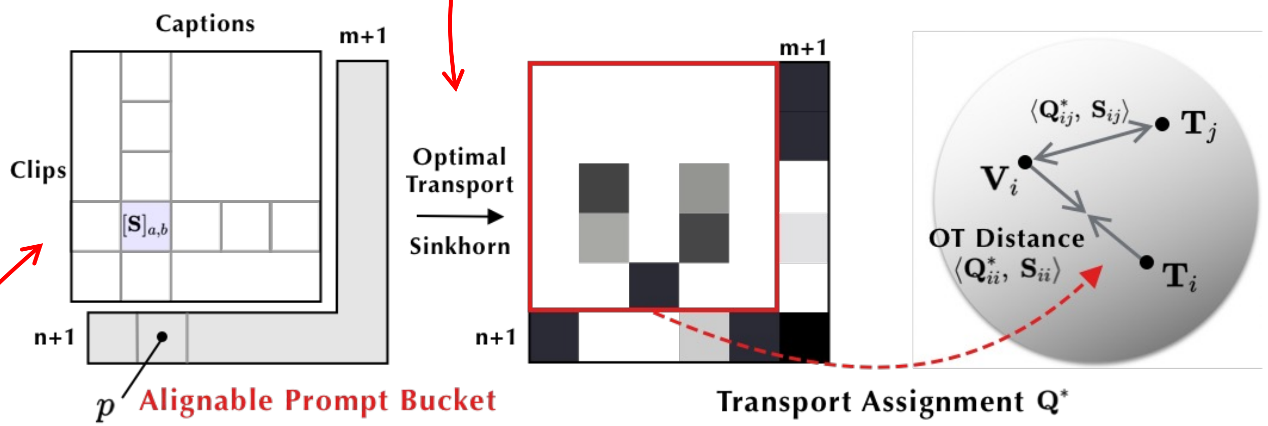
$$[\mathbf{S}]_{a,b} = \frac{1}{2}\left(\frac{1}{f}\sum_{i=1}^{f}\alpha\log\left(\sum_{j=1}^{w}\exp\left(\frac{\mathbf{v}_a^i\cdot\mathbf{t}_b^j}{\alpha}\right)\right) + \frac{1}{w}\sum_{i=1}^{w}\alpha\log\left(\sum_{j=1}^{f}\exp\left(\frac{\mathbf{t}_b^i\cdot\mathbf{v}_a^j}{\alpha}\right)\right)\right)$$

Method

Coarse-grained Alignment – Alignable Prompt Bucket on Optimal Transport

- Optimal transport naturally addressing asynchronous and one-to-many alignment
- Alignable prompt bucket filters irrelevant clips/captions, serving as a similarity margin that distinguishes between alignable and unalignable clips and captions
- Seamlessly integrated in Sinkhorn iterations

$$\max_{\mathbf{Q}\in\mathcal{Q}} \quad \langle \mathbf{Q},\, \mathbf{S}\rangle + \varepsilon H(\mathbf{Q})$$
$$\text{s.t.} \quad \mathcal{Q} = \left\{\mathbf{Q}\in\mathbb{R}_{+}^{n\times m} \mid \mathbf{Q}\mathbf{1}_m = \boldsymbol{\mu}, \mathbf{Q}^{\top}\mathbf{1}_n = \boldsymbol{\nu}\right\}.$$



$$[\bar{\mathbf{S}}]_{a,m+1} = [\bar{\mathbf{S}}]_{n+1,b} = [\bar{\mathbf{S}}]_{n+1,m+1} = p,$$
$$[\bar{\mathbf{S}}]_{a,b} = [\mathbf{S}]_{a,b},\ \forall a \in [1,n],\ b \in [1,m].$$

Method

Coarse-grained Alignment – Alignable Prompt Bucket on Optimal Transport

- Sinkhorn iterations

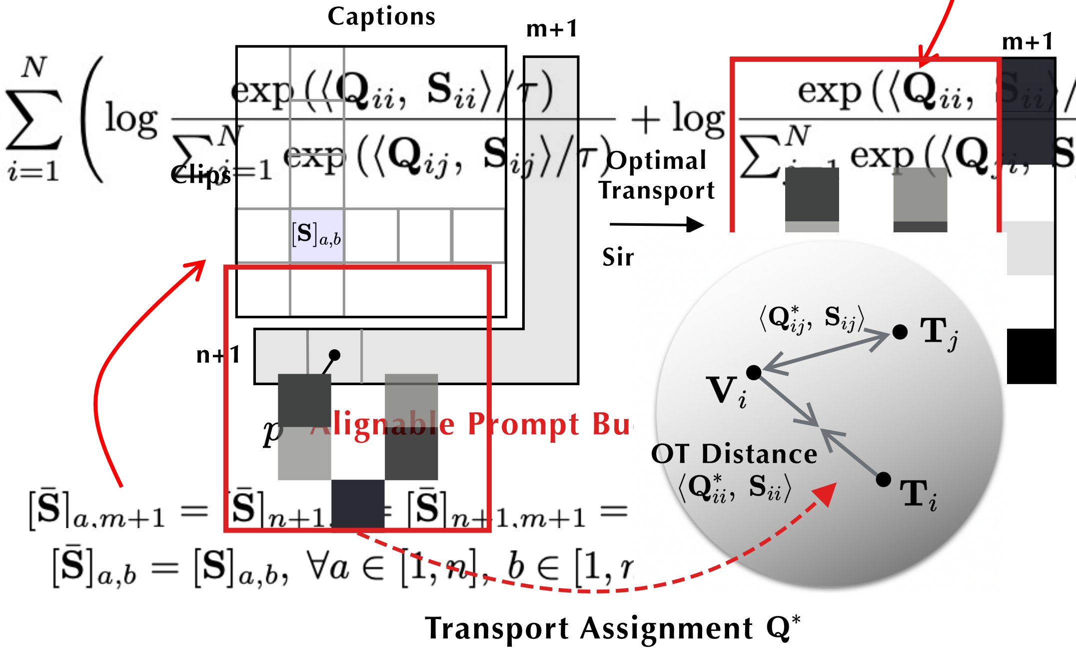
$$\max_{\mathbf{Q}\in\mathcal{Q}} \quad \langle \mathbf{Q},\, \mathbf{S}\rangle + \varepsilon H(\mathbf{Q})$$
$$\text{s.t.} \quad \mathcal{Q} = \left\{\mathbf{Q}\in\mathbb{R}_{+}^{n\times m} \mid \mathbf{Q}\mathbf{1}_m = \boldsymbol{\mu}, \mathbf{Q}^{\top}\mathbf{1}_n = \boldsymbol{\nu}\right\}.$$

$$\mathbf{Q}^{*} = \mathrm{Diag}(\boldsymbol{\kappa}_1)\exp\left(\mathbf{S}/\varepsilon\right)\mathrm{Diag}(\boldsymbol{\kappa}_2),$$
with iteratively updated $\boldsymbol{\kappa}_1 \leftarrow \boldsymbol{\mu}./\left(\exp\left(\mathbf{S}/\varepsilon\right)\boldsymbol{\kappa}_2\right),\ \boldsymbol{\kappa}_2 \leftarrow \boldsymbol{\nu}./\left(\exp\left(\mathbf{S}^{\top}/\varepsilon\right)\boldsymbol{\kappa}_1\right),$

$$\mathbf{Q}^{*} = \mathbf{Q}^{*}_{1:n,1:m}$$

- Video-paragraph contrastive loss

$$\mathcal{L}_{\text{video}} = -\sum_{i=1}^{N}\left(\log\frac{\exp\left(\langle \mathbf{Q}_{ii},\, \mathbf{S}_{ii}\rangle/\tau\right)}{\sum_{j=1}^{N}\exp\left(\langle \mathbf{Q}_{ij},\, \mathbf{S}_{ij}\rangle/\tau\right)} + \log\frac{\exp\left(\langle \mathbf{Q}_{ii},\, \mathbf{S}_{ii}\rangle/\tau\right)}{\sum_{j=1}^{N}\exp\left(\langle \mathbf{Q}_{ji},\, \mathbf{S}_{ji}\rangle/\tau\right)}\right),$$



Method

Faulty Negative Exploitation (Clip-level)

pairs of similar semantic are
WRONGLY regraded as negative

- Identify sematic within-batch clip-caption similarity matrix through optimal transport

$$\max_{\hat{\mathbf{Q}}\in\hat{\mathcal{Q}}} \quad \langle \hat{\mathbf{Q}},\, \hat{\mathbf{S}} \rangle + \varepsilon H(\hat{\mathbf{Q}}) \quad \text{s.t.} \;\; \hat{\mathcal{Q}} = \left\{ \hat{\mathbf{Q}} \in \mathbb{R}_{+}^{B\times B} \mid \hat{\mathbf{Q}}\mathbf{1}_B = \frac{1}{B}\mathbf{1}_B, \hat{\mathbf{Q}}^{\top}\mathbf{1}_B = \frac{1}{B}\mathbf{1}_B \right\},$$

- Rectify the one-hot target **T** of clip-caption contrastive loss based on the transport assignment

$$\mathcal{L}_{\text{clip}} = -\sum_{i=1}^{B}\sum_{j=1}^{B} [\mathbf{T}]_{i,j} \left(\log \frac{\exp\left([\hat{\mathbf{S}}]_{i,j}/\tau\right)}{\sum_{k=1}^{B}\exp\left([\hat{\mathbf{S}}]_{i,k}/\tau\right)} + \log \frac{\exp\left([\hat{\mathbf{S}}]_{i,j}/\tau\right)}{\sum_{k=1}^{B}\exp\left([\hat{\mathbf{S}}]_{k,j}/\tau\right)} \right), \mathbf{T} = (1-\beta)\,\mathbf{I}_B + \beta\hat{\mathbf{Q}}^{*}$$

Experiments

Task1 Long video Retrieval – YoucookII

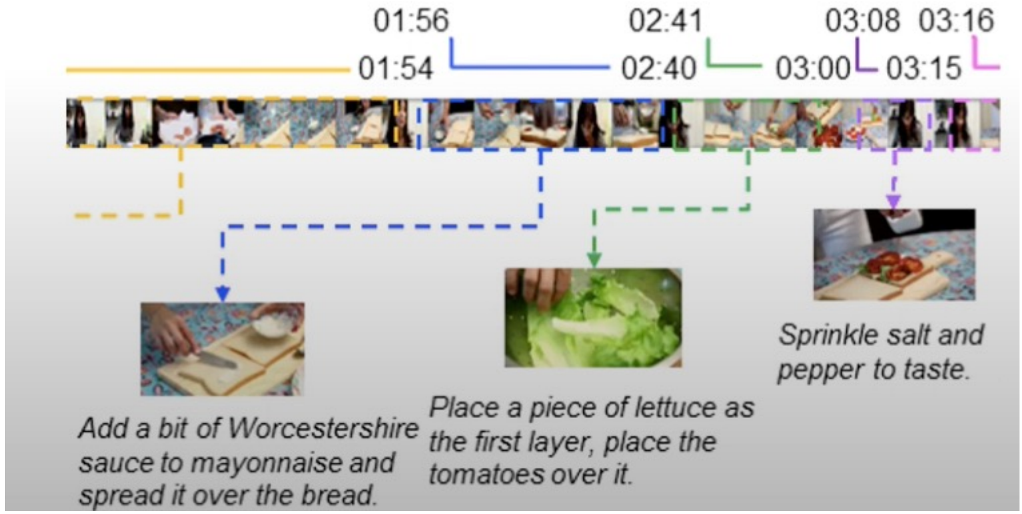
- **Cap. Avg.** matches one clip for each caption and retrieves the video with the most matched clips.
- **DTW and OTAM** calculate the sequence distance by accumulating the clip-caption distance based on chronological order.

Table 1: Video-paragraph retrieval on YouCookII (*Background Removed*). The best and second-best results are **bold** and underlined, respectively.

Approach	Measure	R@1	R@5	R@10
MIL-NCE (Miech et al., 2020)	Cap. Avg.	43.1	68.6	79.1
HT100M (Miech et al., 2019)	Cap. Avg.	46.6	74.3	83.7
MCN (Chen et al., 2021)	Cap. Avg.	53.4	75.0	81.4
VideoCLIP (Xu et al., 2021)	Cap. Avg.	74.5	94.5	**97.9**
TempCLR (Yang et al., 2023b)	Cap. Avg.	74.5	94.6	97.0
Norton (Ours)	Cap. Avg.	**75.5**	**95.0**	97.7
VideoCLIP (Xu et al., 2021)	DTW	56.0	89.9	96.3
TempCLR (Yang et al., 2023b)	DTW	83.5	97.2	99.3
Norton (Ours)	DTW	**88.7**	**98.8**	**99.5**
VideoCLIP (Xu et al., 2021)	OTAM	52.8	89.2	95.0
TempCLR (Yang et al., 2023b)	OTAM	84.9	97.9	99.3
Norton (Ours)	OTAM	**88.9**	**98.4**	**99.5**

Table 2: Video-paragraph retrieval on YouCookII (*Background Kept*).

Approach	R@1	R@5	R@10
Cap. Avg.			
VideoCLIP	73.6	**94.7**	**98.4**
TempCLR	71.7	94.5	97.9
Norton (Ours)	**74.8**	**94.7**	**98.4**
DTW			
VideoCLIP	55.7	93.1	**98.9**
TempCLR	70.4	93.8	97.9
Norton (Ours)	**76.1**	**95.0**	98.4
OTAM			
VideoCLIP	56.6	92.8	**98.9**
TempCLR	72.2	94.5	97.7
Norton (Ours)	**73.6**	**94.7**	97.7



Experiments

Task2 Various Downstream Tasks

- Text-to-video Retrieval: YoucookII, MSR-VTT
- Action Segmentation: COIN
- Video QA: MSR-VTT

Table 5: Text-to-video retrieval on MSR-VTT.

Superivsed	R@1	R@5	R@10
SupportSet (Patrick et al., 2021)	30.1	58.5	69.3
Frozen (Bain et al., 2021)	31.0	59.5	70.5
MMFT (Shvetsova et al., 2022)	23.7	52.1	63.7
VideoCLIP (Xu et al., 2021)	30.9	55.4	**66.8**
TempCLR (Yang et al., 2023b)	30.6	55.1	65.5
Norton (Ours)	**31.2**	**55.7**	**66.8**
Zero-shot	R@1	R@5	R@10
SupportSet (Patrick et al., 2021)	8.7	23.0	31.1
Frozen (Bain et al., 2021)	23.2	44.6	56.6
MIL-NCE (Miech et al., 2020)	9.9	24.0	32.4
MMFT (Shvetsova et al., 2022)	9.9	24.0	**32.6**
VT-TWINS (Ko et al., 2022)	9.4	23.4	31.6
VideoCLIP (Xu et al., 2021)	10.4	22.2	30.0
TempCLR (Yang et al., 2023b)	10.1	22.2	29.4
Norton (Ours)	**10.7**	**24.1**	31.6

Table 3: Clip-caption retrieval on YouCookII.

Approach	Feature	R@1	R@5	R@10
ActBERT (Zhu & Yang, 2020)	R101+Res3D	9.6	26.7	38.0
MIL-NCE (Miech et al., 2020)	S3D-G	15.1	38.0	51.2
MCN (Chen et al., 2021)	R152+RX101	18.1	35.5	45.2
TACo (Yang et al., 2021a)	S3D-G	19.9	43.2	55.7
VT-TWINS (Ko et al., 2022)	S3D-G	9.7	27.0	38.8
MMFT (Shvetsova et al., 2022)	S3D-G	19.8	42.9	55.1
TAN (Han et al., 2022)	S3D-G	20.1	45.5	59.5
VideoCLIP (Xu et al., 2021)	S3D-G	22.7	50.4	63.1
TempCLR (Yang et al., 2023b)	S3D-G	23.3	51.0	**64.5**
Norton (Ours)	S3D-G	**24.2**	**51.9**	64.1

Table 4: Action segmentation on COIN.

Approach	Frame Accuracy
VAVA (Liu et al., 2022)	47.3
ActBERT (Zhu & Yang, 2020)	57.0
Drop-DTW (Dvornik et al., 2021)	59.6
MIL-NCE (Miech et al., 2020)	61.0
ClipBERT (Lei et al., 2021)	65.4
TACo (Yang et al., 2021a)	68.4
VideoCLIP (Xu et al., 2021)	68.7
TempCLR (Yang et al., 2023b)	68.7
Norton (Ours)	**69.8**

Table 6: VideoQA on MSR-VTT.

Superivsed	Accuracy
EITanque (Kaufman et al., 2017)	65.5
MLB(Kim et al., 2016)	76.1
JSFusion (Yu et al., 2018)	83.4
ActBERT (Zhu & Yang, 2020)	85.7
ClipBERT (Lei et al., 2021)	88.2
MERLOT (Zellers et al., 2021)	90.9
VideoCLIP (Xu et al., 2021)	92.1
TempCLR (Yang et al., 2023b)	92.2
Norton (Ours)	**92.7**
Zero-shot	Accuracy
VideoCLIP (Xu et al., 2021)	73.9
TempCLR (Yang et al., 2023b)	74.4
Norton (Ours)	**77.1**

Experiments

Task3 Effectiveness on noisy correspondence – Ablation Study

- Long video retrieval (with background)
- long video retrieval (without backgound)
- Short clip retrieval

Table 7: **Ablation experiments** evaluated on YouCookII, where "Clip" is short for clip-caption retrieval, "Video" for video-paragraph retrieval, "B" for video backgrounds, and "FNE" for faulty negative exploitation. We report the DTW measurement for video-paragraph retrieval.

OT Outperform DTW

Our fine-grained Better than FILIP[1]

Basic Setting				Clip		Video (w/o B)		Video (w B)	
Model	FNE	Soft-max α	APB p	R@1	R@5	R@1	R@5	R@1	R@5
VideoCLIP (Xu et al., 2021)	–	–	–	22.7	50.4	56.0	89.9	55.7	93.1
TempCLR (Yang et al., 2023b)	–	–	–	23.3	51.0	83.5	97.2	70.4	93.8
A (w/o $\mathcal{L}_{\text{video}}$)		–	–	22.8	50.1	56.7	89.0	56.4	91.8
B (w/o $\mathcal{L}_{\text{video}}$)	✓	–	–	23.4	50.8	63.3	93.3	65.1	92.4
C	✓	Mean average	–	23.1	50.1	84.2	97.3	74.3	**94.7**
D	✓	(Yao et al., 2022)	–	23.5	50.5	86.9	98.6	74.1	94.6
E	✓	0.1	–	23.8	51.7	88.1	98.6	74.2	**94.7**
F	✓	0.2	–	**24.0**	**51.8**	88.2	98.6	74.9	94.4
G	✓	1	–	**24.0**	**51.8**	**88.4**	**98.8**	**75.2**	**94.7**
H	✓	1	10%	**24.2**	51.8	88.4	**98.8**	75.9	94.9
I	✓	1	50%	**24.2**	51.9	88.4	98.6	75.9	94.9
J (Norton)	✓	1	30%	**24.2**	**51.9**	**88.7**	**98.8**	**76.1**	**95.0**

Ref:

1. FILIP: Fine-grained Interactive Language-Image Pre-Training, ICLR 2022

Experiments

Task3 Effectiveness on noisy correspondence – HTM-Align[1]

HTM- Align is a subset of the HowTo100M dataset, manually annotated to rectify the alignment in the presence of noisy correspondence.

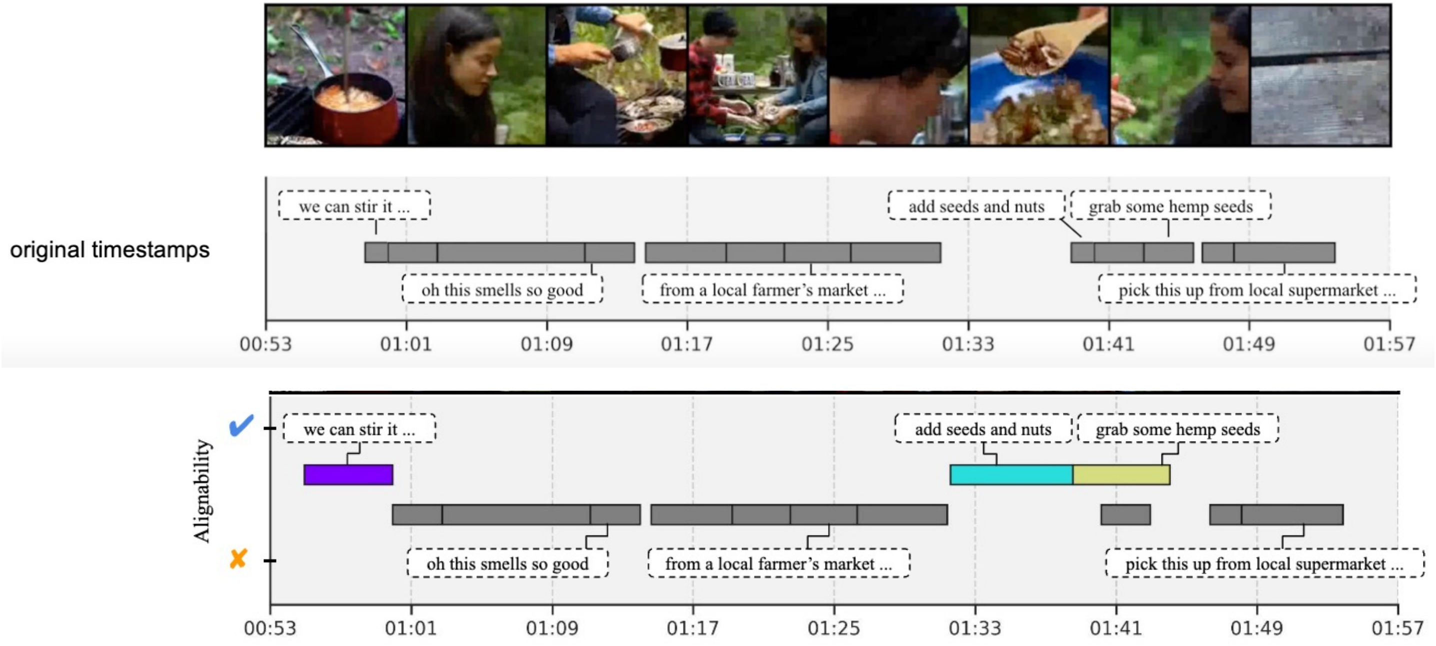


Table 9: Alignment results on the HTM-Align datasets.

Approach	Recall
CLIP (ViT-B/32) (Radford et al., 2021)	17.5
MIL-NCE (Miech et al., 2020)	34.2
TAN (Han et al., 2022) - 32 frame	41.1
TAN (Han et al., 2022) - 64 frame	49.2
VideoCLIP (Xu et al., 2021)	44.4
TempCLR (Yang et al., 2023b)	44.1
Norton (Ours)	**46.9**

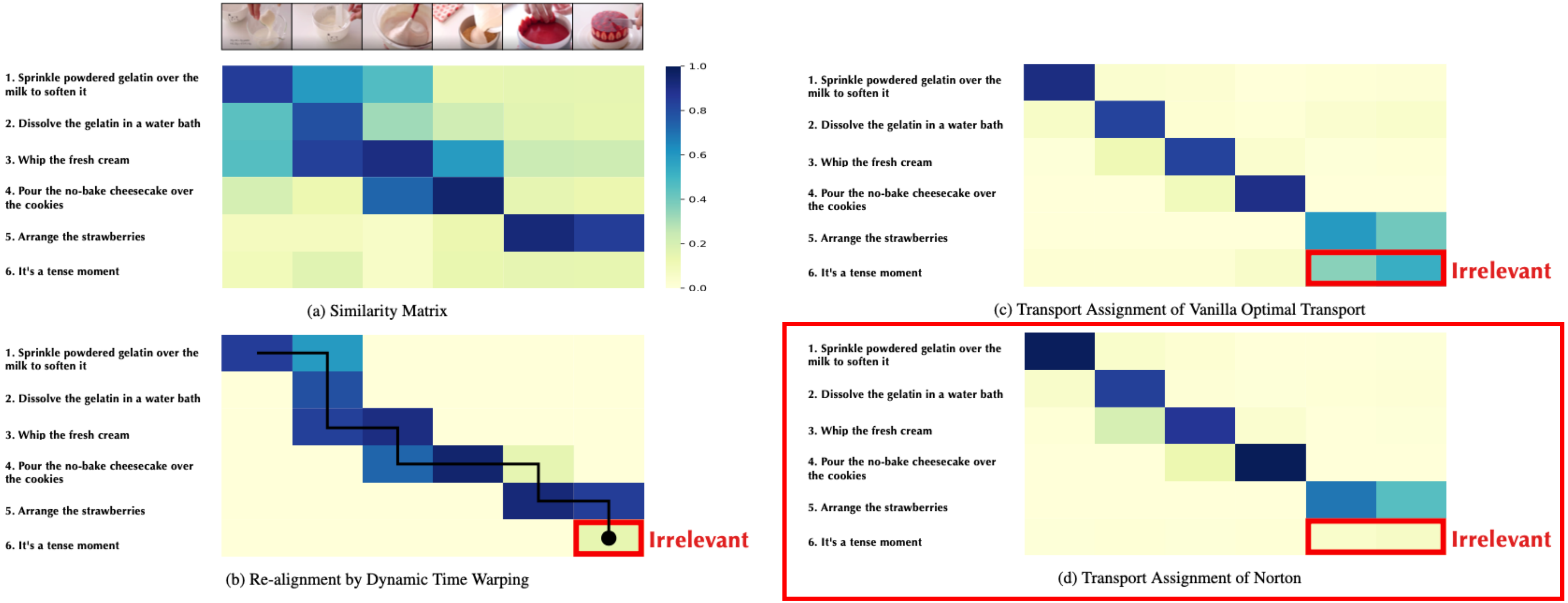
We tend not to fit noise

Ref:

1. Temporal Alignment Networks for Long-term Video, CVPR 2022

Experiments

Task3 Effectiveness on noisy correspondence – Visualization



(a) Similarity Matrix

(b) Re-alignment by Dynamic Time Warping

(c) Transport Assignment of Vanilla Optimal Transport

(d) Transport Assignment of Norton

Ours

Experiments

Task4 Training Efficiency

Table 8: **Training time per epoch.** 'f' denotes the sampled frame for a video clip. We use the time cost of clip-caption contrastive learning (Line 1) as the base value for comparison in the third column. The default setting is marked in gray.

Line	Approach		Time Cost
1	Clip-caption Contrast (16f)		87min (×1.000)
2	+ Faulty Negative Exploitation		92min (×1.057)
3	+ Video-paragraph Contrast (16f×8)	128 frame	142min (×1.632)
4	+ Fine-grained Soft-maximum Operator (16f×8)		146min (×1.678)
5	Clip-caption Contrast (32f)	32 frame	172min (×1.977)
6	Sinkhorn iteration in $\mathcal{L}_{\text{clip}}$		2.4min (×0.027)
7	Sinkhorn iteration in $\mathcal{L}_{\text{video}}$		2.6min (×0.029)

Negligible

Similar time cost but 4x sequence length

Future work

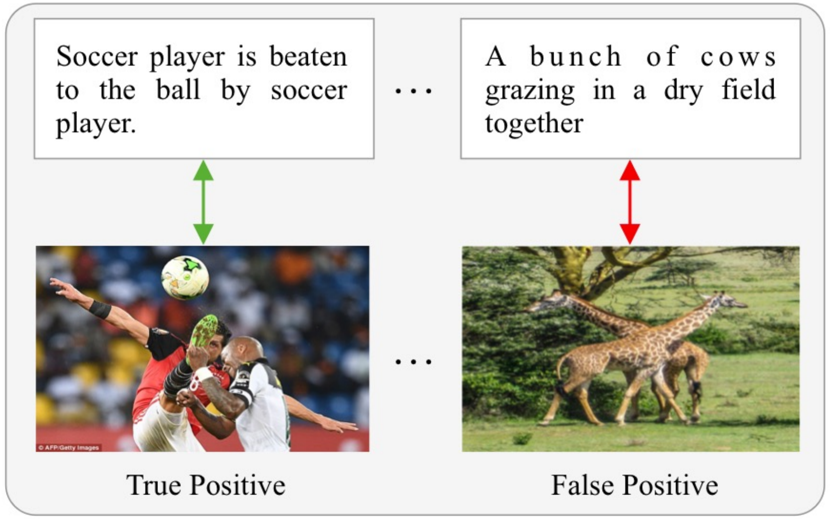
- **Multi-modal scenarios ($\geq$3, plus audio etc.).** Addressing multi-modal noisy correspondence presents an open challenge, given the quadratic growth in combinations concerning the number of modalities.

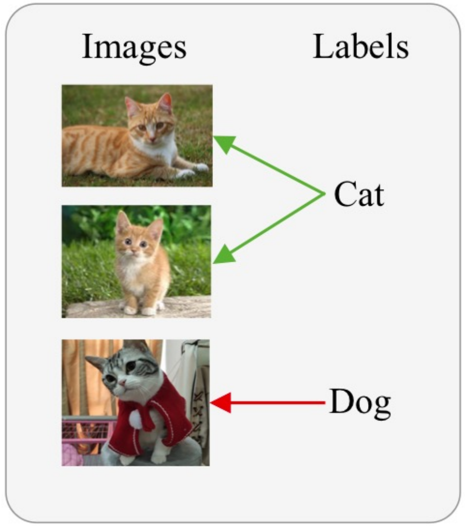
- **Utilization of Noise.** An intriguing question arises regarding whether these noisy samples could be utilized as an incentive for training.

- **All-in-one solution of Noisy Correspondence.** Is it feasible to propose a unified solution that addresses all types of noisy correspondence?

Conclusions

Noisy Correspondence vs Noisy Label





- Study a new paradigm for the noisy labels, i.e., noisy correspondence which is totally different from existing noisy label learning;
- Noisy correspondence is general to many intelligent techniques, including but not limited to multi-agent synchronization, cross-modal retrieval, VQA, visual grounding, visual navigation, tracking, Re-ID, and so on;

Noisy Correspondence Learning

Visit Our Poster @ Halle B #110

Project Page

http://lin-yijie.github.io/projects/Norton/

Multi-granularity Correspondence Learning from Long-term Noisy Videos

Yijie Lin, Jie Zhang, Zhenyu Huang, Jia Liu, Zujie Wen, Xi Peng

ICLR 2024 (Oral)

Paper | Code

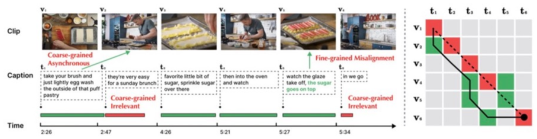


Figure 1: Our observation on multi-granularity noisy correspondence (MNC) in video understanding. (*Left*) The green timeline denotes the alignable captions while the red timeline indicates the unalignable captions. The green text in $\mathbf{t}_5$ denotes partially correlated words w.r.t $\mathbf{v}_5$. (*Right*) The dashed line represents the original alignment according to timestamps and the red block indicates the misaligned clip-caption pair. The green block denotes the ground-truth alignment. The solid line denotes the re-alignment by Dynamic Time Warping (Müller, 2007) which struggles to handle noisy correspondence well.

Related works on Noisy Correspondence

https://github.com/XLearning-SCU/Awesome-Noisy-Correspondence

Noisy-Correspondence Learning Summary (Updating)

A new research **direction** of label noise learning. Noisy correspondence learning aims to eliminate the negative impact of the mismatched pairs (e.g., false positives/negatives) instead of annotation errors in several tasks.

We mark works contributed by ourselves with ⭐.

This repository now is maintained by Mouxing Yang, Yijie Lin, and Yang Qin. We hope more AI-workers join us and thank all contributors!

Tasks

Image-Text Matching/Retrieval	Vision-Language Pre-training
Re-identification	Video-Text Learning
Image Captioning	Image Contrastive Learning
Graph Matching	Visual-Audio Learning
Machine Reading Comprehension	Dense Retrieval
Multi-View Clustering	