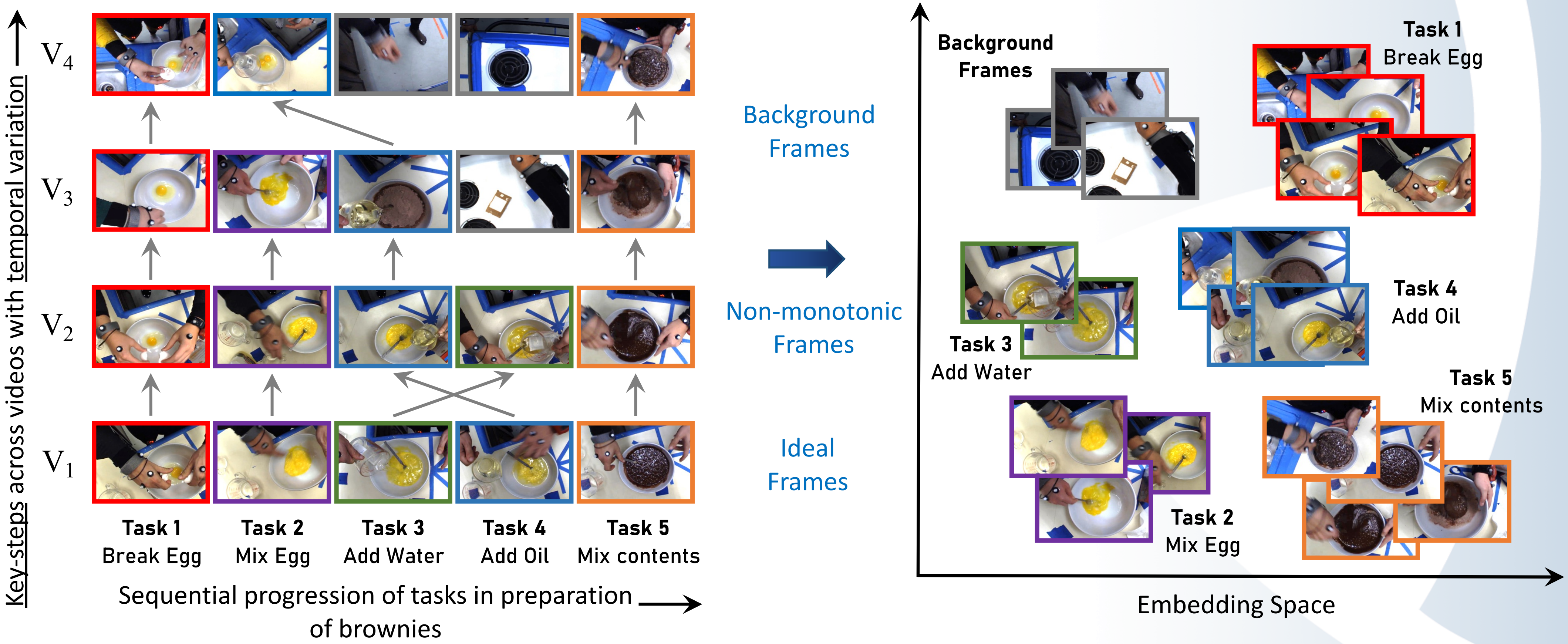
OPEL: Optimal Transport Guided ProcedurE Learning

Sayeed Shafayet Chowdhury, Soumyadeep Chandra, and Kaushik Roy
Purdue University

Acknowledgment: Center for Co-design of Cognitive Systems (CoCoSys)

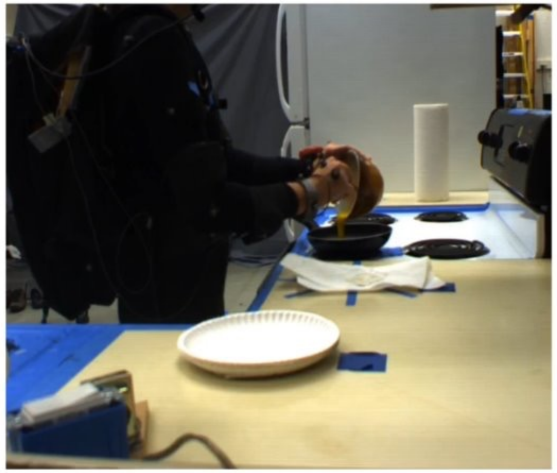
What is Procedure Learning (PL)?



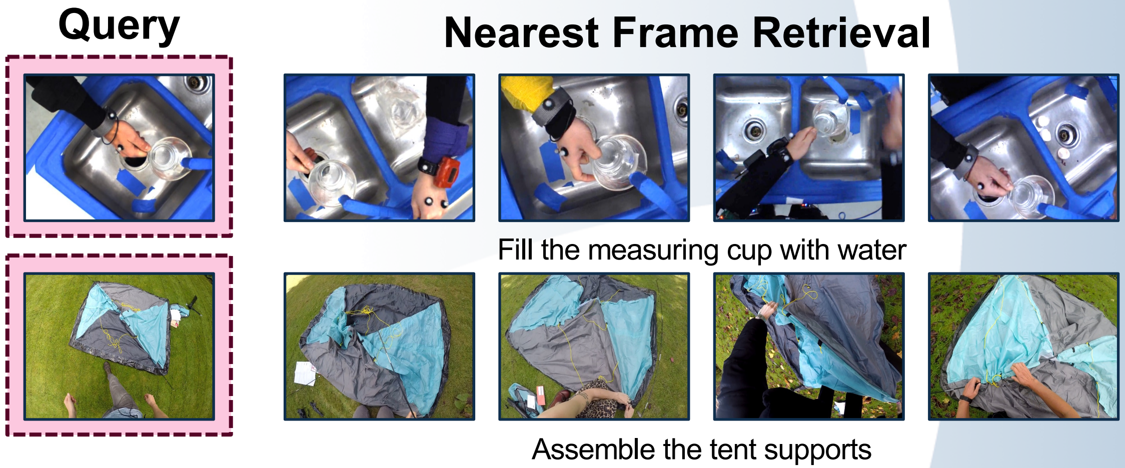
- Given multiple unlabeled videos of the same task,
  - Cluster the subtasks (key-steps) together in an embedding space
  - Determine their sequential ordering (*proper syntax*, but for videos)

Motivation

Human Demonstration

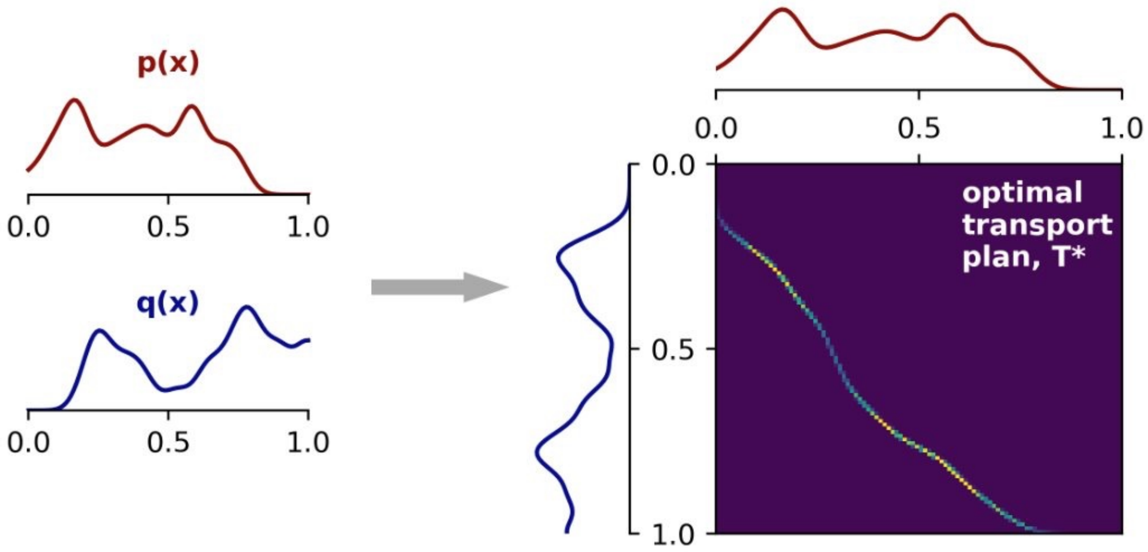


Robot learning and doing



- ➢ Unsupervised Robotic Learning
- ➢ Nearest Frame Retrieval
- ➢ Anomaly detection ensures the proper sequence of tasks, such as jacking up a car before accessing the wheel during a tire change

Background: Optimal Transport (OT)



Goal: **optimal alignment between two distributions**

Background: Optimal Transport (OT)

Sub-optimal Transport Plan

Transport Matrix, T

	0	1	2	3
0	.02	.01	.07	.9
1	.12	.8	.05	.03
2	.06	.87	.02	.05
3	.88	.04	.07	.06

○

Cost Matrix, D

	0	1	2	3
0	.05	2.3	1.1	3.4
1	.08	1.9	3.2	7.5
2	2.5	3.2	.03	1.7
3	9.8	4.3	2.4	.06

=

Local Cost

.0001	.023	.077	3.06
.0096	1.52	.16	.225
.15	2.784	.0006	.085
.0862	.172	.168	.0036

$\xrightarrow{\sum}$ Global Cost, <*T*, *D*> ↑

8.524

Desired Transport Plan

	0	1	2	3
0	.9	.01	.07	.02
1	.8	.12	.05	.03
2	.06	.02	.87	.05
3	.01	.04	.07	.88

○

	0	1	2	3
0	.05	2.3	1.1	3.4
1	.08	1.9	3.2	7.5
2	2.5	3.2	.03	1.7
3	9.8	4.3	2.4	.06

=

.045	.023	.077	.068
.064	.228	.16	.225
.15	.064	.026	.085
.098	.172	.168	.053

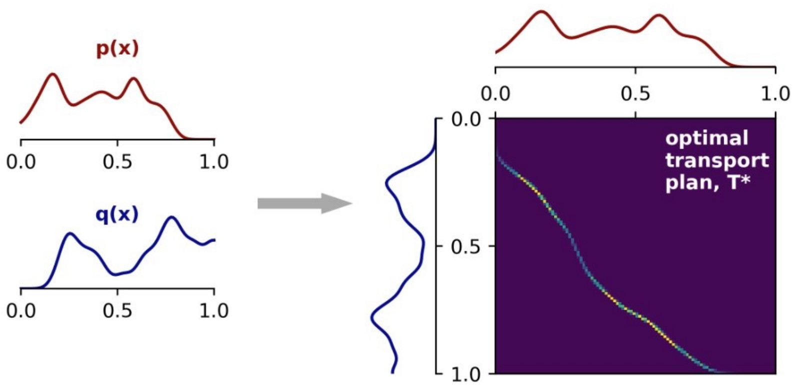
$\xrightarrow{\sum}$ Global Cost, <*T*, *D*> ↓

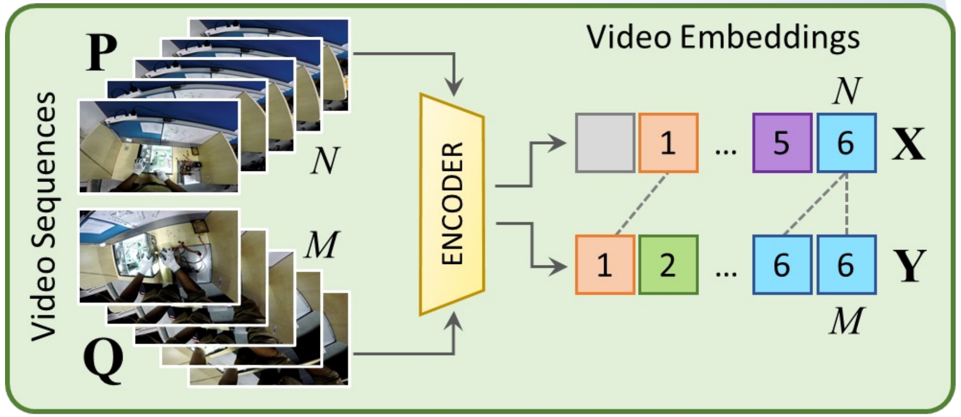
1.706

Objective: minimize <*T*,*D*>

https://www.pragmatic.ml/sparse-sinkhorn-attention/

Proposed Approach: Optimal Transport (OT)



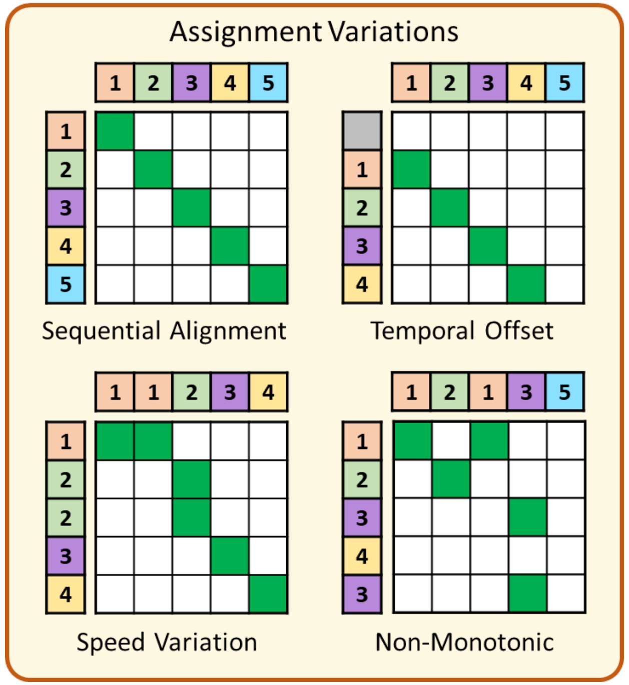


$$l_\lambda^S(\alpha, \beta, \boldsymbol{D}) = \langle \boldsymbol{T}_\lambda, \boldsymbol{D} \rangle$$

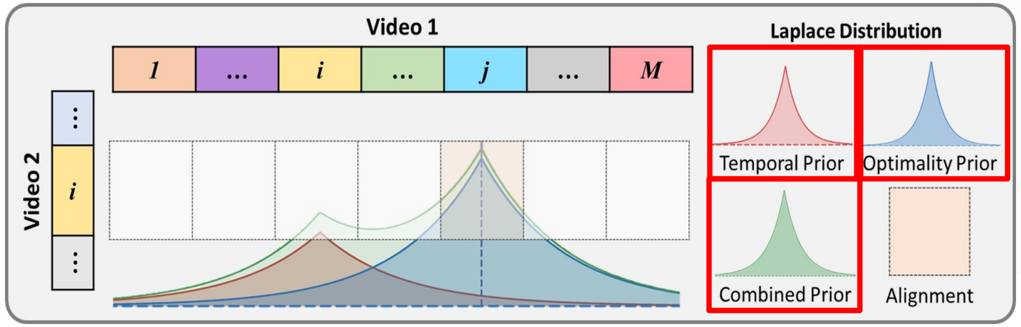
$$\boldsymbol{T}_\lambda = \arg \min_{T \in U(\alpha,\beta)} \langle \boldsymbol{T}, \boldsymbol{D} \rangle - \frac{1}{\lambda} h(\boldsymbol{T})$$

- $l_\lambda^S -$ *Sinkhorn Distance*
- $\boldsymbol{\alpha}_i = 1/N \;\; ; \; \boldsymbol{\beta}_j = 1/M$
- $\boldsymbol{D}$ $-$ *Distance matrix containing*: $d(\boldsymbol{x}_i, \boldsymbol{y}_j) = |\boldsymbol{x}_i - \boldsymbol{y}_j|$
- $\boldsymbol{T}$ $-$ *Transport matrix*: $t_{ij} \propto$ *probablity* $\boldsymbol{x}_i \Leftrightarrow \boldsymbol{y}_j$
- *regularization*, $h(\boldsymbol{T}) -$ *Entropy of* $\boldsymbol{T} = -\sum_{i=1}^{N} \sum_{j=1}^{M} t_{ij} \log t_{ij}$

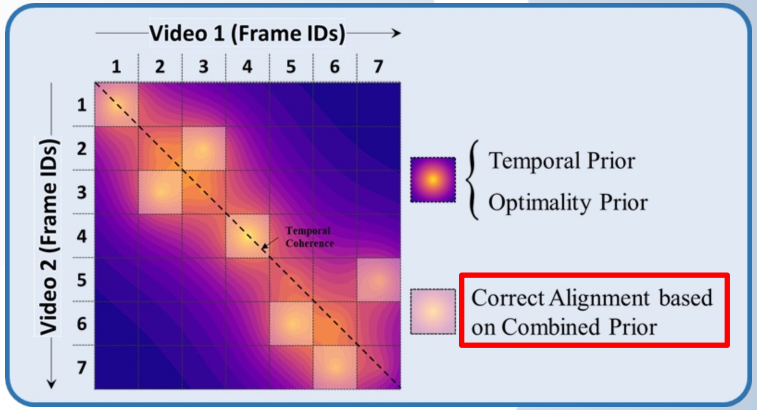
Priors



i and j are temporal frame idx of Video 2 and Video 1, respectively



1-D illustration



2-D depiction

- ➢ To address these variations:
  - ➢ Optimality Prior (handles non-monotonicity, speed variations etc.)
  - ➢ Temporal Prior (promotes temporal coherence)
  - ➢ Virtual frame in ***T*** (to manage background frames)

$$\boldsymbol{Q}_o(i,j) = \frac{1}{2b}e^{-\frac{|d_o(i,j)|}{b}}, \quad d_o(i,j) = \frac{|i/N - i_o/N| + |j/M - j_o/M|}{2\sqrt{1/N^2 + 1/M^2}}$$

$$\boldsymbol{Q}_t(i,j) = \frac{1}{2b}e^{-\frac{|d_t(i,j)|}{b}}, \quad d_t(i,j) = \frac{|i/N - j/M|}{\sqrt{1/N^2 + 1/M^2}}$$

$$Combined\ Prior\colon \boldsymbol{Q}(i,j) = \phi\boldsymbol{Q}_t(i,j) + (1-\phi)\boldsymbol{Q}_o(i,j)$$

Differentiable Formulation

Regularizations on Optimal Transport Matrix ($\widehat{\boldsymbol{T}}$)

$$M_o(\widehat{\boldsymbol{T}}) = \sum_{i=1}^{N+1}\sum_{j=1}^{M+1} \frac{t_{ij}}{\frac{1}{2}d_m + 1} \quad ; \quad d_m = \left(\frac{i - i_o}{N+1}\right)^2 + \left(\frac{j - j_o}{M+1}\right)^2$$

$$M_t(\widehat{\boldsymbol{T}}) = \sum_{i=1}^{N+1}\sum_{j=1}^{M+1} \frac{t_{ij}}{\left(\frac{i}{N+1} - \frac{j}{M+1}\right)^2 + 1}$$

Inverse Difference Moment **(IDM)** Regularization

$$M(\widehat{\boldsymbol{T}}) = \phi M_t(\widehat{\boldsymbol{T}}) + (1 - \phi) M_o(\widehat{\boldsymbol{T}}).$$

Desired: $M(\widehat{\boldsymbol{T}}) \geq \xi_1$ …. (i) $KL(\widehat{\boldsymbol{T}} \parallel \widehat{\boldsymbol{Q}}) \leq \xi_2$ …. (ii)

Using Lagrangian Duality: $l^R_{\lambda_1,\lambda_2}(X, Y) \coloneqq \langle \widehat{\boldsymbol{T}}_{\lambda_1,\lambda_2}, \boldsymbol{D} \rangle,$

$l^R_{\lambda_1,\lambda_2}$ – Regularized Sinkhorn distance

$$\widehat{\boldsymbol{T}}_{\lambda_1,\lambda_2} = \arg\min_{\widehat{T} \in U(\alpha,\beta)} \langle \widehat{\boldsymbol{T}}_{\lambda_1,\lambda_2}, \boldsymbol{D} \rangle - \lambda_1 M(\widehat{\boldsymbol{T}}) + \lambda_2 \mathrm{KL}(\widehat{\boldsymbol{T}} \parallel \widehat{\boldsymbol{Q}})$$

* Derivations are provided in the paper

Loss Functions

Intra-Video Contrastive-Inverse Difference Moment (**C-IDM**) Loss

$$I(\boldsymbol{X}) = \sum_{i=1}^{N+1} \sum_{j=1}^{M+1} \left(1 - \mathcal{N}(i,j)\right)\gamma(i,j) \max\left(0, \lambda_3 - d(i,j)\right) + \mathcal{N}(i,j) \frac{d(i,j)}{\gamma(i,j)}$$

$$\gamma(i,j) = (i-j)^2 + 1; \qquad d(i,j) = \left|\boldsymbol{x}_i - \boldsymbol{x}_j\right|; \qquad \mathcal{N}(i,j) = 1, \text{ if } |i-j| \leq \delta \text{ and } 0 \text{ otherwise}$$

$$best\_distance = \frac{1}{\text{temperature}} \cdot \left(\frac{1}{N} \sum_{i=1}^{N} \left\| \boldsymbol{x}_i - \boldsymbol{y}_{x_{best}(i)} \right\|^2 + \frac{1}{M} \sum_{j=1}^{M} \left\| \boldsymbol{y}_j - \boldsymbol{x}_{y_{best}(j)} \right\|^2 \right)$$

$$worst\_distance = \frac{1}{\text{temperature}} \cdot \left(\frac{1}{N} \sum_{i=1}^{N} \left\| \boldsymbol{x}_i - \boldsymbol{y}_{x_{worst}(i)} \right\|^2 + \frac{1}{M} \sum_{j=1}^{M} \left\| \boldsymbol{y}_j - \boldsymbol{x}_{y_{worst}(j)} \right\|^2 \right)$$

Inter-Video Contrastive Loss

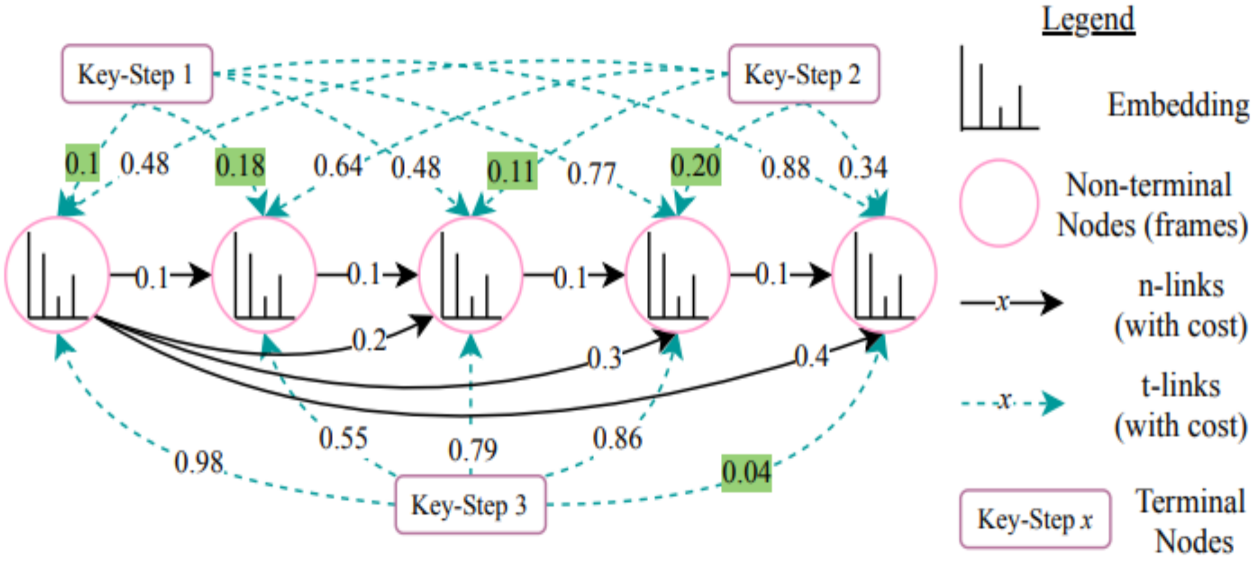
$$loss\_inter = F_{cross\_entropy}\left(\begin{bmatrix} best\_distance \\ worst\_distance \end{bmatrix}, \begin{bmatrix} 0 \\ 1 \end{bmatrix}\right)$$

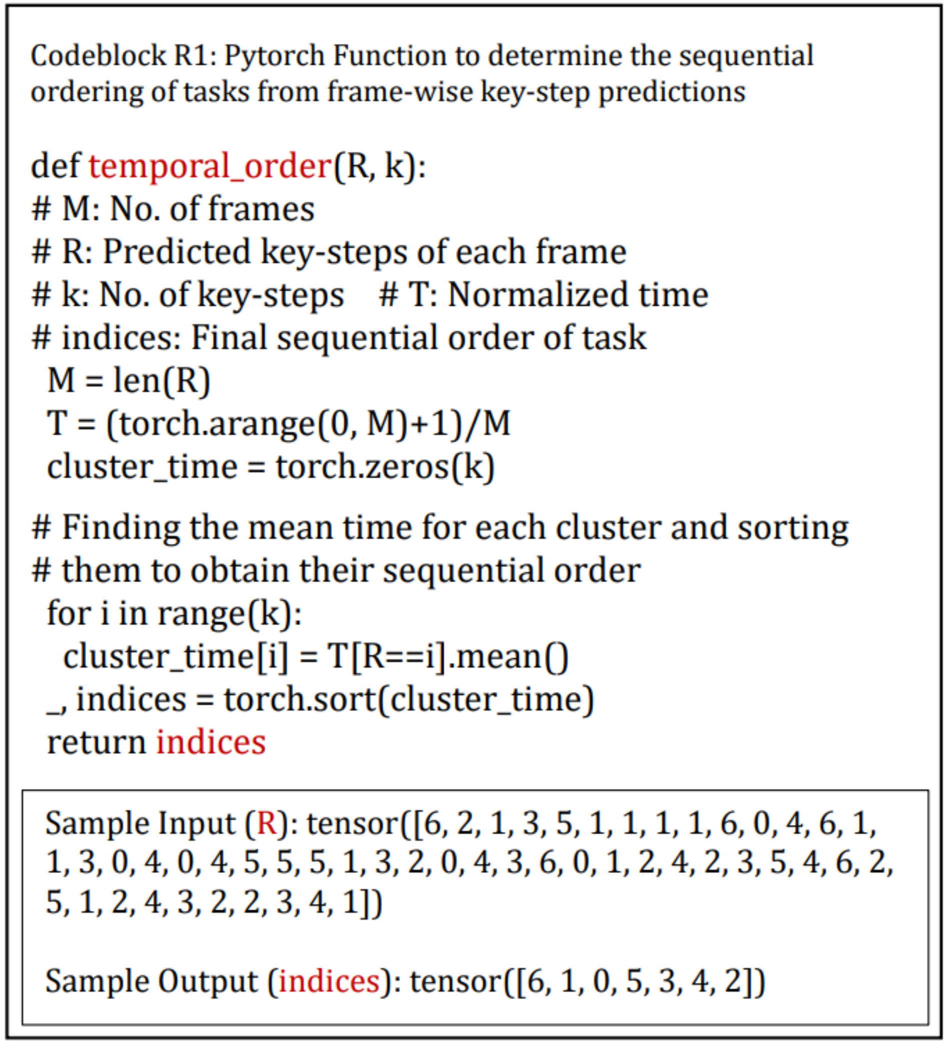
Overall **OPEL** Loss: $L_{OPEL}(X,Y) = c_1 * l^R_{\lambda_1,\lambda_2}(X,Y) + c_2 * \left(I(X) + I(Y)\right) + c_3 * \text{loss\_inter}$

Clustering done using multi-level graph-cut segmentation. Clusters are sequenced by averaging normalized times of frames in each cluster and ordering them to outline the video's key-step sequence.

Clustering and Ordering

- Multi-label graphcut segmentation





Codeblock R1: Pytorch Function to determine the sequential ordering of tasks from frame-wise key-step predictions

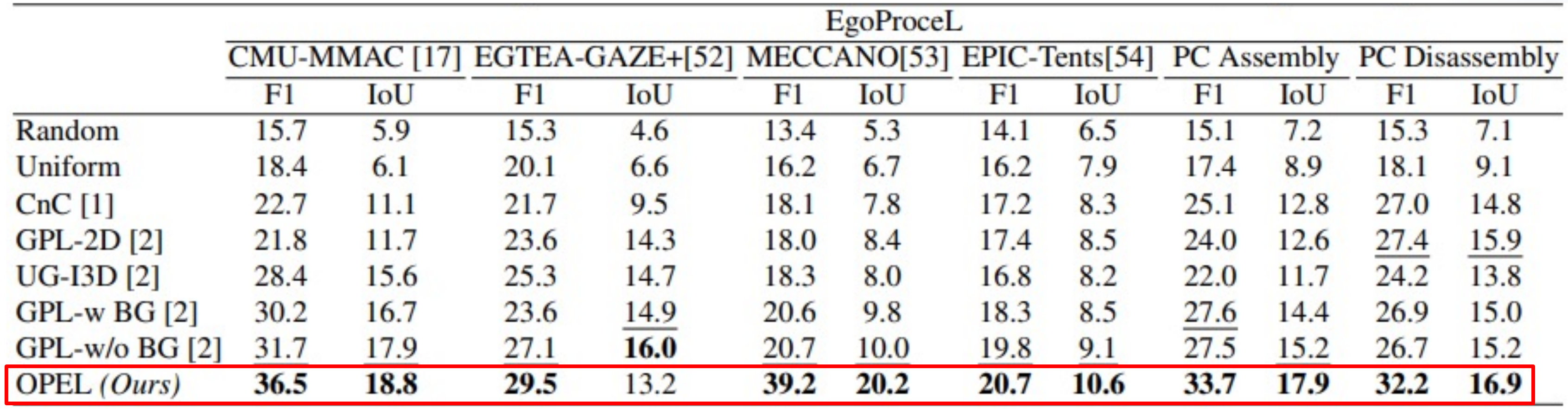
```
def temporal_order(R, k):
# M: No. of frames
# R: Predicted key-steps of each frame
# k: No. of key-steps    # T: Normalized time
# indices: Final sequential order of task
 M = len(R)
 T = (torch.arange(0, M)+1)/M
 cluster_time = torch.zeros(k)

# Finding the mean time for each cluster and sorting
# them to obtain their sequential order
 for i in range(k):
  cluster_time[i] = T[R==i].mean()
 _, indices = torch.sort(cluster_time)
 return indices
```

Sample Input (R): tensor([6, 2, 1, 3, 5, 1, 1, 1, 1, 6, 0, 4, 6, 1, 1, 3, 0, 4, 0, 4, 5, 5, 5, 1, 3, 2, 0, 4, 3, 6, 0, 1, 2, 4, 2, 3, 5, 4, 6, 2, 5, 1, 2, 4, 3, 2, 2, 3, 4, 1])

Sample Output (indices): tensor([6, 1, 0, 5, 3, 4, 2])

Quantitative Results

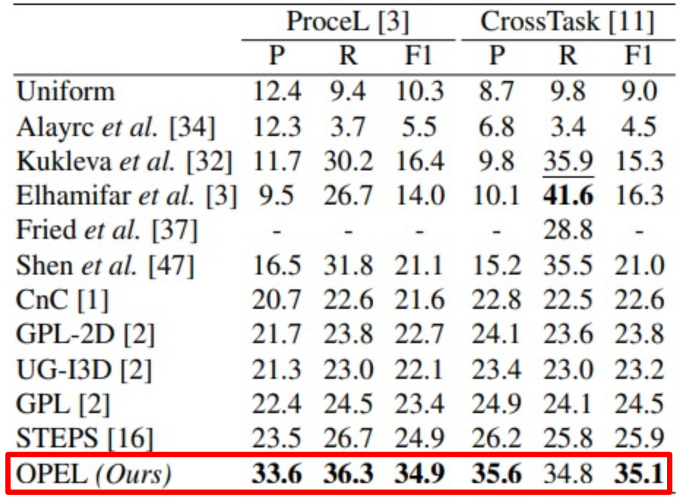
First-person (Egocentric) Videos



	EgoProceL											
	CMU-MMAC [17]		EGTEA-GAZE+[52]		MECCANO[53]		EPIC-Tents[54]		PC Assembly		PC Disassembly	
	F1	IoU	F1	IoU	F1	IoU	F1	IoU	F1	IoU	F1	IoU
Random	15.7	5.9	15.3	4.6	13.4	5.3	14.1	6.5	15.1	7.2	15.3	7.1
Uniform	18.4	6.1	20.1	6.6	16.2	6.7	16.2	7.9	17.4	8.9	18.1	9.1
CnC [1]	22.7	11.1	21.7	9.5	18.1	7.8	17.2	8.3	25.1	12.8	27.0	14.8
GPL-2D [2]	21.8	11.7	23.6	14.3	18.0	8.4	17.4	8.5	24.0	12.6	27.4	15.9
UG-I3D [2]	28.4	15.6	25.3	14.7	18.3	8.0	16.8	8.2	22.0	11.7	24.2	13.8
GPL-w BG [2]	30.2	16.7	23.6	14.9	20.6	9.8	18.3	8.5	27.6	14.4	26.9	15.0
GPL-w/o BG [2]	31.7	17.9	27.1	**16.0**	20.7	10.0	19.8	9.1	27.5	15.2	26.7	15.2
OPEL *(Ours)*	**36.5**	**18.8**	**29.5**	13.2	**39.2**	**20.2**	**20.7**	**10.6**	**33.7**	**17.9**	**32.2**	**16.9**

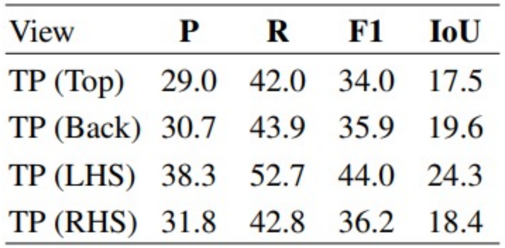
22.4% (IoU) and **26.9%** (F1) average improvement compared to current SOTA

Third-person (TP) Videos



	ProceL [3]			CrossTask [11]		
	P	R	F1	P	R	F1
Uniform	12.4	9.4	10.3	8.7	9.8	9.0
Alayrc *et al.* [34]	12.3	3.7	5.5	6.8	3.4	4.5
Kukleva *et al.* [32]	11.7	30.2	16.4	9.8	35.9	15.3
Elhamifar *et al.* [3]	9.5	26.7	14.0	10.1	**41.6**	16.3
Fried *et al.* [37]	-	-	-	-	28.8	-
Shen *et al.* [47]	16.5	31.8	21.1	15.2	35.5	21.0
CnC [1]	20.7	22.6	21.6	22.8	22.5	22.6
GPL-2D [2]	21.7	23.8	22.7	24.1	23.6	23.8
UG-I3D [2]	21.3	23.0	22.1	23.4	23.0	23.2
GPL [2]	22.4	24.5	23.4	24.9	24.1	24.5
STEPS [16]	23.5	26.7	24.9	26.2	25.8	25.9
OPEL *(Ours)*	**33.6**	**36.3**	**34.9**	**35.6**	34.8	**35.1**

TP Views of CMU-MMAC



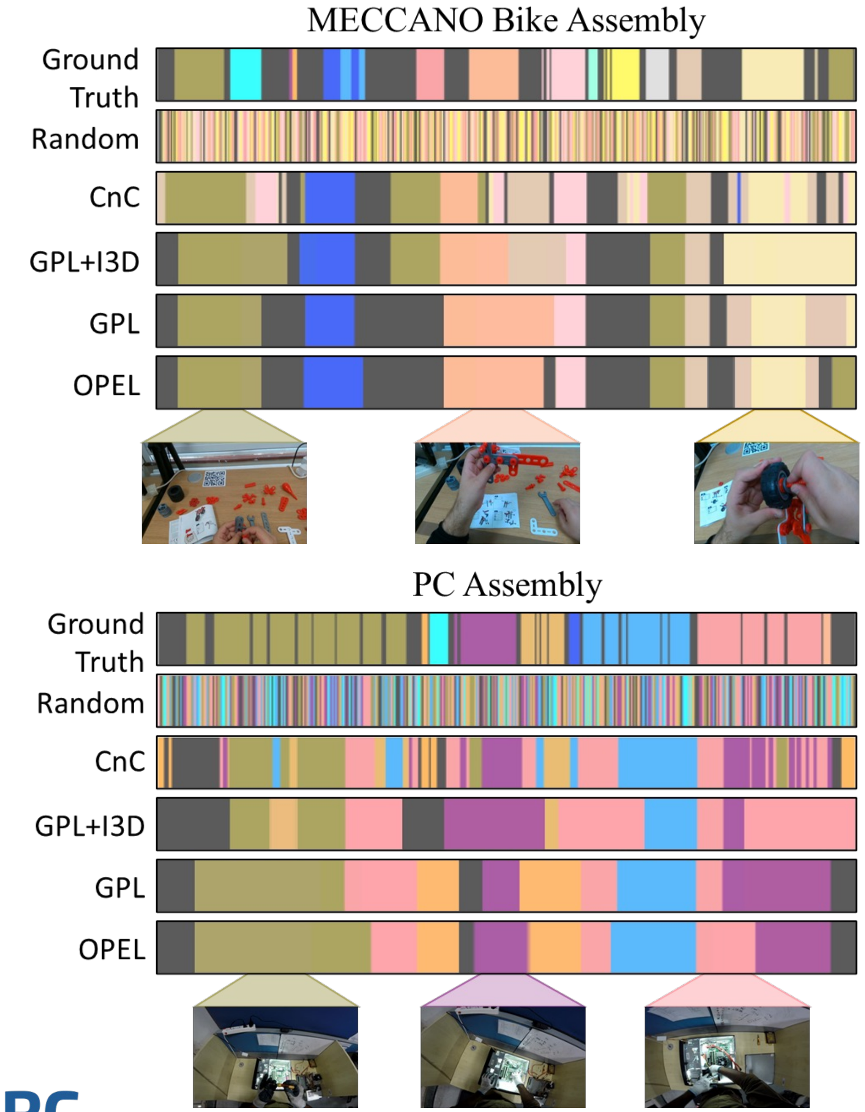
View	**P**	**R**	**F1**	**IoU**
TP (Top)	29.0	42.0	34.0	17.5
TP (Back)	30.7	43.9	35.9	19.6
TP (LHS)	38.3	52.7	44.0	24.3
TP (RHS)	31.8	42.8	36.2	18.4

46.2% (F1) average improvement compared to current SOTA

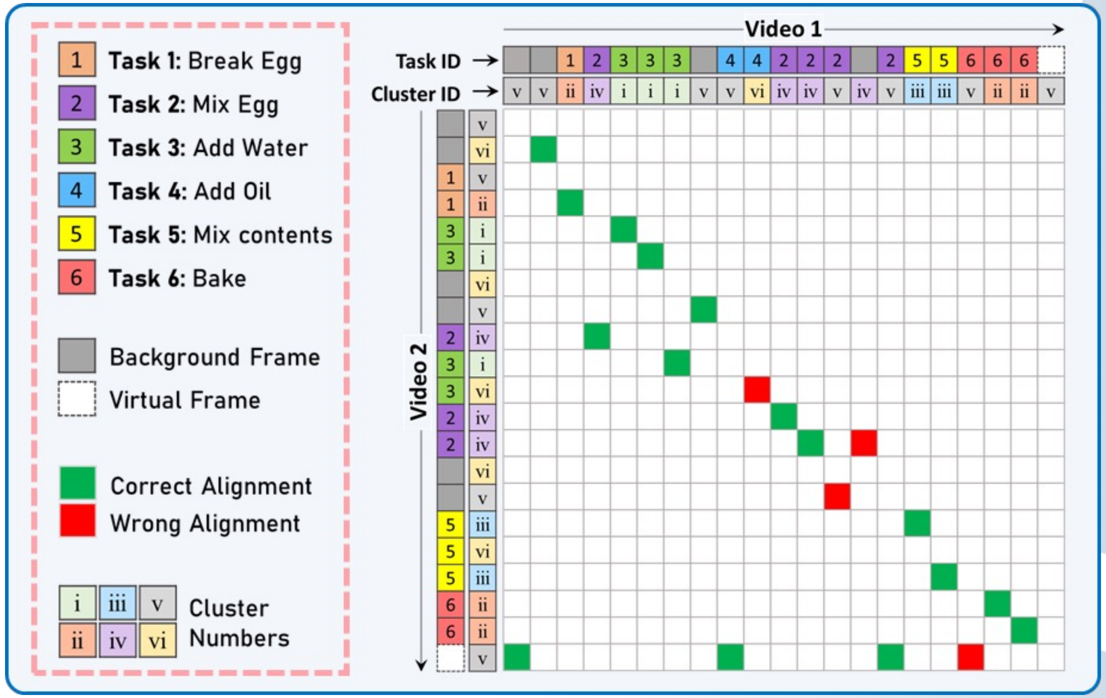
- SOTA on all benchmarks

SRC

Qualitative Results



- Higher overlap with Ground Truth compared to State-of-the-art
- Accurate alignment despite temporal variations


1 Task 1: Break Egg

2 Task 2: Mix Egg

3 Task 3: Add Water

4 Task 4: Add Oil

5 Task 5: Mix contents

6 Task 6: Bake

Background Frame

Virtual Frame

🟩 Correct Alignment

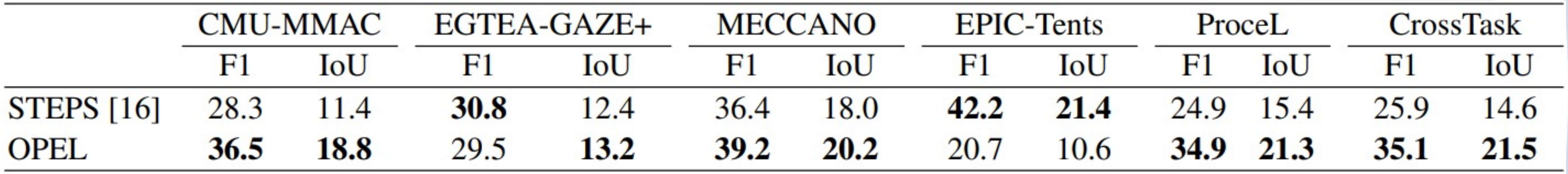
🟥 Wrong Alignment

i iii v / ii iv vi Cluster Numbers

Video 1
Task ID →
1 2 3 3 3 4 4 2 2 2 2 5 5 6 6 6
Cluster ID →
v v ii iv i i i v v vi iv iv v iv v iii iii v ii ii v
Video 2
1 1 3 3 2 3 3 2 2 5 5 5 6 6
v vi v ii i i vi v iv i vi iv iv vi v iii vi iii ii ii v

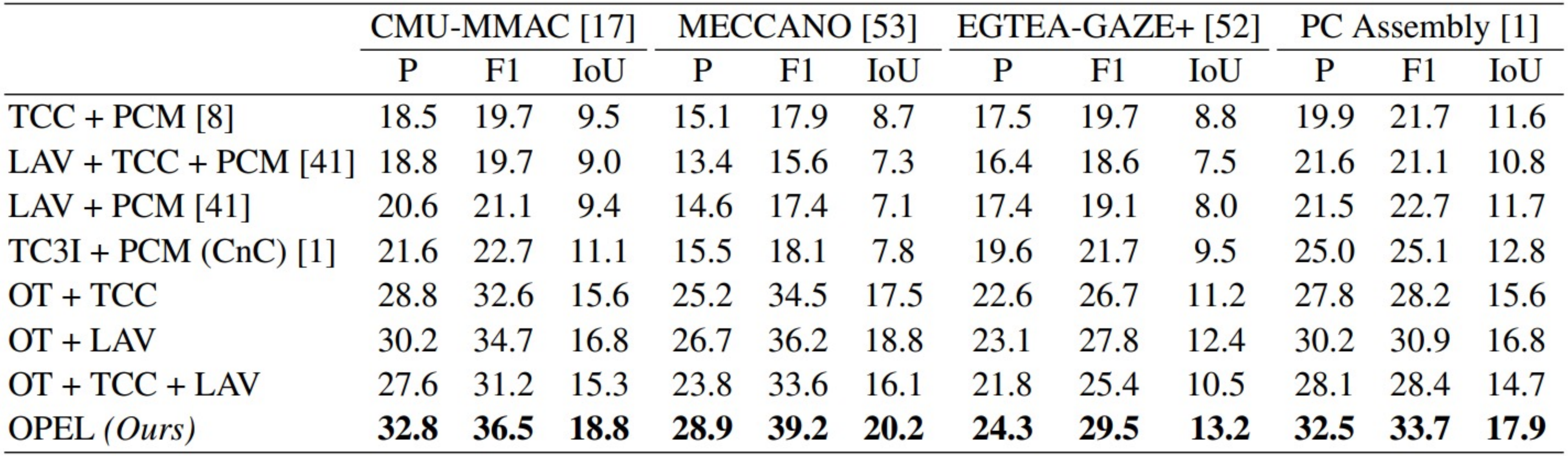
Additional Results

Better results than Multi-modal SOTA



	CMU-MMAC		EGTEA-GAZE+		MECCANO		EPIC-Tents		ProceL		CrossTask	
	F1	IoU	F1	IoU	F1	IoU	F1	IoU	F1	IoU	F1	IoU
STEPS [16]	28.3	11.4	**30.8**	12.4	36.4	18.0	**42.2**	**21.4**	24.9	15.4	25.9	14.6
OPEL	**36.5**	**18.8**	29.5	**13.2**	**39.2**	**20.2**	20.7	10.6	**34.9**	**21.3**	**35.1**	**21.5**

Effectiveness of L_{OPEL}

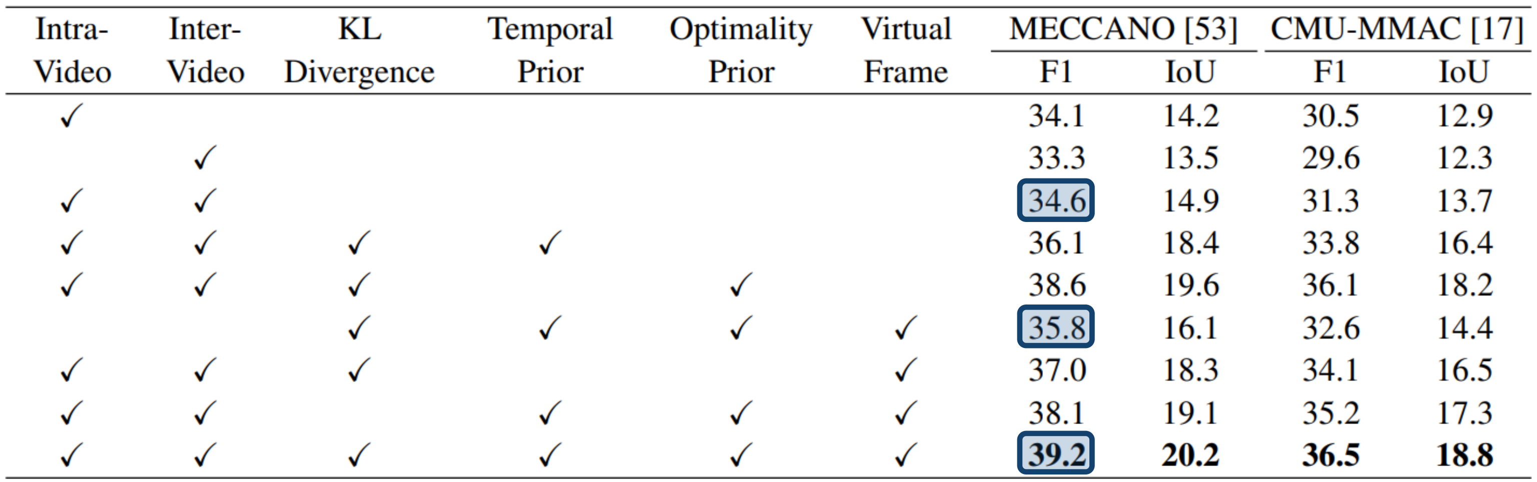


	CMU-MMAC [17]			MECCANO [53]			EGTEA-GAZE+ [52]			PC Assembly [1]		
	P	F1	IoU	P	F1	IoU	P	F1	IoU	P	F1	IoU
TCC + PCM [8]	18.5	19.7	9.5	15.1	17.9	8.7	17.5	19.7	8.8	19.9	21.7	11.6
LAV + TCC + PCM [41]	18.8	19.7	9.0	13.4	15.6	7.3	16.4	18.6	7.5	21.6	21.1	10.8
LAV + PCM [41]	20.6	21.1	9.4	14.6	17.4	7.1	17.4	19.1	8.0	21.5	22.7	11.7
TC3I + PCM (CnC) [1]	21.6	22.7	11.1	15.5	18.1	7.8	19.6	21.7	9.5	25.0	25.1	12.8
OT + TCC	28.8	32.6	15.6	25.2	34.5	17.5	22.6	26.7	11.2	27.8	28.2	15.6
OT + LAV	30.2	34.7	16.8	26.7	36.2	18.8	23.1	27.8	12.4	30.2	30.9	16.8
OT + TCC + LAV	27.6	31.2	15.3	23.8	33.6	16.1	21.8	25.4	10.5	28.1	28.4	14.7
OPEL *(Ours)*	**32.8**	**36.5**	**18.8**	**28.9**	**39.2**	**20.2**	**24.3**	**29.5**	**13.2**	**32.5**	**33.7**	**17.9**

- OPEL loss performs better compared to other existing

Ablation Studies

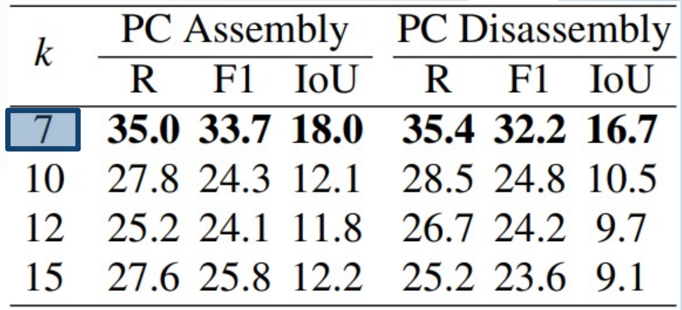
Impact of each term of L_{OPEL}



Intra-Video	Inter-Video	KL Divergence	Temporal Prior	Optimality Prior	Virtual Frame	MECCANO [53]		CMU-MMAC [17]	
						F1	IoU	F1	IoU
✓						34.1	14.2	30.5	12.9
	✓					33.3	13.5	29.6	12.3
✓	✓					34.6	14.9	31.3	13.7
✓	✓	✓	✓			36.1	18.4	33.8	16.4
✓	✓	✓		✓		38.6	19.6	36.1	18.2
		✓	✓	✓	✓	35.8	16.1	32.6	14.4
✓	✓	✓			✓	37.0	18.3	34.1	16.5
✓	✓		✓	✓	✓	38.1	19.1	35.2	17.3
✓	✓	✓	✓	✓	✓	**39.2**	**20.2**	**36.5**	**18.8**

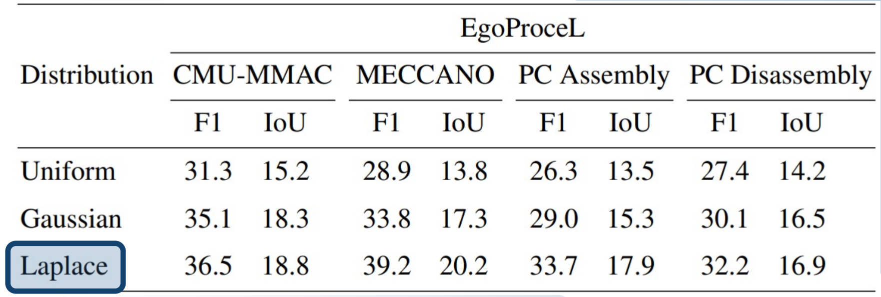
- All terms enhance performance – priors ~5pts, contrastive losses ~ 3.5 pts

Number of clusters



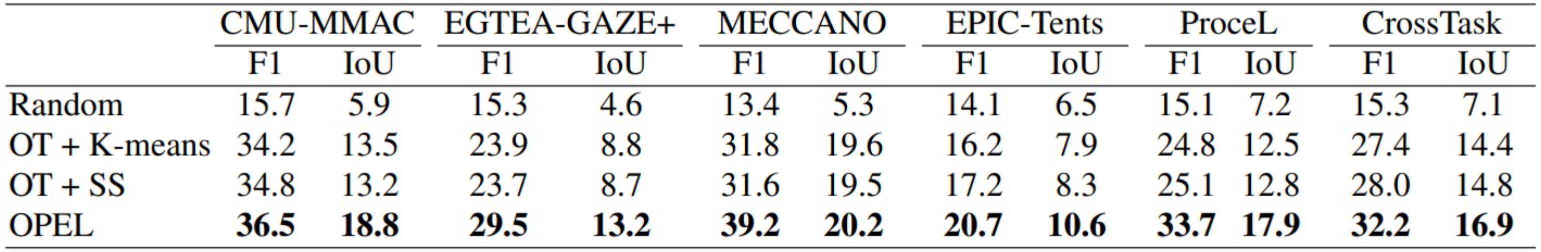
k	PC Assembly			PC Disassembly		
	R	F1	IoU	R	F1	IoU
7	**35.0**	**33.7**	**18.0**	**35.4**	**32.2**	**16.7**
10	27.8	24.3	12.1	28.5	24.8	10.5
12	25.2	24.1	11.8	26.7	24.2	9.7
15	27.6	25.8	12.2	25.2	23.6	9.1

Distribution of Priors



Distribution	EgoProceL							
	CMU-MMAC		MECCANO		PC Assembly		PC Disassembly	
	F1	IoU	F1	IoU	F1	IoU	F1	IoU
Uniform	31.3	15.2	28.9	13.8	26.3	13.5	27.4	14.2
Gaussian	35.1	18.3	33.8	17.3	29.0	15.3	30.1	16.5
Laplace	36.5	18.8	39.2	20.2	33.7	17.9	32.2	16.9

Clustering Algorithms



	CMU-MMAC		EGTEA-GAZE+		MECCANO		EPIC-Tents		ProceL		CrossTask	
	F1	IoU	F1	IoU	F1	IoU	F1	IoU	F1	IoU	F1	IoU
Random	15.7	5.9	15.3	4.6	13.4	5.3	14.1	6.5	15.1	7.2	15.3	7.1
OT + K-means	34.2	13.5	23.9	8.8	31.8	19.6	16.2	7.9	24.8	12.5	27.4	14.4
OT + SS	34.8	13.2	23.7	8.7	31.6	19.5	17.2	8.3	25.1	12.8	28.0	14.8
OPEL	**36.5**	**18.8**	**29.5**	**13.2**	**39.2**	**20.2**	**20.7**	**10.6**	**33.7**	**17.9**	**32.2**	**16.9**

- OT+graphcut segmentation (OPEL) performs best

Summary

- Contributions –
  - A novel OT-guided unsupervised procedure learning framework
  - SOTA results on all benchmarks (1st person as well as 3rd person)
- Limitation – assumption that subjects utilize similar objects for identical key-steps
- Future work – integration of additional contextual and semantic features within the OT framework, extending this framework to other domains of video understanding

THANK YOU!

Questions?