D^3TW : Discriminative Differentiable Dynamic Time Warping for Weakly-Supervised Action Alignment and Segmentation

Chien-Yi Chang, De-An Huang, Yanan Sui, Li Fei-Fei, Juan Carlos Niebles
Stanford Vision and Learning Lab, Stanford University

CVPR LONG BEACH CALIFORNIA
June 16-20, 2019

Summary

Problem:
Given an untrimmed video, our goal is to predict action labels for every frame.

Weakly-Supervised Learning:
We train our model using only action ordering (a.k.a action *transcript*).

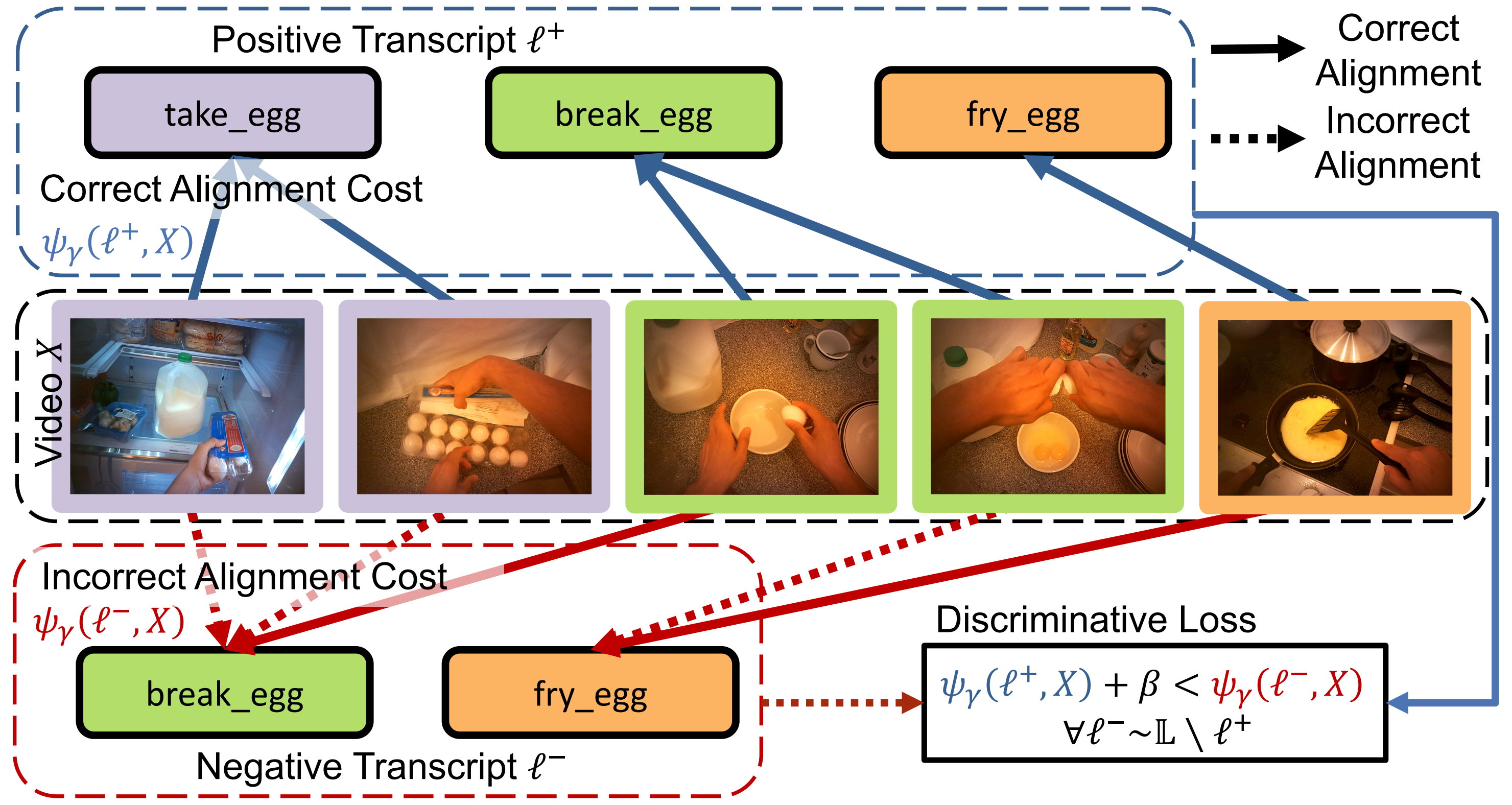
Key contributions:

(i) We introduce the first discriminative model for ordering supervision to address the degenerate sequence problem.

(ii) We propose D^3TW, a novel framework that incorporates the advantage of discriminative modeling and end-to-end training for structural sequence prediction with weak supervision.

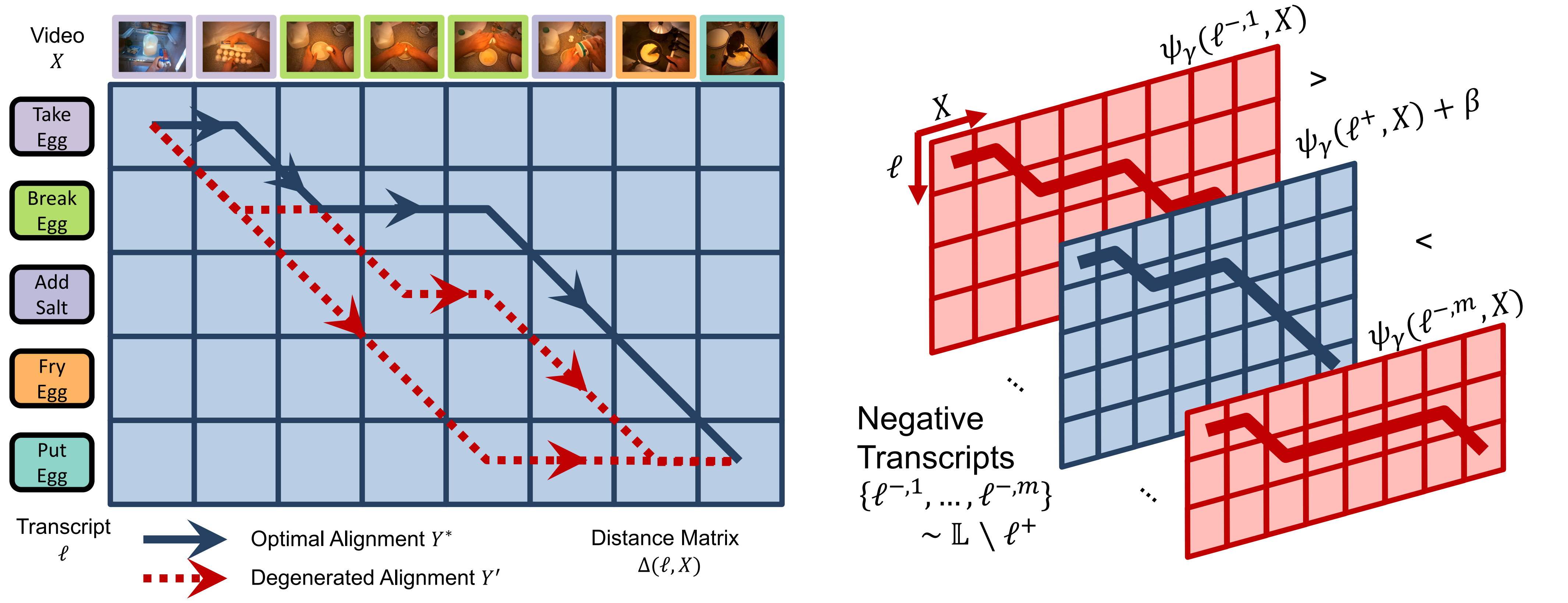
(iii) We apply our method in two challenging real-world video datasets and show that it achieves state-of-the-art for both weakly-supervised action segmentation and alignment.

Weakly Supervised Action Alignment Formalism

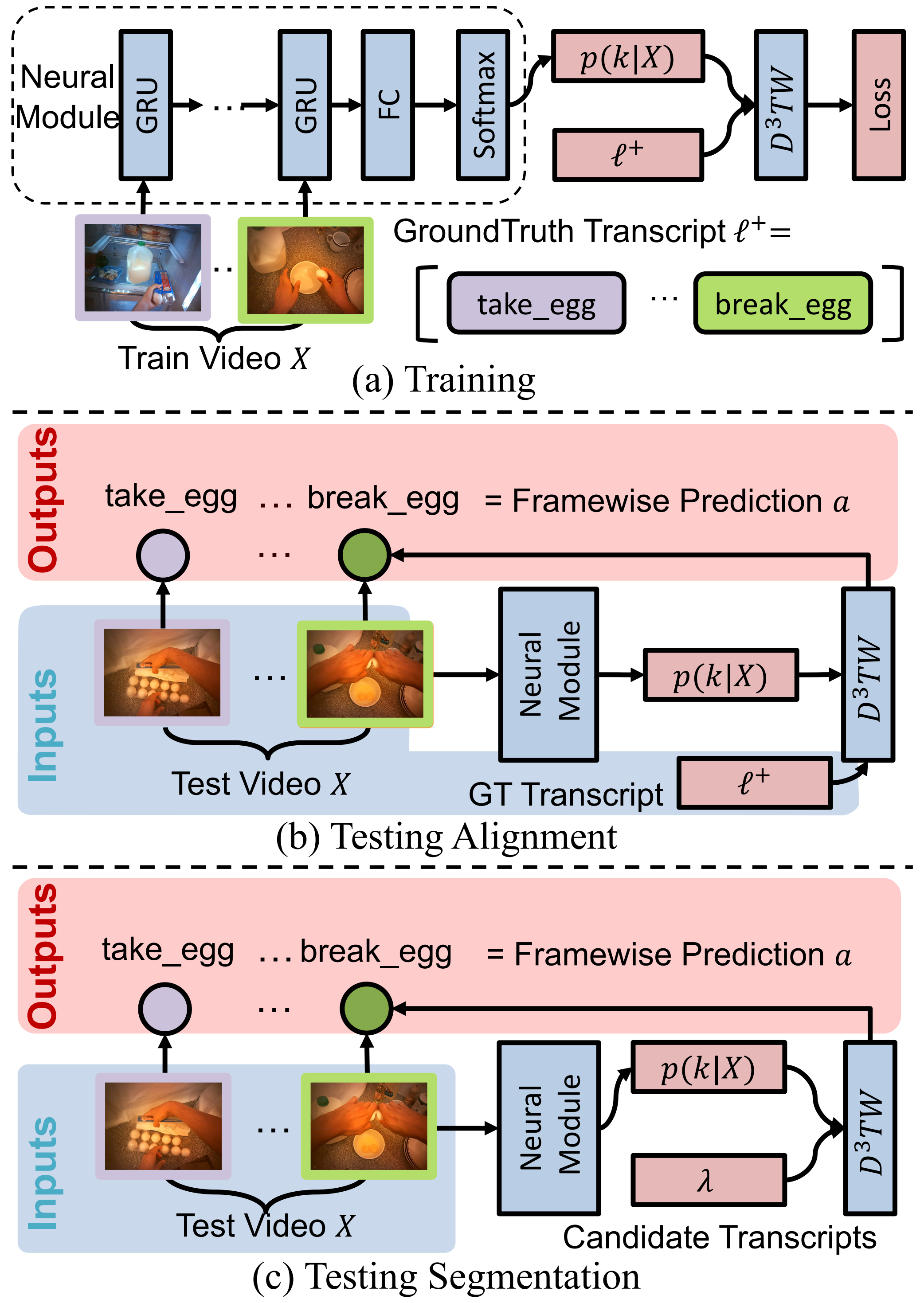


$$\psi_\gamma(\ell^+, X) + \beta < \psi_\gamma(\ell^-, X)$$
$$\forall \ell^- \sim \mathbb{L} \setminus \ell^+$$

D^3TW: Discriminative Differentiable Dynamic Time Warping


Video X
Take Egg
Break Egg
Add Salt
Fry Egg
Put Egg
Transcript ℓ
Optimal Alignment Y^*
Degenerated Alignment Y'
Distance Matrix $\Delta(\ell, X)$
$\psi_\gamma(\ell^{-,1}, X)$
>
$\psi_\gamma(\ell^+, X) + \beta$
<
$\psi_\gamma(\ell^{-,m}, X)$
X
ℓ
Negative Transcripts
$\{\ell^{-,1}, \dots, \ell^{-,m}\}$
$\sim \mathbb{L} \setminus \ell^+$

Model Overview

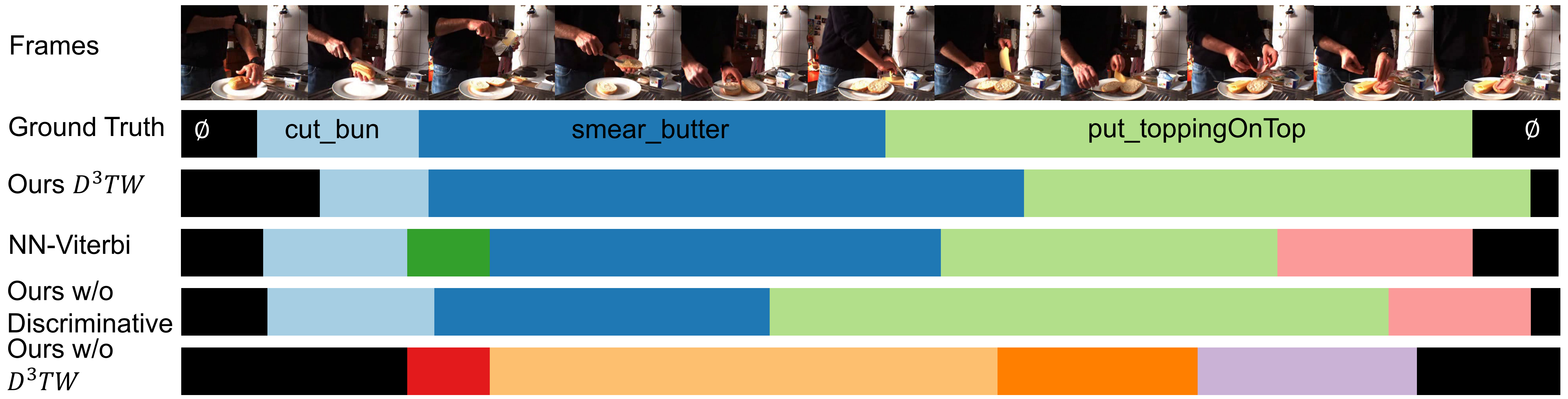


(a) Training

(b) Testing Alignment

(c) Testing Segmentation

Evaluating Weakly Supervised Action Segmentation Results on Breakfast Dataset



Recipe	Delta Facc.
Sandwich	+24.7%
Cereals	+19.9%
Pancake	+0.2%
Scrambled Egg	-0.8%

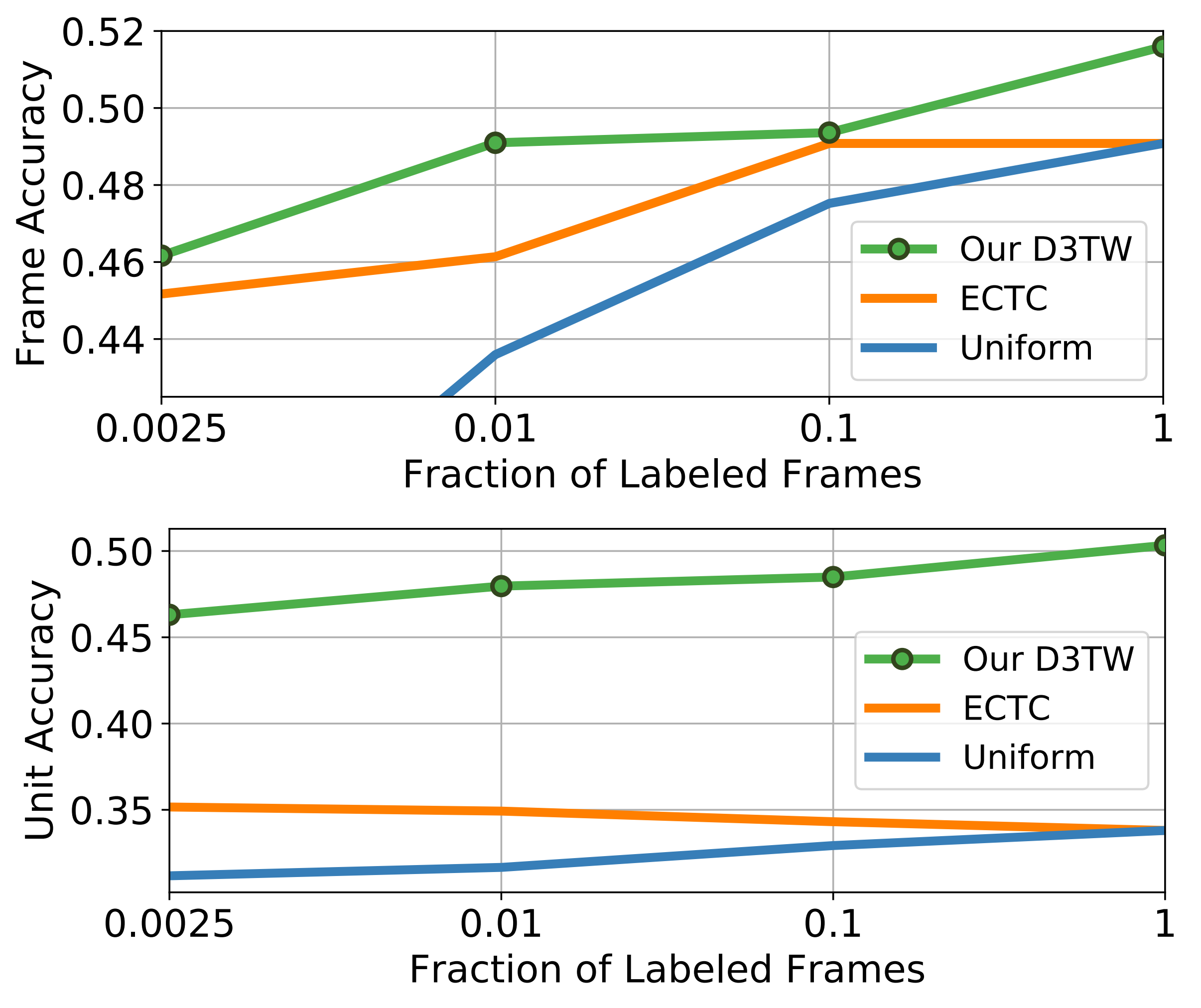
Experimental Results

	Breakfast		Hollywood	
	Facc.	Uacc.	Facc.	Uacc.
ECTC[15]	27.7	35.6	-	-
GRU reest.[28]	33.3	-	-	-
TCFPN[7]	38.4	-	28.7	-
NN-Viterbi[29]	43.0	-	-	-
Ours w/o D^3TW	34.9	36.1	25.9	24.3
Ours w/o Discriminative	38.0	38.4	30.0	28.3
Ours (D^3TW)	**45.7**	**47.4**	**33.6**	**30.5**

Weakly-supervised Action Segmentation

	Breakfast		Hollywood	
	Facc.	IoD	Facc.	IoD
ECTC[15]	~35	~45	-	~41
GRU reest.[28]	-	47.3	-	46.3
TCFPN[7]	53.5	52.3	57.4	39.6
NN-Viterbi[29]	-	-	-	48.7
Ours w/o D^3TW	42.8	49.5	51.2	47.2
Ours w/o Discriminative	52.3	47.6	51.8	46.9
Ours (D^3TW)	**57.0**	**56.3**	**59.4**	**50.9**

Weakly-supervised Action Alignment



Semi-supervised Action Segmentation

Acknowledgement

This work was partially funded by Toyota Research Institute (TRI). This article solely reflects the opinions and conclusions of its authors and not TRI or any other Toyota entity.

Chien-Yi Chang, De-An Huang, Yanan Sui, Li Fei-Fei, and Juan Carlos Niebles. D3TW: Discriminative differentiable dynamic time warping for weakly-supervised action alignment and segmentation. In *CVPR*, 2019