

Text-Visual Prompting for Efficient 2D Temporal Video Grounding

Yimeng Zhang^{1,2}, Xin Chen², Jinghan Jia¹, Sijia Liu¹, Ke Ding²

¹OPTML lab, Michigan State University, ²Applied ML, Intel

Introduction

What is temporal video grounding (TVG) ?

TVG is to predict the **starting/ending time points** of moments described by a text sentence within a long untrimmed video.

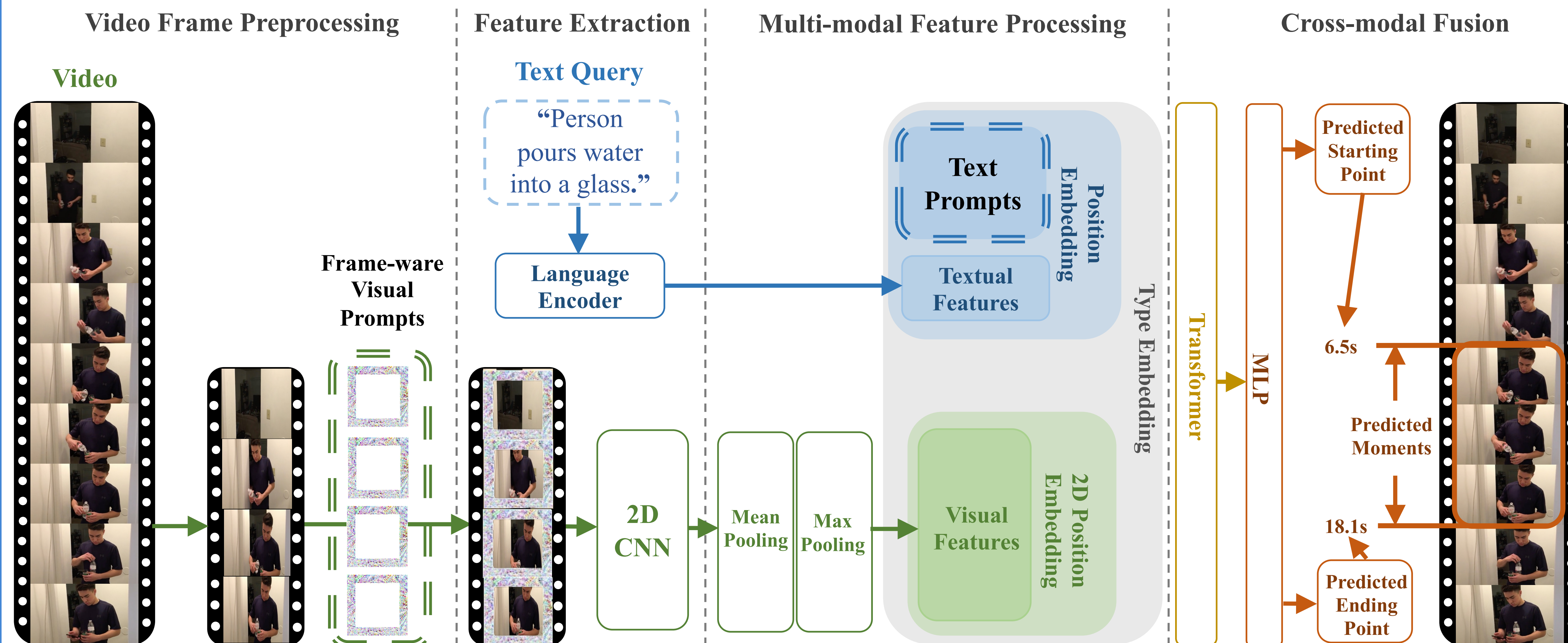
Motivation

High complexity of 3D CNNs makes extracting dense 3D visual features time-consuming, which calls for intensive memory and computing resources.

Challenges

How to advance 2D TVG methods so as to achieve comparable results to 3D TVG methods?

Text-Visual Prompting (TVP) Framework for TVG



Loss Function: TD-IoU Loss

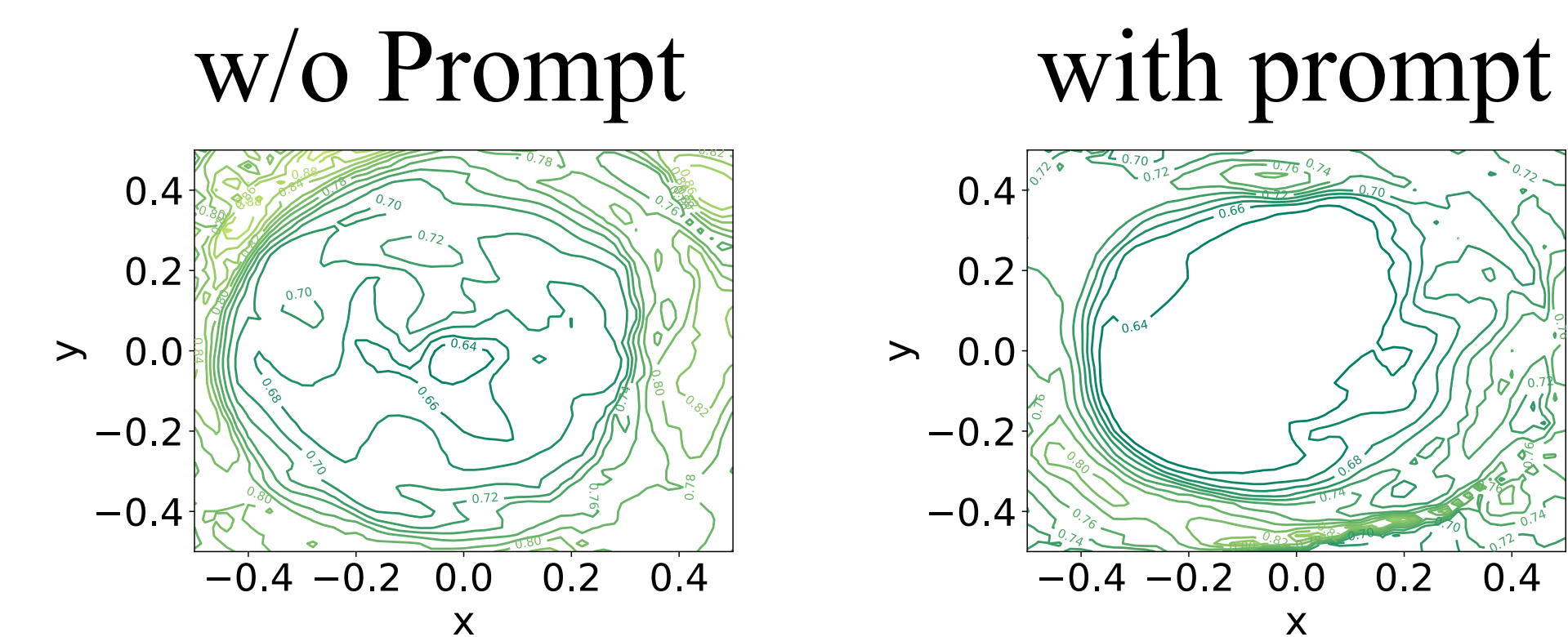
$$\mathcal{L} = \mathcal{L}_{\text{tIoU}} + \beta_1 \mathcal{L}_{\text{dis}} + \beta_2 \mathcal{L}_{\text{dur}}$$

$$\mathcal{L}_{\text{tIoU}} = \left(1 - \frac{|\hat{\mathbf{T}}(\theta) \cap \mathbf{T}|}{|\hat{\mathbf{T}}(\theta) \cup \mathbf{T}|} \right)$$

$$\mathcal{L}_{\text{dur}} = \max \left(\frac{|\mathbf{T} - \hat{\mathbf{T}}(\theta)|}{|\mathbf{T}|}, \alpha_2 \right)$$

$$\mathcal{L}_{\text{dis}} = \max \left(\frac{|(t_{\text{sta}} + t_{\text{end}})/2 - (\hat{t}_{\text{sta}} + \hat{t}_{\text{end}})/2|}{|\hat{\mathbf{T}} \cup \mathbf{T}|}, \alpha_1 \right)$$

Loss Landscape Analysis



Overall Results

Charades-STA

Type	Method	Visual Feature	Acc(R@1, IoU=m)		
			m=0.3	m=0.5	m=0.7
3D TVG	CTRL [14]	C3D	-	23.63	8.89
	ABLR [67]	C3D	-	24.36	9.01
	BPNet [62]	C3D	55.46	38.25	20.51
	LPNet [61]	C3D	59.14	40.94	21.13
	QSPN [64]	C3D	54.70	35.60	15.80
	TSP-PRL [60]	C3D	-	45.45	24.75
	TripNet [18]	C3D	54.64	38.29	16.07
	DRN [69]	C3D	-	45.40	26.40
	CPNet [34]	C3D	-	40.32	22.47
	DEBUG [43]	C3D	54.95	37.39	17.92
	ExCL [16]	I3D	61.50	44.1	22.40
	VSLNet [73]	I3D	64.30	47.31	30.19
	MAN [71]	I3D	-	46.53	22.72
2D TVG	MCN [1]	VGG	-	17.46	8.01
	SAP [7]	VGG	-	27.42	13.36
Ours					
TVP-Based 2D TVG	Base w/o prompts	ResNet	61.29	40.43	19.89
	Base + Visual Prompts		65.38	44.31	20.22
	Base + Text Prompts		65.81	43.44	20.65
	Base + Both Prompts		65.92	44.39	21.51

ActivityNet Captions

Type	Method	Visual Feature	Acc(R@1, IoU=m)		
			m=0.3	m=0.5	m=0.7
3D TVG	CTRL [14]	C3D	28.70	14.00	-
	BPNet [62]	C3D	59.98	42.07	24.69
	LPNet [61]	C3D	64.29	45.92	25.39
	QSPN [64]	C3D	45.30	27.70	13.60
	TSP-PRL [60]	C3D	56.02	38.83	-
	TripNet [18]	C3D	48.42	32.19	13.93
	DRN [69]	C3D	-	45.45	24.36
	CPNet [34]	C3D	-	40.56	21.63
	ABLR [67]	C3D	55.67	36.79	-
	DEBUG [43]	C3D	55.91	39.72	-
	ExCL [16]	C3D	63.00	43.60	24.10
	VSLNet [73]	C3D	63.16	43.22	26.16
Ours					
TVP-Based 2D TVG	Base w/o prompts	ResNet	57.20	40.16	19.14
	Base + Visual Prompts		60.12	43.39	23.71
	Base + Text Prompts		60.48	42.58	24.39
	Base + Both Prompts		60.71	43.44	25.03