Text-Visual Prompting for Efficient 2D Temporal Video Grounding (WED-PM-233)

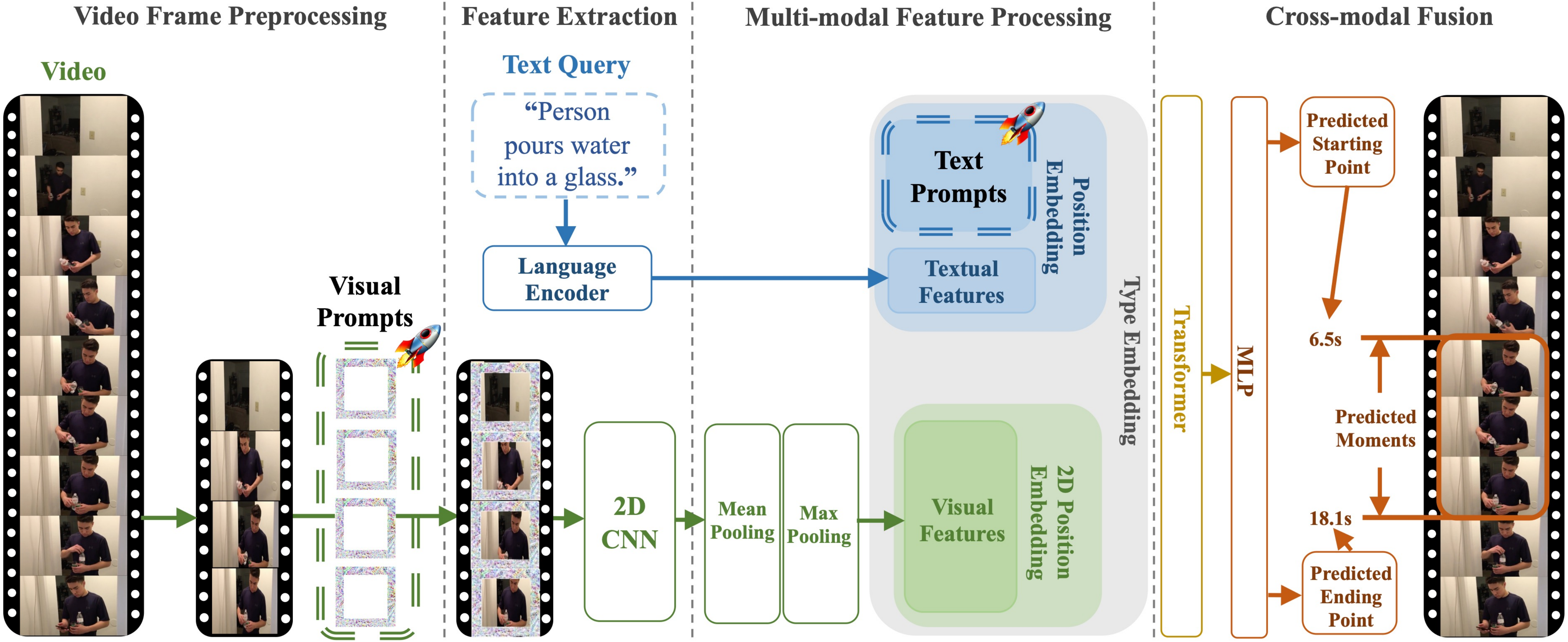
Yimeng Zhang[1,2], Xin Chen[2], Jinghan Jia[1], Sijia Liu[1], Ke Ding[2]

[1] OPTML Lab, Michigan State University
[2] Applied ML, Intel

intel

OPTML
WE MAKE THE GLOBAL OPTIMAL

1. [Summary]
Text-Visual Prompting (TVP) for Efficient 2D Temporal Video Grounding (TVG)



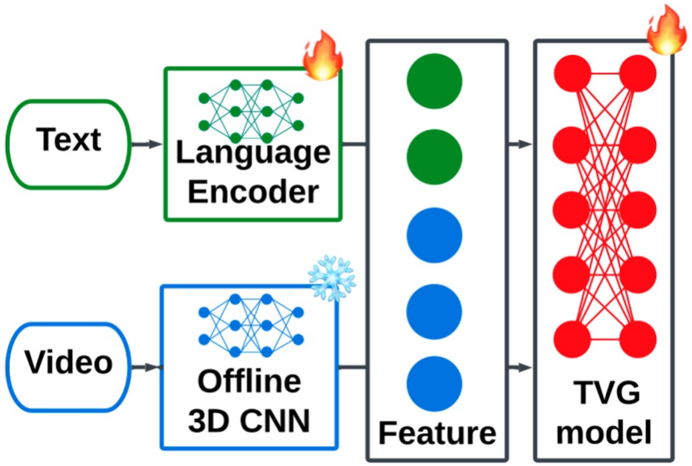
2. Temporal Video Grounding (TVG) Method Comparison

2. Temporal Video Grounding (TVG) Method Comparison

Trainable Modules

Frozen Modules

Performance Booster

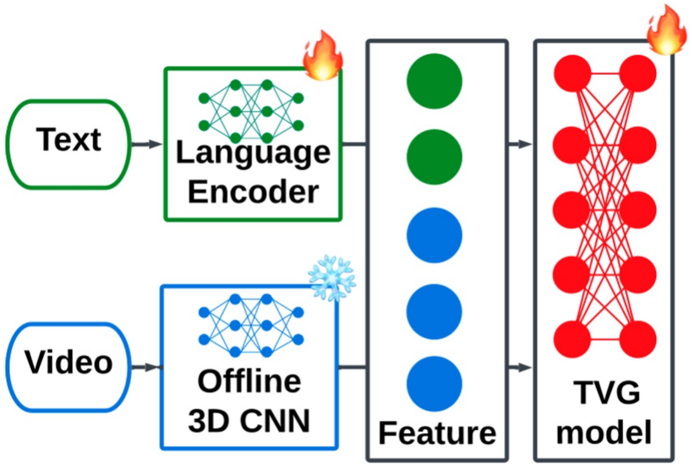


(a) 3D TVG

3D TVG

2. Temporal Video Grounding (TVG) Method Comparison

🔥 Trainable Modules ❄️ Frozen Modules 🚀 Performance Booster



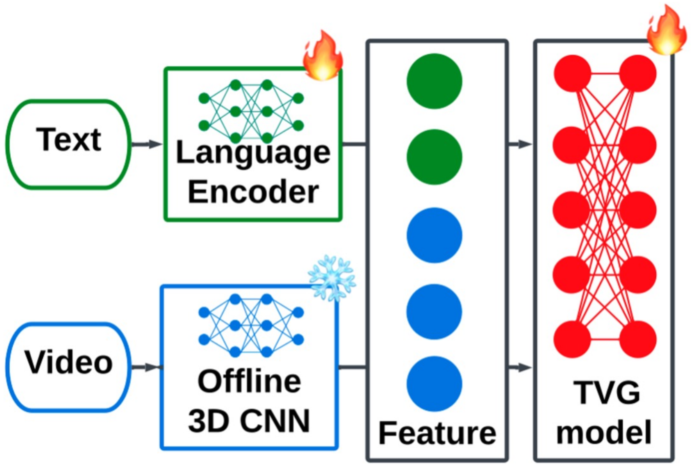
(a) 3D TVG

3D TVG

➢ Using <u>offline 3D CNN</u> as the video encoder.

2. Temporal Video Grounding (TVG) Method Comparison

Trainable Modules Frozen Modules Performance Booster



(a) 3D TVG

3D TVG

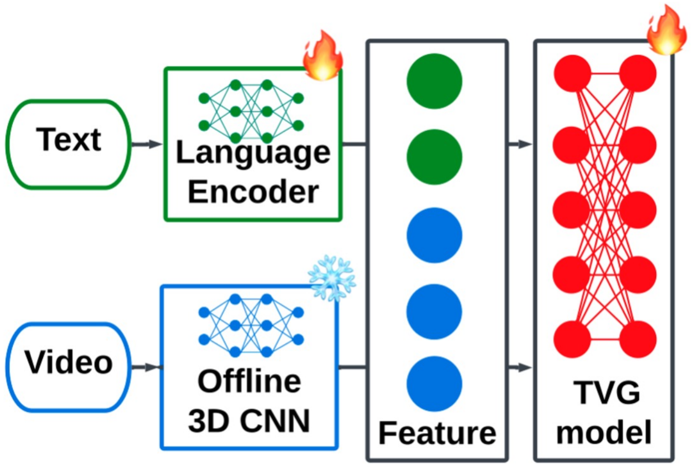
- ➢ Using offline 3D CNN as the video encoder.
- ➢ During training, **3D-CNN parameters are fixed**, which means modules for text and video processing **cannot be co-trained** for better multimodal feature fusion.

2. Temporal Video Grounding (TVG) Method Comparison

Trainable Modules

Frozen Modules

Performance Booster



(a) 3D TVG

3D TVG

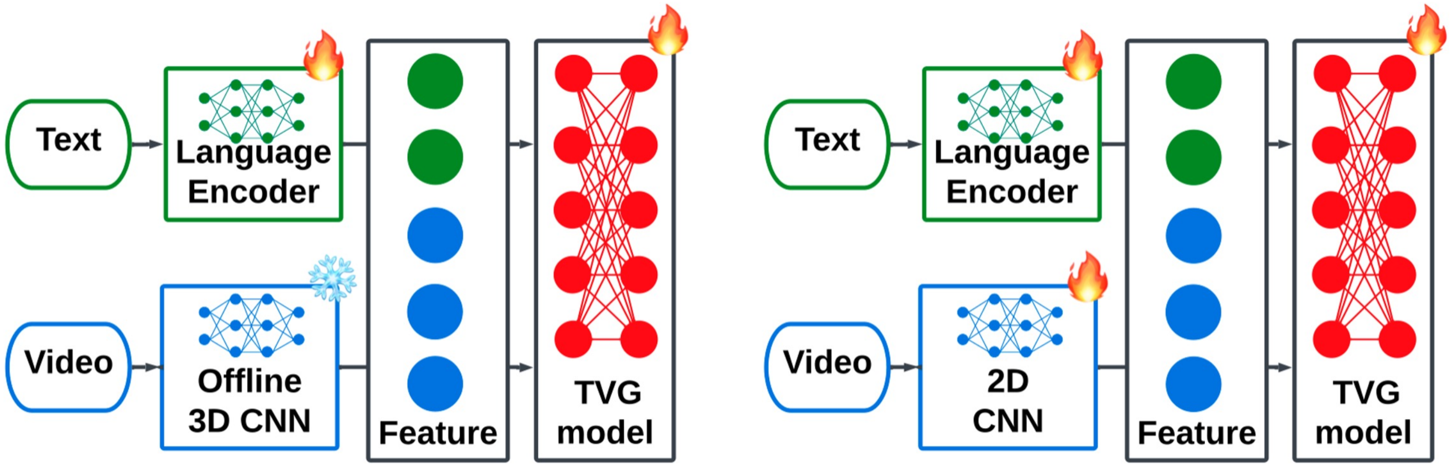
- ➢ Using offline 3D CNN as the video encoder.
- ➢ During training, **3D-CNN parameters are fixed**, which means modules for text and video processing **cannot be co-trained** for better multimodal feature fusion.
- ➢ It is challenging to train 3D-CNNs, which is why most methods **do not involve 3D-CNNs during training** and **directly utilize the video features** extracted by offline 3D-CNNs as the video input.

2. Temporal Video Grounding (TVG) Method Comparison

Trainable Modules

Frozen Modules

Performance Booster



(a) 3D TVG

(b) 2D TVG

3D TVG

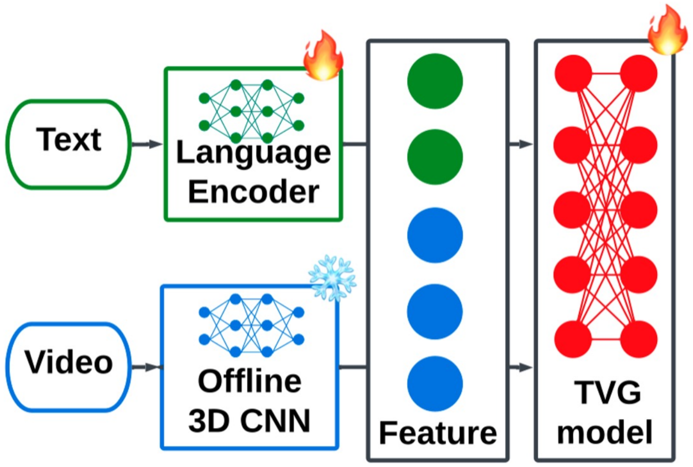
- ➢ Using offline 3D CNN as the video encoder.
- ➢ During training, **3D-CNN parameters are fixed**, which means modules for text and video processing **cannot be co-trained** for better multimodal feature fusion.
- ➢ It is challenging to train 3D-CNNs, which is why most methods **do not involve 3D-CNNs during training** and **directly utilize the video features** extracted by offline 3D-CNNs as the video input.

2D TVG

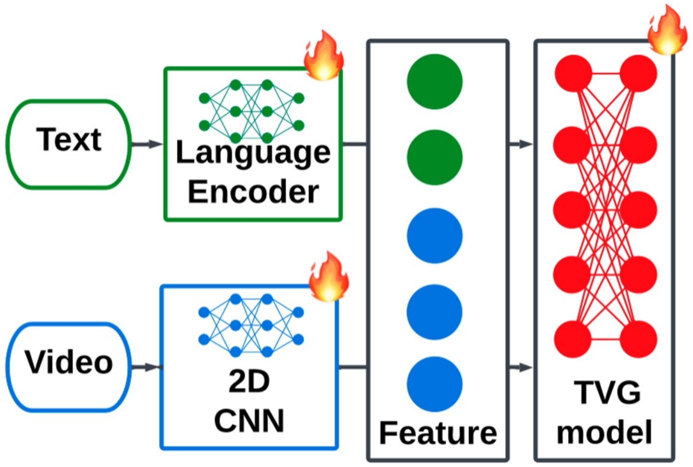
- ➢ Using 2D CNN as the video encoder.

2. Temporal Video Grounding (TVG) Method Comparison

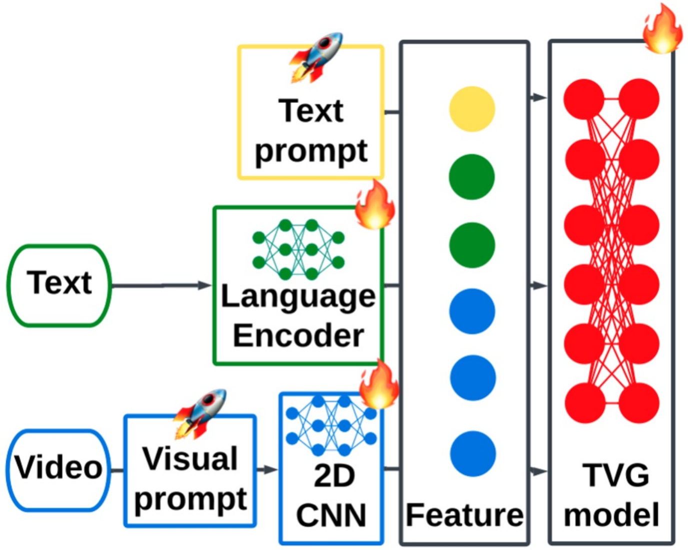
Trainable Modules Frozen Modules Performance Booster



(a) 3D TVG



(b) 2D TVG



(c) TVP-based 2D TVG (Ours)

3D TVG

- Using offline 3D CNN as the video encoder.
- During training, **3D-CNN parameters are fixed**, which means modules for text and video processing **cannot be co-trained** for better multimodal feature fusion.
- It is challenging to train 3D-CNNs, which is why most methods **do not involve 3D-CNNs during training** and **directly utilize the video features** extracted by offline 3D-CNNs as the video input.

2D TVG

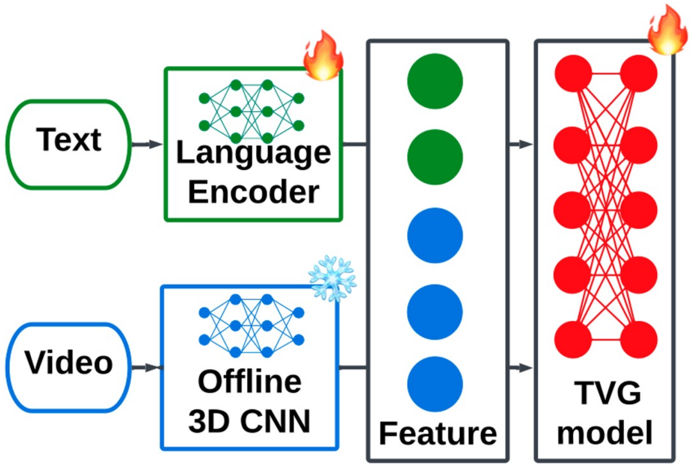
- Using 2D CNN as the video encoder.

TVP-Based 2D TVG

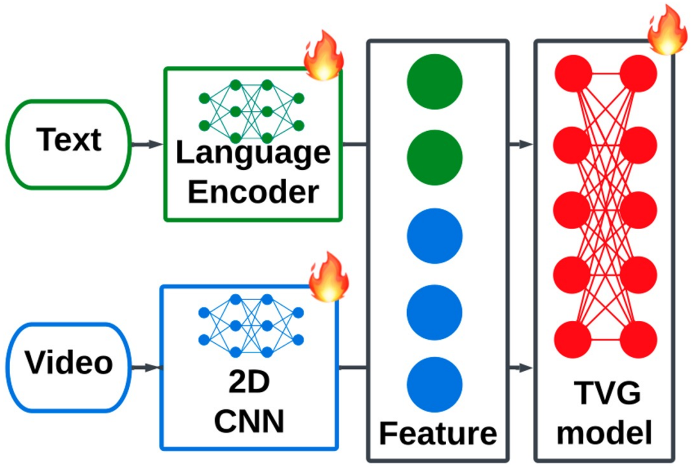
- The proposed **text-visual prompts (TVP)** compensate for the lack of spatiotemporal information in 2D CNNs for visual feature extraction.

2. Temporal Video Grounding (TVG) Method Comparison

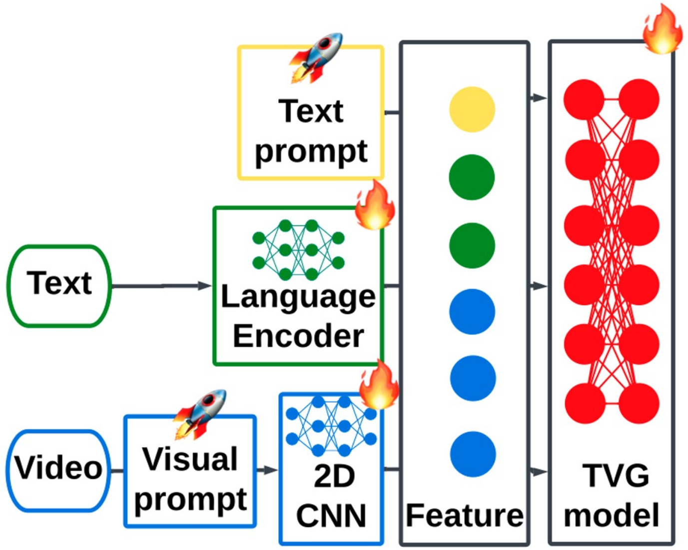
Trainable Modules   Frozen Modules   Performance Booster



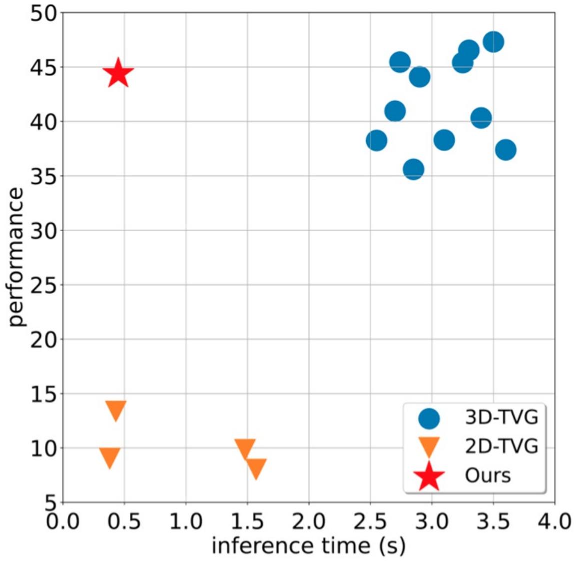
(a) 3D TVG



(b) 2D TVG



(c) TVP-based 2D TVG (Ours)



(d) Overall performance comparison

3D TVG

- ➢ Using <u>offline 3D CNN</u> as the video encoder
- ➢ During training, **3D-CNN parameters are fixed**, which means modules for text and video processing **cannot be co-trained** for better multimodal feature fusion.
- ➢ It is challenging to train 3D-CNNs, which is why most methods **do not involve 3D-CNNs during training** and **directly utilize the video features** extracted by offline 3D-CNNs as the video input.

2D TVG

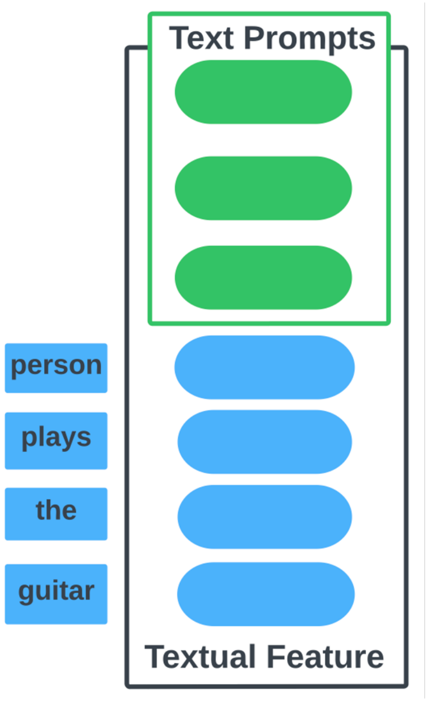
- ➢ Using <u>2D CNN</u> as the video encoder.

TVP-Based 2D TVG

- ➢ The proposed **text-visual prompts (TVP)** compensate for the lack of spatiotemporal information in 2D CNNs for visual feature extraction.

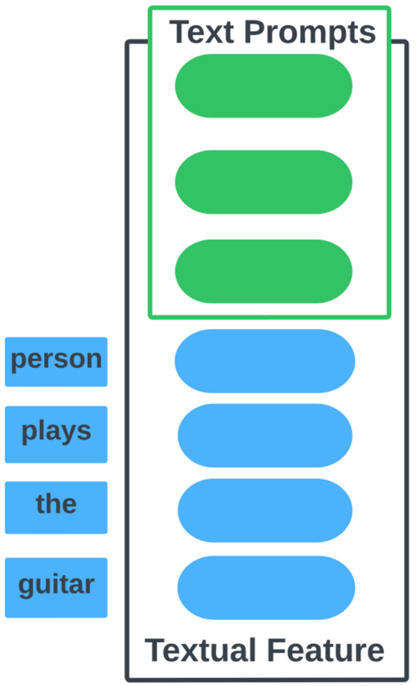
3. Text Prompts and Visual Prompts

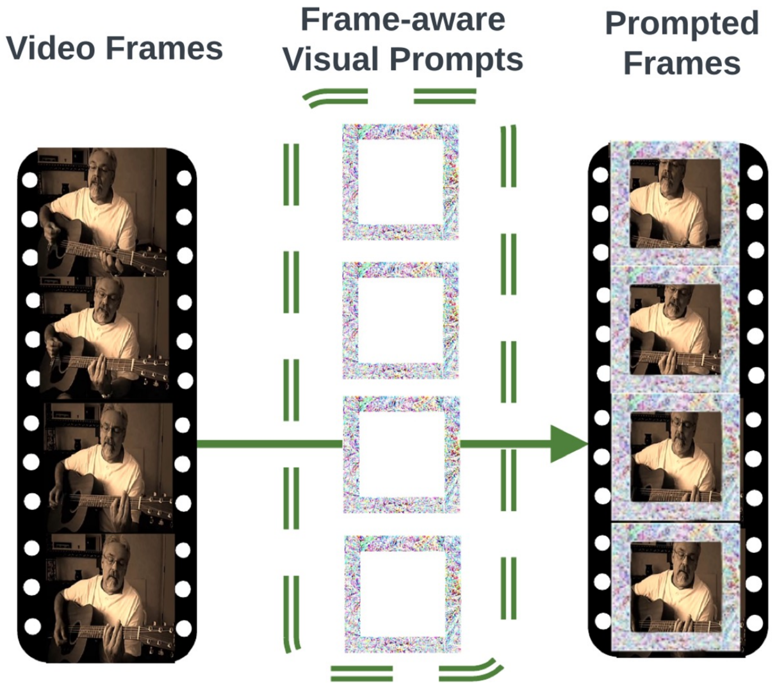
3. Text Prompts and Visual Prompts



➢ Text prompts are directly applied in the feature space.

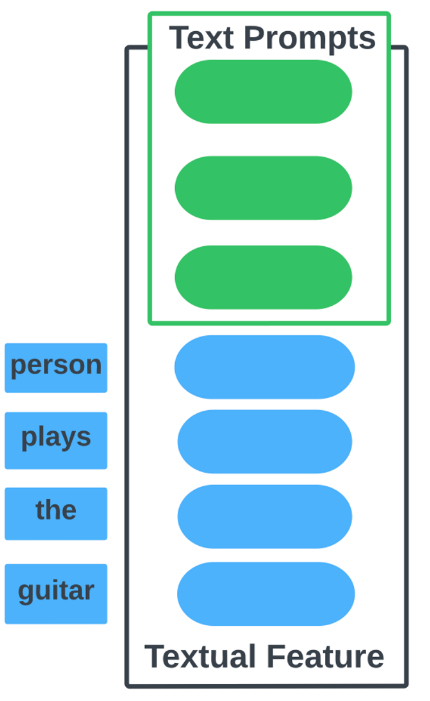
3. Text Prompts and Visual Prompts

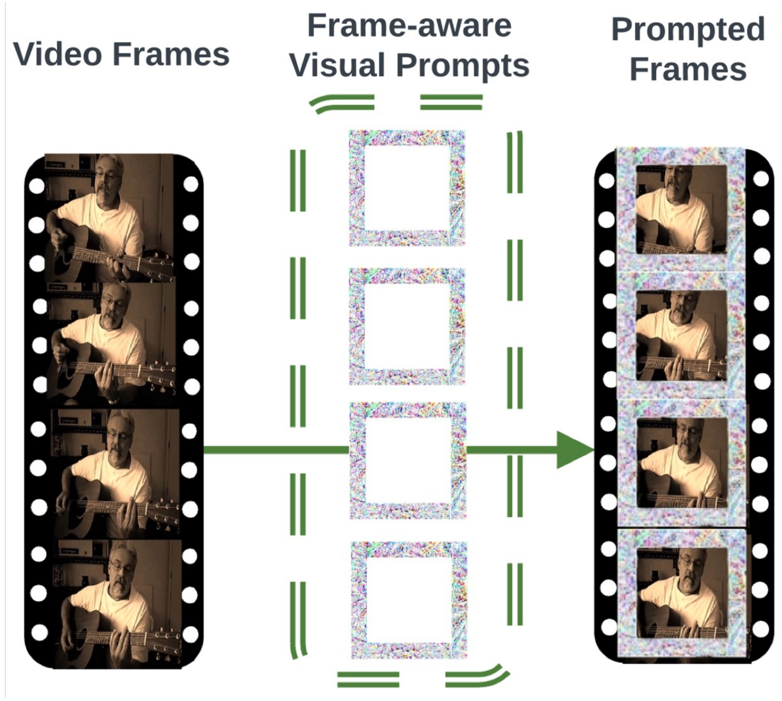




- ➢ Text prompts are directly applied in the feature space.
- ➢ A set of frame-aware visual prompts are applied to pixel space of video frames in order.

3. Text Prompts and Visual Prompts

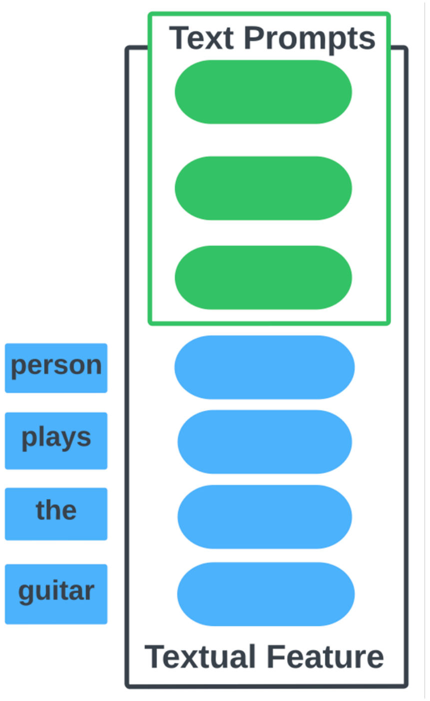


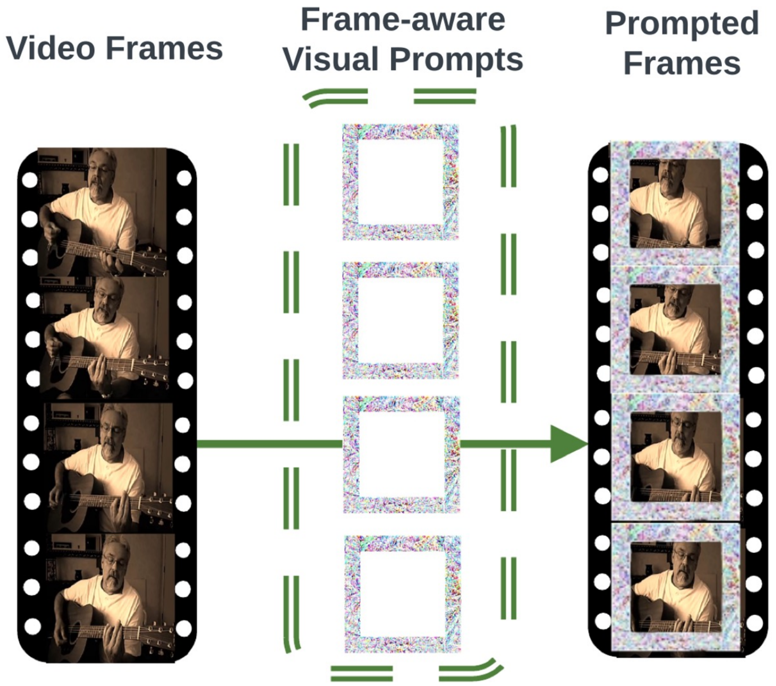


- ➢ Text prompts are directly applied in the feature space.
- ➢ A set of frame-aware visual prompts are applied to pixel space of video frames in order.

- ➢ During training, only the set of visual prompts and text prompts are updated through backpropagation.

3. Text Prompts and Visual Prompts

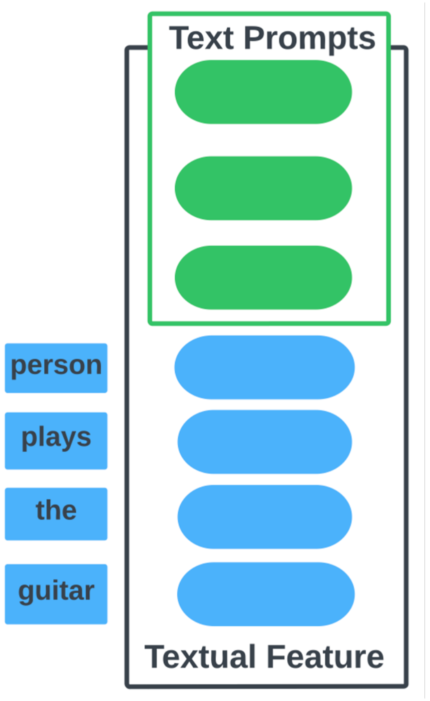


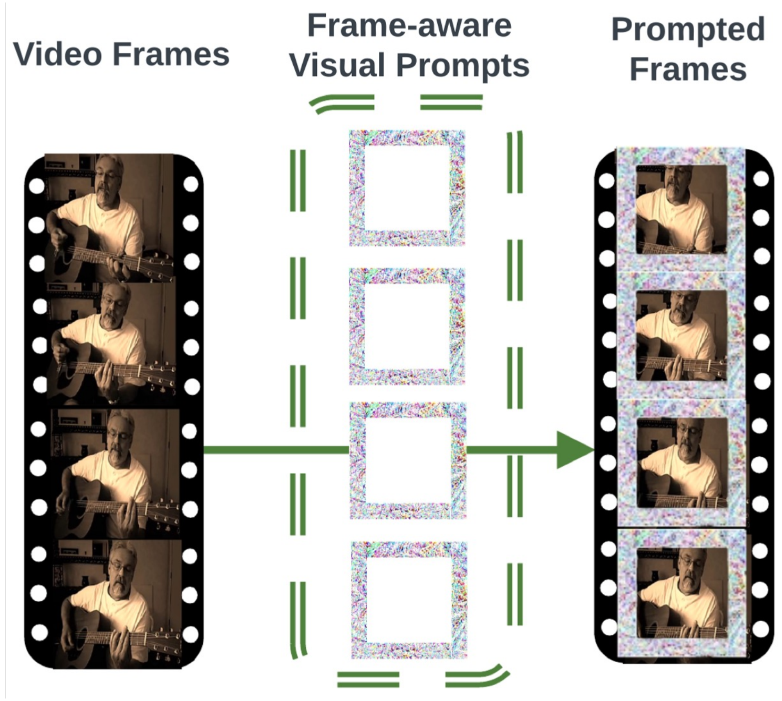


- ➢ Text prompts are directly applied in the feature space.
- ➢ A set of frame-aware visual prompts are applied to pixel space of video frames in order.

- ➢ During <u>training</u>, only the set of visual prompts and text prompts are updated through backpropagation.
- ➢ During <u>finetuning</u>, prompts are frozen, and the parameters of the TVG model and encoders are updated.

3. Text Prompts and Visual Prompts

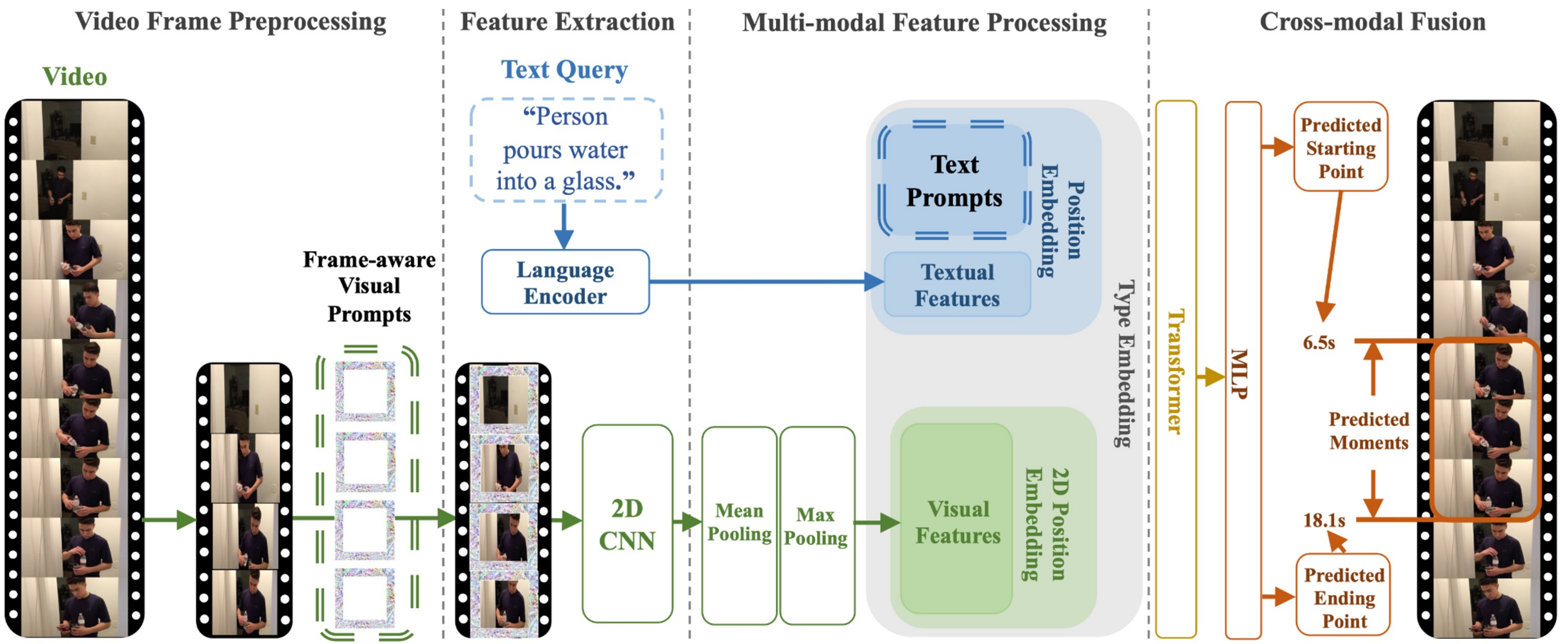




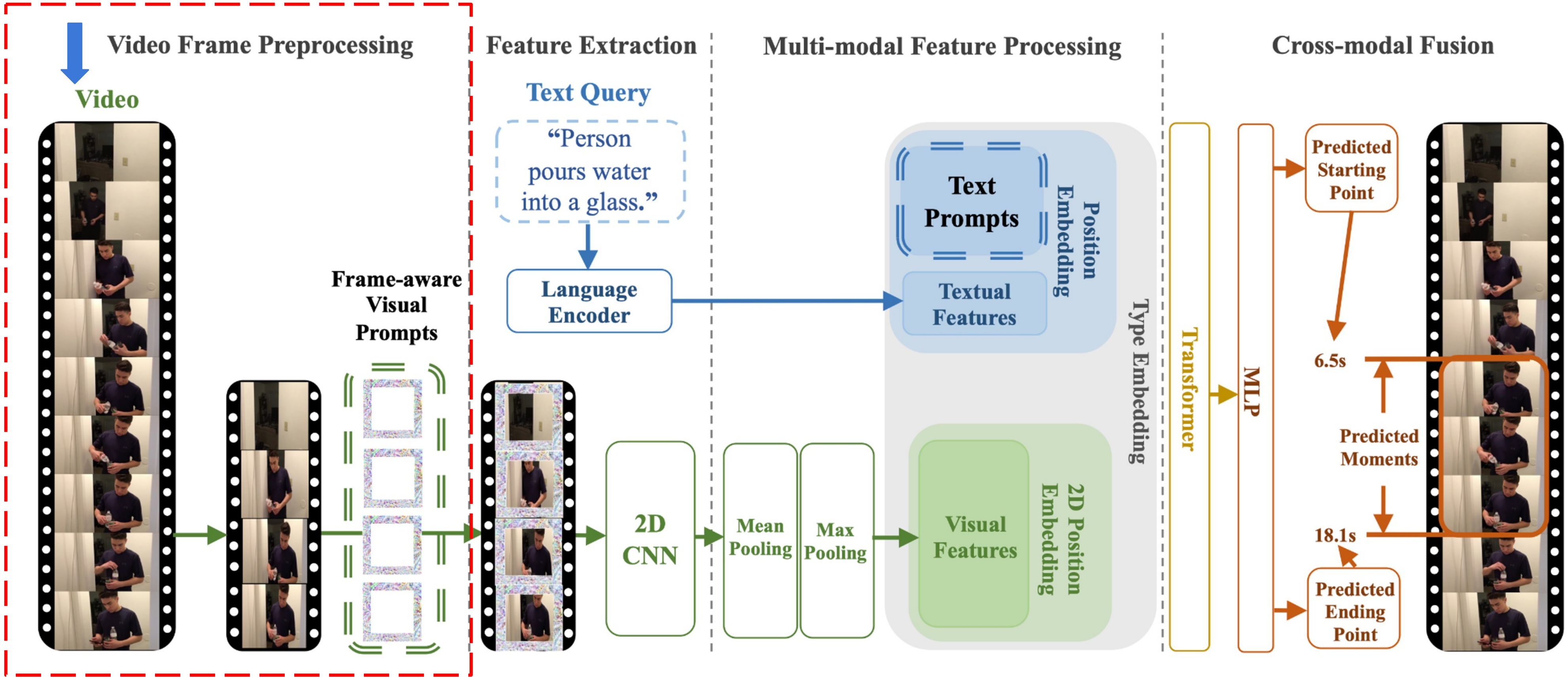
- ➢ Text prompts are directly applied in the feature space.
- ➢ A set of frame-aware visual prompts are applied to pixel space of video frames in order.

- ➢ During <u>training</u>, only the set of visual prompts and text prompts are updated through backpropagation.
- ➢ During <u>finetuning</u>, prompts are frozen, and the parameters of the TVG model and encoders are updated.
- ➢ During <u>testing</u>, the set of optimized visual prompts and the optimized text prompts are **applied to all test-time video-query pairs.**

4. Text-Visual Prompt for 2D TVG

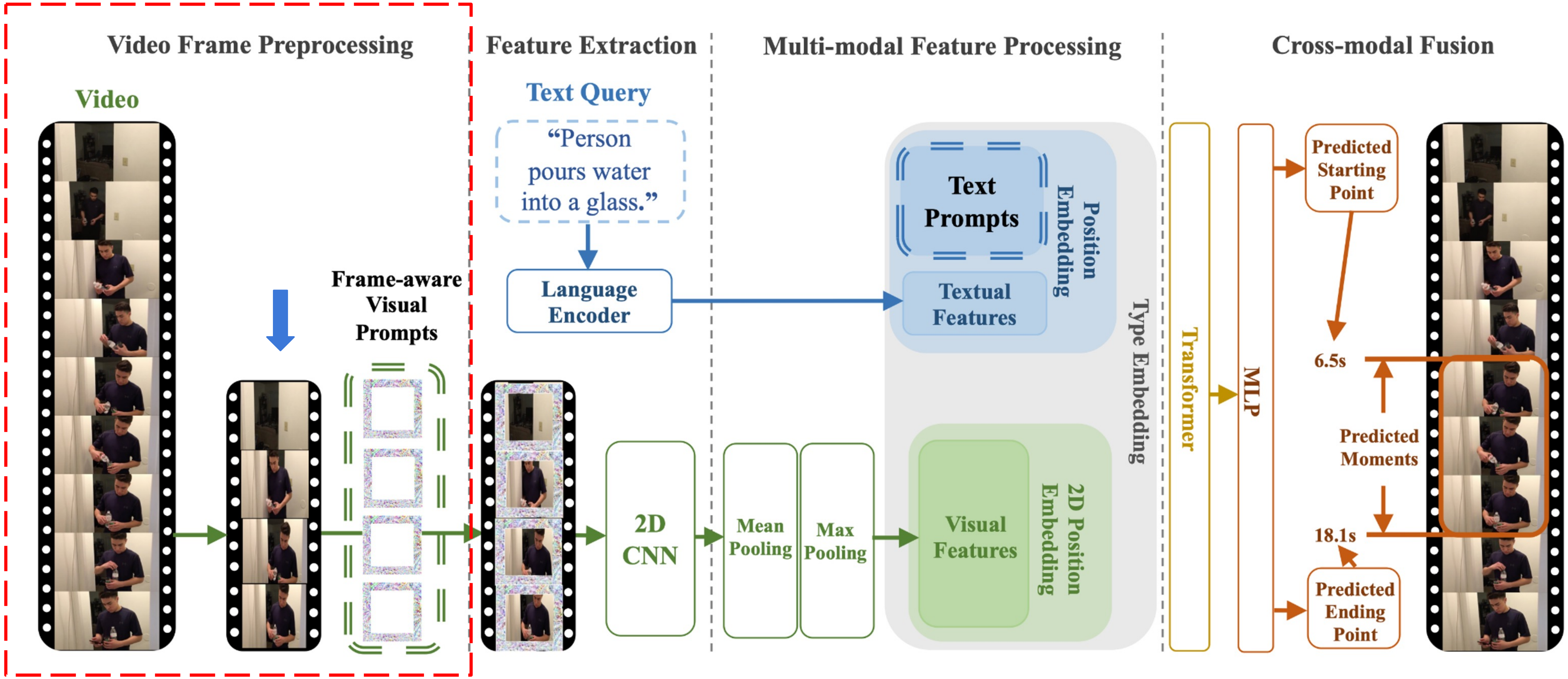


4. Text-Visual Prompt for 2D TVG



Video frame preprocessing

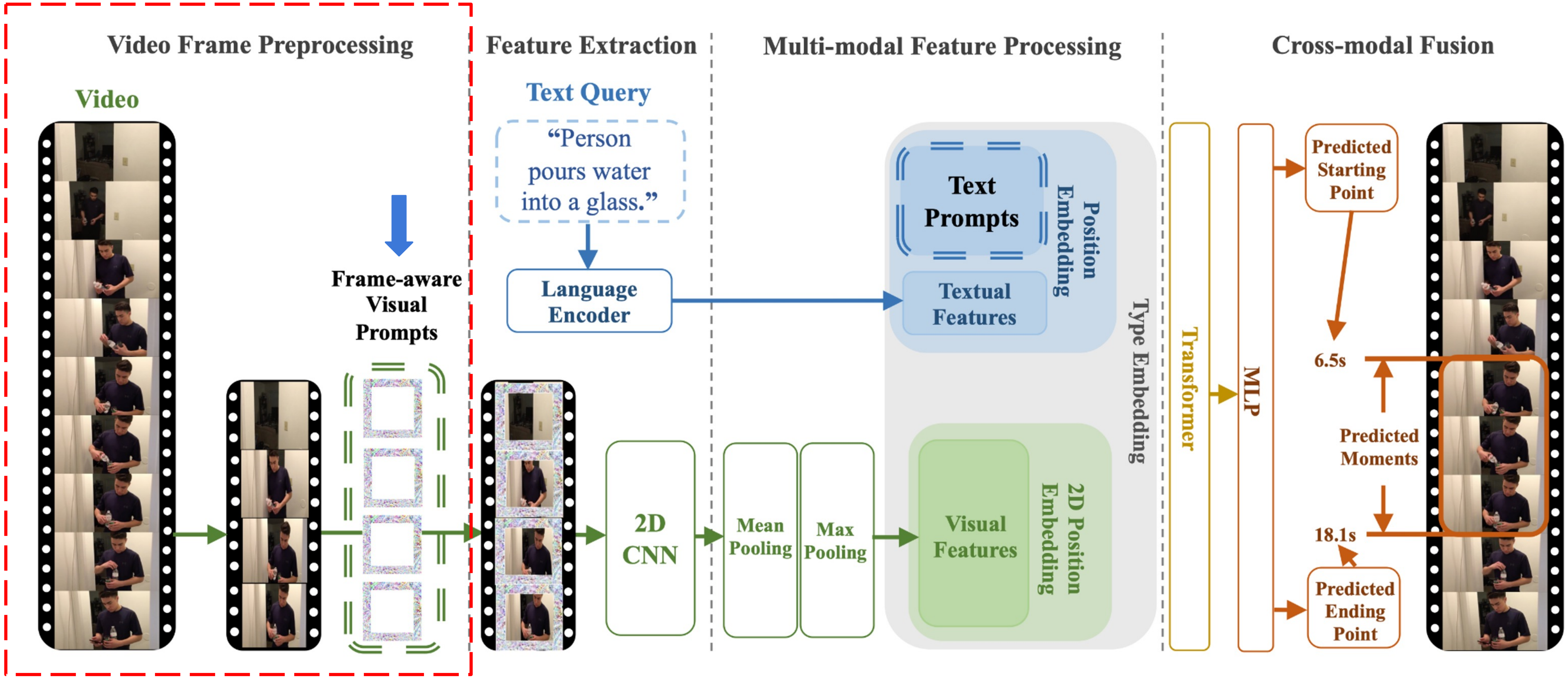
4. Text-Visual Prompt for 2D TVG



Video frame preprocessing

1) Uniformly sample frames from input video.

4. Text-Visual Prompt for 2D TVG

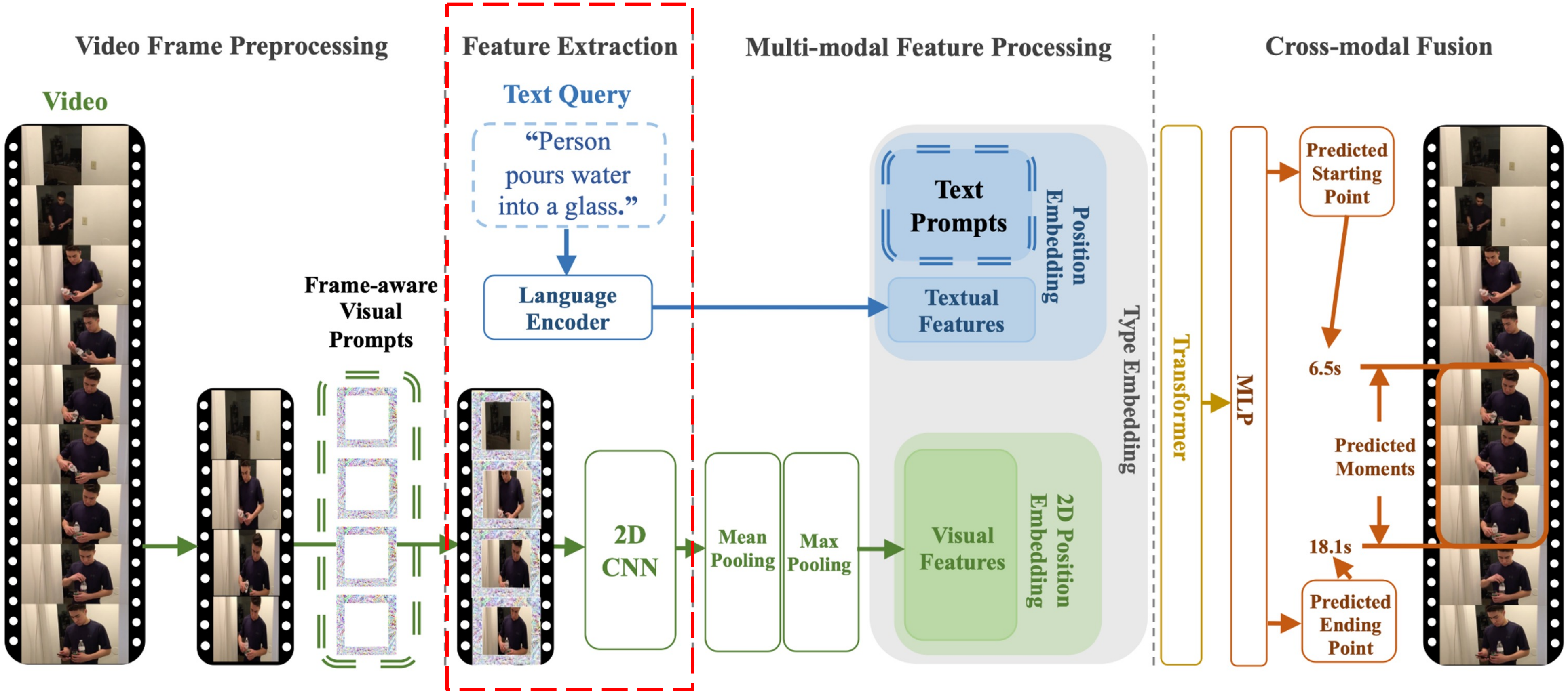


Video frame preprocessing

1) Uniformly sample frames from input video.

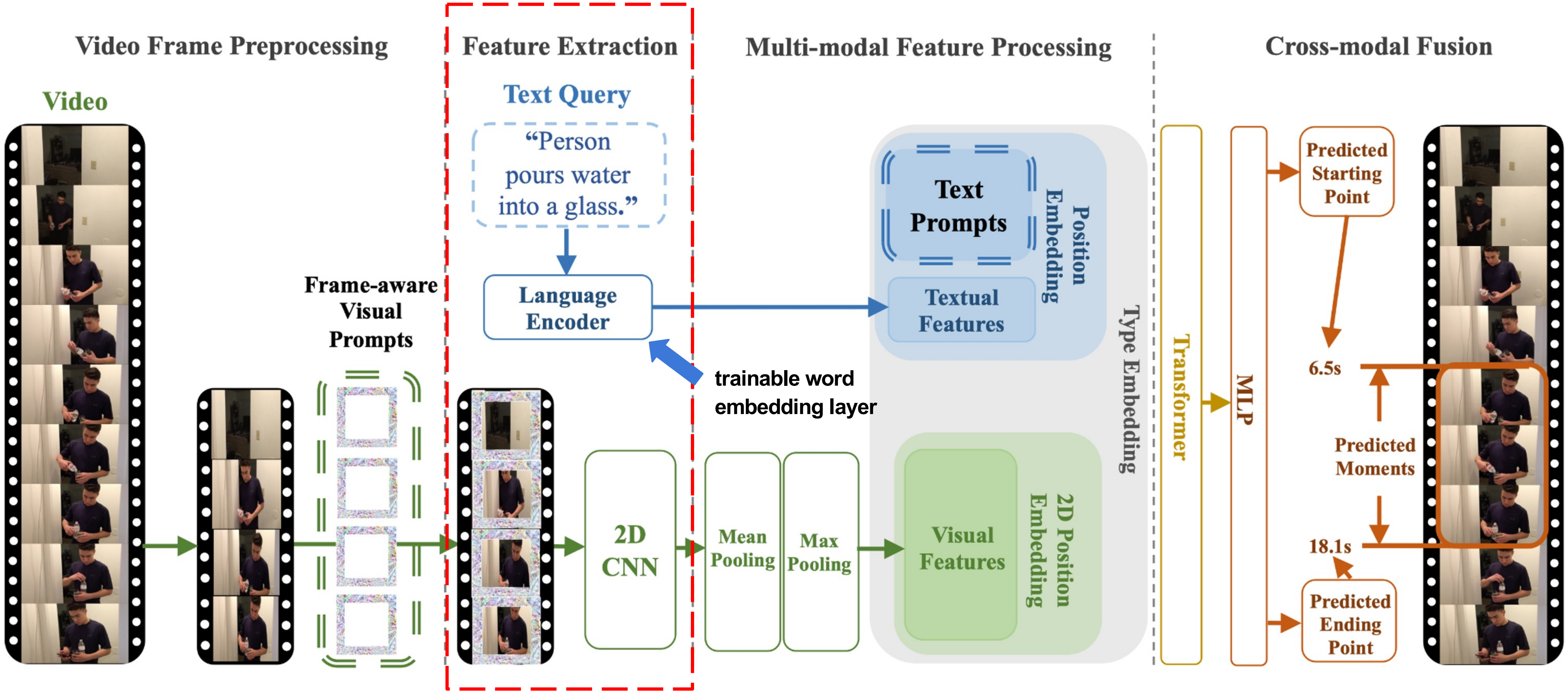
2) Apply an set of frame-aware visual prompts to the sampled frames in order.

4. Text-Visual Prompt for 2D TVG



Feature extraction

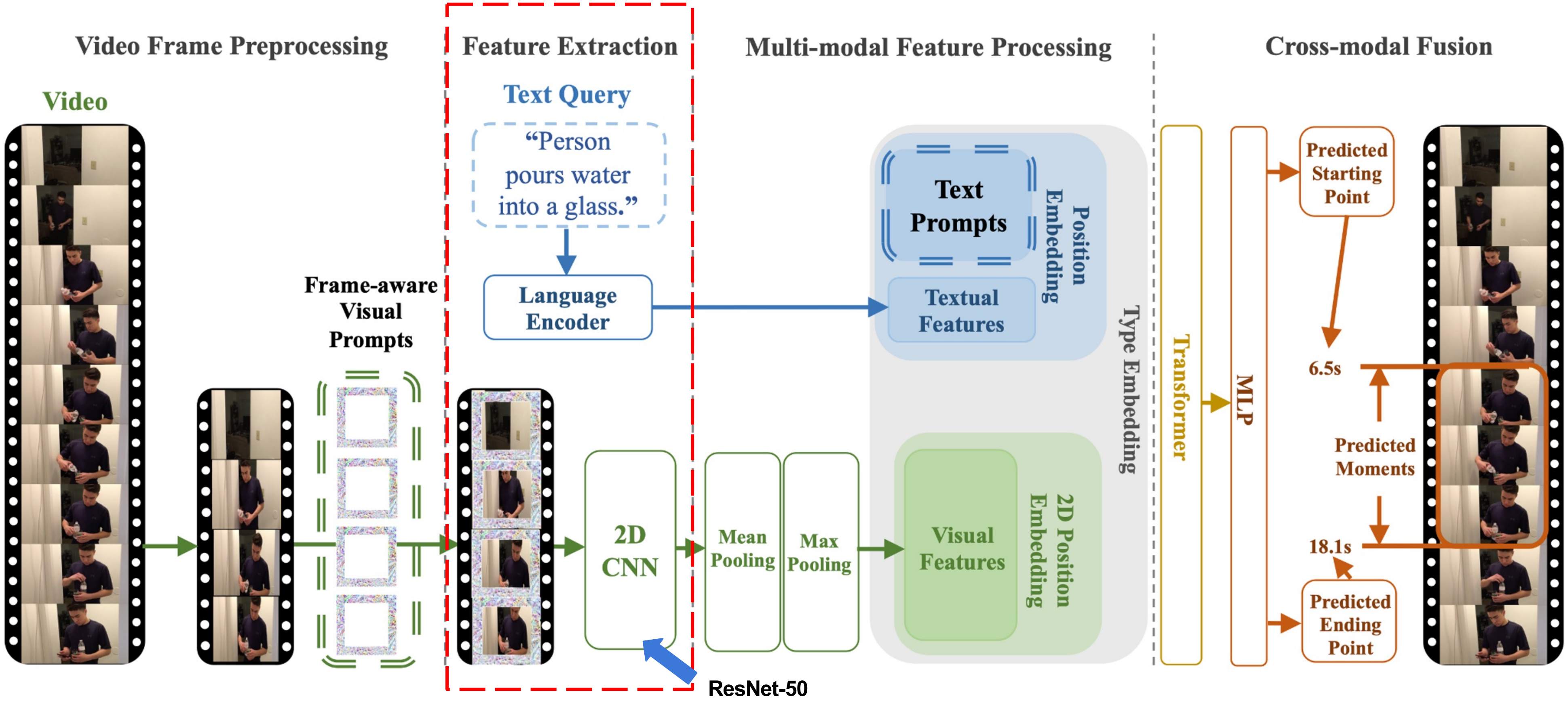
4. Text-Visual Prompt for 2D TVG



Feature extraction

1) The <u>language encoder</u> extracts textual features.

4. Text-Visual Prompt for 2D TVG

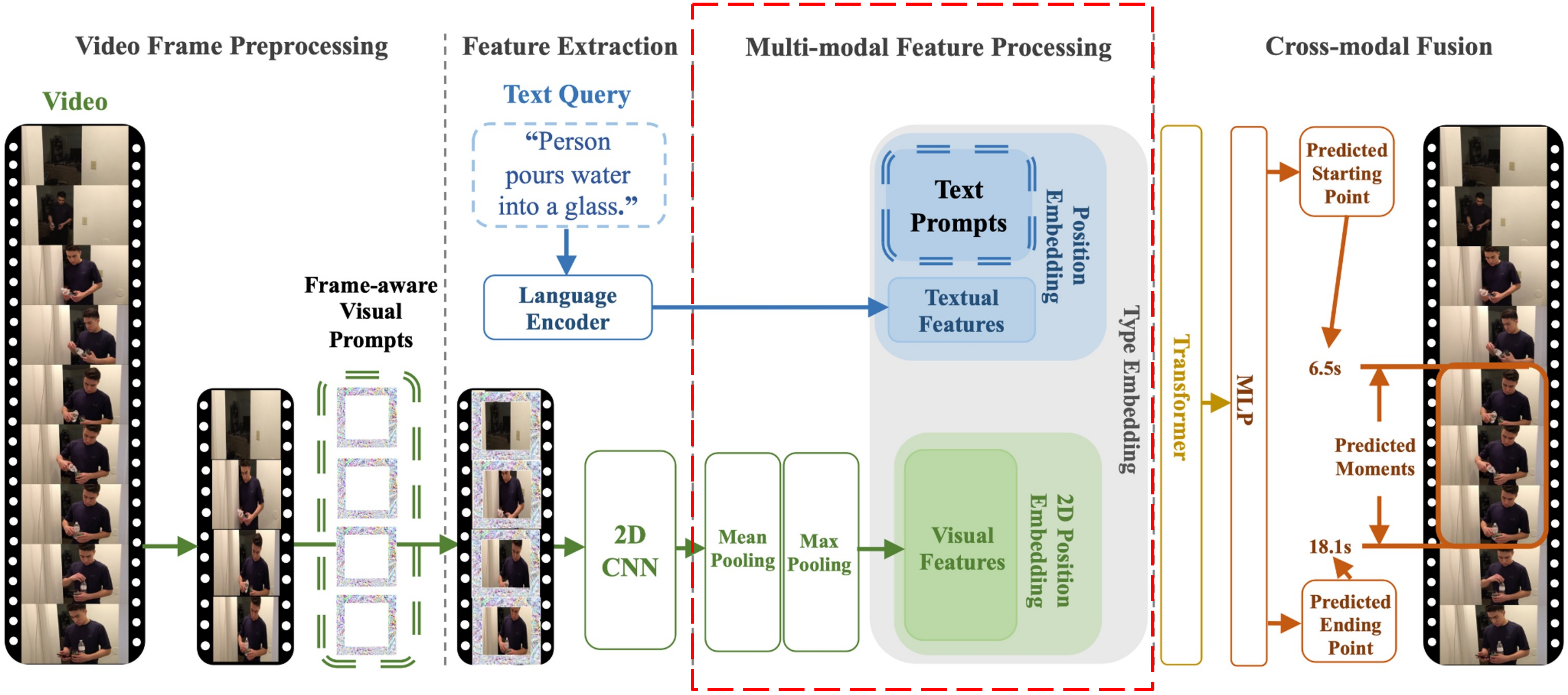


Feature extraction

1) The language encoder extracts textual features.

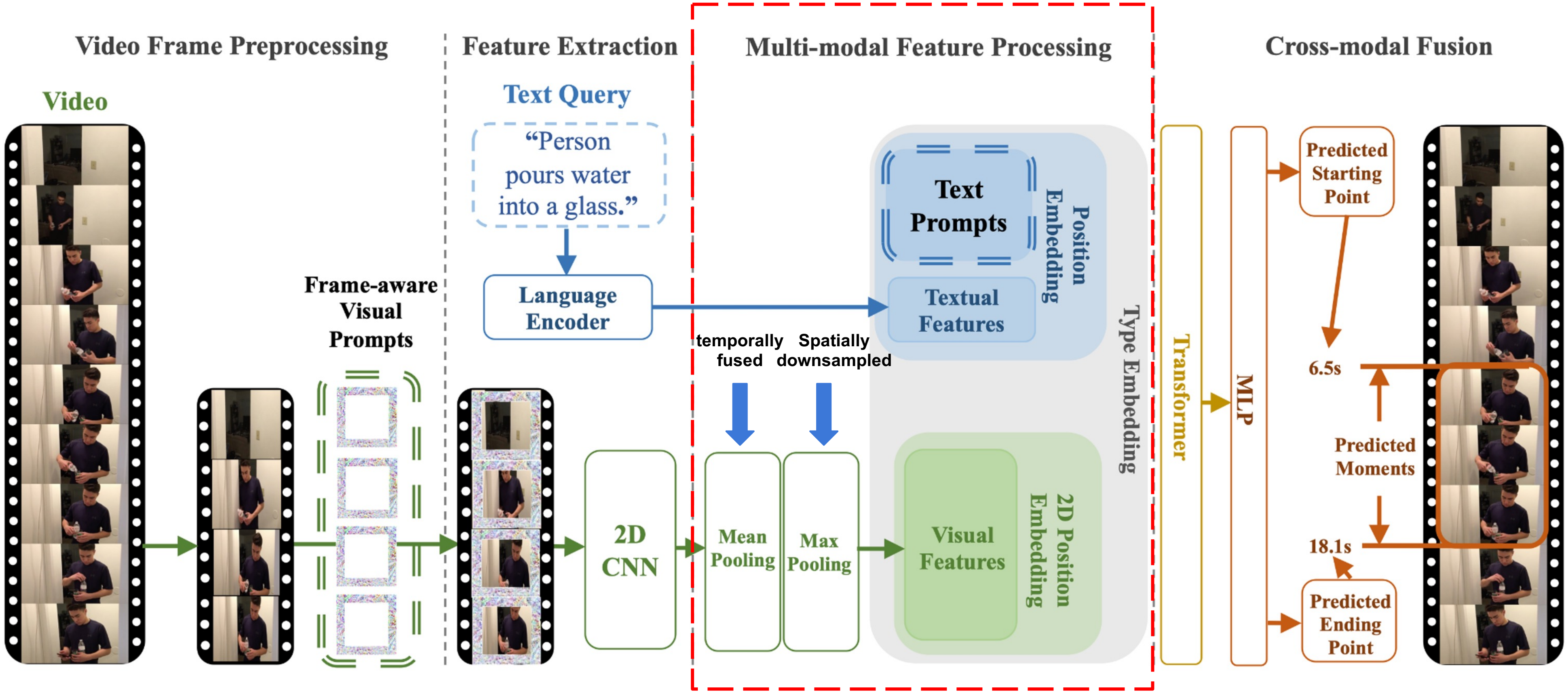
2) 2D CNN extracts features from sampled video frames with visual prompts.

4. Text-Visual Prompt for 2D TVG



Multimodal feature processing

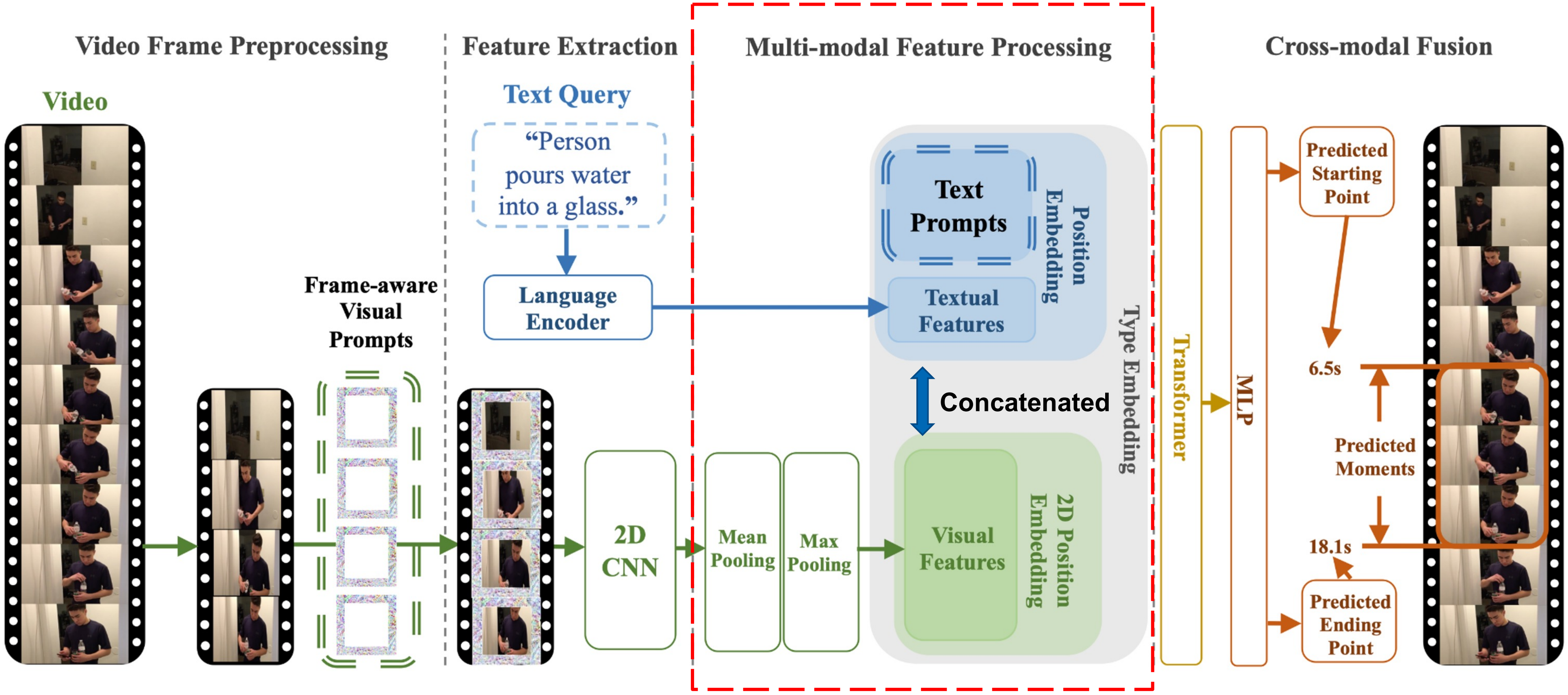
4. Text-Visual Prompt for 2D TVG



Multimodal feature processing

1) Visual features would be <u>temporally fused and spatially downsampled</u> by mean pooling and max pooling, respectively.

4. Text-Visual Prompt for 2D TVG

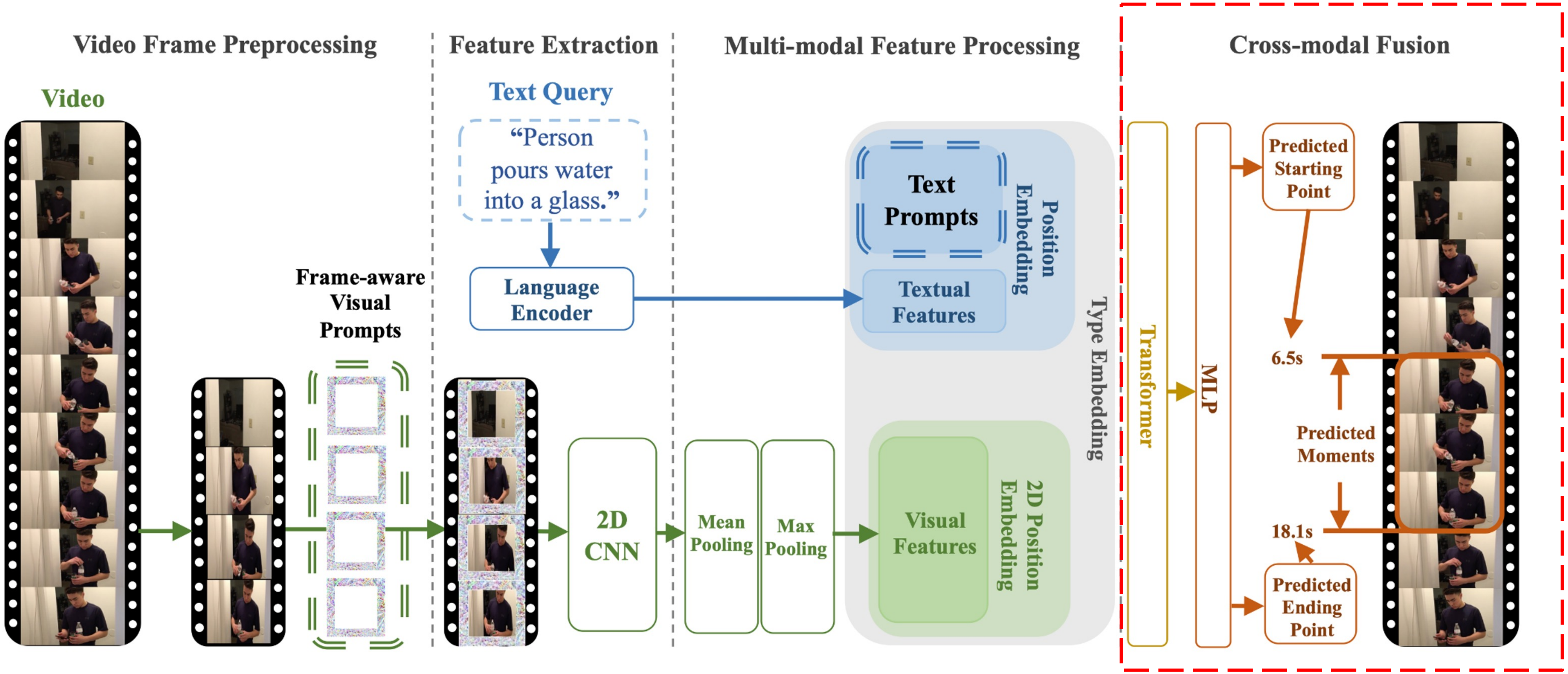


Multimodal feature processing

1) Visual features would be <u>temporally fused and spatially downsampled</u> by mean pooling and max pooling, respectively.

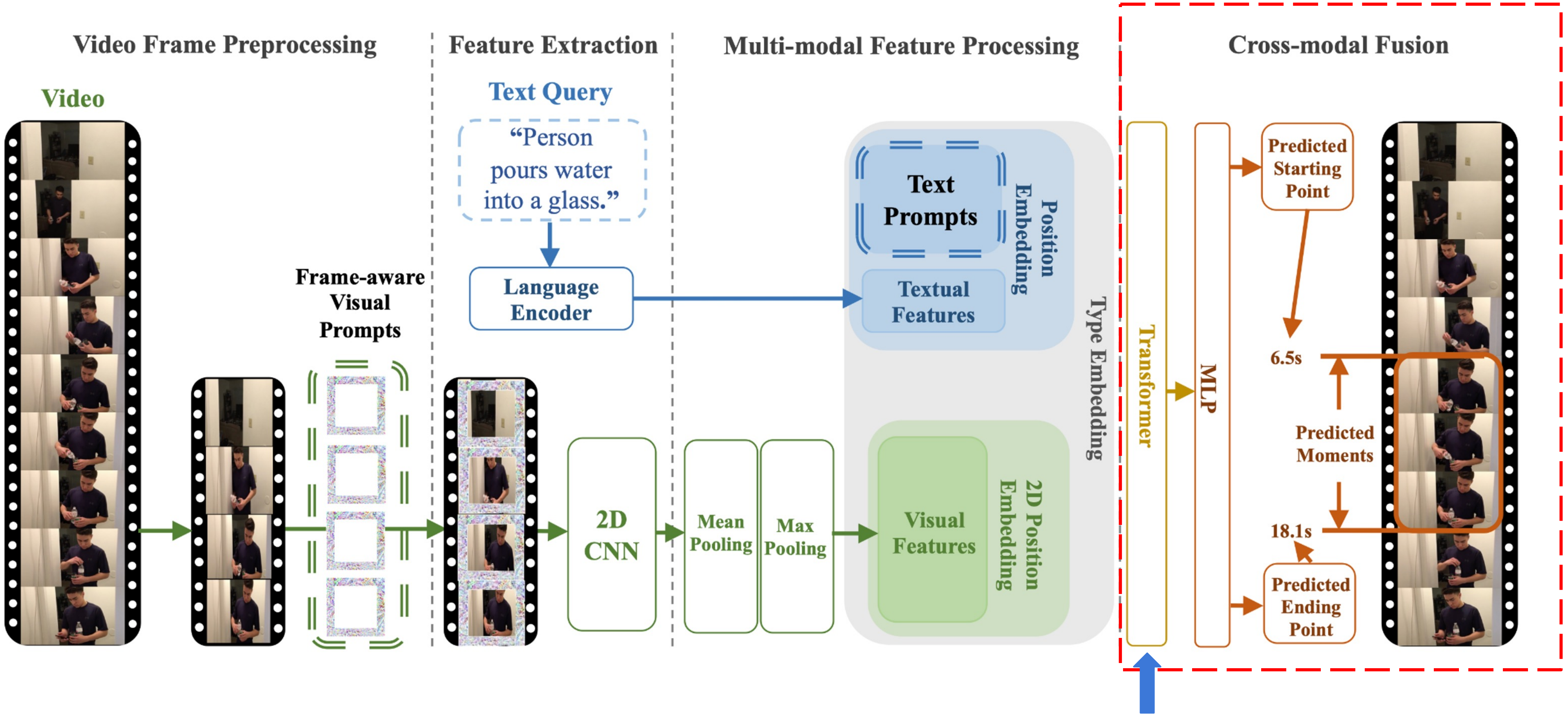
2) The 2D visual features would be <u>concatenated</u> with textual features and text prompts.

4. Text-Visual Prompt for 2D TVG



Crossmodal fusion

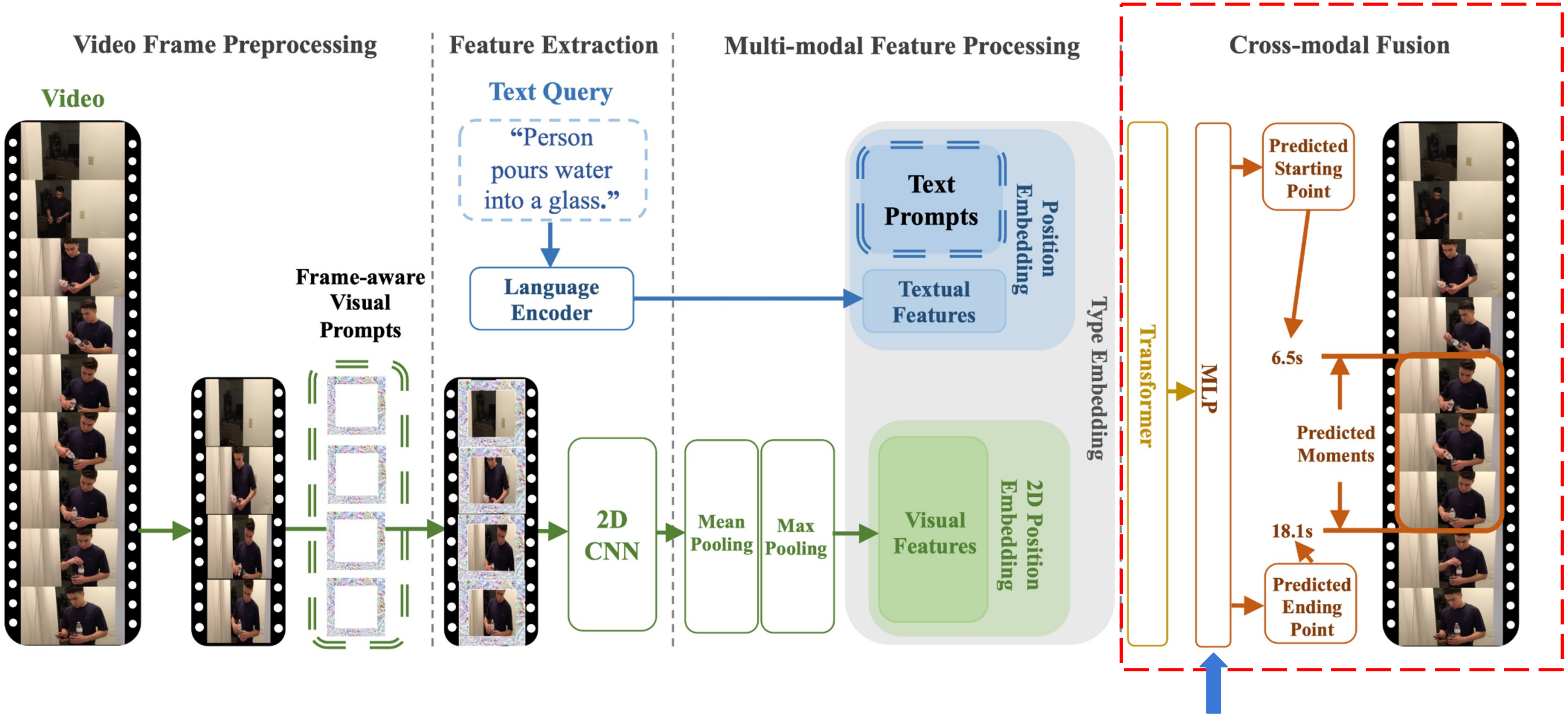
4. Text-Visual Prompt for 2D TVG



Crossmodal fusion

1) The multimodal features would be processed by a 12-layer transformer encoder,

4. Text-Visual Prompt for 2D TVG



Crossmodal fusion

1) The multimodal features would be processed by a 12-layer transformer encoder,

2) MLP would predict the starting/ending time points of the target moment.

6. Training Pipeline

6. Training Pipeline

a) **Cross-modal pretraining** on large-scale image-text datasets.
(COCO Captions and Visual Genome Captions)

6. Training Pipeline

a) **Cross-modal pretraining** on large-scale image-text datasets.
(COCO Captions and Visual Genome Captions)

b) **Base model training** on the target dataset.

6. Training Pipeline

a) **Cross-modal pretraining** on large-scale image-text datasets.
(COCO Captions and Visual Genome Captions)

b) **Base model training** on the target dataset.

c) **Prompt training.** ← Base model parameter are frozen !

6. Training Pipeline

a) **Cross-modal pretraining** on large-scale image-text datasets.
(COCO Captions and Visual Genome Captions)

b) **Base model training** on the target dataset.

c) **Prompt training.** ← Base model parameter are frozen !

d) **Base model finetuning.** ← Text-Visual Prompts are frozen !

7. Dataset

Dataset	Charades-STA	ActivityNet Captions
Domain	Indoor Activity	Indoor/Outdoor Activity
# Videos	$6,672$	$14,926$
Avg. Video Length (*second*)	30.6	117.6
# Moments	$11,767$	$71,953$
Avg. Moment Length (*second*)	8.1	37.1
Vocabulary Size	$1,303$	$15,505$
# Queries	$16,124$	$71,953$
Avg. Query Length (*word*)	7.2	14.4

Table 1. Statics of temporal video grounding benchmark datasets (Charades-STA and ActivityNet Captions datasets).

8. Evaluation Metric

$$Acc(R@1, IoU=m)$$

8. Evaluation Metric

Acc(R@1, IoU=m)

The percentage of predicted moments
<u>achieving IoU higher than *m*</u>
with the groundtruth moment.

8. Experimental Results

Table 2. Performance comparison of different thresholds m on the Charades-STA dataset.

Type	Method	Visual Feature	Acc(R@1, IoU=m) m=0.3	m=0.5	m=0.7
3D TVG	BPNet [53]	C3D	55.46	38.25	20.51
	LPNet [52]	C3D	59.14	40.94	21.13
	QSPN [55]	C3D	54.70	35.60	15.80
	TSP-PRL [51]	C3D	-	45.45	24.75
	TripNet [15]	C3D	54.64	38.29	16.07
	DRN [59]	C3D	-	45.40	26.40
	CPNet [28]	C3D	-	40.32	22.47
	DEBUG [34]	C3D	54.95	37.39	17.92
	ExCL [14]	I3D	61.50	44.1	22.40
	VSLNet [63]	I3D	64.30	**47.31**	**30.19**
	MAN [61]	I3D	-	46.53	22.72
2D TVG	CTRL [12]	VGG	13.5	9.82	-
	MCN [1]	VGG	17.46	8.01	-
	ABLR [58]	VGG	24.36	9.01	-
	SAP [5]	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

Table 3. Performance comparison of different thresholds m on the ActivityNet Captions dataset.

Type	Method	Visual Feature	Acc(R@1, IoU=m) m=0.3	m=0.5	m=0.7
3D TVG	CTRL [12]	C3D	28.70	14.00	-
	BPNet [53]	C3D	59.98	42.07	24.69
	LPNet [52]	C3D	**64.29**	**45.92**	25.39
	QSPN [55]	C3D	45.30	27.70	13.60
	TSP-PRL [51]	C3D	56.02	38.83	-
	TripNet [15]	C3D	48.42	32.19	13.93
	DRN [59]	C3D	-	45.45	24.36
	CPNet [28]	C3D	-	40.56	21.63
	ABLR [58]	C3D	55.67	36.79	-
	DEBUG [34]	C3D	55.91	39.72	-
	ExCL [14]	C3D	63.00	43.60	24.10
	VSLNet [63]	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

Thanks for watching! !

More Details on Project Website