CVPR JUNE 19-24 2022 NEW ORLEANS LOUISIANA

TubeDETR: Spatio-Temporal Video Grounding with Transformers

Antoine Yang, Antoine Miech, Josef Sivic, Ivan Laptev, Cordelia Schmid

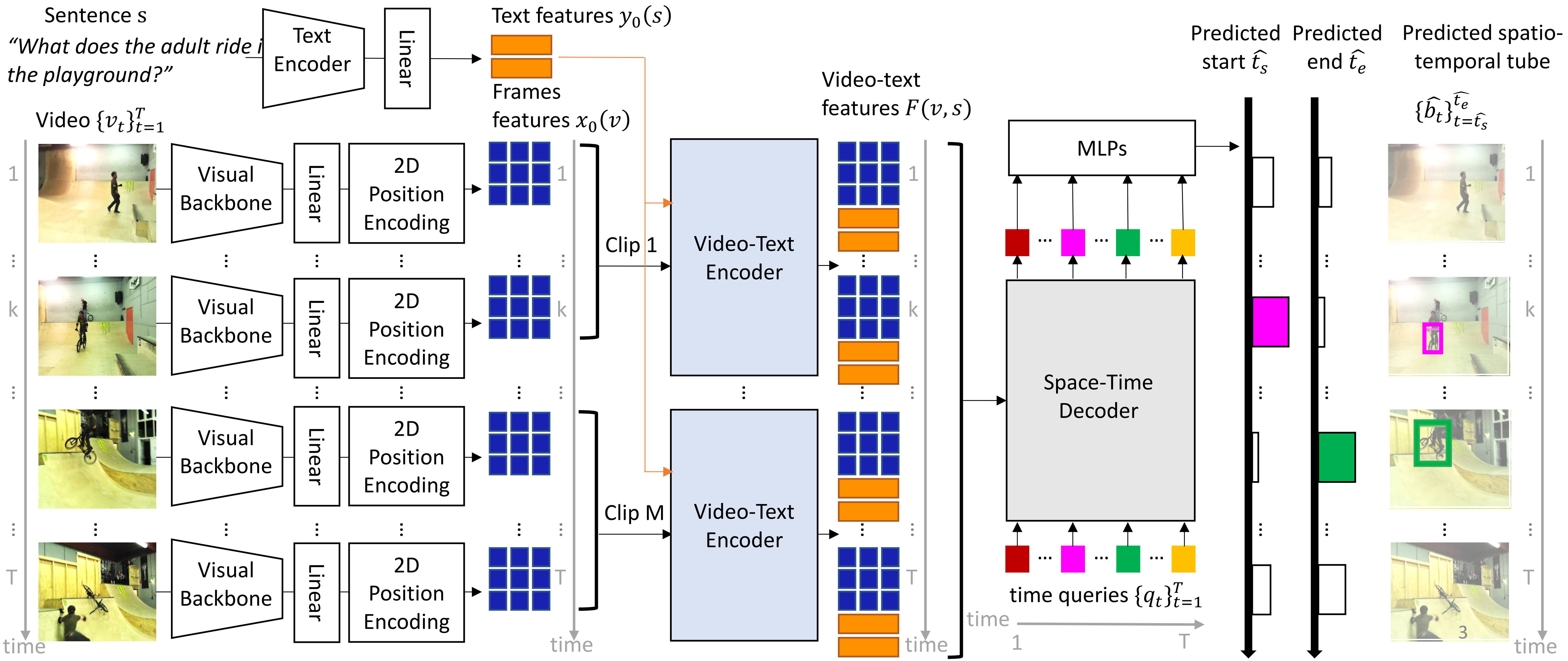
Project page: https://antoyang.github.io/tubedetr.html

Paper: https://arxiv.org/abs/2203.16434

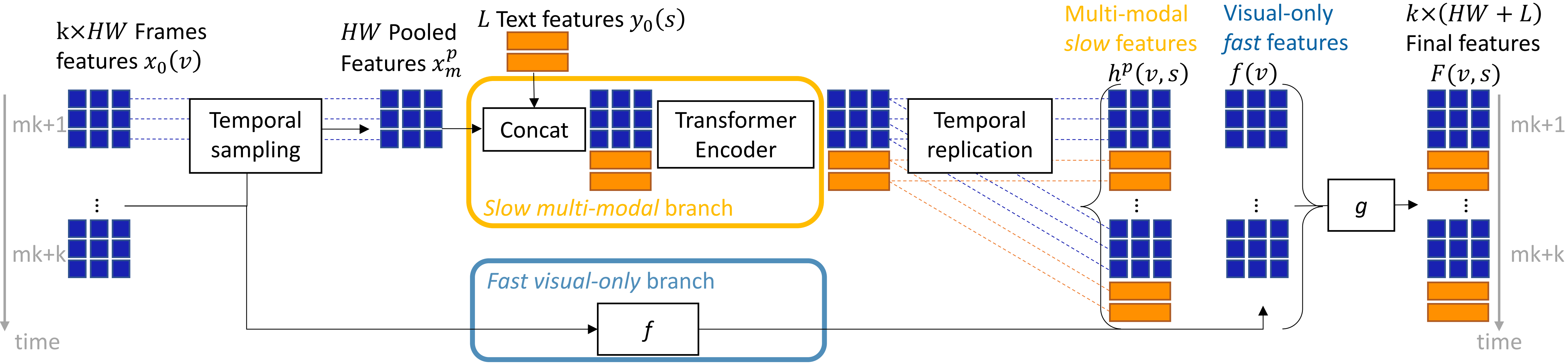
Spatio-Temporal Video Grounding

- **Input text query:** What does the adult ride in the playground?
- **Output spatio-temporal tube:**

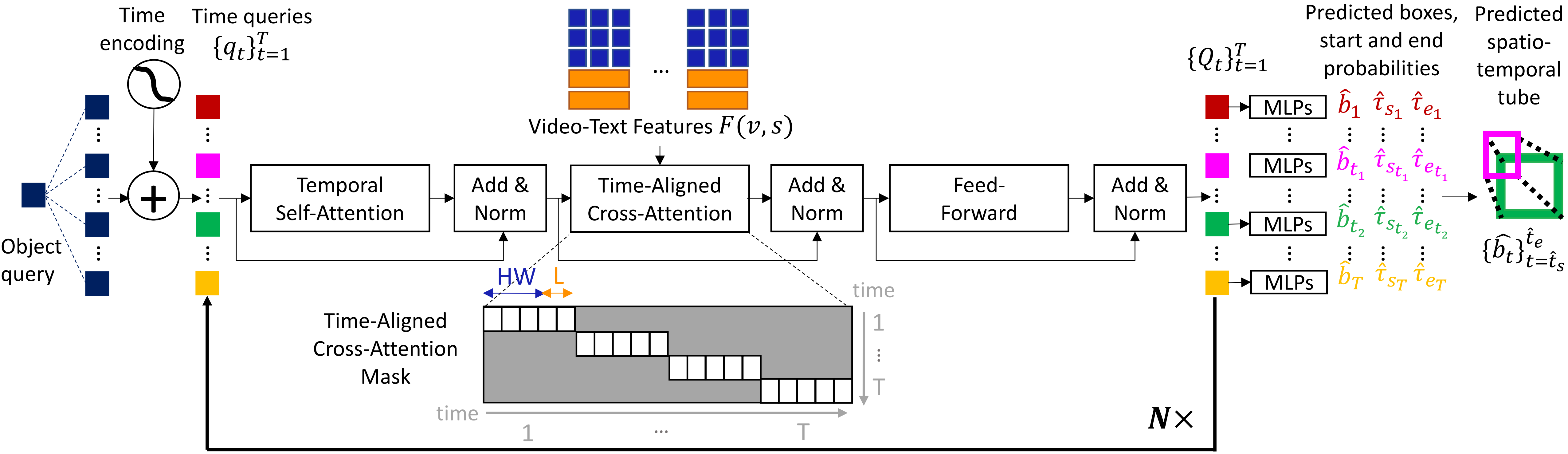
TubeDETR Architecture Overview


Sentence s
"What does the adult ride i the playground?"
Text Encoder
Linear
Text features $y_0(s)$
Frames features $x_0(v)$
Video $\{v_t\}_{t=1}^{T}$
Visual Backbone
Linear
2D Position Encoding
1
k
T
time
Clip 1
Clip M
Video-Text Encoder
Video-text features $F(v,s)$
MLPs
Space-Time Decoder
time queries $\{q_t\}_{t=1}^{T}$
time
1
T
Predicted start $\hat{t_s}$
Predicted end $\hat{t_e}$
Predicted spatio-temporal tube
$\{\hat{b_t}\}_{t=\hat{t_s}}^{\hat{t_e}}$
3

Video-Text Encoder


k×HW Frames features $x_0(v)$
HW Pooled Features x_m^p
L Text features $y_0(s)$
Temporal sampling
Concat
Transformer Encoder
Slow multi-modal branch
Fast visual-only branch
f
Temporal replication
Multi-modal *slow* features $h^p(v,s)$
Visual-only *fast* features $f(v)$
g
k×$(HW + L)$ Final features $F(v,s)$
mk+1
mk+k
time
mk+1
mk+k
time

Space-Time Decoder


Time encoding
Time queries $\{q_t\}_{t=1}^{T}$
Object query
Video-Text Features $F(v,s)$
Temporal Self-Attention
Add & Norm
Time-Aligned Cross-Attention
Add & Norm
Feed-Forward
Add & Norm
$\{Q_t\}_{t=1}^{T}$
MLPs
MLPs
MLPs
MLPs
Predicted boxes, start and end probabilities
$\hat{b}_1$ $\hat{\tau}_{s_1}$ $\hat{\tau}_{e_1}$
$\hat{b}_{t_1}$ $\hat{\tau}_{s_{t_1}}$ $\hat{\tau}_{e_{t_1}}$
$\hat{b}_{t_2}$ $\hat{\tau}_{s_{t_2}}$ $\hat{\tau}_{e_{t_2}}$
$\hat{b}_T$ $\hat{\tau}_{s_T}$ $\hat{\tau}_{e_T}$
Predicted spatio-temporal tube
$\{\hat{b}_t\}_{t=\hat{t}_s}^{\hat{t}_e}$
HW
L
Time-Aligned Cross-Attention Mask
time
1
⋮
T
time
1
...
T
$N\times$

Training

- **Loss:** Combination of spatial localization ($\mathcal{L}_1$, $gIoU$) and temporal localization (KL , att) objectives

$$\mathcal{L} = \lambda_{\mathcal{L}_1}\mathcal{L}_{\mathcal{L}_1}(\hat{b}, b) + \lambda_{gIoU}\mathcal{L}_{gIoU}(\hat{b}, b) + \lambda_{KL}\mathcal{L}_{KL}(\hat{\tau}_s, \hat{\tau}_e, \tau_s, \tau_e) + \lambda_{att}\mathcal{L}_{att}(A)$$

$\lambda_\bullet$: scalar weights of the individual losses

$\hat{b}$ and b : predicted and ground truth boxes

$\hat{\tau}_s$ and τ_s : predicted and ground truth start probability distribution

$\hat{\tau}_e$ and τ_e : predicted and ground truth end probability distribution

A : temporal self-attention matrix

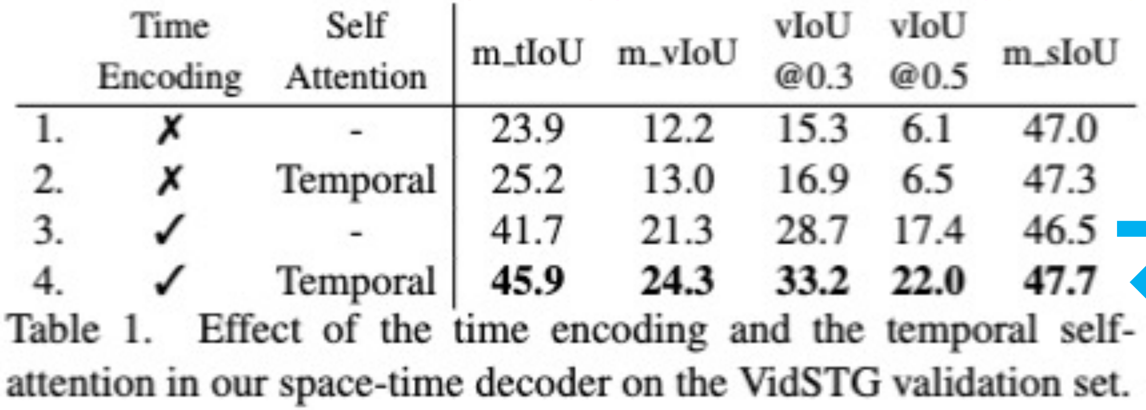
- **Initialization:** from MDETR weights pretrained on Visual Genome, COCO and Flickr

Ablations: Space-Time Decoder

Temporal IoU

Video IoU

Spatial IoU



	Time Encoding	Self Attention	m_tIoU	m_vIoU	vIoU @0.3	vIoU @0.5	m_sIoU
1.	✗	-	23.9	12.2	15.3	6.1	47.0
2.	✗	Temporal	25.2	13.0	16.9	6.5	47.3
3.	✓	-	41.7	21.3	28.7	17.4	46.5
4.	✓	Temporal	**45.9**	**24.3**	**33.2**	**22.0**	**47.7**

Table 1. Effect of the time encoding and the temporal self-attention in our space-time decoder on the VidSTG validation set.

Time encoding matters.

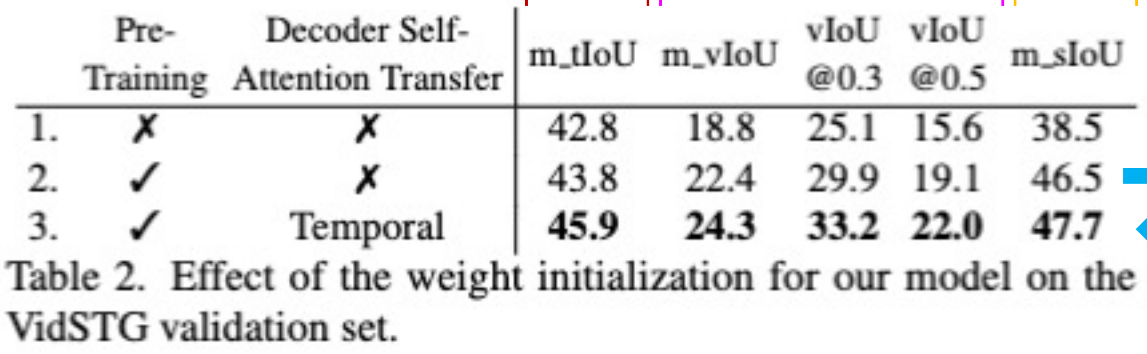
Temporal self-attention helps.

Ablations: Weights initialization

Temporal IoU

Video IoU

Spatial IoU



	Pre-Training	Decoder Self-Attention Transfer	m_tIoU	m_vIoU	vIoU @0.3	vIoU @0.5	m_sIoU
1.	✗	✗	42.8	18.8	25.1	15.6	38.5
2.	✓	✗	43.8	22.4	29.9	19.1	46.5
3.	✓	Temporal	**45.9**	**24.3**	**33.2**	**22.0**	**47.7**

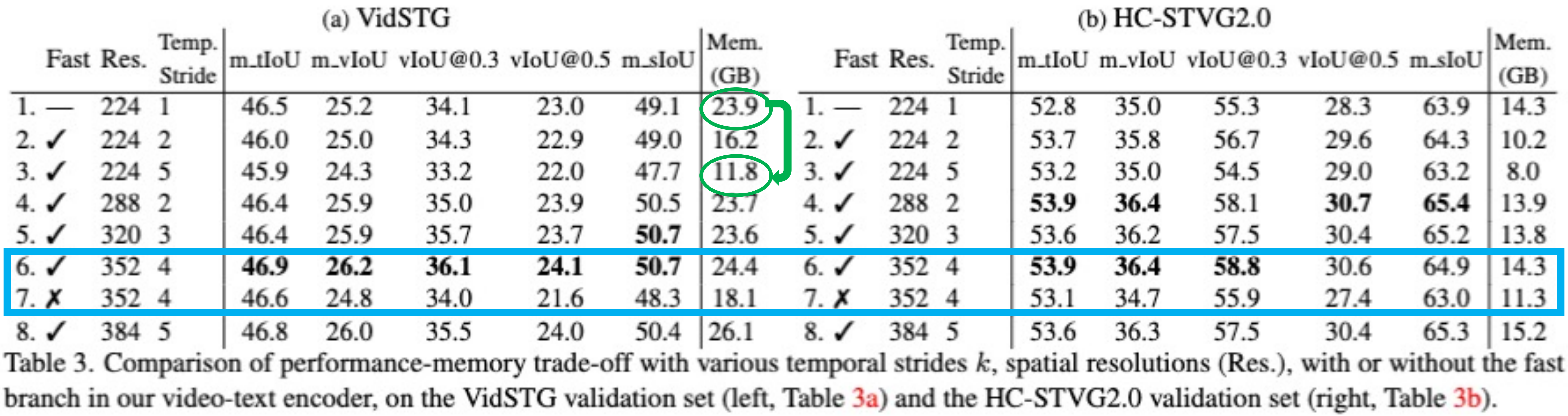
Table 2. Effect of the weight initialization for our model on the VidSTG validation set.

MDETR pretraining matters.

Transferring spatial self-attention to temporal self-attention helps.

Ablations: Video-Text Encoder

- Our encoder is memory-efficient.
- Fast branch matters.



(a) VidSTG

	Fast	Res.	Temp. Stride	m_tIoU	m_vIoU	vIoU@0.3	vIoU@0.5	m_sIoU	Mem. (GB)
1.	—	224	1	46.5	25.2	34.1	23.0	49.1	23.9
2.	✓	224	2	46.0	25.0	34.3	22.9	49.0	16.2
3.	✓	224	5	45.9	24.3	33.2	22.0	47.7	11.8
4.	✓	288	2	46.4	25.9	35.0	23.9	50.5	23.7
5.	✓	320	3	46.4	25.9	35.7	23.7	**50.7**	23.6
6.	✓	352	4	**46.9**	**26.2**	**36.1**	**24.1**	**50.7**	24.4
7.	✗	352	4	46.6	24.8	34.0	21.6	48.3	18.1
8.	✓	384	5	46.8	26.0	35.5	24.0	50.4	26.1

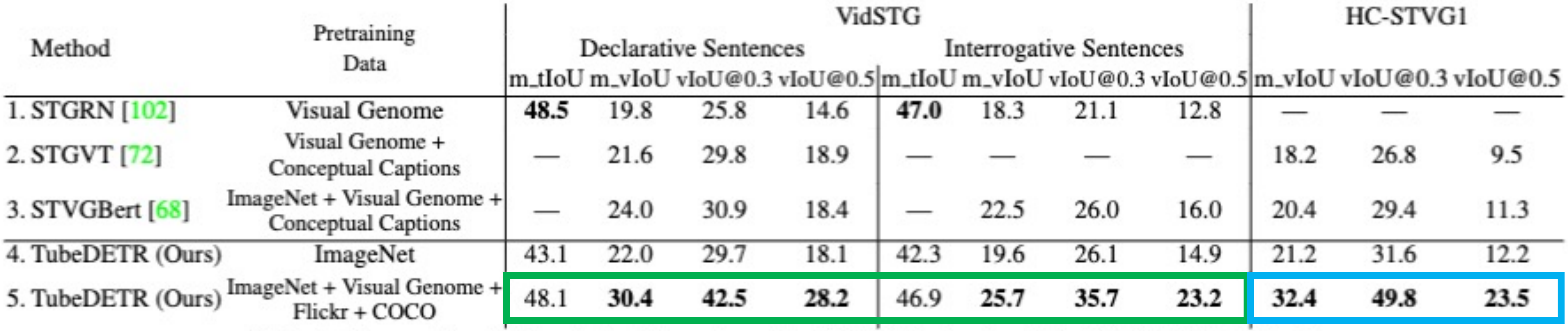
(b) HC-STVG2.0

	Fast	Res.	Temp. Stride	m_tIoU	m_vIoU	vIoU@0.3	vIoU@0.5	m_sIoU	Mem. (GB)
1.	—	224	1	52.8	35.0	55.3	28.3	63.9	14.3
2.	✓	224	2	53.7	35.8	56.7	29.6	64.3	10.2
3.	✓	224	5	53.2	35.0	54.5	29.0	63.2	8.0
4.	✓	288	2	**53.9**	**36.4**	58.1	**30.7**	**65.4**	13.9
5.	✓	320	3	53.6	36.2	57.5	30.4	65.2	13.8
6.	✓	352	4	**53.9**	**36.4**	**58.8**	30.6	64.9	14.3
7.	✗	352	4	53.1	34.7	55.9	27.4	63.0	11.3
8.	✓	384	5	53.6	36.3	57.5	30.4	65.3	15.2

Table 3. Comparison of performance-memory trade-off with various temporal strides k, spatial resolutions (Res.), with or without the fast branch in our video-text encoder, on the VidSTG validation set (left, Table 3a) and the HC-STVG2.0 validation set (right, Table 3b).

Comparison with state of the art

- State-of-the-art results on: VidSTG and HC-STVG.



Method	Pretraining Data	VidSTG								HC-STVG1		
		Declarative Sentences				Interrogative Sentences						
		m_tIoU	m_vIoU	vIoU@0.3	vIoU@0.5	m_tIoU	m_vIoU	vIoU@0.3	vIoU@0.5	m_vIoU	vIoU@0.3	vIoU@0.5
1. STGRN [102]	Visual Genome	**48.5**	19.8	25.8	14.6	**47.0**	18.3	21.1	12.8	—	—	—
2. STGVT [72]	Visual Genome + Conceptual Captions	—	21.6	29.8	18.9	—	—	—	—	18.2	26.8	9.5
3. STVGBert [68]	ImageNet + Visual Genome + Conceptual Captions	—	24.0	30.9	18.4	—	22.5	26.0	16.0	20.4	29.4	11.3
4. TubeDETR (Ours)	ImageNet	43.1	22.0	29.7	18.1	42.3	19.6	26.1	14.9	21.2	31.6	12.2
5. TubeDETR (Ours)	ImageNet + Visual Genome + Flickr + COCO	48.1	**30.4**	**42.5**	**28.2**	46.9	**25.7**	**35.7**	**23.2**	**32.4**	**49.8**	**23.5**

Table 4. Comparison to the state of the art on the VidSTG test set and the HC-STVG1 test set.

Qualitative results

- **Interactive Demo:** http://stvg.paris.inria.fr/
- **Query:** What is beneath the adult in the snow?

🟨 TubeDETR

🟩 Ground Truth