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:**

![Image of a playground with a man riding a ramp]
TubeDETR Architecture Overview

Sentence s
“What does the adult ride in the playground?”

Video \(\{v_t\}_{t=1}^T\)

Text Encoder

Text features \(y_0(s)\)

Frames features \(x_0(v)\)

Video Text Encoder

Video-text features \(F(v, s)\)

MLPs

Predicted start \(t_s\)

Predicted end \(t_e\)

Predicted spatio-temporal tube

MLPs

Video Text Encoder

Clip 1

Clip M

Video Text Encoder

MLPs

Space-Time Decoder

Time queries \(\{q_t\}_{t=1}^T\)
Video-Text Encoder

$k \times HW$ Frames features $x_0(v)$

$HW$ Pooled Features $x_p$ $m$

$L$ Text features $y_0(s)$

Temporal sampling

Concat

Transformer Encoder

Temporal replication

Multi-modal slow features $h^p(v, s)$

Visual-only fast features $f(v)$

$k \times (HW + L)$ Final features $F(v, s)$

$F(v)$

$g$

$mk+1$

$mk+k$

time

Fast visual-only branch

$F(v)$

$g$

$mk+1$

$mk+k$

time

慢多模态分支

Fast visual-only branch

时间
Space-Time Decoder

Video-Text Features $F(v, s)$

Temporal Self-Attention

Add & Norm

Time-Aligned Cross-Attention

Add & Norm

Feed-Forward

Add & Norm

Predicted boxes, start and end probabilities

Predicted spatio-temporal tube

Object query

Time queries $\{q_t\}_{t=1}^T$

Time encoding

Video-Text Features $F(v, s)$

Time-Aligned Cross-Attention Mask

MLPs

MLPs

MLPs

MLPs

$\{Q_t\}_{t=1}^T$

$\{\hat{t}_e\}_{t=1}^{T}$

$\{\hat{t}_{s}^t, \hat{t}_{e}^t\}$

$\{\hat{b}_t\}_{t=1}^{T}$

${\hat{b}}_t$, $\hat{t}_s$, $\hat{t}_e$

$N \times T$
Training

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

$$\mathcal{L} = \lambda_{\mathcal{L}_1} \mathcal{L}_{\mathcal{L}_1}(\hat{b}, b) + \lambda_{gIoU} \mathcal{L}_{gIoU}(\hat{b}, b) + \lambda_{KL} \mathcal{L}_{KL}(\hat{r}_s, \hat{r}_e, r_s, r_e) + \lambda_{att} \mathcal{L}_{att}(\hat{A})$$

$\lambda_i$ : scalar weights of the individual losses
$\hat{b}$ and $b$ : predicted and ground truth boxes
$\hat{r}_s$ and $r_s$ : predicted and ground truth start probability distribution
$\hat{r}_e$ and $r_e$ : predicted and ground truth end probability distribution
$\hat{A}$ : temporal self-attention matrix

- **Initialization:** from MDETR weights pretrained on Visual Genome, COCO and Flickr
Ablations: Space-Time Decoder

| Time Encoding | Self Attention | m_tIoU | m_vIoU | vIoU@0.3 | vIoU@0.5 | m_sIoU |
|---------------|----------------|--------|--------|----------|----------|--------|
| 1.            | X              | 23.9   | 12.2   | 15.3     | 6.1      | 47.0   |
| 2.            | X              | 25.2   | 13.0   | 16.9     | 6.5      | 47.3   |
| 3.            | ✓              | 41.7   | 21.3   | 28.7     | 17.4     | 46.5   |
| 4.            | ✓              | 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

MDETR pretraining matters.

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

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

| Pre-Training | Decoder Self-Attention Transfer | m_tIoU | m_vIoU | vIoU @0.3 | vIoU @0.5 | m_sIoU |
|--------------|---------------------------------|--------|--------|-----------|-----------|--------|
| 1.           | X                               | 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** |
### Ablations: Video-Text Encoder

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

| Fast Res. | Temp. Stride | VidSTG m.tIoU | VidSTG m.vIoU | VidSTG vIoU@0.3 | VidSTG vIoU@0.5 | HC-STV2.0 m.tIoU | HC-STV2.0 m.vIoU | HC-STV2.0 vIoU@0.3 | HC-STV2.0 vIoU@0.5 | HC-STV2.0 m.sIoU | Mem. (GB) |
|-----------|--------------|---------------|---------------|-----------------|-----------------|-----------------|-----------------|-----------------|-----------------|-----------------|----------|
| 1. —      | 224 1        | 46.5          | 25.2          | 34.1            | 23.0            | 23.9            | 52.8            | 35.0            | 55.3            | 28.3            | 14.3     |
| 2. ✓      | 224 2        | 46.0          | 25.0          | 34.3            | 22.9            | 49.0            | 53.7            | 35.8            | 56.7            | 29.6            | 10.2     |
| 3. ✓      | 224 5        | 45.9          | 24.3          | 33.2            | 22.0            | 47.7            | 53.2            | 35.0            | 54.5            | 29.0            | 8.0      |
| 4. ✓      | 288 2        | 46.4          | 25.9          | 35.0            | 23.9            | 50.5            | 53.9            | 36.4            | 58.1            | 30.7            | 13.9     |
| 5. ✓      | 320 3        | 46.4          | 25.9          | 35.7            | 23.7            | 50.7            | 53.6            | 36.2            | 57.5            | 30.4            | 13.8     |

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-STV2.0 validation set (right, Table 3b).
Comparison with state of the art

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

| Method         | Pretraining Data          | Declarative Sentences | Interrogative Sentences | HC-STVG1                        |
|----------------|---------------------------|-----------------------|-------------------------|---------------------------------|
| 1. STGRN [102] | Visual Genome             | 48.5                  | 19.8 25.8 14.6          |                                 |
| 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          | 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/](http://stvg.paris.inria.fr/)
- **Query:** What is beneath the adult in the snow?

---

**TubeDETR**

**Ground Truth**