International Conference On Learning Representations

Knowing Your Target 🎯: Target-Aware Transformer Makes Better Spatio-Temporal Video Grounding

Xin Gu, Yaojie Shen, Chenxi Luo, Tiejian Luo, Yan Huang, Yuewei Lin, Heng Fan*, Libo Zhang*

(*equal advising and co-last authors)

Code and model:
https://github.com/HengLan/TA-STVG

University of Chinese Academy of Sciences
ISCAS
LA TROBE UNIVERSITY
UNT UNIVERSITY OF NORTH TEXAS
Brookhaven National Laboratory

What is Spatio-Temporal Video Grounding (STVG)?

- STVG aims to localize the object of interest in an untrimmed video with a spatio-temporal tube given a free-form textual query

Input text query: What does the adult ride in the playground?
Output spatio-temporal tube from an untrimmed video:

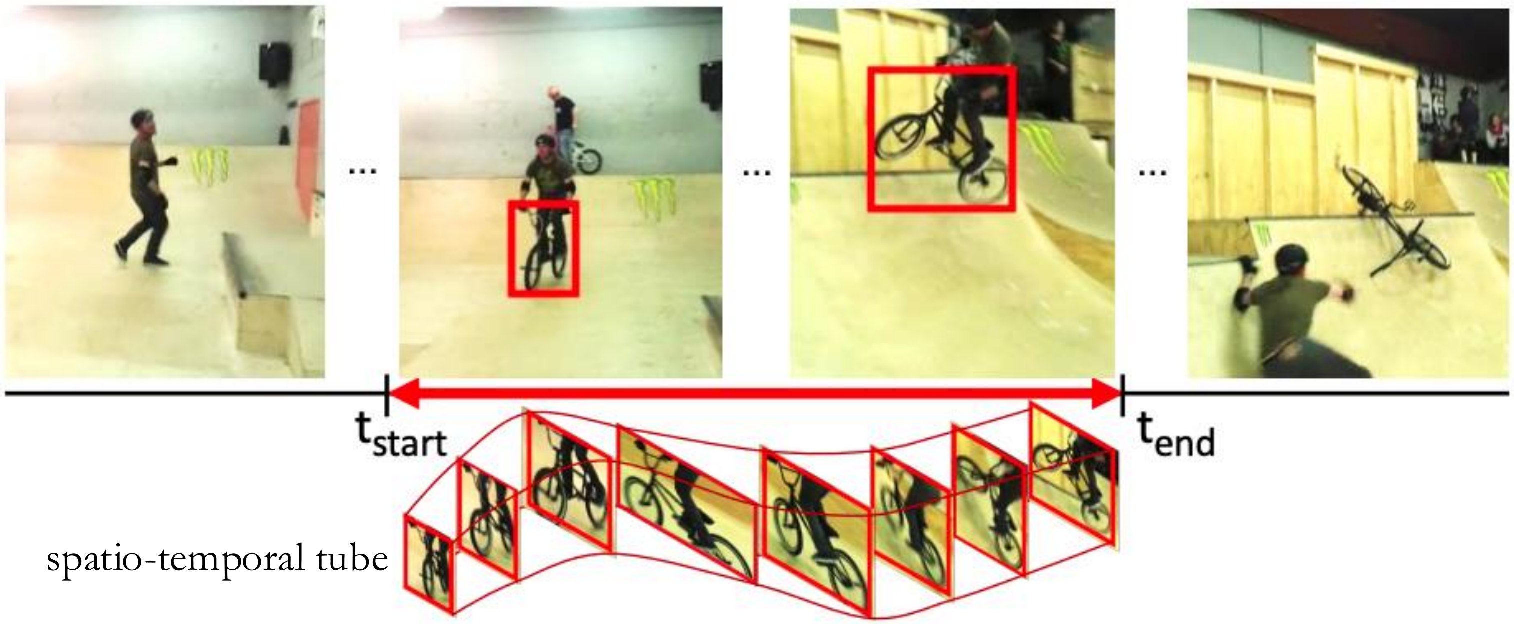
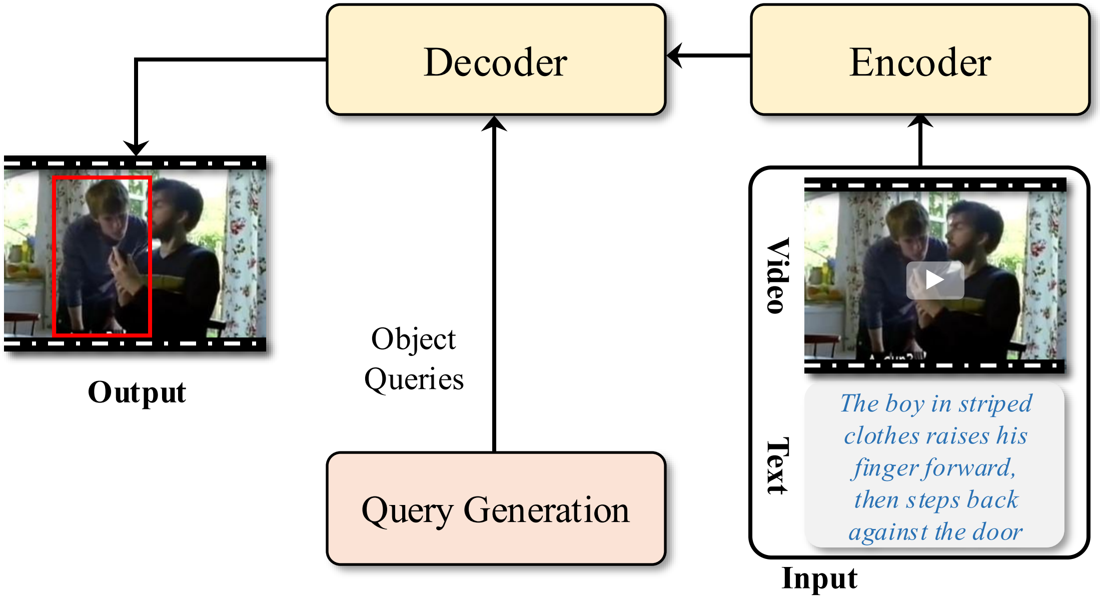


Image courtesy
Yang *et al*, CVPR'2022

Existing Transformer-based STVG Methods

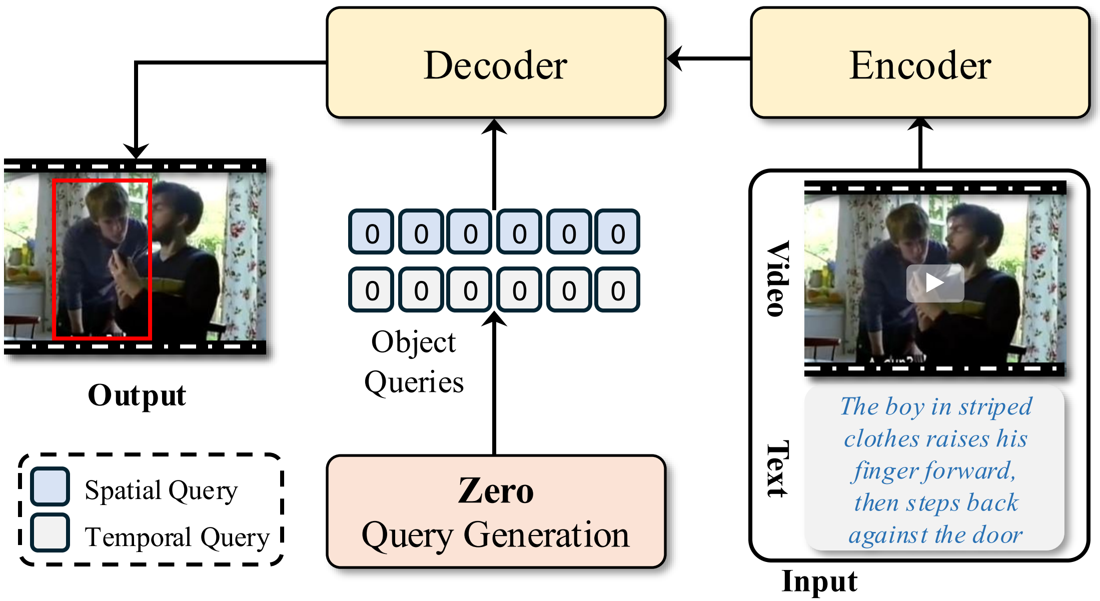
- Current Transformer-based STVG methods [Yang et al, CVPR'2022; Jin et al, NeurIPS'2022, Gu, et al, CVPR' 2024, etc] inspired by the DETR [Carion et al, ECCV, 2020]



- Multimodal Encoder
  - visual and textual feature fusion
- Decoder:
  - learning target position in queries from video and text

Existing Transformer-based STVG Methods

- Current Transformer-based STVG methods [Yang et al, CVPR'2022; Jin et al, NeurIPS'2022, Gu, et al, CVPR' 2024, etc] inspired by the DETR [Carion et al, ECCV, 2020]



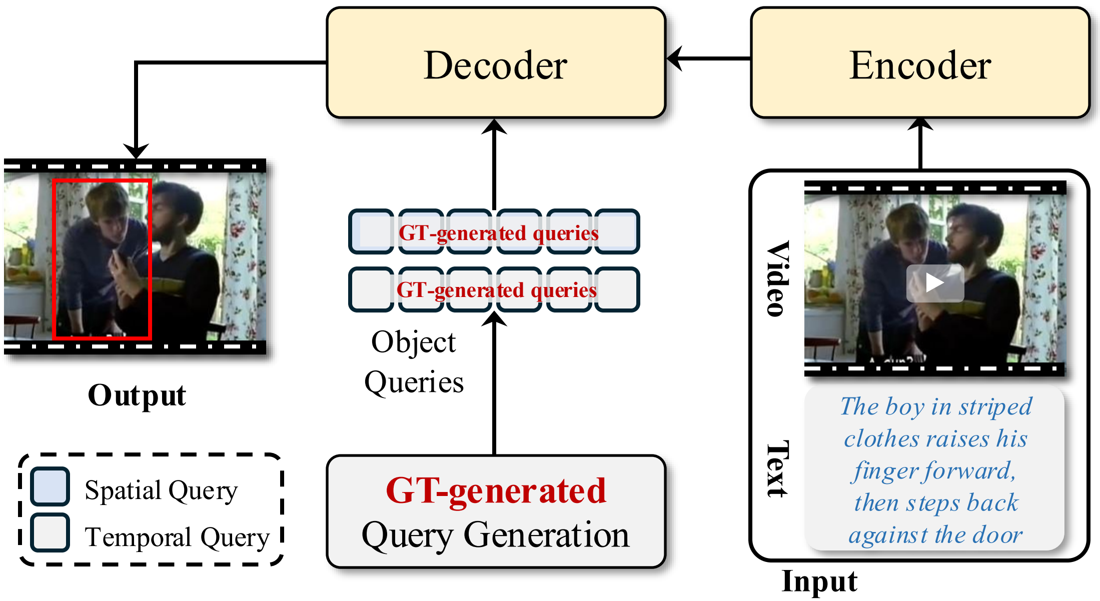
Zero query generation:
- Current STVG methods simply utilize **zeros** to initialize queries

Problem:
- Zero-queries are difficult to learn discriminative target position information in complicated scenarios due **lacking effective target-specific semantic cues**

- Multimodal Encoder
  - visual and textual feature fusion
- Decoder:
  - learning target position in queries from video and text

Motivation

- Target-specific cues as a prior to guide object query learning
  - If object queries know the target from the very beginning, i.e., **they know what to learn**, they can better interact with the multimodal features for more accurate localization.

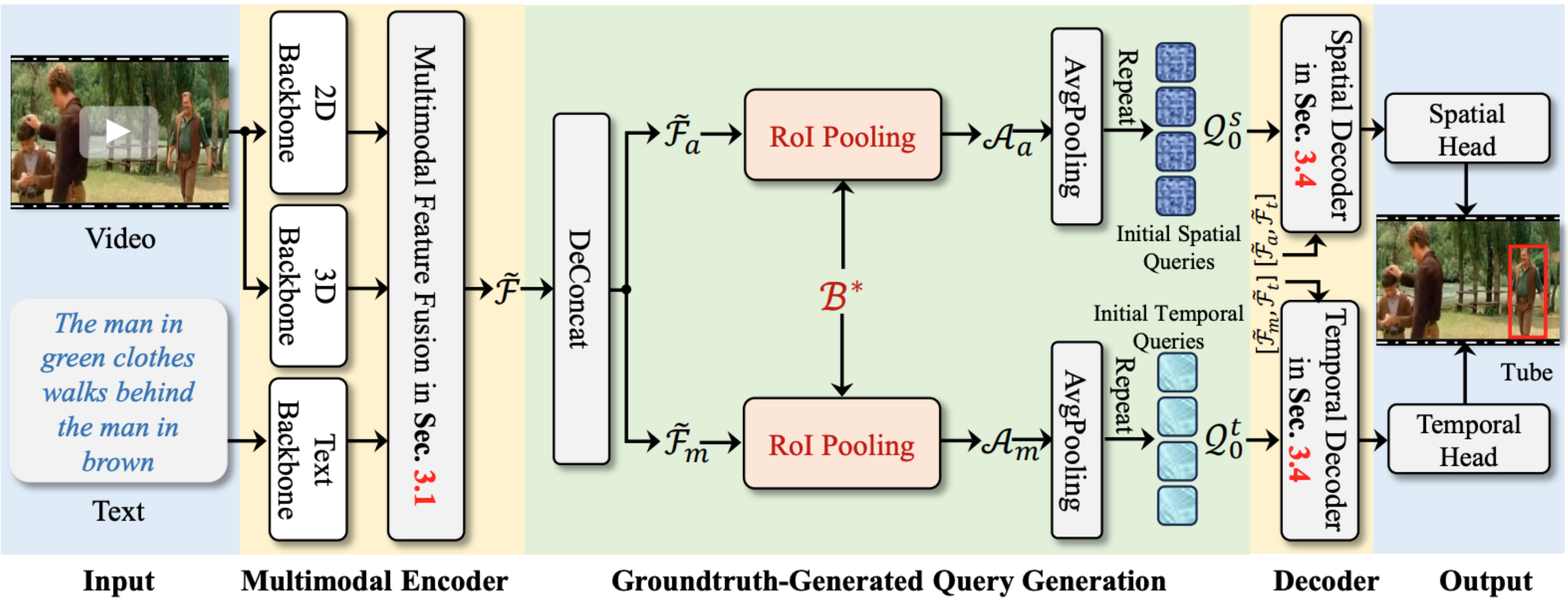


GT-generated queries:
- Apply **groundtruth (GT)** to initialize queries

- Multimodal Encoder
  - visual and textual feature fusion
- Decoder:
  - learning target position in queries from video and text

Motivation

- Oracle Experiments
  - Apply groundtruth-generated object queries for STVG.

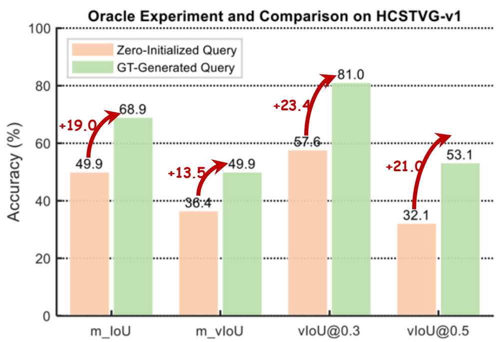


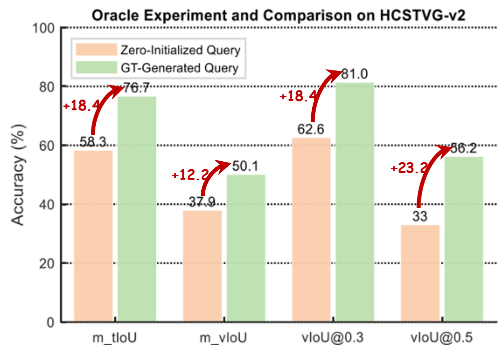
$\mathcal{B}^*$ groundtruth bounding box used to generate object queries

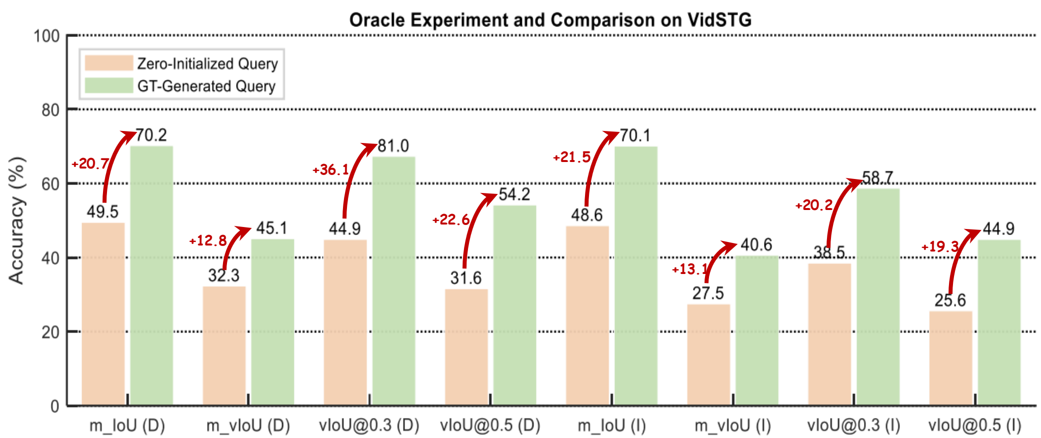
Motivation

❑ Oracle Experiments

- ○ Comparison of performance using zero-initialized object queries and groundtruth-generated object queries for STVG on three popular benchmarks





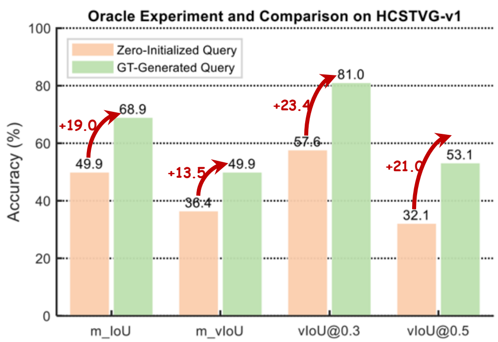


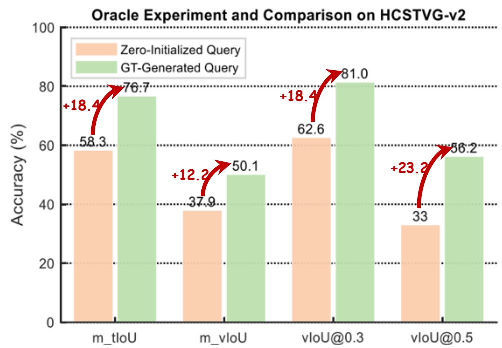
Observation:
- ○ Introduction of target-specific information from groundtruth to initialize object queries significantly improves STVG performance.

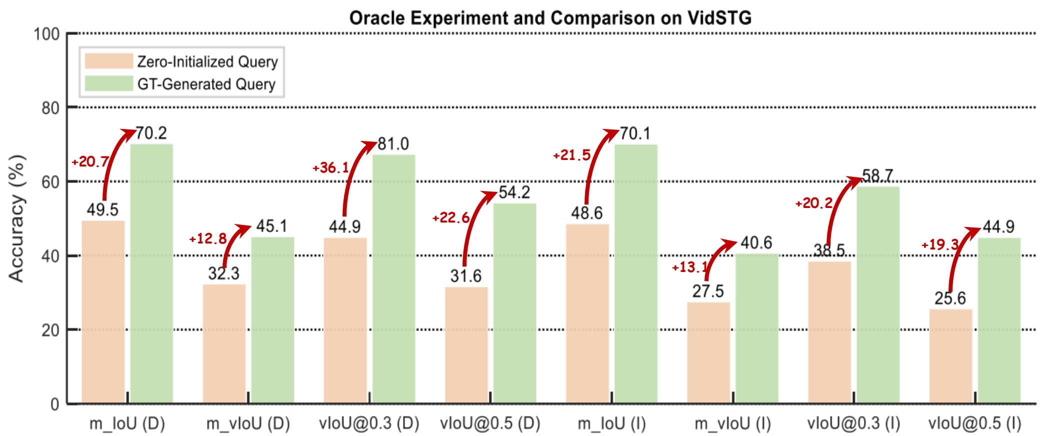
Motivation

❑ Oracle Experiments

- Comparison of performance using zero-initialized object queries and groundtruth-generated object queries for STVG on three popular benchmarks





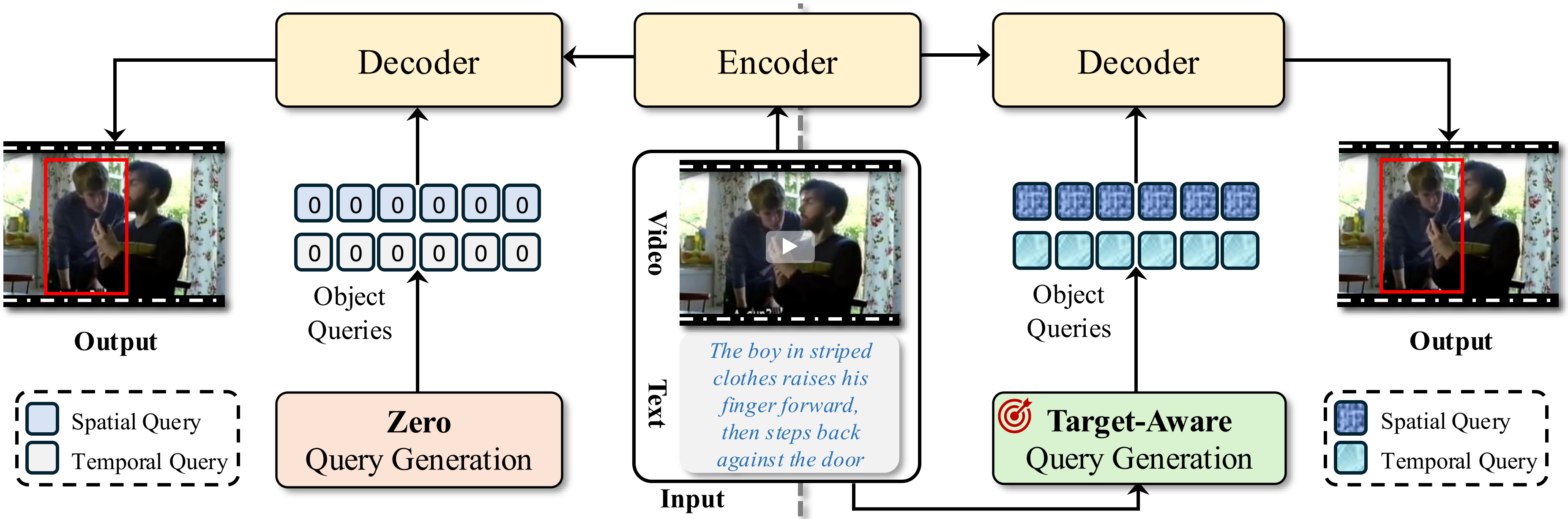


Observation:
- Introduction of target-specific information from groundtruth to initialize object queries significantly improves STVG performance.

Motivation:
- Exploring target-cues from the video to initialize the object queries in Transformer-based STVG

The Proposed TA-STVG Approach

- The proposed Target-Aware Transformer-based STVG generating queries with target-aware cues from video and text for STVG
  - Comparison between existing methods and our approach

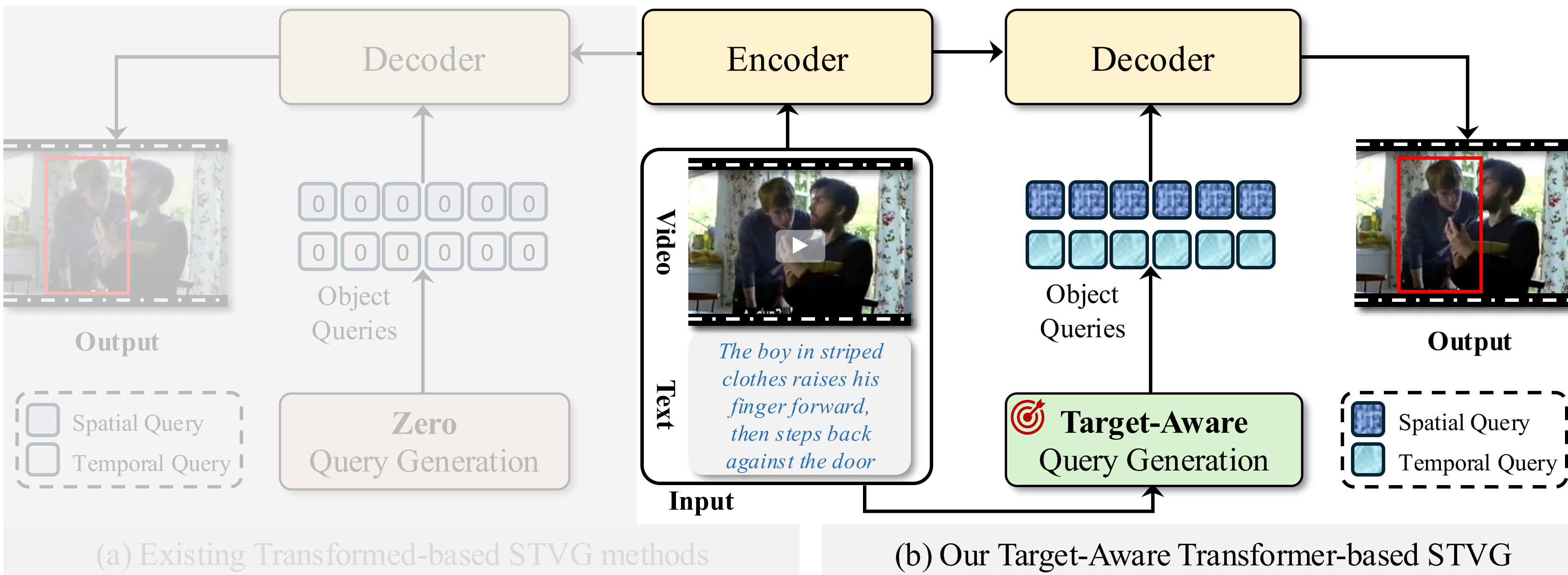


(a) Existing Transformed-based STVG methods

(b) Our Target-Aware Transformer-based STVG

The Proposed TA-STVG Approach

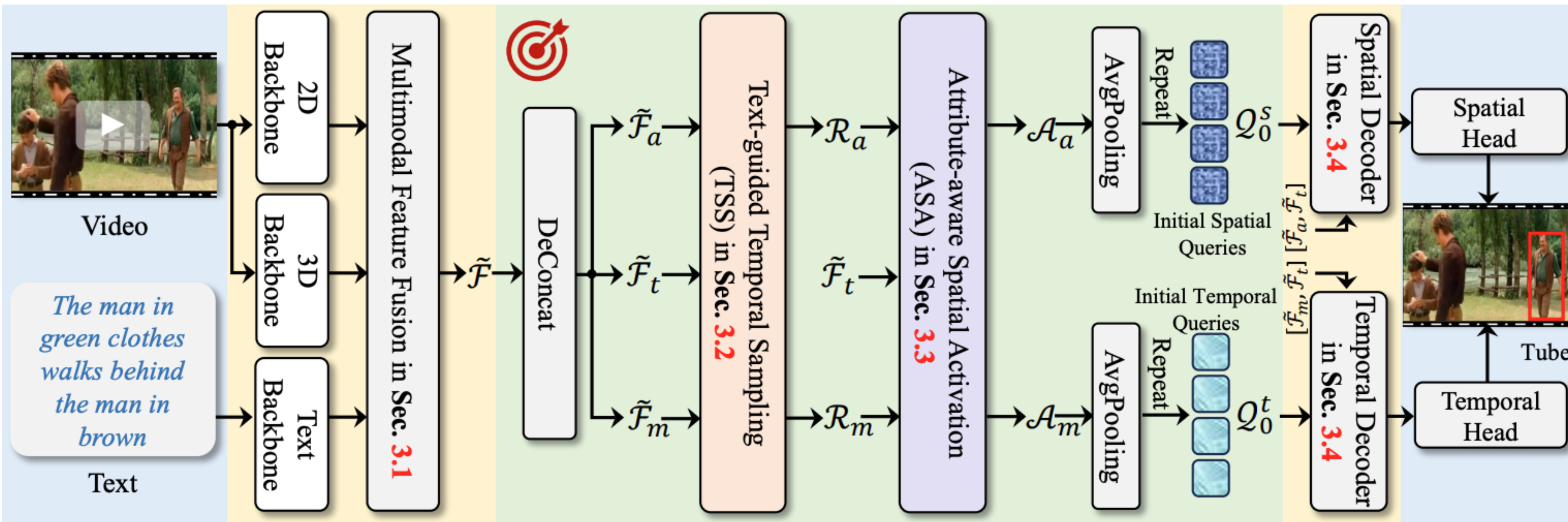
- The proposed Target-Aware Transformer-based STVG generating queries with target-aware cues from video and text for STVG
  - Comparison between existing methods and our approach



Core:
- learning target-aware queries directly from the given video-text pair
- Queries naturally carrying target-specific cues

The Proposed TA-STVG Approach

❑ The proposed Target-Aware Transformer-based STVG generating queries with target-aware cues from video and text for STVG

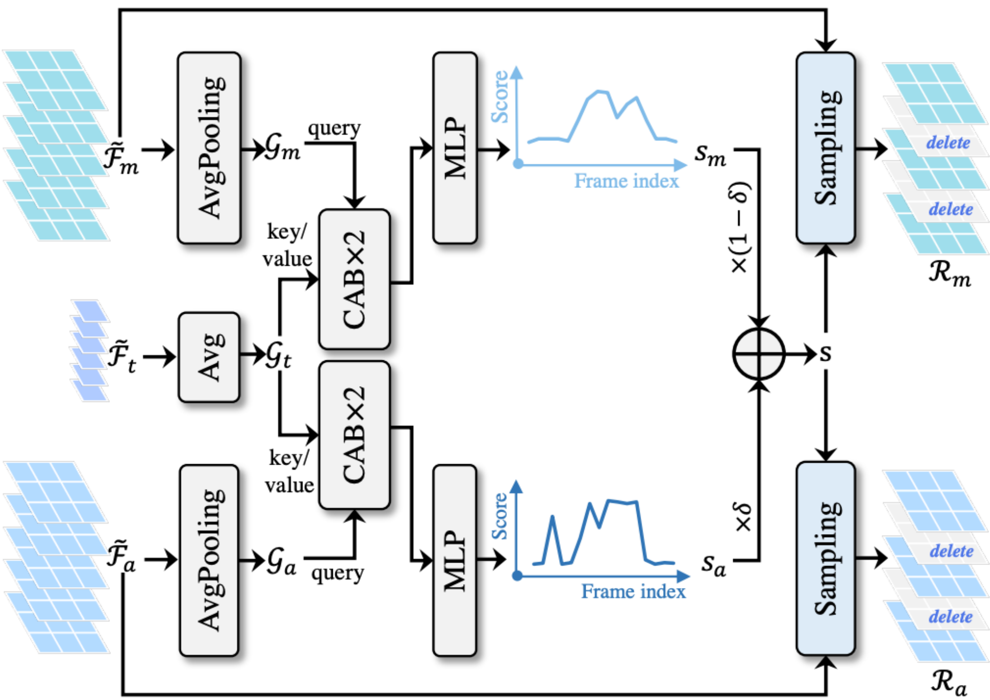


Overview of TA-STVG, which exploits target-specific information for STVG.

- Multimodal Encoder: visual and textual feature fusion
- Target-aware Query Generation: learning object queries from the video
  - Text-guided Temporal Sampling (TTS)
  - Attribute-aware Spatial Activation (ASA)
- Decoder: learning target position in queries from video and text

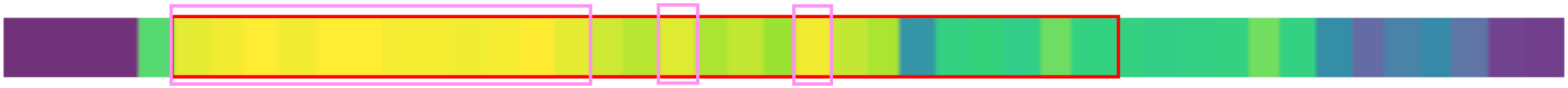
The Proposed TA-STVG Approach

- Text-guided Temporal Sampling (TTS)



- Identify and sample frames relevant to the target guided by holistic textual features
  - Consider both motion and appearance information
  - Predict relevance scores for each frame
  - Predict sampled target-relevant temporal appearance and motion features

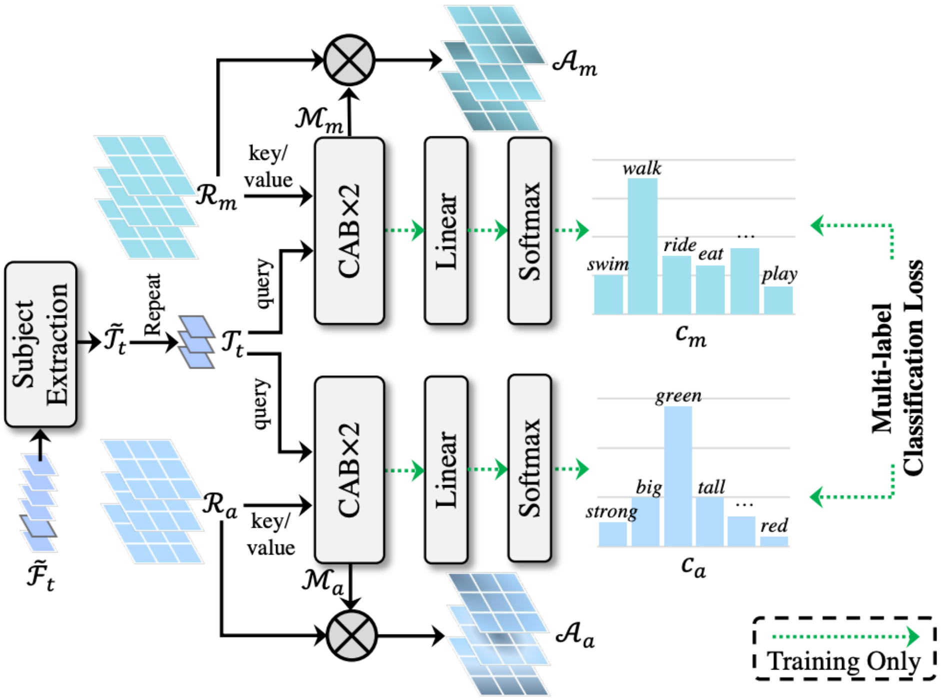
- Analysis of TTS



(a) Temporal relevance score s predicted by TTS (red rectangle: groundtruth; pink rectangles: sampled frames)

The Proposed TA-STVG Approach

❑ Attribute-aware Spatial Activation (ASA)



- Mine fine-grained visual semantic information
  - Consider both motion and appearance attribute
  - Use attention maps as attribute-specific spatial activation
  - Learn the target-specific attribute features

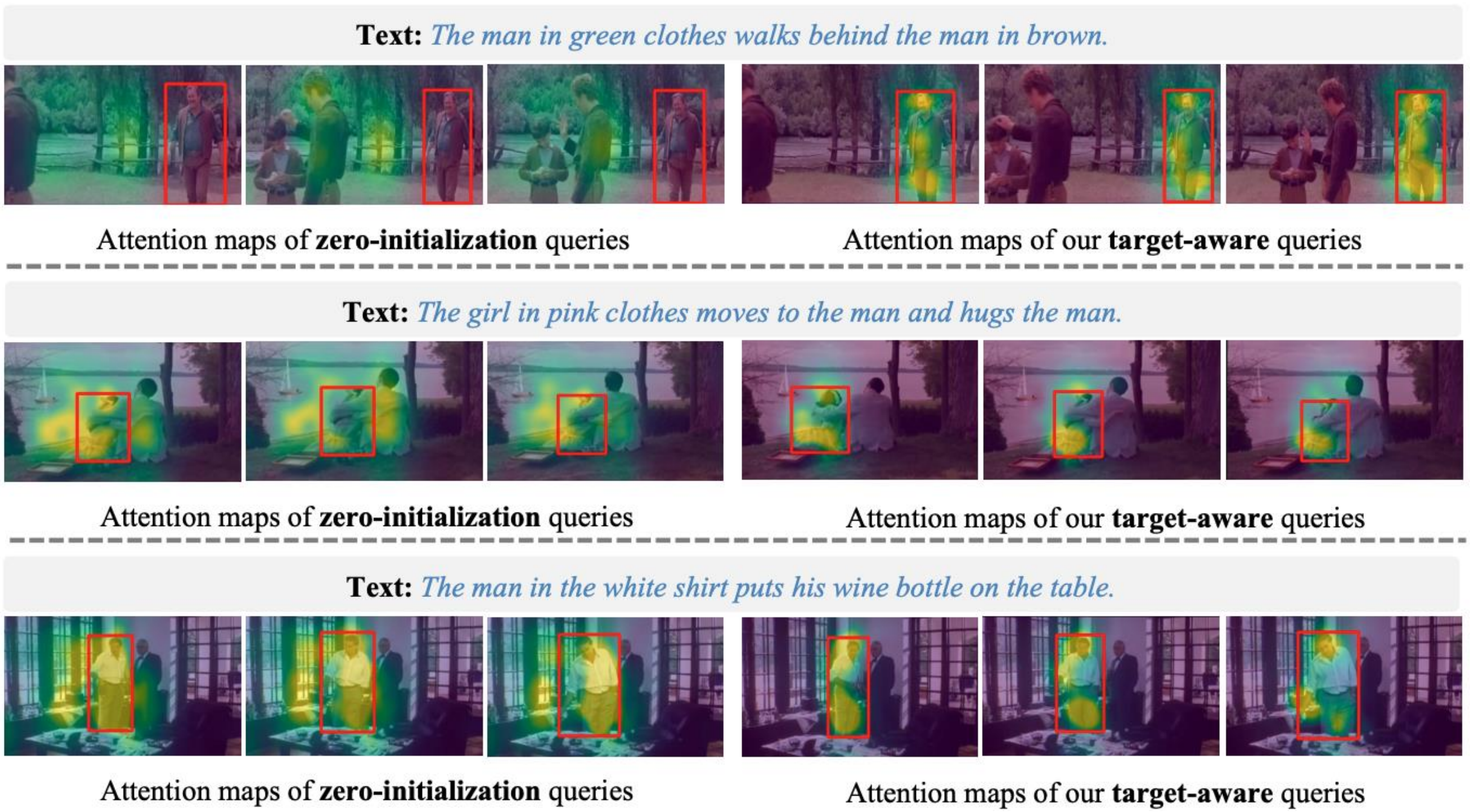
○ Analysis of ASA

Red box: target of interest

(b) Appearance ("yellow") activation $\mathcal{M}_a$ by ASA (c) Motion ("walks" and "stops") activation $\mathcal{M}_m$ by ASA

The Proposed TA-STVG Approach

- Comparison of attention maps for zero-initialized and our target-aware object queries in video frames in the spatial decoder



Red box: target of interest

Target-queries focus better on target regions, which benefits target localization!

The Proposed TA-STVG Approach

❑ Experiments – State-of-the-art Comparison

Table 1: Comparison on HCSTVG-v1 (%).

Methods	m_tIoU	m_vIoU	vIoU@0.3	vIoU@0.5
STVGBert (Su et al., 2021)	-	20.4	29.4	11.3
TubeDETR (Yang et al., 2022a)	43.7	32.4	49.8	23.5
STCAT (Jin et al., 2022)	49.4	35.1	57.7	30.1
SGFDN (Wang et al., 2023c)	46.9	35.8	56.3	37.1
STVGFormer (Lin et al., 2023b)	-	36.9	62.2	34.8
CG-STVG (Gu et al., 2024a)	52.8	38.4	61.5	36.3
Baseline (ours)	49.9	36.4	57.6	32.1
TA-STVG (ours)	**53.0** (+3.1)	**39.1** (+2.7)	**63.1** (+5.5)	**36.8** (+4.7)

Table 2: Comparison on HCSTVG-v2 (%).

Methods	m_tIoU	m_vIoU	vIoU@0.3	vIoU@0.5
PCC (Yu et al., 2021)	-	30.0	-	-
2D-Tan (Tan et al., 2021)	-	30.4	50.4	18.8
MMN (Wang et al., 2022)	-	30.3	49.0	25.6
TubeDETR (Yang et al., 2022a)	53.9	36.4	58.8	30.6
STVGFormer (Lin et al., 2023b)	58.1	38.7	65.5	33.8
CG-STVG (Gu et al., 2024a)	60.0	39.5	64.5	36.3
Baseline (ours)	58.3	37.9	62.6	33.0
TA-STVG (ours)	**60.4** (+2.1)	**40.2** (+2.3)	**65.8** (+3.2)	**36.7** (+3.7)

Table 3: Comparison with existing state-of-the-art methods on VidSTG (%).

Methods	Declarative Sentences				Interrogative Sentences			
	m_tIoU	m_vIoU	vIoU@0.3	vIoU@0.5	m_tIoU	m_vIoU	vIoU@0.3	vIoU@0.5
STGRN (Zhang et al., 2020b)	48.5	19.8	25.8	14.6	47.0	18.3	21.1	12.8
OMRN (Zhang et al., 2020a)	50.7	23.1	32.6	16.4	49.2	20.6	28.4	14.1
STGVT (Tang et al., 2021)	-	21.6	29.8	18.9	-	-	-	-
STVGBert (Su et al., 2021)	-	24.0	30.9	18.4	-	22.5	26.0	16.0
TubeDETR (Yang et al., 2022a)	48.1	30.4	42.5	28.2	46.9	25.7	35.7	23.2
STCAT (Jin et al., 2022)	50.8	33.1	46.2	32.6	49.7	28.2	39.2	26.6
SGFDN (Wang et al., 2023c)	45.1	28.3	41.7	29.1	44.8	25.8	36.9	23.9
STVGFormer (Lin et al., 2023b)	-	33.7	47.2	32.8	-	28.5	39.9	26.2
CG-STVG (Gu et al., 2024a)	51.4	34.0	47.7	33.1	49.9	29.0	40.5	27.5
Baseline (ours)	49.5	32.3	44.9	31.6	48.6	27.5	38.5	25.6
TA-STVG (ours)	**51.7** (+2.2)	**34.4** (+2.1)	**48.2** (+3.3)	**33.5** (+1.9)	**50.2** (+1.6)	**29.5** (+2.0)	**41.5** (+3.0)	**28.0** (+2.4)

Observations:

- State-of-the-art by outperforming other methods
- Significantly improving the baseline method using zero-initialized queries

Target-Aware Transformer makes better STVG!

The Proposed TA-STVG Approach

❑ Experiments – Key Ablations (see more in the paper)

Table 4: Ablations of TTS and ASA.

	TTS	ASA	m_tIoU	m_vIoU	vIoU@0.3	vIoU@0.5
❶	-	-	49.9	36.4	57.6	32.1
❷	✓	-	52.2	38.4	61.7	36.2
❸	-	✓	51.4	38.0	60.4	34.1
❹	✓	✓	**53.0**	**39.1**	**63.1**	**36.8**

Table 5: Ablations of branches in TTS. "TG", "AB", and "MB" are the text-guided, appearance and motion branches, respectively.

	TG	AB	MB	m_tIoU	m_vIoU	vIoU@0.3	vIoU@0.5
❶	-	-	-	51.4	38.0	60.4	34.1
❷	✓	✓	-	51.8	38.4	61.1	36.3
❸	✓	-	✓	52.3	38.3	62.0	36.5
❹	-	✓	✓	51.8	38.5	62.0	36.9
❺	✓	✓	✓	**53.0**	**39.1**	**63.1**	**36.8**

Table 6: Ablations of attributes in ASA. "SG", "AA", and "MA" are the subject-guided, appearance and motion attributes, respectively.

	SG	AA	MA	m_tIoU	m_vIoU	vIoU@0.3	vIoU@0.5
❶	-	-	-	52.2	38.4	61.7	36.2
❷	✓	✓	-	52.3	38.6	62.4	36.2
❸	✓	-	✓	52.7	38.6	61.3	36.6
❹	-	✓	✓	52.6	38.8	61.9	**36.8**
❺	✓	✓	✓	**53.0**	**39.1**	**63.1**	**36.8**

❑ Experiments – Validation of Generality

- Apply our TTS and ASA modules on two popular frameworks, including TubeDETR [Yang et al, CVPR'2022] and STCAT [Jin et al, NeurIPS'2022]

Table 10: Incorporate the TTS and ASA modules into different methods on HCSTVG-v1 (%). ♦: results from the original paper. ♦: retrained results.

	Method	TTS + ASA	m_tIoU	m_vIoU	vIoU@0.3	vIoU@0.5
❶	TubeDETR♦	-	43.7	32.4	49.8	23.5
❷	TubeDETR♦	-	43.2	31.6	49.1	25.5
❸	TubeDETR♦	✓	**45.5** (+2.3)	**33.5** (+1.9)	**53.0** (+3.9)	**27.1** (+1.6)
❹	STCAT♦	-	49.4	35.1	57.7	30.1
❺	STCAT♦	-	48.3	34.9	57.2	29.8
❻	STCAT♦	✓	**50.0** (+1.7)	**36.7** (+1.8)	**59.9** (+2.7)	**31.7** (+1.9)

Observation:

- TTS and ASA are general and applicable to other methods for improvements

The Proposed TA-STVG Approach

- ❑ Experiments – Demos

Text: *The woman goes to the man and talks to him.*

Text: *The man turns around and points to the woman in the blue skirt, and takes a few steps to stop.*

Red box: our results;
Green box: groundtruth.

Knowing Your Target 🎯: Target-Aware Transformer Makes Better Spatio-Temporal Video Grounding

Code and model:
https://github.com/HengLan/TA-STVG

Thank You! ☕