Context-Guided Spatio-Temporal Video Grounding

Supplementary Material

In this supplementary material, we present more details and analysis as well as results of our work, as follows,

A. Detailed Architectures of Modules
We display the detailed architectures for `SAEncoder`$(\cdot)$, `SA`$(\cdot)$ and `CA`$(\cdot)$ in the main text. In addition, we present the architectures for different usage of temporal and spatial confidence scores.

B. Additional Ablation on Motion Information
We conduct an extra experiment to ablate motion information in our approach.

C. More Visualization Analysis on Attention Maps
We include more visualization analysis on the attention maps to show the effectiveness of instance context in improving target-awareness for localization.

D. More Qualitative Results
We demonstrate more qualitative results of our method for grounding the target object.

E. Additional Analysis of Efficacy and Complexity
We compared the efficacy and model complexity with other methods.

A. Detailed Architectures of Modules

A.1. Architecture of Self-Attention Encoder

The self-attention encoder module, *i.e.*, `SAEncoder`$(\cdot)$, is to enhance multimodal feature $\mathcal{X}'$ and output $\tilde{\mathcal{X}}$, which is composed of L (L=6) standard self-attention blocks, as depicted in Fig. 1.

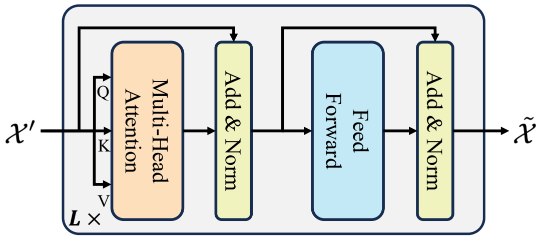
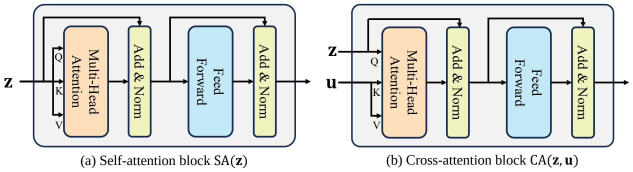


Figure 1. Detailed architecture of `SAEncoder`$(\cdot)$.

A.2. Architectures of Attention Blocks in Decoder

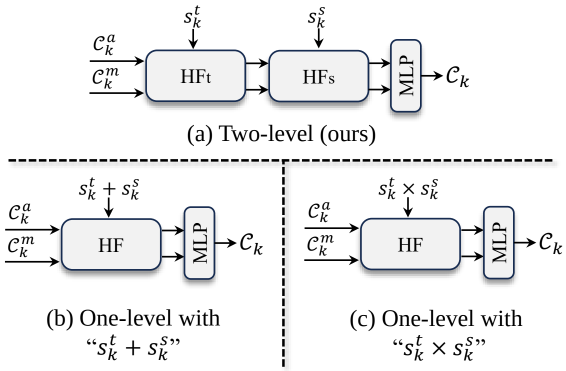
In our context-guided decoder, we employ attention blocks, including the self-attention block, *i.e.*, `SA`$(\mathbf{z})$ and the cross-attention block, *i.e.*, `CA`$(\mathbf{z}, \mathbf{u})$. There architectures are shown in Fig. 2.



(a) Self-attention block SA($\mathbf{z}$)

(b) Cross-attention block CA($\mathbf{z}, \mathbf{u}$)

Figure 2. The architectures `SA`$(\mathbf{z})$ and `CA`$(\mathbf{z}, \mathbf{u})$ in our model.



(a) Two-level (ours)

(b) One-level with “$s_k^t + s_k^s$”

(c) One-level with “$s_k^t \times s_k^s$”

Figure 3. Architectures for different usage of spatial and temporal confidence scores. Image (a) is the proposed two-level architecture, and image (b) and (c) are two one-level variants with “addition” and “multiplication”, respectively.

Motion		m_tIoU	m_vIoU	vIoU@0.3	vIoU@0.5
Feature	Context				
-	-	49.28	35.81	58.36	29.91
✓	-	51.56	37.62	59.57	32.16
✓	✓	**52.84**	**38.42**	**61.47**	**36.29**

Table 1. Ablation of motion information on HCSTVG-v1 (%).

A.3. Architectures for Different Usage of Temporal and Spatial Confidence Scores

In the ablation, we compare our two-level strategy with two additional one-level strategiess for exploiting the temporal and spatial confidence scores. The structures of these three mechanisms are compared and illustrated in Fig. 3.

B. Ablation on Motion Information

Motion information is complementary to appearance cues and can benefit STVG. Specifically, it provides a few extra advantages: (1) Motion features contain the movement details of the target object, which are crucial for the STVG task; (2) Motion cues can provide useful temporal information to some extent even when the appearance partially invisible; (3) Motion features can better comprehend the spatial relationships between objects in the video, such as distance and relative position. Thus inspired, we utilize both appearance and motion features in our CG-STVG, as in many other STVG methods. To study the impact of motion information in our CG-STVG, we provide additional ablation results in Tab. 1. As shown in Tab. 1, with the help of the motion features $\tilde{\mathcal{X}}_m$ in multimodal feature $\tilde{\mathcal{X}}$, the m_vIoU increases by 1.81, achieving 37.62, which shows that motion features can provide the necessary action information for STVG. After integrating motion context extracted from motion features, the m_vIoU score has improved to 38.42, demonstrating the effectiveness of motion context.

C. More Visualization on Attention Maps

In order to analyze the role of instance context, we compare the attention maps of spatial queries in the spatial-decoding block (SDB), with and without using instance context, as in Fig. 4. From Fig. 4, we can clearly see that, the queries, without being enhanced by the instance context, are unable to focus on the foreground object across different frames. However, when employing instance context, the queries are significantly enhanced by gaining more target-awareness knowledge to focus on the foreground regions, which benefits accurate localization of the target object and thus improves the STVG performance.

D. More Qualitative Results

To further validate the effectiveness of our method (with instance context), we provide additional examples of grounding results compared to the baseline method (without instance context) on the HCSTVG dataset in Fig. 5. From the shown visualizations, the baseline model struggles to locate the target object accurately within the video frames. However, when employing the mined instance visual context, our method is able to localize the target object with better temporal and spatial accuracy.

In detail, for the second example, the key words in the text "The woman in the white skirt adjusts her skirt and walks slowly to the other woman" are "white" and "walks". However, since there are three women wearing white in the video, the information of "white" may not be discriminative and useful. Therefore, we can merely rely on other information such as "walks" and only the fourth frame of the video contains information about "walks". As a result, our method and the baseline method can both accurately locate the target in the fourth frame. However, since the baseline method does not have context guidance, there is no reliable information to use in the remaining frames, leading to errors in time and space localization. Our method, on the other hand, accurately locates the target by using the mined

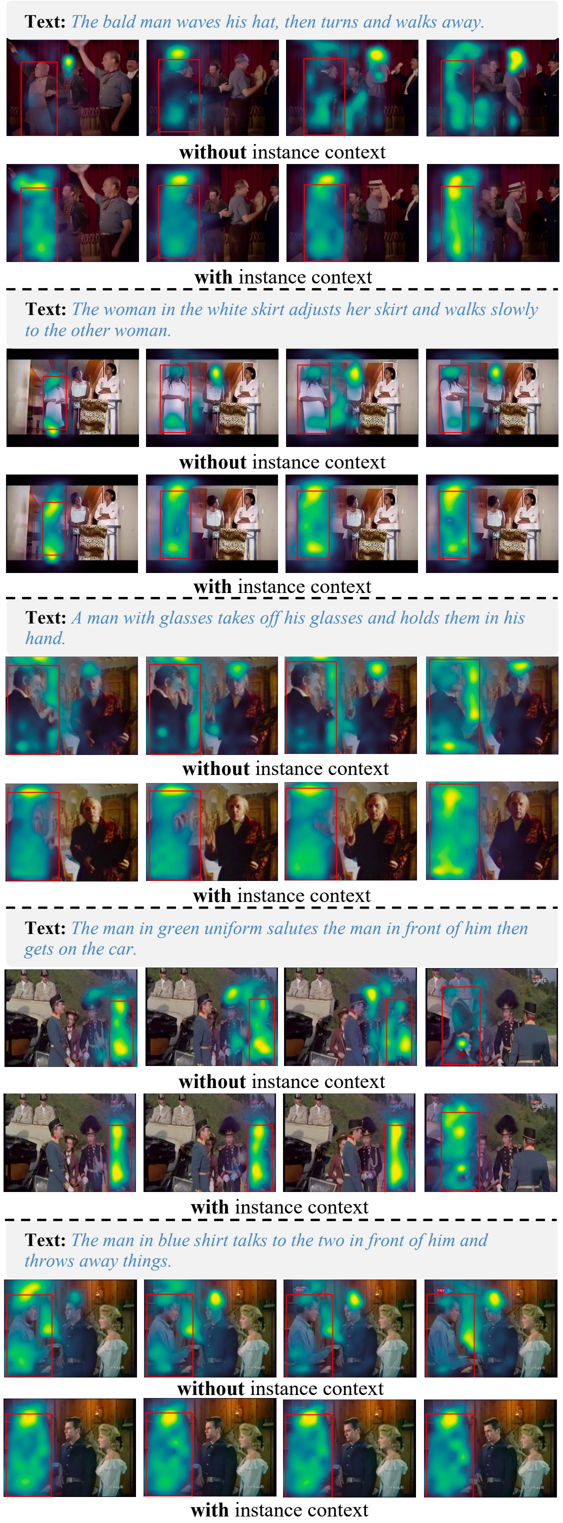

Text: *The bald man waves his hat, then turns and walks away.*
without instance context
with instance context
Text: *The woman in the white skirt adjusts her skirt and walks slowly to the other woman.*
without instance context
with instance context
Text: *A man with glasses takes off his glasses and holds them in his hand.*
without instance context
with instance context
Text: *The man in green uniform salutes the man in front of him then gets on the car.*
without instance context
with instance context
Text: *The man in blue shirt talks to the two in front of him and throws away things.*
without instance context
with instance context

Figure 4. Attention maps for spatial queries in video frames in the spatial-decoding block *without* and *with* our proposed instance context. The red boxes indicate the foreground object to localize.

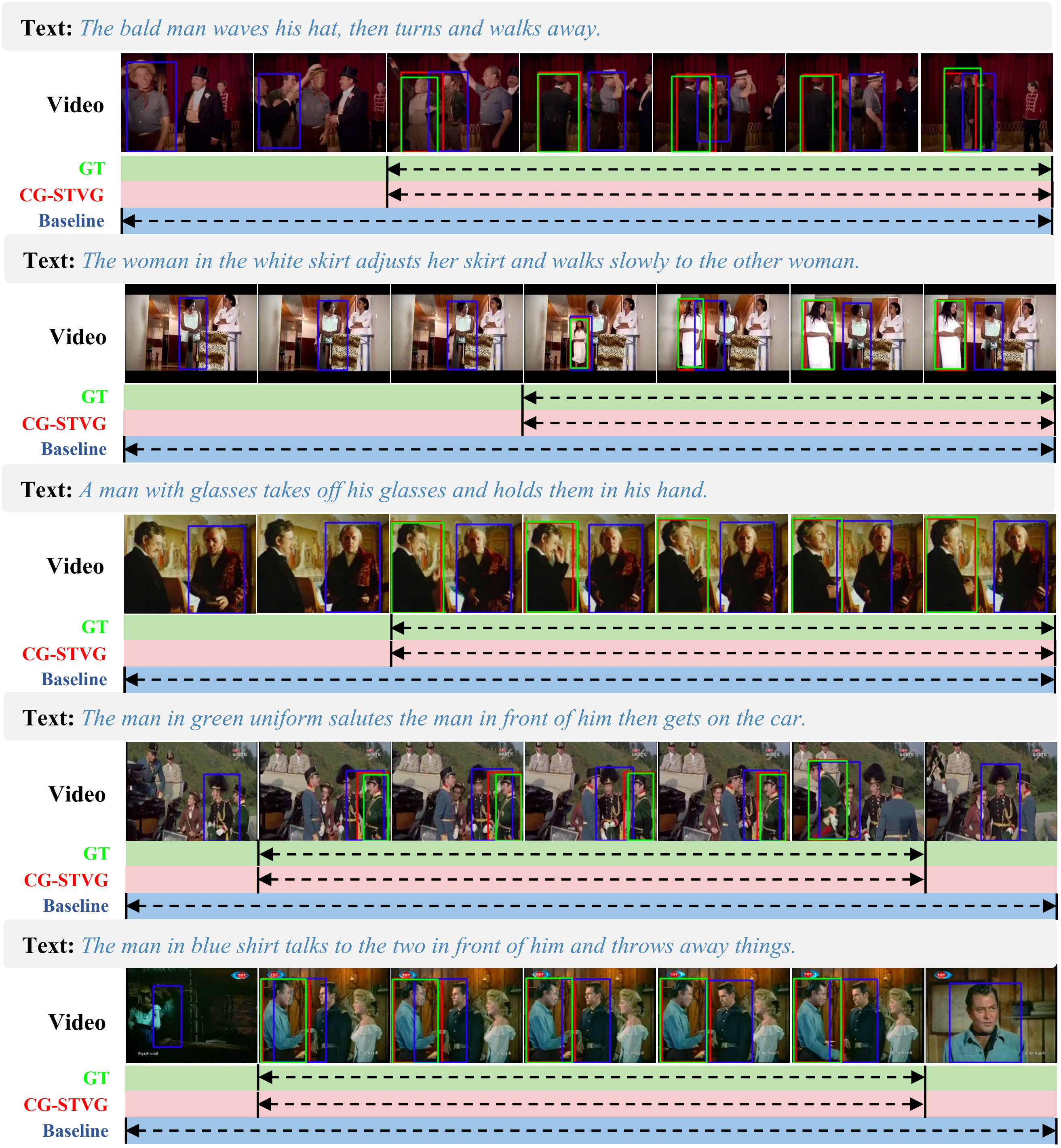

Text: The bald man waves his hat, then turns and walks away.
Video
GT
CG-STVG
Baseline
Text: The woman in the white skirt adjusts her skirt and walks slowly to the other woman.
Video
GT
CG-STVG
Baseline
Text: A man with glasses takes off his glasses and holds them in his hand.
Video
GT
CG-STVG
Baseline
Text: The man in green uniform salutes the man in front of him then gets on the car.
Video
GT
CG-STVG
Baseline
Text: The man in blue shirt talks to the two in front of him and throws away things.
Video
GT
CG-STVG
Baseline

Figure 5. Qualitative results of our method (red), our baseline method (blue) and ground truth (green). In all examples, our method shows better localization than the baseline, which proves the importance of instance context.

instance context from the fourth frame to assist in locating the target in the remaining frames. The same comparison can also be observed in the fourth example. These examples further demonstrate the importance of instance context in guiding target localization.

Methods	Params		Training			Inference		FLOPs	m_vIoU
	Trainable	Total	Time	GPU Mem	GPU Num	Time	GPU Mem		
TubeDETR [4]	185	185	48 h	29.9	16 V100	0.40 s	24.4	1.45 T	30.4
STCAT [1]	207	207	12 h	39.2	32 A100	0.51 s	29.4	2.85 T	33.1
CSDVL [2]	-	-	∼ 48 h	-	8 A6000	-	-	-	33.7
Baseline	200	228	12 h	41.2	32 A100	0.53 s	29.6	2.89 T	32.4
CG-STVG	203	231	13.6 h	43.9	32 A100	0.61 s	29.7	3.03 T	34.0

Table 2. Complexity comparison on VidSTG [3]. An A100-GPU is used for all inferences. Note, key information of CSDVL is not shown due to no available code, and its GPU number and training time are from its supplementary material. Please notice, since VidSwin-T is frozen in training, the number of trainable parameters are less than the total number of parameters.

E. Efficacy and Complexity of the Model

We compare our method and other models (**note**, CSDVL does **not** provide code) as in Tab. 2. We can see our method has ***similar*** efficacy and complexity with recent SOTAs (*e.g.*, STCAT) and baseline, but ***better*** m_vIoU of 34.0%, showing it effectiveness.

References

[1] Yang Jin, Yongzhi Li, Zehuan Yuan, and Yadong Mu. Embracing consistency: A one-stage approach for spatio-temporal video grounding. In *NeurIPS*, 2022. 4

[2] Zihang Lin, Chaolei Tan, Jian-Fang Hu, Zhi Jin, Tiancai Ye, and Wei-Shi Zheng. Collaborative static and dynamic vision-language streams for spatio-temporal video grounding. In *CVPR*, 2023. 4

[3] Ze Liu, Jia Ning, Yue Cao, Yixuan Wei, Zheng Zhang, Stephen Lin, and Han Hu. Video swin transformer. In *CVPR*, 2022. 4

[4] Antoine Yang, Antoine Miech, Josef Sivic, Ivan Laptev, and Cordelia Schmid. Tubedetr: Spatio-temporal video grounding with transformers. In *CVPR*, 2022. 4