Spatio-Temporal Video Grounding of Human Actions for Human-in-the-Loop Artificial Intelligence

Hans Tiwari[1], Selen Pehlivan[2], Jorma Laaksonen[1]

[1]Aalto University School of Science, Espoo, Finland
[2]VTT Technical Research Centre of Finland, Oulu, Finland
hans.tiwari@aalto.fi, selen.pehlivantort@vtt.fi, jorma.laaksonen@aalto.fi

Abstract

In this paper, we study *spatio-temporal video grounding*, that is a recently introduced *computer vision* task, for *human-in-the-loop* artificial intelligence setups. We hypothesize an AI expert system capable of giving instructions for building or servicing technical apparatus. The task of the spatio-temporal video grounding subsystem in the setup is to observe the actions of the human operator, ground the given instructions in the video frames, and provide this information back to the AI system for verification and preparation of the next instruction. Our experiments were carried out with the large HC-STVG dataset of human action videos. The results of the experiments show that our proposed enhancements to the state-of-the-art STCAT architecture provide improved performance, especially in cases where the operated objects are small in their spatial size. Spatio-temporal video grounding will prove to be a necessary building block in future assistive AI systems that relate to human actions and their automated observation.

1 Introduction

Human-in-the-loop or collaborative AI systems are a solution to needs where neither a human alone nor an artificial intelligent system alone can perform some required task. In such systems, the role of the human is typically either to make the final judgement based on the suggestions given by the machine, or to perform physical tasks that the expert system is not equipped to carry out. In our current work we are interested in the latter category of human-in-the-loop AI setups.

Figure 1 depicts our hypothesized setup that consists of 1) an AI expert system capable of giving building and servicing instructions to human operators, 2) a camera that observes the operator, and 3) a spatio-temporal video grounding subsystem that grounds the given instructions in the recorded video frames. The AI system will then receive the results of the video grounding stage, deduce whether the action was correctly performed, and prepare the next instructions based on the outcomes of the previous ones.

Video grounding refers to the task of locating a specific segment within a video that corresponds to a given textual description. This can be considered as a crucial intersection between computer vision and natural language processing, with applications in video indexing, retrieval, and analysis (Zhang et al. 2019). The generic video grounding setup comes with two specialized tasks: *temporal* video grounding and *spatio-temporal* video grounding.

In temporal video grounding (TVG), the objective is to identify the start and end times of a segment where a described action or event takes place. It essentially narrows down the time window where the event happens but does not provide specific spatial information about where in the frames the action or object is located.

Spatio-temporal video grounding (STVG) takes the objective a step further. While retaining the temporal grounding aspect, it introduces the spatial dimension into the equation. Instead of just returning a time segment, STVG pinpoints the exact spatial locations within that time segment where the described action or object is present. This is usually represented as a series of bounding boxes across the identified frames, which creates a tube-like structure.

In this paper we study a number of improvements on the state-of-the-art STCAT (Jin et al. 2022) model for STVG. The results of the experiments on the large HC-STVG (Tang et al. 2021) dataset show that our proposed enhancements provide improved performance. Further, we discuss the applicability and necessity of spatio-temporal video grounding for human-in-the-loop or collaborative AI systems.

2 Background

Several approaches have been proposed for temporal and spatio-temporal video grounding, with advancements often revolving around improved feature representations, attention mechanisms, and training strategies. These include:

1. **Cross-modal Interaction Frameworks:** Such methods aim at fostering a rich interaction between the video and text representations. The MAN (Memory Augmented Network) model (Zhang et al. 2019) is an exemplar, utilizing memory networks to capture the cross-modal dynamics.
2. **Rank-based Approaches:** These approaches generate multiple candidate segments and rank them based on their relevance to the textual description (Gao et al. 2020).
3. **Use of Pre-trained Models:** Leveraging pre-trained models from both vision (e.g., ResNet (He et al. 2016)) and text (e.g., BERT (Devlin et al. 2019)) domains to

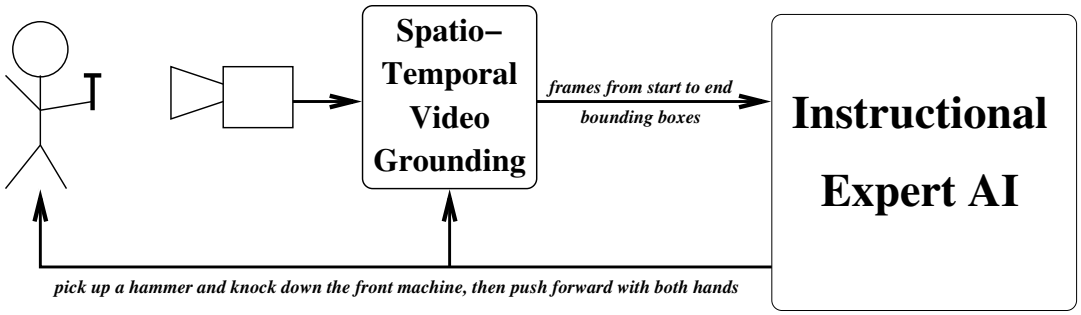


Figure 1: The block diagram of the hypothesized instructional AI system with a spatio-temporal video grounding subsystem.

extract richer features has been a recent trend, further boosted by fine-tuning on the grounding task.

4. **Self-attention Mechanisms:** Building upon the success of Transformers in NLP (Vaswani et al. 2017), recent approaches have started employing self-attention mechanisms to selectively focus on video frames that are most relevant to a given textual description. This allows models to automatically weigh the significance of different segments in a video.
5. **Cross-modal Embeddings:** The development of unified embedding spaces where video frames and textual tokens coexist has been another trend. These shared spaces facilitate better understanding and alignment between the two modalities (Zhang et al. 2020b).
6. **Few-shot Learning:** With the vast diversity in video content and textual descriptions, few-shot learning techniques are being employed to adapt models to new tasks with limited labeled data (Zhou et al. 2020).

Despite the remarkable progress first in temporal and then also in spatio-temporal video grounding, several challenges have still persisted, including:

1. **Ambiguity in Descriptions:** Textual descriptions can sometimes be vague or have multiple valid groundings in a video, making it challenging to identify the correct segment.
2. **Diverse Temporal Structures:** Actions in videos do not always follow a linear progression. Some might have long build-ups, while others can be instantaneous. This variability demands flexible models that can adapt to diverse temporal structures.
3. **Real-time Applications:** For use-cases like live video streaming, there is a need for models that can perform grounding tasks in real-time, which is computationally challenging given the complexity of state-of-the-art models.

While these challenges underscore the intricacies of video grounding, they also highlight the avenues for future research and innovation, emphasizing the dynamic and evolving nature of this domain.

In addition to the STCAT model (Jin et al. 2022), which we have chosen for our experiments and will address in full detail in the next section, also a few other notable STVG model have emerged recently. These include the STVG-BERT model (Su, Yu, and Xu 2021), which employs separate branches for spatial and temporal grounding tasks and merges tjem in subsequent processing. Another remarkable model has been the TubeDETR model (Yang et al. 2022), based on the Transformer-based DETR (Carion et al. 2020) end-to-end object detector network.

3 STCAT Model

In the exploration of STVG, the paper titled *"Embracing Consistency: A One-Stage Approach for Spatio-Temporal Video Grounding"* (STCAT) (Jin et al. 2022) has offered a novel perspective on the subject. This work introduces a one-stage methodology that emphasizes the importance of consistency in the STVG task. Building upon existing foundational models, the paper provides a comprehensive approach that aims to further refine and enhance the accuracy and efficiency of video grounding techniques.

Earlier to STCAT, many proposed methods have approached STVG as if it were a parallel frame-grounding problem. This approach, however, has led to numerous challenges, most notably to issues related to feature and prediction inconsistencies. These can summarized as follows:

1. **Feature Alignment Inconsistency:** In the realm of STVG, the significance of global context modeling cannot be overstated. To elucidate, pinpointing the subject **"adult"** from a query such as **"An adult in blue grabs a ball on the basketball court"** mandates a holistic grasp of the video narrative. Yet, several established methodologies, as discussed earlier, tend to confine cross-modal fusion to immediate, short-span video contexts. Even contemporary strategies predominantly adhere to frame-centric alignment.
2. **Prediction Inconsistency:** A notable fraction of extant approaches conceptualize STVG as a parallel, frame-by-frame grounding endeavor. These techniques delineate a bounding box for every individual frame, correlating it with the query, but often neglect inter-frame consistency. Considering the primary goal of STVG is to spatially and temporally anchor a distinct target instance as described by the textual narrative, it is imperative that the grounding remains coherent across frames, irrespective of potential variations in the target's visual representation due to elements like camera transitions or dynamic scenes.

Both of these inconsistencies alone – and especially when occurring together – can hinder the accuracy and reliability of the grounding process, making it imperative to seek alternative strategies. To address the challenges, STCAT (Jin et al. 2022) introduces an innovative end-to-end one-stage framework. Its essence lies in its ability to ensure consistent grounding across video frames. It achieves this by employing a novel multi-modal template as a global objective. This template serves to constrict the grounding region, ensuring that predictions are consistently associated across different video frames. A visual representation of this architecture, depicted at a coarse scale, is shown in Figure 2.

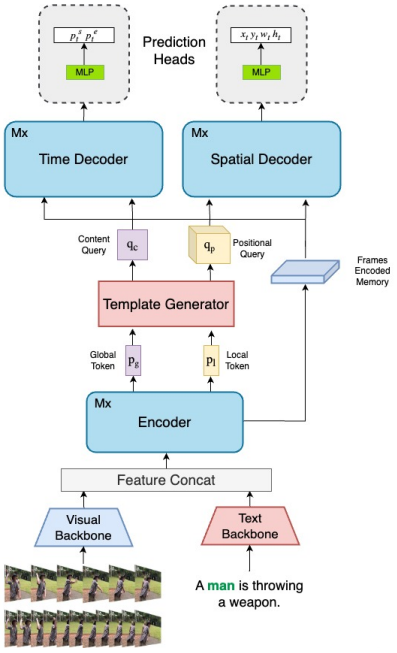
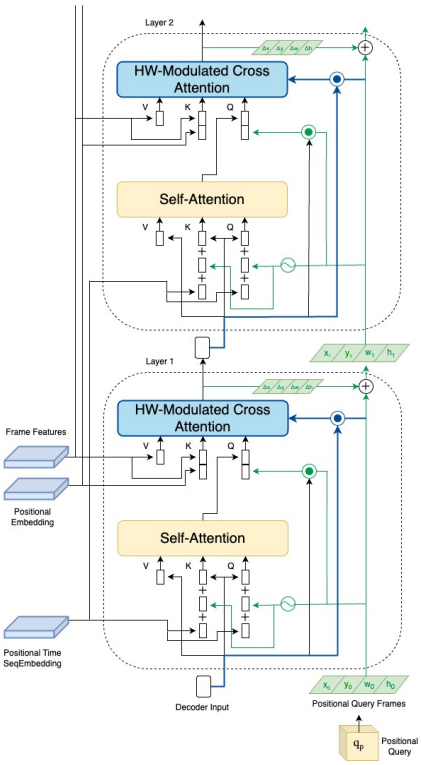


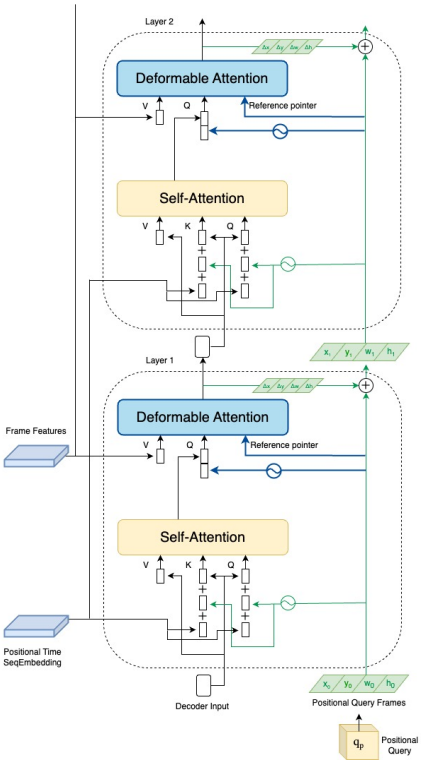
Figure 2: The STCAT architecture (Jin et al. 2022) shown at a coarse scale.

The attention unit in Transformers, particularly the *cross-attention unit* (Vaswani et al. 2017), plays a crucial role in capturing dependencies between different parts of the input data. In essence, the cross-attention mechanism allows a Transformer to focus on specific parts of the input when producing an output, enabling it to handle long-range dependencies and contextual relationships effectively. This becomes especially vital in tasks like machine translation, where understanding the context from both the source and target sequences is paramount.

In the context of STVG, the attention unit's functionality and its variants are of significant interest. As the STCAT model in the time of writing this paper stands as the state-of-the-art in STVG tasks, we have based our video grounding model on it and additionally proposed various modifications in its architecture, especially in its spatial decoder's attention unit. We call these modifications as *Width-Height Modulation* and *Deformable Attention Transformer* and address them in detail in the following.



(a) Width-Height Modulation.



(b) Deformable Attention.

Figure 3: Two proposed variants of the spatial decoder's attention unit for the STCAT model.

Width-Height Modulation

The concept of width-height modulation emerges as an enhancement in the STCAT's decoder's attention unit as shown in Figure 3a. Traditional positional attention maps, often visualized as Gaussian-like priors, have been conventionally assumed to be isotropic with a fixed size for all objects. This assumption inadvertently neglects the scale information, specifically the width and height of objects.

To address this limitation and enhance the positional prior, a novel approach is proposed in DAB-DETR (Liu et al. 2022), which is being leveraged here: the integration of scale information directly into the attention maps. In the conventional positional attention map, the query-to-key similarity is computed as:

$$Attn((x,y),(x_{ref},y_{ref})) = \frac{PE(x)\cdot PE(x_{ref}) + PE(y)\cdot PE(y_{ref})}{\sqrt{D}}, \tag{1}$$

where $PE(\cdot)$ stands for the sinusoidal position encoding and the $\frac{1}{\sqrt{D}}$ factor serves as a rescaling term, as suggested by Vaswani et al. (2018). To better accommodate objects of varying scales, the positional attention maps can be modulated by dividing the relative anchor width and height from its x and y components, respectively. This modulation can be represented as:

$$Modulate_x = PE(x)\cdot PE(x_{ref})\cdot \frac{w_{q,ref}}{w_q}, \tag{2}$$

$$Modulate_y = PE(y)\cdot PE(y_{ref})\cdot \frac{h_{q,ref}}{h_q}, \tag{3}$$

$$ModulateAttn((x,y),(x_{ref},y_{ref})) = \frac{Modulate_x + Modulate_y}{\sqrt{D}}, \tag{4}$$

where w_q and h_q denote the width and height of the anchor A_q, and $w_{q,ref}$ and $h_{q,ref}$ represent the reference width and height. These two are computed from the positional query C_q as:

$$w_{q,ref}, h_{q,ref} = \sigma(MLP(C_q)). \tag{5}$$

This modulation in the positional attention facilitates robust extraction of features from objects with diverse widths and heights.

Deformable Attention Transformer

The Deformable Attention Transformer (Zhu et al. 2021) introduces a novel approach to address the challenges of applying Transformer attention on image feature maps. Traditional Transformer attention would consider all possible spatial locations, leading to potential inefficiencies. The deformable attention module, inspired by deformable convolution, focuses only on a select set of key sampling points around a reference point. This approach, regardless of the spatial size of the feature maps, can mitigate issues related to convergence and spatial resolution of features.

Given an input feature map $x \in R^{C\times H\times W}$, the deformable attention feature is formulated as:

$$DeformAttn(z_q,p_q,x) = \sum_{m=1}^{M} W_m \sum_{k=1}^{K} A_{mqk}\cdot W_m^0 x(p_q + \Delta p_{mqk}), \tag{6}$$

where m and k index the attention head and the sampled keys respectively. The attention weight A_{mqk} is normalized and the sampling offset Δp_{mqk} is obtained through linear projection over the query feature z_q.

In the STCAT framework, the positional query output from its Template Generator is utilized as an input reference point for the deformable attention as shown in Figure 3b. One of the challenges when adapting Deformable Attention to STCAT is the integration of textual features directly into its implementation. To address this, several experiments were conducted. In one approach, the text was not fed directly; instead, the textual context was conveyed through the positional query. In another approach, harnessing the adaptability of deformable attention to accommodate multi-scale features, rather than merging multiple visual feature scales, this approach employs a singular visual feature scale and appends the textual feature to it in a comparable manner. In the experiments reported in this paper, the textual feature is fed directly.

Training Objectives

STCAT employs a set of pivotal loss functions to optimize the model's performance. The models reported in this paper also utilize the same objective with the same loss components. The primary loss components utilized are:

L1 Loss (L_{L1}). Measures the absolute differences between the true and predicted values. It is used in the spatial localization of the model to compute the box prediction loss L_{bbox}.

gIoU Loss (L_{giou}). Generalized Intersection over Union loss, which is particularly useful for object detection tasks. It is also used in the spatial localization of the model as part of the L_{bbox} computation.

KL Divergence Loss (L_s and L_e). Measures the difference between two probability distributions. In the context of temporal localization, it is used to compute the divergence between target and predicted distributions for starting and ending positions.

Binary Cross Entropy (L_{seg}). Used for binary classification tasks to measure the error between true and predicted values. In this work, it is used to predict whether a frame belongs to the ground-truth segment.

Guided Attention Loss (L_{att}). The attention mechanism produces an attention map, which represents the model's focus on different parts of the input. Ideally, the model should focus on the most relevant parts of the input, the action part, to make accurate predictions. This loss penalizes the model for focusing on not-so-relevant information. It does so by computing the negative logarithm of the attention weights corresponding to these regions. To ensure that the loss is not dominated by the sheer number of these, it is normalized by the number of relevant samples in the attention map.

The losses are combined together to compute the composite loss L defined as:

$$L = L_{bbox} + \lambda_{temp}L_{temp} + \lambda_{seg}L_{seg} + \lambda_{att}L_{att}. \tag{7}$$

4 Dataset

The *Human-centric Spatio-Temporal Video Grounding* (HC-STVG) dataset (Tang et al. 2021) is a specialized collection focusing solely on humans in videos. This dataset provides 16,500 annotation-video pairs from various movie scenes. Each video clip is accompanied by a descriptive statement and trajectories of the corresponding person, represented as a series of bounding boxes. Notably, all clips include multiple individuals, enhancing the challenge of video comprehension.The primary objective of the STVG task, as detailed in the previous sections, is to accurately localize the spatio-temporal segment or tube of an untrimmed video that corresponds to a provided sentence describing an object.

Figure 4 shows frames from one training set video from the HC-STVG dataset together with the sentence grounded in them. One can notice the obvious difficulty of the data.

Figure 4: A selection of frames used in grounding *"The man in the brown hat picks up a hammer and knocks down the front machine, then pushes forward with both hands."*

The primary objective of the HC-STVG benchmark task is to spatio-temporally localize the spatio-temporal section of an untrimmed video that corresponds to a provided sentence describing actions of an individual. This dataset offers a benchmark with spatio-temporal annotations related to target persons in intricate multi-person scenarios, complete with comprehensive interaction and action details. The training split of the dataset contains 10131, the validation split 2000, and the test split 4413 videos, respectively. The dataset ensures that test and training samples are not derived from the same raw video. Key features of the initial HC-STVG dataset include:

1. Human-centric annotations with precise spatio-temporal details and textual descriptions for the person of interest.
2. Videos captured in complex multi-person scenes, with 57.2% of the videos containing more than three individuals.
3. Rich descriptive sentences detailing human-human or human-object interactions, with 56.1% of the descriptions encompassing both interaction types.

The construction process has encompassed five stages (Tang et al. 2021):

1. **Raw Video Preparation:** Videos are filtered using the AVA dataset (Gu et al. 2018), sourced from YouTube, ensuring diversity and high quality.
2. **Video Span Selection:** Annotators mark suitable describable spans from untrimmed videos, focusing on multi-person scenes.
3. **Video Description:** Annotators provide descriptions for the target person, emphasizing visual attributes, actions, and relations.
4. **Bounding Box Annotation:** Frame-level bounding box annotations are created using a combination of manual keyframe annotations and automatic tracking.
5. **Video Span Extension:** Video clips are temporally extended to achieve a fixed length, ensuring no ambiguity in referring.

Throughout our research, the HC-STVG dataset has been predominantly utilized for the majority of testing and comparisons. This preference stems from its comprehensive public availability and its lightweight nature in comparison to the other publicly available STVG datasets.

5 Performance Measures

In this paper and the referred STVG task, a set of evaluation metrics are adopted to rigorously assess the performance of our models. These metrics have been introduced in (Zhang et al. 2020a; Chen et al. 2019; Ging et al. 2020) and are detailed as follows.

Mean Temporal Intersection over Union (m_tIoU)

This metric evaluates the temporal localization performance of the model. Specifically, it calculates the average temporal Intersection-over-Union (tIoU) between the predicted video clips and the ground truth clips. The tIoU is defined as the ratio of the intersection to the union of the predicted and ground truth clips, mathematically represented as,

$$tIoU = \frac{|Ti|}{|Tu|}, \tag{8}$$

where Ti and Tu are the intersection and union between the temporal locations of predicted tube and ground-truth tube, respectively.

Mean Visual Intersection over Union (m_vIoU)

This metric provides an average of the vIoU scores across all testing videos. The vIoU is computed as:

$$vIoU = \frac{1}{|Su|} \sum_{t \in Si} IoU(\hat{bt}, bt), \tag{9}$$

where $\hat{bt}$ and bt are the detected and ground-truth bounding boxes at frame t, respectively, and IoU is the Intersection-over-Union between these bounding boxes. Si and Su represent the union between the predicted and ground-truth tubes.

vIoU@0.3 and vIoU@0.5

These metrics measure the proportion of samples for which the vIoU score exceeds a certain threshold (0.3 and 0.5, respectively). Specifically, $vIoU@R$ represents the ratio of samples with $vIoU > R$ in the testing subset.

Mean Ground Truth vIoU (m_gt_vIoU).
The metric m_gt_vIoU is computed analogously to $vIoU$, with a distinct difference in its focus. While the standard $vIoU$ calculates the IoU between predicted boxes of the frames where actions are predicted, m_gt_vIoU evaluates the IoU between the ground truth boxes and the predicted frames specifically the ground-truth temporal segment of the action. This subtle yet crucial distinction allows for a more nuanced evaluation of the model's spatial accuracy. The extended metrics, $m\_gt\_vIoU@0.3$ and $m\_gt\_vIoU@0.5$, are derived in a similar manner, setting IoU thresholds at 0.3 and 0.5, respectively. The primary significance of employing this metric is to segregate the spatial performance evaluations from other metrics, ensuring a more isolated and focused assessment of the bounding box predictions.

These evaluation criteria are pivotal in understanding both the temporal and spatial accuracy of the model in the context of STVG task. As crucial as the above metrics are, to gain a deeper understanding of the model's performance across different scales of bounding boxes within the dataset, several variations of these metrics in this paper are introduced. These metrics extend the current spatial evaluation metrics, such as m_vIoU and its thresholded versions, to specific subsets of the test set. Given the variability in width and height for each sample in the dataset, these subsets are delineated based on the threshold of the metric as follows:

$$\textbf{BBoxRatio} = \frac{1}{N}\sum_{t=1}^{N}\frac{GTBBoxArea_t}{FrameTotalArea_t}, \qquad (10)$$

where t indexes the ground truth frames in the video from 1 to N covering the action occurrence and GTBBox respresenting the ground truth bounding boxes in the corresponding video.

Subsequently, the videos are categorized into *Small*, *Medium*, and *Large* subsets, each containing a varying number of samples based on the thresholded BBoxRatio in Equation (10). In the HC-STVG dataset, observing the histogram of the BBoxRatio values in the test set in Figure 5, the used thresholds are as follows:

1. **Small:** The threshold is defined as $0 < BBoxRatio \leq 0.25$, resulting in 709 samples.

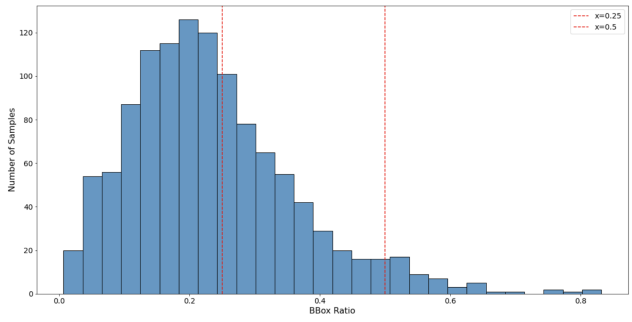


Figure 5: The HC-STVG Test set Histogram to show the distribution of bounding-box Ratios (BBox Ratio).

2. **Medium:** The threshold is defined as $0.25 < BBoxRatio \leq 0.5$, resulting in 401 samples.
3. **Large:** The threshold is defined as $0.5 < BBoxRatio \leq 1.0$, also resulting in 50 samples.

These thresholds categorize the samples based on the bounding-box ratio given in Equation (10), providing a structured approach to evaluate the model's performance across different scales of boxes in the HC-STVG dataset.

6 Results

The results of a subset of our experiments are shown in Table 1. They relate to the published results of the original STCAT model, our replication of it and our improvements on the model. We can see that the best results were obtained with the model where the original spatial decoder had been replaced with the Deformable Attention Transformer.

Method	Best Epoch	m_tiou	m_viou	$viou@0.3$	$viou@0.5$	m_gt_viou	$gt\_viou@0.3$	$gt\_viou@0.5$
STCAT published (Jin et al. 2022)	90/90	49.44	35.09	57.67	30.09	-	-	-
STCAT replicated	57/90	48.00	32.95	54.48	24.22	66.08	90.26	79.83
H&W Modulation: Default parameters	74/90	**49.55**	34.86	56.98	29.48	67.99	90.52	80.60
Deformable Attention: Single Scale + Text Feature V1	67/90	49.52	**35.31**	**58.28**	**30.17**	**69.90**	**91.90**	**83.62**
Deformable Attention: Single Scale + Text Feature V2	60/90	47.85	33.56	54.48	27.16	68.47	91.21	82.24
Deformable Attention: Multi-Scale + Text Feature V1	74/90	41.33	21.24	29.74	11.90	50.92	71.21	57.24

Table 1: Best epoch result for a variety of STCAT based models on the HC-STVG test set.

In another series of experiments we studied the models' performances on different bounding box sizes. A selection of these results are shown in Table 2. As can be seen, the Deformable Attention variant was again superior to the other studied models. Most importantly, it performed best for all object sizes. It can be also observed that the performances of all the models are worse for the small objects than for the medium-sized ones.

Methods	Small		Medium		Large	
	vIoU@0.3	vIoU@0.5	vIoU@0.3	vIoU@0.5	vIoU@0.3	vIoU@0.5
STCAT replicated	53.17	26.94	61.10	32.17	60.00	38.00
Deformable Attention: Single Scale + Text Feature V1	**53.74**	**28.07**	**61.85**	**36.91**	**62.00**	**44.00**
Deformable Attention: Single Scale + Text Feature V2	53.17	24.82	59.35	34.66	42.00	36.00
Deformable Attention: Multi-Scale + Text Feature V1	26.80	10.30	35.16	15.96	32.00	14.00

Table 2: Performance evaluation of the Bounding Box Size Variants on the HC-STVG test set.

7 Discussion

The results of our experiments showed that we could slightly improve the performance of the state-of-the-art STCAT model for spatio-temporal video grounding on the demanding HC-STVG dataset. It cannot yet be said, whether the obtained accuracy would be high enough for the STVG subsystem to be useful in a human-in-the-loop collaborative AI system. In order to get a proper answer to that question, one should experiment with a complete instructional AI system with fully-implemented action–reaction abilities and a number of genuine humans in the loop.

Our experiments with the HC-STVG dataset were biased from the hypothesized human-in-the-loop AI setup as only a fraction of the videos in the dataset were instructional ones. Nevertheless, all of the videos depicted human actions and therefore the HC-STVG data could in any case be used for pre-training a model that would then be fine-tuned with a smaller more task-specific video collection.

There also exist already a number of well-curated instructional video datasets, e.g., How2 (Sanabria et al. 2018), YouCook2 (Zhou, Xu, and Corso 2018), and HowTo100M (Miech et al. 2019). Similarly, many first-person or egocentric video datasets exist, and such a vision setup could also be used in our proposed human-in-the-loop AI setting. The applicability of all these datasets to the STVG task is, however, still limited as long as they have not been annotated for that purpose with action time-codes and bounding boxes. Would such annotations become available, both the third- and first-person vision datasets of instructional videos could easily lend themselves for our hypothesized human-in-the-loop collaborative AI setup.

8 Conclusions

In this paper we presented improvements to the state-of-the-art STCAT model for spatio-temporal video grounding and evaluated them with the demanding HC-STVG dataset. We also proposed and elaborated a hypothetical human-in-the-loop collaborative AI setup, where an AI expert system is capable of giving building and servicing instructions to human operators. The actions of the human operators are then recorded and the instructions grounded in the frames of the captured videos. By feeding the results of grounding back to the AI system, the human–computer action–reaction loop can be closed. More suitable data and further research are required in the future for evaluating the performance of our studied models in a genuine human-in-the-loop AI setting.

References

Carion, N.; Massa, F.; Synnaeve, G.; Usunier, N.; Kirillov, A.; and Zagoruyko, S. 2020. End-to-End Object Detection with Transformers.

Chen, Z.; Ma, L.; Luo, W.; and Wong, K.-Y. K. 2019. Weakly-Supervised Spatio-Temporally Grounding Natural Sentence in Video.

Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding.

Gao, J.; et al. 2020. Fast, accurate, and lightweight temporal action proposal generation. In *Proceedings of the European Conference on Computer Vision (ECCV)*.

Ging, S.; Zolfaghari, M.; Pirsiavash, H.; and Brox, T. 2020. COOT: Cooperative Hierarchical Transformer for Video-Text Representation Learning. In Larochelle, H.; Ranzato, M.; Hadsell, R.; Balcan, M.; and Lin, H., eds., *Advances in Neural Information Processing Systems*, volume 33, 22605–22618. Curran Associates, Inc.

Gu, C.; Sun, C.; Ross, D. A.; Vondrick, C.; Pantofaru, C.; Li, Y.; Vijayanarasimhan, S.; Toderici, G.; Ricco, S.; Sukthankar, R.; Schmid, C.; and Malik, J. 2018. AVA: A Video Dataset of Spatio-temporally Localized Atomic Visual Actions.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 770–778.

Jin, Y.; Li, Y.; Yuan, Z.; and Mu, Y. 2022. Embracing Consistency: A One-Stage Approach for Spatio-Temporal Video Grounding.

Liu, S.; Li, F.; Zhang, H.; Yang, X.; Qi, X.; Su, H.; Zhu, J.; and Zhang, L. 2022. DAB-DETR: Dynamic Anchor Boxes are Better Queries for DETR. *CoRR*, abs/2201.12329.

Miech, A.; Zhukov, D.; Alayrac, J.-B.; Tapaswi, M.; Laptev, I.; and Sivic, J. 2019. HowTo100M: Learning a Text-Video Embedding by Watching Hundred Million Narrated Video Clips. In *ICCV*.

Sanabria, R.; Çağlayan, O.; Palaskar, S.; Elliott, D.; Barrault, L.; Specia, L.; and Metze, F. 2018. How2: A Large-scale Dataset for Multimodal Language Understanding. In *Proceedings of NeurIPS*.

Su, R.; Yu, Q.; and Xu, D. 2021. STVGBert: A Visual-linguistic Transformer based Framework for Spatio-temporal Video Grounding. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, 1513–1522.

Tang, Z.; Liao, Y.; Liu, S.; Li, G.; Jin, X.; Jiang, H.; Yu, Q.; and Xu, D. 2021. Human-centric Spatio-Temporal Video Grounding With Visual Transformers.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, ; and Polosukhin, I. 2017. Attention is All You Need. In *Advances in Neural Information Processing Systems*.

Vaswani, A.; et al. 2018. The Transformer – Attention is All You Need. *The Journal of Machine Learning Research*.

Yang, A.; Miech, A.; Sivic, J.; Laptev, I.; and Schmid, C. 2022. TubeDETR: Spatio-Temporal Video Grounding with Transformers.

Zhang, Z.; Zhao, Z.; Zhao, Y.; Wang, Q.; Liu, H.; and Gao, L. 2020a. Where Does It Exist: Spatio-Temporal Video Grounding for Multi-Form Sentences. In *CVPR*.

Zhang, Z.; et al. 2019. MAN: Moment alignment network for natural language moment retrieval via iterative graph adjustment. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 1247–1255.

Zhang, Z.; et al. 2020b. Cross-modal interaction networks for query-based moment retrieval in videos. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, 655–664.

Zhou, L.; Xu, C.; and Corso, J. J. 2018. Towards Automatic Learning of Procedures From Web Instructional Videos. In *AAAI Conference on Artificial Intelligence*, 7590–7598.

Zhou, L.; et al. 2020. Few-shot temporal localization via query-adaptive contrastive learning.

Zhu, X.; Su, W.; Lu, L.; Li, B.; Wang, X.; and Dai, J. 2021. Deformable DETR: Deformable Transformers for End-to-End Object Detection.