STPro: Spatial and Temporal Progressive Learning for Weakly Supervised Spatio-Temporal Grounding

Supplementary Material

Here, we provide more details about our approach, additional results, and visual analysis. We also include and expand tables we could not include in the main paper due to space limitations. In summary the supplementary includes the following:

- Section 1 sheds light on the challenges of using detector and tracker to generate proposal tubelets. We also explore methods to denoise training data and results of the same.
- Section 2 explores the construction of datasets for SA-TCL & CG-SCL accompanied by ablations to demonstrate their effectiveness.
- Section 3 shows ablation study on HCSTVG-2 dataset.
- Section 4 compares STPro with fully-supervised approaches across all datasets.
- Section 5 presents ablations on the filtering criteria used in joint spatio-temporal inference of STPro. We also measure the effects of SLF at inference time and explain the challenges in extending our inference method directly to VidSTG.
- Section 6 discusses how Soft-Label Filtering (SLF), a component of CG-SCL is implemented.
- Section 7 shows the prompts used for the extraction of `POS` and sub-action phrases from original captions via an LLM. It also captures some failure cases.
- Section 8 visually demonstrates qualitative improvements brought about by STPro's individual components (SA-TCL & CG-SCL).

1. Pre-processing Challenges

Upper-Bound Analysis: STVG datasets present considerable challenges, with detection and tracking often falling short, as evidenced by the maximum upper bounds in Table 1. Datasets such as HCSTVG and VidSTG feature rapid zooms, scene shifts, occlusions, and defocusing—conditions under which even state-of-the-art detectors struggle to perform reliably. Moreover, the pre-processing step of tracking detections to generate tubelets introduces additional noise. Challenges such as person crossover, abrupt scene transitions (where large bounding box displacements lead to ID mismatches), viewpoint changes, and instances where only part of the body is visible further complicate the task.

Table 1 presents an upper bound analysis on TRG following two schemes. To the left, we present the upper bound analysis by finding the maximum overlapping detection from Grounding-DINO to the ground-truth bounding box for every frame within the ground-truth temporal boundary. Hence, we consider maximum spatial and temporal overlap. To the right, we present the upper bound analysis by similarly finding which proposed tubelet (as obtained after detection and tracking) has maximum vIoU with the ground-truth tubelet and consider it to be the solution for this sample. There's some decline in performance since tracker adds another level of complexity. We find that our upper bound surpasses the performance of current state-of-the-art fully-supervised approaches by significant margins.

Dataset	m_tIoU	m_vIoU	vIoU@0.5	m_tIoU	m_vIoU	vIoU@0.5
	Detection			Post-Tracking		
HCSTVG-v1	92.1	69.3	86.2	83.6	65.0	76.3
HCSTVG-v2	95.1	68.8	87.0	81.5	60.0	68.0
VidSTG-D	83.8	52.7	60.9	72.5	45.4	50.6
VidSTG-I	83.7	48.8	54.7	72.7	41.6	44.6

Table 1. **Upper-bound analysis** on HCSTVG-1, HCSTVG-2, and VidSTG utilizing only per-frame detections (left) and tubelets (right) as proposals.

Tubelets Pre-processing In this section, we detail the pre-processing steps to clean the training data used in SRM for learning referral subject grounding. A significant issue identified during this process is the prevalence of soft-label switching within tubelets. Specifically, a tubelet often contains multiple detections, each with a different soft-label generated by Grounding-DINO.

Figure 1 illustrates some common combinations of soft-labels observed in tubelets. Examples like *man+woman* and *girl+man* indicate that some tubelets do not consistently track a single individual but instead switch subjects temporally. Such label inconsistencies can degrade training performance, especially when subjects belong to different categories. For instance, if a tubelet switches between *man* and *woman*, the model learns to ground both categories simultaneously, despite the referred subject being explicitly a *man*. This creates a noisy training signal, undermining the model's ability to learn accurate grounding.

To address this, we first quantify the extent of soft-label switching in Section 1.1. Subsequently, in Section 1.2, we propose leveraging soft-labels in conjunction with Intersection over Union (IoU) metrics to detect cases where subject switching actually occurs within tubelets. In Section 1.2 we analyze the temporal duration over which such switches take place within a parent tubelet. Finally in Section 1.3 (based on analysis in the previous two sections) we devise and employ dataset denoising strategies aimed at improving training data quality and, consequently, model grounding

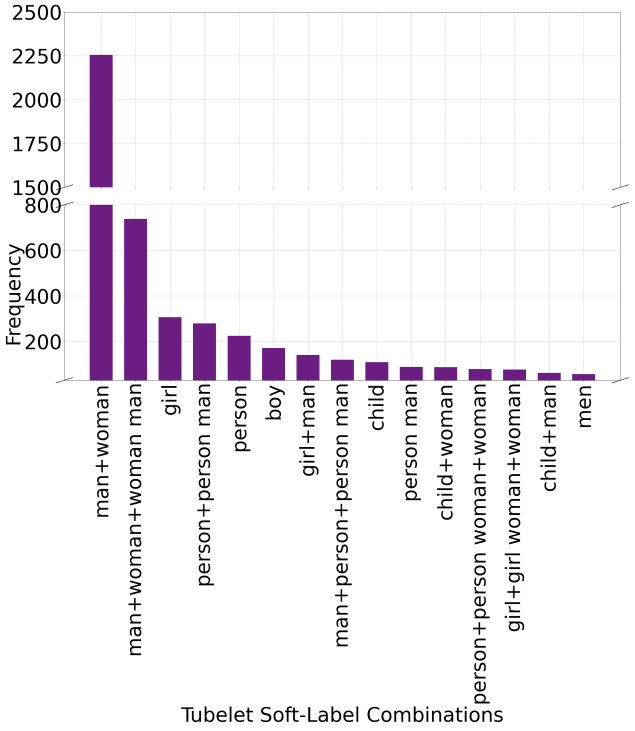

Figure 1. Distribution of frequently co-occurring soft-labels in tubelets for HCSTVG-1. Grounding-DINO soft labels can be more than one class for a given detection (e.g., person man).

capabilities.

1.1. Percentage Switching in Train

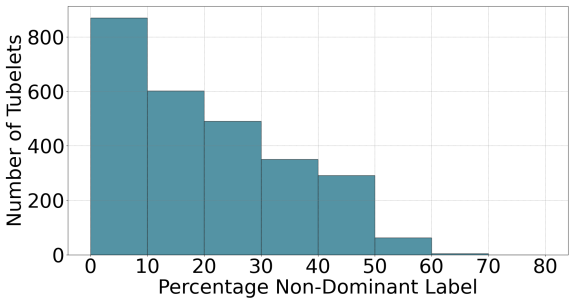

Figure 2. Distribution of the percentage of non-dominant soft-label detections in tubelets for all tubelets with conflicting combinations (e.g., man + woman) in HCSTVG-1.

In Figure 1, we observe several instances where soft-label switching occurs within tubelets. For tubelets with conflicting combinations (e.g., *man+woman*), Figure 2 illustrates the percentage of detections in each tubelet that do not correspond to the most frequently occurring soft-label. Our analysis reveals that while the majority of tubelets with conflicting combinations exhibit less than 10% switching, a significant portion contains more than 30% switching. This

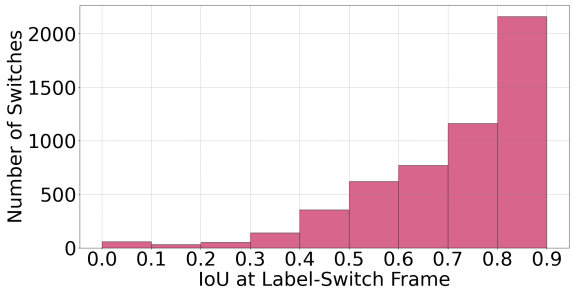

Figure 3. Distribution of the IoU between detections on the frame before and after faulty soft-label switch in HCSTVG-1.

highlights the presence of considerable noise in the training dataset, underscoring the need for data cleaning to ensure a high-quality training signal for the model.

1.2. Label Switched Sub-Tubelet Analysis

Fig 2 highlights the extent of switching in tubelets with conflicting label combinations. We hypothesize that in cases where the tracked subject switches, the IoU overlap between the detection of the subject in the frame preceding the label switch and the detection at the switching frame should be low. To investigate whether IoU can reliably identify instances of visual subject-switching as opposed to faulty soft-label switches, we analyze the distribution of IoU values at these switching points for all tubelets with conflicting label combinations. This distribution is presented in Figure 3.

The results indicate that a substantial number of switching points exhibit IoU values below 0.50, suggesting that the tracked subject may have visually switched, rather than the switch being solely a result of faulty soft-labeling. However, a significant number of high IoU values at switch points indicate that Grounding-DINO may sometimes produce faulty label switches when subjects overlap, even while the original subject continues to be correctly tracked. Differentiating between faulty soft-label switches and visual-subject switches in the case of overlapping subjects, however, cannot reliably be performed using this technique.

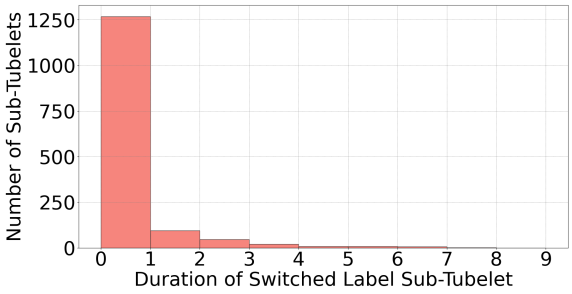

Figure 4. Distribution of durations (in seconds) for soft-label switched sub-tubelets for tubelets with conflicting combinations in HCSTVG-1. For each tubelet we find continuous sub-sections within the tubelet where soft-label switch occurs and find its duration.

Our analysis reveals that when a label switch occurs, the sub-tubelet associated with the new soft label is typically less than 1 second in duration, as shown in Figure 4. This indicates that label switches often manifest as brief, intermittent segments within the parent tubelet, rather than representing a permanent change in the tracked subject. Such transient switches highlight the need for targeted strategies to address these short-lived inconsistencies.

1.3. Training Set Denoising Strategies

To address the primary challenges, we propose two de-noising strategies based on the heuristics from Section 1.2.

Switch-Addition: This strategy focuses on sub-tubelets with durations exceeding 1 second. Sub-tubelets meeting this threshold are extracted and converted into independent tubelets, while the remaining non-mode labeled detections are eliminated. This allows sub-tubelets to be grounded independently. A consequence of this choice is that some samples have an increased number of tubelets to distinguish over possibly making training more challenging.

Switch-Dropping: This strategy eliminates all detections within a tubelet (conflicting combination tubelets only) that do not correspond to the mode label. The goal of this strategy is to remove maximum cases in which visual-switching occurs. However this strategy may result in the elimination of certain continuous stretches of switched label (termed sub-tubelet) which correspond to the optimal tubelet to be grounded for that sample.

We employ both of these de-noising strategies in our TRG training pipeline and summarize the results in Table 2. Both strategies improve the performance over the baseline. Switch-dropping outperforms switch-addition strategy.

Method	m_vIoU	vIoU@0.1	vIoU@0.3	vIoU@0.5
Baseline	10.40	30.69	12.67	5.00
Switch-Addition	9.74	29.31	11.81	4.40
Switch-Dropping	**10.81**	**31.47**	**13.19**	**5.43**

Table 2. Comparison and analysis of train set denoising strategies for TRG on HCSTVG-1.

2. SA-TCL & CG-SCL Stage-wise Construction

In Section 2.1, we first analyze SRM model performance based on different configurations of pairwise average temporal IoU (between tubelets) to find the ideal hyper-parameters for stage-wise CGS construction. In Section 2.2 we ponder over our temporal curriculum (SA-TCL) to better understand why training benefits from individual actions to action-combinations and not the opposite. Finally, we explore whether it is beneficial to have cumulative training stages (each stage comprises all training samples from preceding stages) for CG-SCL and SA-TCL in Section 2.3. The resulting stage-wise dataset statistics based on the below analysis can be found in Figure 7 and Figure 8.

2.1. Interval Selection for CGS

Stages	IoU/Stage	m_tIoU	m_vIoU	vIoU@0.3	vIoU@0.5
Low to High Temporal IoU overlap					
7	0.14	17.63	**9.30**	**12.16**	**4.14**
5	0.20	**18.89**	9.24	11.64	**4.14**
3	0.33	18.51	9.08	11.47	4.31
High to Low Temporal IoU overlap					
7	0.14	**21.88**	**11.10**	**14.05**	**5.60**
5	0.20	21.13	11.01	13.36	5.43
3	0.33	20.14	10.36	12.67	5.00

Table 3. Ablation on number of curriculum stages at different temporal IoU overlap/stage in both low-to-high and high-to-low settings for HCSTVG-1.

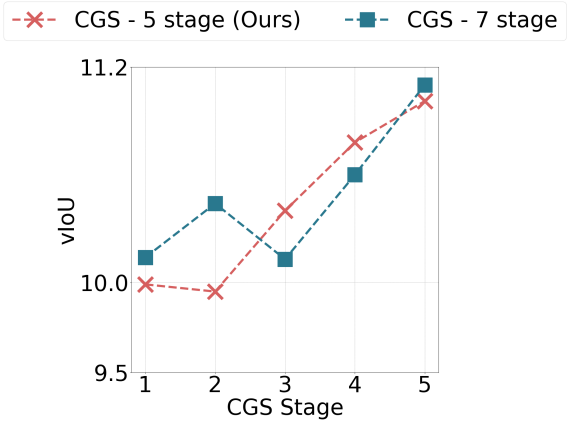

Figure 5. Per-stage performance gain for 5-stage and 7-stage (high-to-low) CGS for HCSTVG-1.

Intuitively, using temporally spaced candidate tubelets aids in learning referral subject localization by making it easier to distinguish attributes and actor types due to their diversity in features. Conversely, less diverse candidates might help the model focus on actor differentiation while minimizing background variation, promoting the gradual divergence of unrelated features in the joint semantic space. To evaluate these hypotheses, we conduct a systematic study of two settings. We tested two strategies: gradually increasing the per-stage temporal IoU threshold from low-to-high and high-to-low (Ours CGS method). From Table 3, we observe that progressing from high-to-low temporal IoU improves the model's referral capabilities. We also observe that using 5 stages with a per-stage delta of 0.20 IoU results in consistent improvements at each stage, as expected by the curriculum. Although using 7 stages increases the model's overall capability, the per-stage progress is less consistent (see Figure 5), and the training time increases substantially.

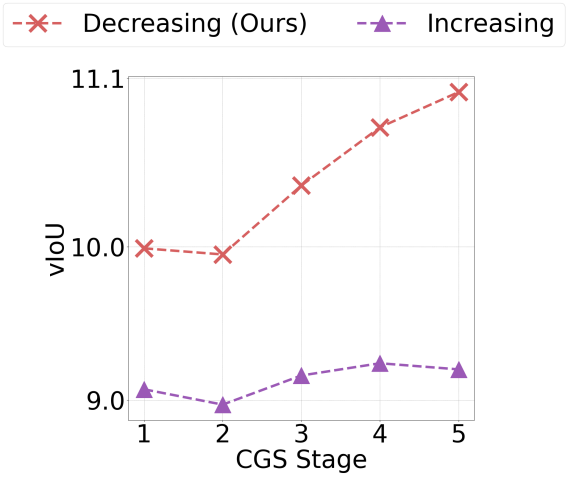

Figure 6. Per-stage SRM performance improvement in 5-stage low-to-high and high-to-low CGS for HCSTVG-1. In the low-to-high CGS, each stage yields minor increments

Furthermore, while progressing from low to high IoU shows some relative improvement, Figure 6 highlights a notable difference: decreasing the IoU (high-to-low) leads to a consistent improvement in model capabilities at each stage. This pattern is absent when increasing IoU (low-to-high), suggesting that the model progressively learns to differentiate actors in similar background contexts before handling more diverse contexts.

2.2. Utility of Extracted Actions for SA-TCL

We first examine the benefits of deconstructing complex captions and using their sub-components for training in SA-TCL. Specifically, we compare the temporal localization performance of TRM using SA-TCL with a curriculum where we progressively increase the number of actions in the original captions without extracting sub-action phrases. As shown in Table 4 , the model benefits from the extracted sub-captions, gaining compositional capabilities that are not learned by simply increasing the number of actions.

Method	m_tIoU	tIoU@0.1	tIoU@0.3	tIoU@0.5
Baseline	32.99	78.43	56.43	22.52
Original Captions	33.05	77.48	57.03	22.78
Extracted Captions	**33.92**	**78.77**	**57.72**	**24.68**

Table 4. Ablation on the use of extracted sub-actions compared to original captions for increasing action SA-TCL (HCSTVG-1).

Next, we investigate whether construction or deconstruction is more effective for learning action composition in SA-TCL. We reverse the order of sub-action phrases during training, starting with compound phrases and gradually adding individual actions (Dec in Table 5). In contrast, the original SA-TCL method (Inc in Table 5) increases action composition progressively. Our results show that construction (Inc) significantly outperforms deconstruction (Dec) in understanding action compositionality. We hypothesize that this is due to the model's difficulty in reducing its predicted temporal span in later training stages.

Curriculum	m_tIoU	tIoU@0.1	tIoU@0.3	tIoU@0.5
Dec	31.41	74.03	53.15	22.69
Baseline	32.99	78.43	56.43	22.52
Inc	**33.92**	**78.77**	**57.72**	**24.68**

Table 5. Comparing re-construction (Inc) and de-construction (Dec) using extracted sub-actions for SA-TCL (HCSTVG-1).

2.3. Merit of Overlapping Sub-stages

Table 6 compares overlapping substages in SA-TCL and CGS (in CG-SCL), with the first row showing results for TRM and SRM without any curriculum. Our analysis reveals that overlapping substages, by incorporating training samples from previous stages (stages already trained over), helps prevent forgetting and offer benefits.

Curriculum	Cumulative	m_tIoU	tIoU@0.1	tIoU@0.3	tIoU@0.5
SA-TCL (Inc)	-	32.99	78.43	56.43	22.52
SA-TCL (Inc)		33.74	80.41	57.64	23.81
SA-TCL (Inc)	✓	**33.92**	**78.77**	**57.72**	**24.68**
Curriculum	Cumulative	m_vIoU	vIoU@0.1	vIoU@0.3	vIoU@0.5
CGS (Dec)	-	10.40	30.69	12.67	5.00
CGS (Dec)		10.79	31.29	13.19	5.43
CGS (Dec)	✓	**11.01**	**32.41**	**13.36**	**5.43**

Table 6. Comparing the use of overlapping (cummulative) substages in SA-TCL & CG-SCL for HCSTVG-1. The first row (in upper and lower) represents the baseline scores from TRM and SRM.

Stage	Additional	Total
Stage-1	356	356
Stage-2	140	496
Stage-3	292	788
Stage-4	1560	2348
Stage-5	2069	4417

(a) **HCSTVG-1:** CGS stage-wise

Stage	Additional	Total
Stage-1	302	302
Stage-2	125	427
Stage-3	255	682
Stage-4	1354	2036
Stage-5	2129	4165

(b) **HCSTVG-1:** CGS + SLF stage-wise

Stage	Additional	Total
Stage-1	882	882
Stage-2	318	1200
Stage-3	707	1907
Stage-4	3548	5491
Stage-5	4604	10095

(c) **HCSTVG-2:** CGS stage-wise

Stage	Additional	Total
Stage-1	721	721
Stage-2	288	1009
Stage-3	617	1626
Stage-4	3057	4683
Stage-5	4825	9508

(d) **HCSTVG-2:** CGS+SLF stage-wise

Figure 7. CG-SCL per-stage and cumulative dataset statistics for HCSTVG-1 & HCSTVG-2.

Stage	Additional	Total
Stage-1	974	974
Stage-2	2068	3042
Stage-3	1251	4293
Stage-4	182	4475

(a) **HCSTVG-1:** Original captions

Stage	Additional	Total
Stage-1	11394	11394
Stage-2	7045	18439
Stage-3	2903	21342
Stage-4	952	22294

(b) **HCSTVG-1:** Extracted captions

Stage	Additional	Total
Stage-1	806	806
Stage-2	4532	5338
Stage-3	3270	8608
Stage-4	1386	9994

(c) **HCSTVG-2:** Original captions

Stage	Additional	Total
Stage-1	26211	26211
Stage-2	18042	44253
Stage-3	6846	51099
Stage-4	2253	53352

(d) **HCSTVG-2:** Extracted captions

Figure 8. SA-TCL per-stage and cummulative statistics for HCSTVG-1 & HCSTVG-2.

3. Ablation study on HCSTVG-v2 dataset

We show an ablation study of the breakdown of the different components of TRG as well as STPro (SA-TCL & CG-SCL) on the HCSTVG-v2 dataset, an extended version of HCSTVG-v1 in Table 7.

Effectiveness of TRG sub-modules Similar to HCSTVG-v1 dataset, we observe SRM, TRM and POS improves the performance over weakly adapted Grounding DINO. Combining all sub-modules, the performance boost is 16.35%, 7.71%, 14.00%, and, 6.37% at mean tIoU, vIoU, vIoU@0.3, and, vIoU@0.5 respectively.

Impact of SA-TCL and CG-SCL Both TA-SCL and CG-SCL improves TRG's performance independently. On combining both, we find that we do not achieve improvement over CG-SCL alone. However, the performance difference is very minimal (< 0.2% at mean vIoU). STPro outperforms W-GDINO by a margin of 10.14% and TRG by a margin of 2.43% at mean vIoU.

Analysis on stages of Curriculum Learning Fig. 9 shows increment of score for successive stages of SA-TCL and CG-SCL.

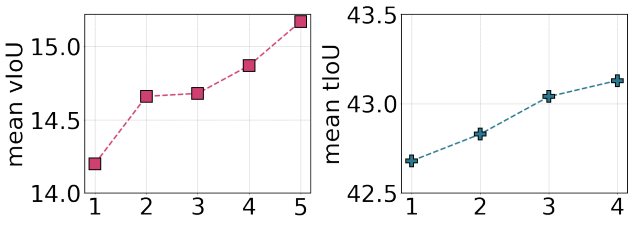

Figure 9. Performance at successive stages of SA-TCL and CG-SCL on HCSTVG-v2 dataset.

SRM	TRM	POS	m_tIoU	m_vIoU	vIoU@0.3	vIoU@0.5
			23.30	9.85	13.30	5.63
✓			30.99	10.44	13.90	6.35
✓		✓	28.97	13.62	19.15	9.00
✓	✓	✓	**39.55**	**17.56**	**27.20**	**12.00**
TRG	CG-SCL	SA-TCL	m_tIoU	m_vIoU	vIoU@0.3	vIoU@0.5
			23.30	9.85	13.30	5.63
✓			**39.55**	17.56	27.20	12.00
✓		✓	39.02	17.62	27.50	12.10
✓	✓		39.00	**19.99**	**31.70**	**14.55**
✓	✓	✓	38.46	19.82	31.35	14.25

Table 7. **Ablation study** on sub-modules of TRG module(upper) and different components of STPro (lower) on HCSTVG-v2 dataset.

4. Comparison with Fully-Supervised Approaches

We show a comparison of STPro with fully-supervised approaches (See Table 8 and Table 9). In comparison to recent approaches[3, 7, 14], our approach performs nearly at 50% on both mean vIoU and tIoU on HCSTVG-1, HCSTVG-2 and VidSTG.

Though WSSTVG involves joint spatio-temporal inference, no existing work shows performance on the tIoU performance metric. We also include our tIoU scores for this task.

5. Joint-Inference Ablations

STPro's joint inference pipeline consists of two stages: First, TRM filters and trims tubelet proposals to a subset of temporall refined candidates; Then, SRM grounds the correct tubelet based on query-tubelet similarity from the selected

Methods	HCSTVG - v1 m_tIoU	m_vIoU	vIoU@0.3	vIoU@0.5	HCSTVG - v2 m_tIoU	m_vIoU	vIoU@0.3	vIoU@0.5
Fully-Supervised								
STGVT [TCSVT20] [13]	-	18.2	26.8	9.5	-	-	-	-
STVGBert [ICCV21] [12]	-	20.4	29.4	11.3	-	-	-	-
TubeDETR [CVPR22] [16]	43.7	32.4	49.8	23.5	53.9	36.4	58.8	30.6
STCAT [NeurIPS22] [5]	49.4	35.1	57.7	30.1	-	-	-	-
CSDVL [CVPR23] [7]	-	36.9	62.2	34.8	58.1	38.7	65.5	33.8
CG-STVG [CVPR24] [3]	52.8	38.4	61.5	36.3	60.0	39.5	64.5	36.3
VGDINO [CVPR24] [14]	-	38.3	62.5	36.1	-	39.9	67.1	34.5
Weakly-Supervised (Two-stage pipelines)								
GroundeR [ECCV16] [10]+LCNet [IEEE17] [17]	-	4.17	3.28	1.05	-	-	-	-
MATN [CVPR18] [19]+LCNet [IEEE17] [17]	-	4.41	3.53	1.12	-	-	-	-
GroundeR [ECCV16] [10]+CPL [CVPR22] [21]	-	5.23	4.18	1.25	-	-	-	-
RAIR [CVPR21] [8]+CPL [CVPR22] [21]	-	6.88	4.87	1.36	-	-	-	-
Weakly-Supervised (Single-stage pipelines)								
WSSTG [ACL19] [2]	-	6.52	4.54	1.27	-	-	-	-
AWGU [ACMMM20] [1]	-	8.20	4.48	0.78	-	-	-	-
Vis-CTX [CVPR19] [11]	-	9.76	6.81	1.03	-	-	-	-
WINNER [CVPR23] [6]	-	14.20	17.24	6.12	-	-	-	-
VCMA [ECCV24] [4]	-	14.64	18.60	5.75	-	-	-	-
W-GDINO (Ours-Baseline)	18.0	9.04	11.56	4.57	23.3	9.85	13.30	5.63
STPro	**30.6**	**17.56**	**26.98**	**12.93**	**39.0**	**19.99**	**31.70**	**14.55**
	(+12.6)	(+2.92)	(+8.38)	(+6.81)	(+15.7)	(+10.14)	(+18.40)	(+8.92)

Table 8. Comparison with existing state-of-the-art weakly and fully-supervised methods on HCSTVG-1 and HCSTVG-2 datasets. **Bold** denotes best and underline denotes second best.

Methods	Declarative Sentences m_tIoU	m_vIoU	vIoU@0.3	vIoU@0.5	Interrogative Sentences m_tIoU	m_vIoU	vIoU@0.3	vIoU@0.5
Fully-Supervised								
Ground-R [ECCV16] [9]	-	9.8	11.0	4.1	-	9.3	11.4	3.2
STPR [CVPR17] [15]	34.6	10.1	12.4	4.3	33.7	10.0	11.7	4.4
WSSTG [ACL19] [2]	-	11.4	14.6	5.9	-	10.7	13.9	5.3
STGRN [CVPR20] [18]	48.5	19.8	25.8	14.6	46.9	18.3	21.1	12.8
STVGBert [ICCV21] [12]	-	24.0	30.9	18.4	-	22.5	26.0	16.0
TubeDETR [CVPR22] [16]	48.1	30.4	42.5	28.2	46.9	25.7	35.7	23.2
STCAT [NeurIPS22] [5]	50.8	33.1	46.2	32.6	49.7	28.2	39.2	26.6
CSDVL [CVPR23] [7]	-	33.7	47.2	32.8	-	28.5	39.9	26.2
CG-STVG [CVPR24] [3]	51.4	34.0	47.7	33.1	49.9	29.0	40.5	27.5
VGDINO [CVPR24] [14]	52.0	34.7	48.1	34.0	50.8	29.9	41.0	27.6
Weakly-Supervised (Two-stage pipelines)								
GroundeR [ECCV16] [10]+LCNet [IEEE17] [17]	-	7.85	7.96	3.02	-	6.43	6.58	2.92
MATN [CVPR18] [19]+LCNet [IEEE17] [17]	-	8.16	8.03	3.59	-	6.97	6.64	3.05
GroundeR [ECCV16] [10]+CPL [CVPR22] [21]	-	8.28	8.35	3.68	-	7.16	7.28	3.23
RAIR [CVPR21] [8]+CPL [CVPR22] [21]	-	8.67	8.72	4.01	-	7.68	7.71	3.58
Weakly-Supervised (Single-stage pipelines)								
WSSTG [ACL19] [2]	-	8.85	8.52	3.87	-	7.12	6.87	2.96
AWGU [ACMM20] [1]	-	8.96	7.86	3.10	-	8.57	6.84	2.88
Vis-CTX [CVPR19] [11]	-	9.34	7.32	3.34	-	8.69	7.18	2.91
WINNER [CVPR23] [6]	-	11.62	14.12	7.40	-	10.23	11.96	5.46
VCMA [ECCV24] [4]	-	14.45	18.57	8.76	-	**13.25**	**16.74**	7.66
W-GDINO (Ours-Baseline)	28.7	10.69	13.02	7.83	29.1	9.87	12.16	6.71
STPro (Ours)	**35.8**	**15.52**	**19.39**	**12.69**	**34.6**	12.56	14.95	**9.29**
	(+7.1)	(+1.07)	(+0.82)	(+3.93)	(+5.5)	(-0.69)	(-1.79)	(+1.63)

Table 9. Comparison with existing state-of-the-art weakly and fully-supervised methods on VidSTG dataset. **Bold** denotes best and underline denotes second best.

candidates. In Section 5.1, we present an ablation on the tubelet filtering criteria used in TRM. CG-SCL employs soft-label filtering (SLF) to obtain more relevant training tubelets for SRM, which can also be applied at inference time in tandem with TRM candidate selection. We explore the impact of this added filtering in Section 5.2. For VidSTG, we find that using TRM for both filtering and trimming reduces the overall referral performance. Hence, we use TRM only as a filter as discussed in Section 5.3.

5.1. TRM for Filtering & Trimming

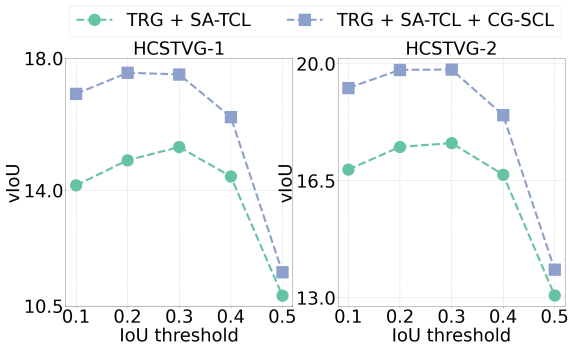

Figure 10. Joint spatio-temporal performance at different IoU thresholds (T_{filt}) for TRM on HCSTVG-1 & HCSTVG-2.

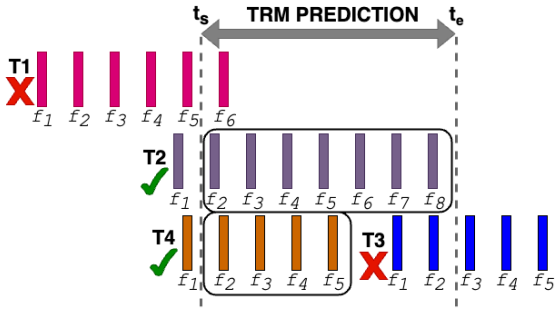

Figure 11. TRM for tubelet filtering & trimming

To select candidate tubelets from all proposals, TRM generates a temporal prediction $[t_s, t_d]$, where t_s is the start time and t_d is the end time. A tubelet is considered a valid candidate if it either fully contains or is fully contained within the TRM prediction. For tubelets that partially overlap with TRM's prediction, we compute their temporal IoU and select those with an IoU above a threshold T_{filt}, ensuring only temporally relevant tubelets are retained, as shown in Figure 11. The joint inference results at different T_{filt} thresholds are shown in Figure 10. Before being used in SRM, selected tubelets undergo nearest-neighbor interpolation to fill missing frames and are trimmed to the predicted temporal boundaries.

We find that $T_{filt} = 0.2$ works best for STPro, while $T_{filt} = 0.3$ is optimal for SA-TCL. As TRM's temporal grounding capabilities improve, we expect the threshold T_{filt} to increase for optimal performance.

5.2. SLF at Inference Time

Dataset	SLF	m_tIoU	m_vIoU	vIoU@0.1	vIoU@0.3	vIoU@0.5
HCSTVG-1		**30.81**	17.37	41.81	26.55	12.67
HCSTVG-1	✓	30.56	**17.56**	**41.90**	**26.98**	**12.93**
HCSTVG-2		**38.99**	19.53	47.45	30.85	13.8
HCSTVG-2	✓	38.46	**19.82**	**47.95**	**31.35**	**14.25**

Table 10. STPro performance with and without soft-label-filtering at inference time on HCSTVG-1 and HCSTVG-2.

We apply SLF, as described in Section 6, as an additional filter to the candidate tubelets selected by TRM. The results with and without test-time SLF are shown in Table 10. We find that SLF at test time provides minimal improvement, and our models outperform previous baselines even without SLF during inference. This suggests that TRM alone effectively filters the candidate tubelets, enabling better referral grounding in SRM. Dataset statistics in Figure 7 further support this, showing that few samples have a large number of temporally overlapping tubelets (stage-1 and stage-2), reducing the need for additional filtering.

5.3. TRM as a Filter for VidSTG

SRM	Filter	Trim	m_tIoU	m_vIoU	vIoU@0.1	vIoU@0.3	vIoU@0.5
			28.79	10.69	25.03	13.02	7.83
✓			34.17	14.93	32.49	18.44	12.26
✓	✓	✓	24.97	11.1	24.97	17.29	4.75
✓	✓		**35.77**	**15.52**	**33.41**	**19.39**	**12.69**
SRM	Filter	Trim	m_tIoU	m_vIoU	vIoU@0.1	vIoU@0.3	vIoU@0.5
			29.10	9.87	23.19	12.16	6.71
✓			33.64	12.29	28.3	14.56	9.05
✓	✓	✓	24.39	8.76	24.86	12.4	2.85
✓	✓		**34.64**	**12.56**	**29.12**	**14.95**	**9.29**

Table 11. Analysis on the use of TRM as filter and trimmer versus only as a filter for joint spatiotemporal inference on VidSTG declarative (upper), VidSTG interrogative (lower).

Using TRM as both a filter and trimmer leads to a performance decline when combined with SRM for both VidSTG declarative and VidSTG interrogative tasks. Analysis shows that VidSTG tubelet proposals, generated through detection and tracking, as well as their ground-truth temporal boundaries, are heavily skewed toward durations under 2 seconds. In contrast, TRM often predicts significantly larger temporal spans. Following [20], instead of imposing an upper limit on temporal width predictions, we avoid further trimming

already short tubelets. We apply nearest-neighbor interpolation but restrict TRM to act only as a filter. The performance decline from trimming and the improvement from using TRM purely as a filter are shown in Table 11.

6. Soft-Label Filtering (SLF) in Practice

To enhance TRM's ability to distinguish referral subjects based on attributes and subject types, we employ Soft Label Filtering (SLF) in CG-SCL. Figures 12 and 13 illustrate the distribution of soft labels obtained from Grounding-DINO detections across all tubelets in the training set for HCSTVG-1 and HCSTVG-2, respectively.

In Figure 13, we observe two distinct types of labels: straightforward labels such as *"man"* and *"woman"*, which represent unambiguous categories, and compound labels such as *"person woman"* and *"man woman"*, which occur frequently. To enable proper tubelet selection, we use *gender* as a filtering criterion.

The filtering rules are as follows:

- For compound labels such as *"woman person"*, the gender is assigned based on the specific component (e.g., *"woman"*).
- For ambiguous labels such as *"child"*, or compound labels like *"man woman"* and *"boy girl"*, we assign a new *neutral gender* to unify all such uncertain or generic labels.
- For each tubelet, the dominant gender across all individual detections is used as its label.

Additionally, the majority of referred subjects in captions belong to categories such as *man*, *woman*, *person*, *child*, *boy*, *girl*, *kid*, and *lady*. While these subjects have identifiable genders, to account for other specific subjects like *singer* or *police officer*, we consider them as being gender neutral. For filtration, we then apply the following rules:

- If the referred subject (in caption) is neutral: Include all tubelets for training.
- If the referred subject (in caption) is specific: Include all tubelets with matching gender as well as those with *neutral gender*.

Extending the Approach

This method is flexible and can be adapted to other datasets:

- Each category can be treated independently.
- If a detection combines parent and child categories (e.g., *"woman person"*), it can be assigned the child label (e.g., *"woman"*).
- For combinations of two or more specific categories (e.g., *"man woman"*), a *neutral label* can be assigned, ensuring these tubelets are always included for training.

7. Data Pre-processing Prompts

7.1. Sub-Action Extraction

The prompt used with GPT 3.5-Turbo for sub-action extraction for each caption in the train and test set.

System: You are a language expert. Your task is to break up a sentence into multiple sub-sentences. Any given sentence may contain multiple verbs/actions. Each sub-sentence will contain some subset of the actions in the original sentence. When splitting the sentence into sub-sentences, you must ensure that the actions are only grouped in the order in which they appear in the original sentence. You cannot skip over any actions in the original sentence; they should always be contiguous when splitting sentences. Your goal is to create all the different combinations of actions possible while maintaining their ordering. If a sentence has 10 actions, then we can create groups of up to 9 continuous actions. You must create all such groups. The keys within the JSON object are the number of actions used to form the group of sub-sentences, and the list within it is the list of sentences containing exactly those many actions. If some pronouns are ambiguous, qualify them with the object or subject they actually refer to. Each sentence must make complete sense on its own.
Example 1
Original sentence: The man in the jacket walks to the woman in red and stops, takes a golf club, gives it to the woman in red, and pushes her.
Answer:

```
{
    "1": [
        "The man in the jacket walks
           to the woman in red",
        "The man in the jacket stops",
        "The man in the jacket takes a
            golf club",
        "The man in the jacket gives
           the golf club to the woman
            in red",
        "The man in the jacket pushes
           her"
    ],
    "2": [
        "The man in the jacket walks
           to the woman in red and
           stops",
        "The man in the jacket stops,
           takes a golf club",
        "The man in the jacket takes a
            golf club, gives it to
```

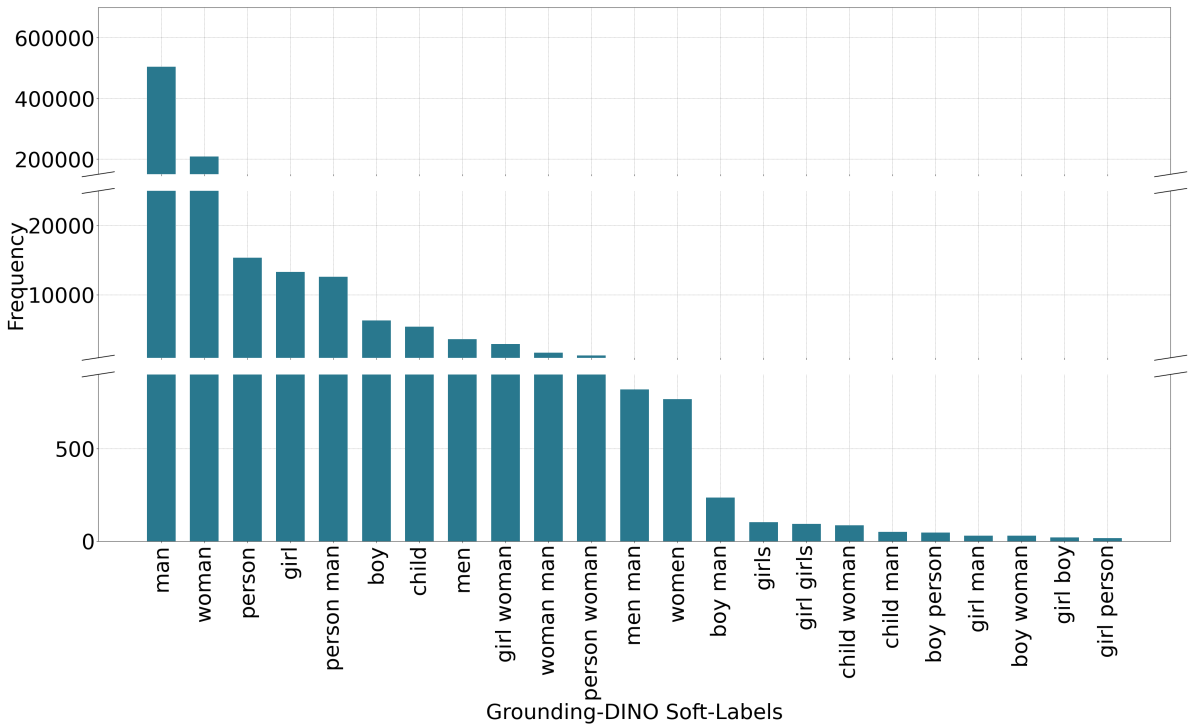

Figure 12. G-DINO soft-label distribution HCSTVG-1.

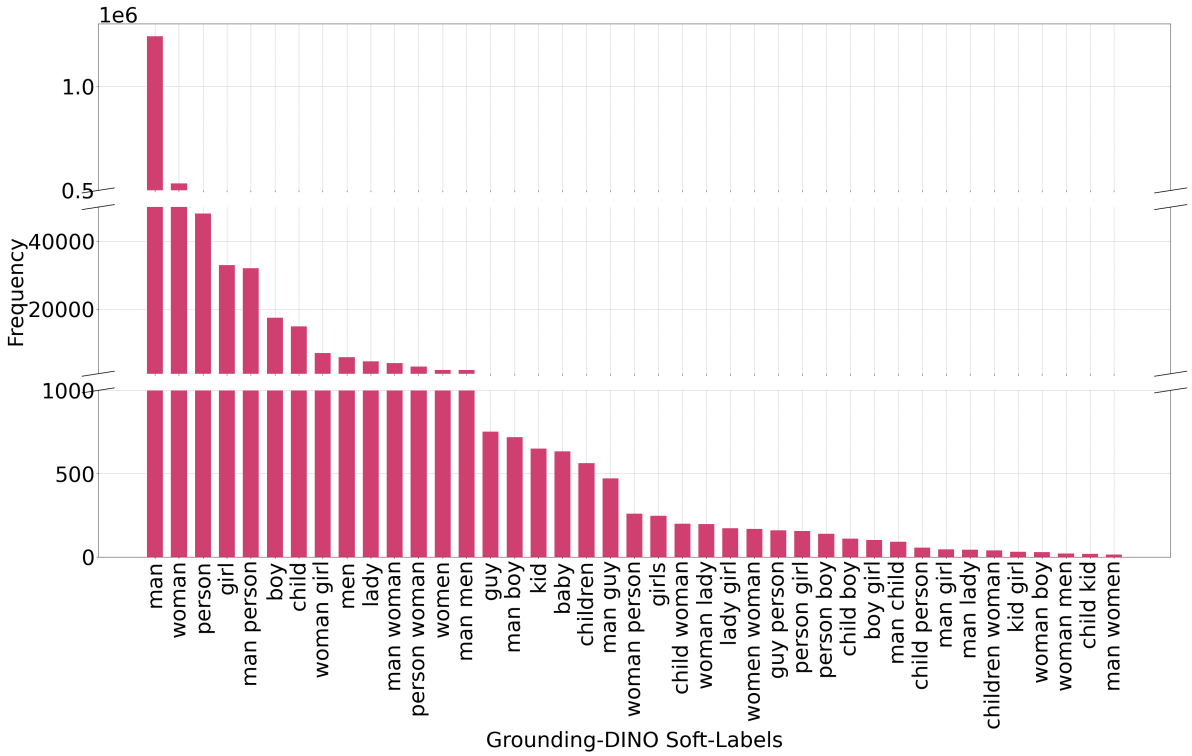

Figure 13. G-DINO soft-label distribution HCSTVG-2.

Figure 14. **Qualitative Impact of SA-TCL:** Red indicates the ground-truth, Blue represents the prediction from TRG, and Violet corresponds to STPro (SA-TCL + CG-SCL applied over TRG). Darker shades denote predictions. The bars below the frames depict the temporal predictions from TRM, while the bounding boxes represent the tubelets grounded by SRM. In the top example, TRM incorrectly localizes the query temporally, leading to a failure in SRM's spatial grounding. In contrast, the bottom example demonstrates how STPro significantly refines TRM's temporal predictions, enabling SRM to correctly ground the referred subject.

Figure 15. **Qualitative Impact of CG-SCL:** Red indicates the ground truth, Blue represents the prediction from TRG, and Violet corresponds to STPro (SA-TCL + CG-SCL applied over TRG). Darker shades denote predictions. The bars below the frames depict the temporal predictions from TRM, while the bounding boxes represent the tubelets grounded by SRM. In the top example, SRM incorrectly localizes the referred subject (*woman in red clothes*) though TRM's prediction is near optimal. In contrast, the bottom example demonstrates how STPro significantly refines SRM to correctly ground the referred subject.

```
            the woman in red",
        "The man in the jacket gives
            the golf club to the woman
             in red, and pushes her"
    ],
    "3": [
        "The man in the jacket walks
            to the woman in red and
            stops, takes a golf club",
```

```
        "The man in the jacket stops,
            takes a golf club, gives
            it to the woman in red",
        "The man in the jacket takes a
             golf club, gives it to
            the woman in red, and
            pushes her"
    ],
    "4": [
```

```
        "The man in the jacket walks
            to the woman in red and
            stops, takes a golf club,
            gives it to the woman in
            red",
        "The man in the jacket stops,
            takes a golf club, gives
            it to the woman in red,
            and pushes her"
    ]
}
```

Example 2
Original sentence: The bald man leaves the room, pulls the door, walks towards the man in the white suit, and then turns to face the white-suit man.
Answer:

•••

Original sentence: <*train/test caption*>
Answer:

Failure case: Though the actions and their relative orders which comprise sub-action phrases are largely correct, at times the number of combinations of extracted sub-action phrases is incorrect. For example :

Q: The man sitting on the left in the first row in brown clothes stands up, waves his hands and talks, turns around, looks down, then sits down again.

Answer:

```
{
    "1": [...],
    "2": [...],
    "3": [...],
    "4": [...]
}
```

In this case, we expect five keys to be present in the answer JSON since we can make two combinations of five actions in order. However, we get only four keys in the response, effectively losing out on some training examples in our devised SA-TCL.

7.2. Descriptor (POS) Extraction

The prompt used with GPT 3.5-Turbo for POS extraction for each caption in the train and test set.

System: Extract the quantifier phrase describing the main person.
Example 1
Original sentence: The man in brown clothes pours the contents of the bag into his hand, and then takes out a piece of paper from the bag and opens it.
Answer: The main person in this sentence is T̈he man in brown clothes.¨

Example 2
Original sentence: The girl gets up, is caught by the opposite man, and pushes her to sit down.
Answer: The main person in this sentence is T̈he girl.¨

Example 3
Original sentence: The man in the hat reaches out and points to the front, then puts his hand down and turns his head.
Answer: The main person in this sentence is T̈he man in the hat.¨

•••

Original sentence: <*train/test caption*>
Answer:

Failure case: The extracted referral subject is nearly always correct, however in some cases, the referral subject itself does not contain any distinguishable attributes. Instead the series of actions performed serves as the distinguishing factor, which our extraction scheme fails to capture. For example :

Q: <prompt> The woman who points her hand at the other woman

Answer: The main person in this sentence is The woman

In this case, the identifying characteristic is the action itself. This would lead to the generation of a noisy training sample where-in any related tubelet might be selected but the incorrect one for the particular sample.

8. Qualitative Analysis (STPro)

STPro improves model performance for weakly supervised temporal video grounding (WSTVG) in two key ways:

Improved Temporal Prediction via SA-TCL: As shown in Figure 14, the original predictions from the temporal rea-

soning module (TRM) fail to include the optimal tubelet of the referred subject. Consequently, the candidate proposals obtained from TRM do not contain the referred subject, leading to an incorrect spatial prediction by the spatial reasoning module (SRM). By applying SA-TCL, the temporal predictions of TRM are refined, ensuring that the correct referral subject is included among the candidate proposals. In this case, the revised temporal boundary contains only one tubelet, corresponding to the referred subject, enabling accurate spatial grounding.

Enhanced Referral Grounding via CG-SCL: In scenarios where the temporal predictions of TRM and SA-TCL are nearly identical and near-optimal (Figure 15), SRM may still misidentify the referred subject due to insufficient grounding of distinguishing features, such as *red*. For instance, the original SRM prediction incorrectly grounds another subject in the spatio-temporal region. CG-SCL directly enhances SRM's referral grounding capabilities, enabling STPro to correctly ground the referred subject based on spatial and temporal cues. By combining temporal refinement through SA-TCL and spatial enhancement via CG-SCL, STPro achieves significant improvements in WSTVG tasks.

References

[1] Junwen Chen, Wentao Bao, and Yu Kong. Activity-driven weakly-supervised spatio-temporal grounding from untrimmed videos. *Proceedings of the 28th ACM International Conference on Multimedia*, 2020. 6

[2] Zhenfang Chen, Lin Ma, Wenhan Luo, and Kwan-Yee Kenneth Wong. Weakly-supervised spatio-temporally grounding natural sentence in video. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1884–1894, Florence, Italy, 2019. Association for Computational Linguistics. 6

[3] Xin Gu, Heng Fan, Yan Huang, Tiejian Luo, and Libo Zhang. Context-guided spatio-temporal video grounding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18330–18339, 2024. 5, 6

[4] Yang Jin and Yadong Mu. Weakly-supervised spatio-temporal video grounding with variational cross-modal alignment. In *Computer Vision–ECCV 2024: 18th European Conference, Milano, Italy, September 29–October 4*, 2024. 6

[5] Yang Jin, Zehuan Yuan, Yadong Mu, et al. Embracing consistency: A one-stage approach for spatio-temporal video grounding. *Advances in Neural Information Processing Systems*, 35:29192–29204, 2022. 6

[6] Mengze Li, Han Wang, Wenqiao Zhang, Jiaxu Miao, Zhou Zhao, Shengyu Zhang, Wei Ji, and Fei Wu. Winner: Weakly-supervised hierarchical decomposition and alignment for spatio-temporal video grounding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 23090–23099, 2023. 6

[7] Z. Lin, C. Tan, J. Hu, Z. Jin, T. Ye, and W. Zheng. Collaborative static and dynamic vision-language streams for spatio-temporal video grounding. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 23100–23109, Los Alamitos, CA, USA, 2023. IEEE Computer Society. 5, 6

[8] Yongfei Liu, Bo Wan, Lin Ma, and Xuming He. Relation-aware instance refinement for weakly supervised visual grounding. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5612–5621, 2021. 6

[9] Anna Rohrbach, Marcus Rohrbach, Ronghang Hu, Trevor Darrell, and Bernt Schiele. Grounding of textual phrases in images by reconstruction. In *European Conference on Computer Vision*, 2015. 6

[10] Anna Rohrbach, Marcus Rohrbach, Ronghang Hu, Trevor Darrell, and Bernt Schiele. Grounding of textual phrases in images by reconstruction. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part I 14*, pages 817–834. Springer, 2016. 6

[11] Jing Shi, Jia Xu, Boqing Gong, and Chenliang Xu. Not all frames are equal: Weakly-supervised video grounding with contextual similarity and visual clustering losses. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10436–10444, 2019. 6

[12] Rui Su, Qian Yu, and Dong Xu. Stvgbert: A visual-linguistic transformer based framework for spatio-temporal video grounding. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 1513–1522, 2021. 6

[13] Zongheng Tang, Yue Liao, Si Liu, Guanbin Li, Xiaojie Jin, Hongxu Jiang, Qian Yu, and Dong Xu. Human-centric spatio-temporal video grounding with visual transformers. *IEEE Transactions on Circuits and Systems for Video Technology*, 32:8238–8249, 2020. 6

[14] Syed Talal Wasim, Muzammal Naseer, Salman Khan, Ming-Hsuan Yang, and Fahad Shahbaz Khan. Videogrounding-dino: Towards open-vocabulary spatio-temporal video grounding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18909–18918, 2024. 5, 6

[15] Masataka Yamaguchi, Kuniaki Saito, Y. Ushiku, and Tatsuya Harada. Spatio-temporal person retrieval via natural language queries. *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 1462–1471, 2017. 6

[16] Antoine Yang, Antoine Miech, Josef Sivic, Ivan Laptev, and Cordelia Schmid. Tubedetr: Spatio-temporal video grounding with transformers. *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 16421–16432, 2022. 6

[17] Wenfei Yang, Tianzhu Zhang, Yongdong Zhang, and Feng Wu. Local correspondence network for weakly supervised temporal sentence grounding. *IEEE Transactions on Image Processing*, 30:3252–3262, 2021. 6

[18] Zhu Zhang, Zhou Zhao, Yang Zhao, Qi Wang, Huasheng Liu, and Lianli Gao. Where does it exist: Spatio-temporal video grounding for multi-form sentences. *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10665–10674, 2020. 6

[19] Fang Zhao, Jianshu Li, Jian Zhao, and Jiashi Feng. Weakly supervised phrase localization with multi-scale anchored transformer network. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5696–5705, 2018. 6

[20] Minghang Zheng, Yanjie Huang, Qingchao Chen, and Yang Liu. Weakly supervised video moment localization with contrastive negative sample mining. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2022. 7

[21] Minghang Zheng, Yanjie Huang, Qingchao Chen, Yuxin Peng, and Yang Liu. Weakly supervised temporal sentence grounding with gaussian-based contrastive proposal learning. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 15534–15543, 2022. 6