LRECCOLING 2024

北京大學
PEKING UNIVERSITY

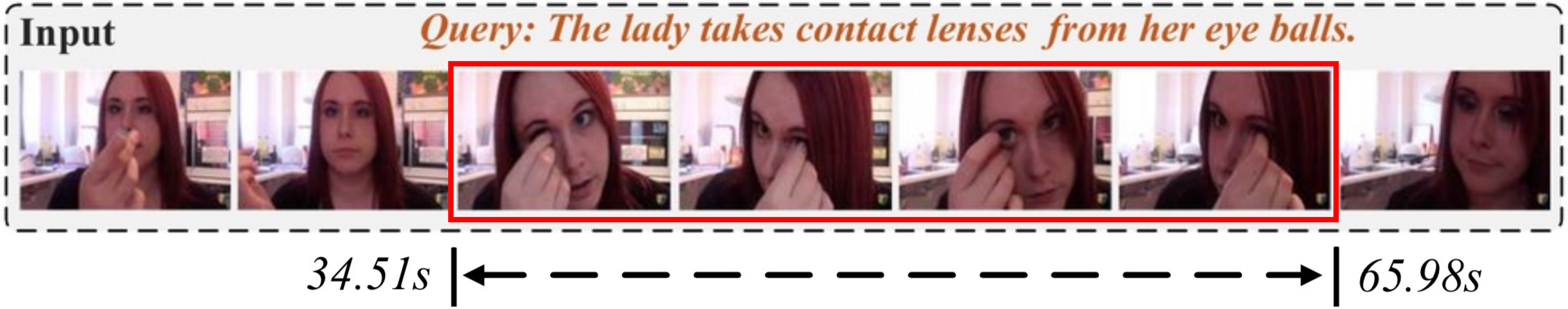
# Towards Robust Temporal Activity Localization Learning with Noisy Labels

Daizong Liu[1]

[1]Wangxuan Institute of Computer Technology, Peking University

# Temporal Sentence Grounding

- Video-text retrieval task for segment localization
- Inputs: an untrimmed video and a sentence query
- Outputs: start and end timestamps of a specific video segment



"Towards Robust Temporal Activity Localization Learning with Noisy Labels" , COLING 2024

# Challenges

- Although recent works have made significant progress in TAL research, almost all of them depend on an implicit data assumption, i.e., the moment boundary labels in training data are correctly annotated. However, in practical scenarios, it is extremely expensive and time-consuming to annotate or collect such dense labeled data.

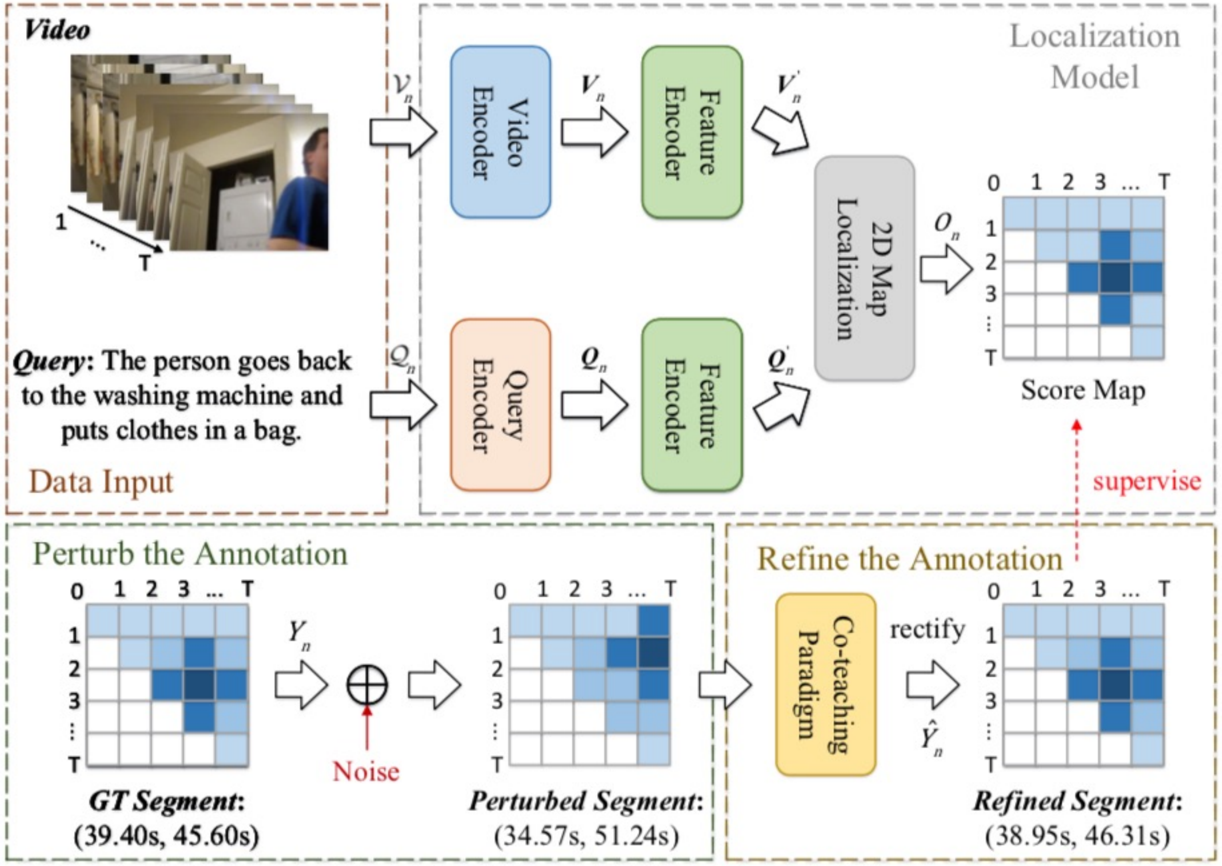
# Motivation

- There are no extra annotations to distinguish the clean and noisy video-query samples. Therefore, it is hard to directly train a robust model in a fully-clean set.
- Noisy samples also provide additional knowledge during the training. How to rectify their labels for assisting the model learning is worth investigating.
- Utilizing a single model to distinguish samples and noisy labels might not be robust enough, since it may prone to specific mistakes during the training process.

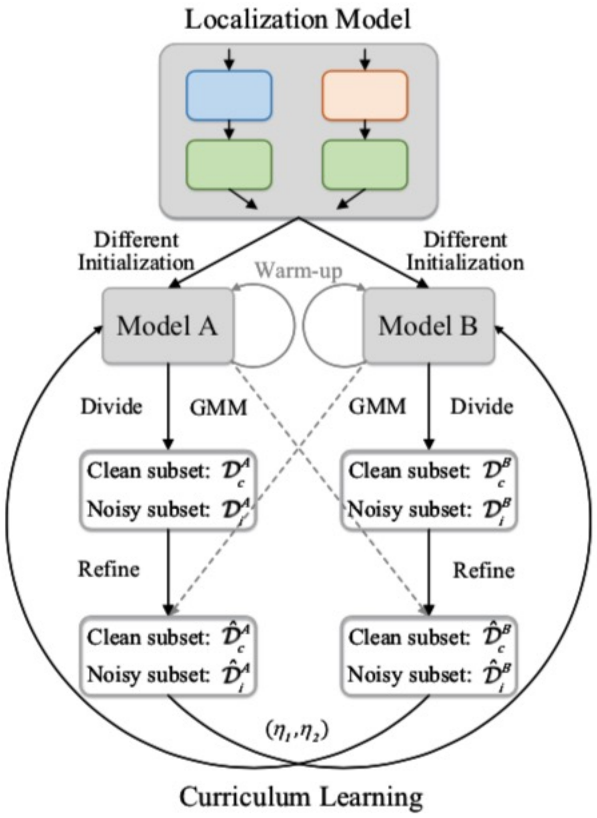
# Motivation

- To tackle the above issues, we propose a novel framework, named Co-Teaching Regularizer (CTR). Our method is based on the memorization effect of DNNs, i.e., DNNs tend to learn the simple patterns before fitting noisy samples.

# Pipeline



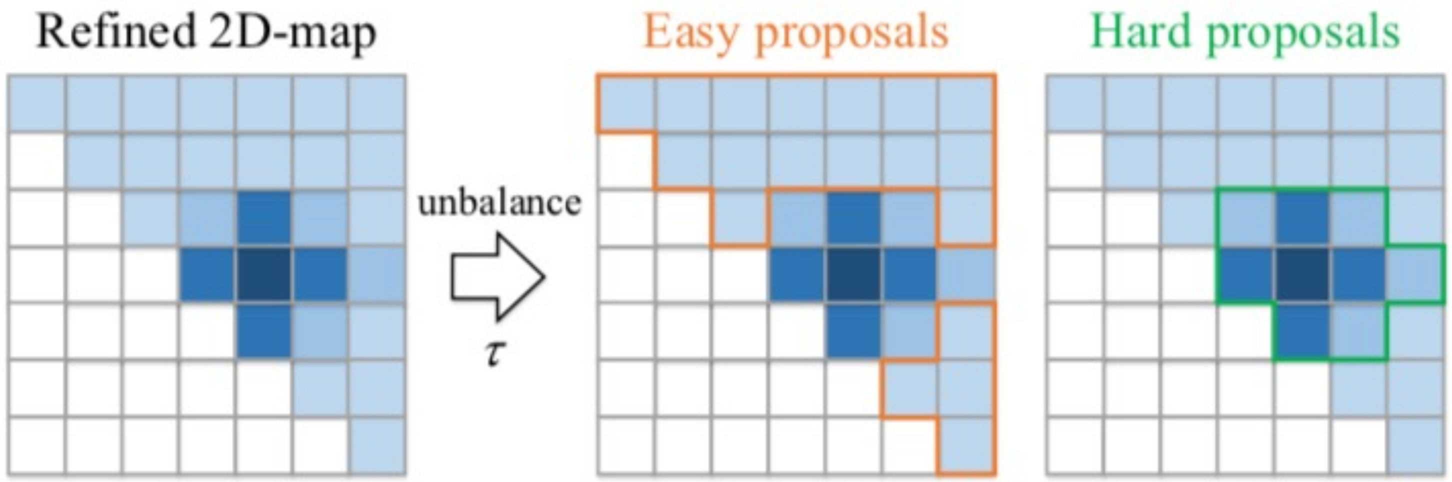
(a) Robust 2D-map based localization network against noise


Localization Model
Different Initialization
Different Initialization
Warm-up
Model A
Model B
Divide
GMM
GMM
Divide
Clean subset: $\mathcal{D}_c^A$
Noisy subset: $\mathcal{D}_i^A$
Clean subset: $\mathcal{D}_c^B$
Noisy subset: $\mathcal{D}_i^B$
Refine
Refine
Clean subset: $\hat{\mathcal{D}}_c^A$
Noisy subset: $\hat{\mathcal{D}}_i^A$
Clean subset: $\hat{\mathcal{D}}_c^B$
Noisy subset: $\hat{\mathcal{D}}_i^B$
$(\eta_1,\eta_2)$
Curriculum Learning

(b) Co-teaching paradigm

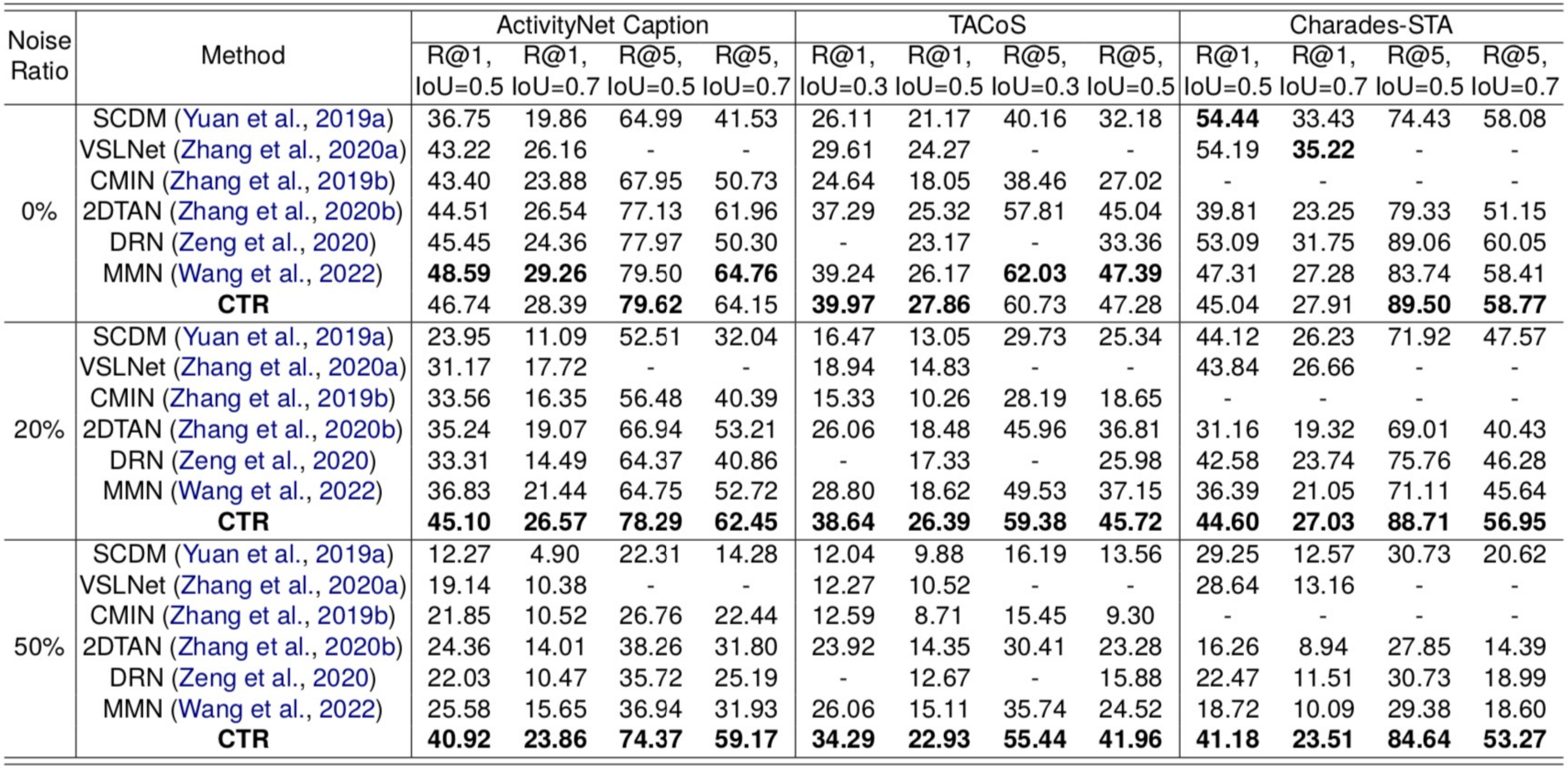
“Towards Robust Temporal Activity Localization Learning with Noisy Labels ” , COLING 2024

# Pipeline



“Towards Robust Temporal Activity Localization Learning with Noisy Labels ” , COLING 2024

# Quantitative Comparison

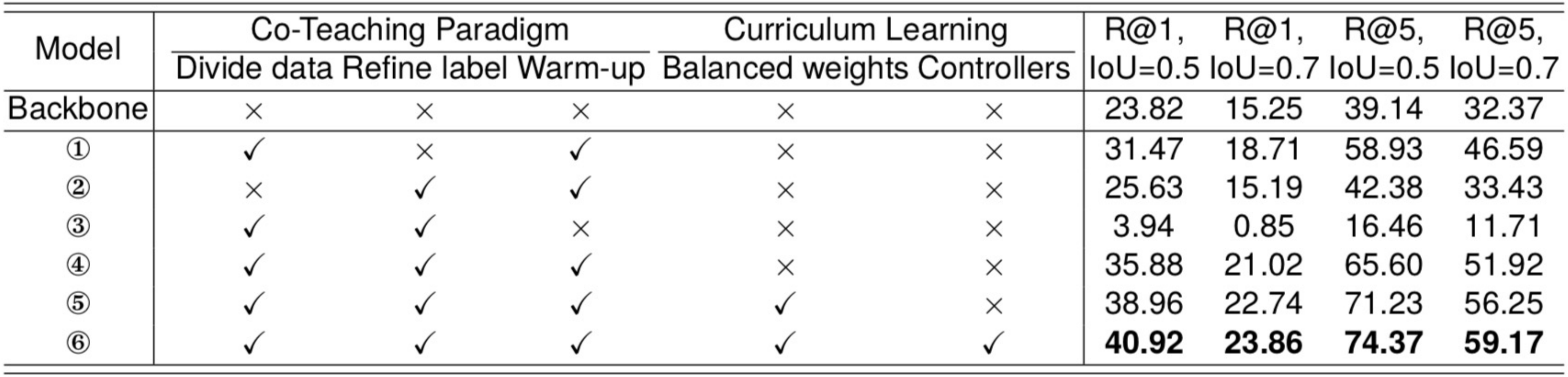


| Noise Ratio | Method | ActivityNet Caption | | | | TACoS | | | | Charades-STA | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | R@1, IoU=0.5 | R@1, IoU=0.7 | R@5, IoU=0.5 | R@5, IoU=0.7 | R@1, IoU=0.3 | R@1, IoU=0.5 | R@5, IoU=0.3 | R@5, IoU=0.5 | R@1, IoU=0.5 | R@1, IoU=0.7 | R@5, IoU=0.5 | R@5, IoU=0.7 |
| 0% | SCDM (Yuan et al., 2019a) | 36.75 | 19.86 | 64.99 | 41.53 | 26.11 | 21.17 | 40.16 | 32.18 | **54.44** | 33.43 | 74.43 | 58.08 |
| | VSLNet (Zhang et al., 2020a) | 43.22 | 26.16 | - | - | 29.61 | 24.27 | - | - | 54.19 | **35.22** | - | - |
| | CMIN (Zhang et al., 2019b) | 43.40 | 23.88 | 67.95 | 50.73 | 24.64 | 18.05 | 38.46 | 27.02 | - | - | - | - |
| | 2DTAN (Zhang et al., 2020b) | 44.51 | 26.54 | 77.13 | 61.96 | 37.29 | 25.32 | 57.81 | 45.04 | 39.81 | 23.25 | 79.33 | 51.15 |
| | DRN (Zeng et al., 2020) | 45.45 | 24.36 | 77.97 | 50.30 | - | 23.17 | - | 33.36 | 53.09 | 31.75 | 89.06 | 60.05 |
| | MMN (Wang et al., 2022) | **48.59** | **29.26** | 79.50 | **64.76** | 39.24 | 26.17 | **62.03** | **47.39** | 47.31 | 27.28 | 83.74 | 58.41 |
| | **CTR** | 46.74 | 28.39 | **79.62** | 64.15 | **39.97** | **27.86** | 60.73 | 47.28 | 45.04 | 27.91 | **89.50** | **58.77** |
| 20% | SCDM (Yuan et al., 2019a) | 23.95 | 11.09 | 52.51 | 32.04 | 16.47 | 13.05 | 29.73 | 25.34 | 44.12 | 26.23 | 71.92 | 47.57 |
| | VSLNet (Zhang et al., 2020a) | 31.17 | 17.72 | - | - | 18.94 | 14.83 | - | - | 43.84 | 26.66 | - | - |
| | CMIN (Zhang et al., 2019b) | 33.56 | 16.35 | 56.48 | 40.39 | 15.33 | 10.26 | 28.19 | 18.65 | - | - | - | - |
| | 2DTAN (Zhang et al., 2020b) | 35.24 | 19.07 | 66.94 | 53.21 | 26.06 | 18.48 | 45.96 | 36.81 | 31.16 | 19.32 | 69.01 | 40.43 |
| | DRN (Zeng et al., 2020) | 33.31 | 14.49 | 64.37 | 40.86 | - | 17.33 | - | 25.98 | 42.58 | 23.74 | 75.76 | 46.28 |
| | MMN (Wang et al., 2022) | 36.83 | 21.44 | 64.75 | 52.72 | 28.80 | 18.62 | 49.53 | 37.15 | 36.39 | 21.05 | 71.11 | 45.64 |
| | **CTR** | **45.10** | **26.57** | **78.29** | **62.45** | **38.64** | **26.39** | **59.38** | **45.72** | **44.60** | **27.03** | **88.71** | **56.95** |
| 50% | SCDM (Yuan et al., 2019a) | 12.27 | 4.90 | 22.31 | 14.28 | 12.04 | 9.88 | 16.19 | 13.56 | 29.25 | 12.57 | 30.73 | 20.62 |
| | VSLNet (Zhang et al., 2020a) | 19.14 | 10.38 | - | - | 12.27 | 10.52 | - | - | 28.64 | 13.16 | - | - |
| | CMIN (Zhang et al., 2019b) | 21.85 | 10.52 | 26.76 | 22.44 | 12.59 | 8.71 | 15.45 | 9.30 | - | - | - | - |
| | 2DTAN (Zhang et al., 2020b) | 24.36 | 14.01 | 38.26 | 31.80 | 23.92 | 14.35 | 30.41 | 23.28 | 16.26 | 8.94 | 27.85 | 14.39 |
| | DRN (Zeng et al., 2020) | 22.03 | 10.47 | 35.72 | 25.19 | - | 12.67 | - | 15.88 | 22.47 | 11.51 | 30.73 | 18.99 |
| | MMN (Wang et al., 2022) | 25.58 | 15.65 | 36.94 | 31.93 | 26.06 | 15.11 | 35.74 | 24.52 | 18.72 | 10.09 | 29.38 | 18.60 |
| | **CTR** | **40.92** | **23.86** | **74.37** | **59.17** | **34.29** | **22.93** | **55.44** | **41.96** | **41.18** | **23.51** | **84.64** | **53.27** |

Table 1: Performance comparison on ActivityNet Caption, TACoS, and Charades-STA datasets.

“Towards Robust Temporal Activity Localization Learning with Noisy Labels ” , COLING 2024

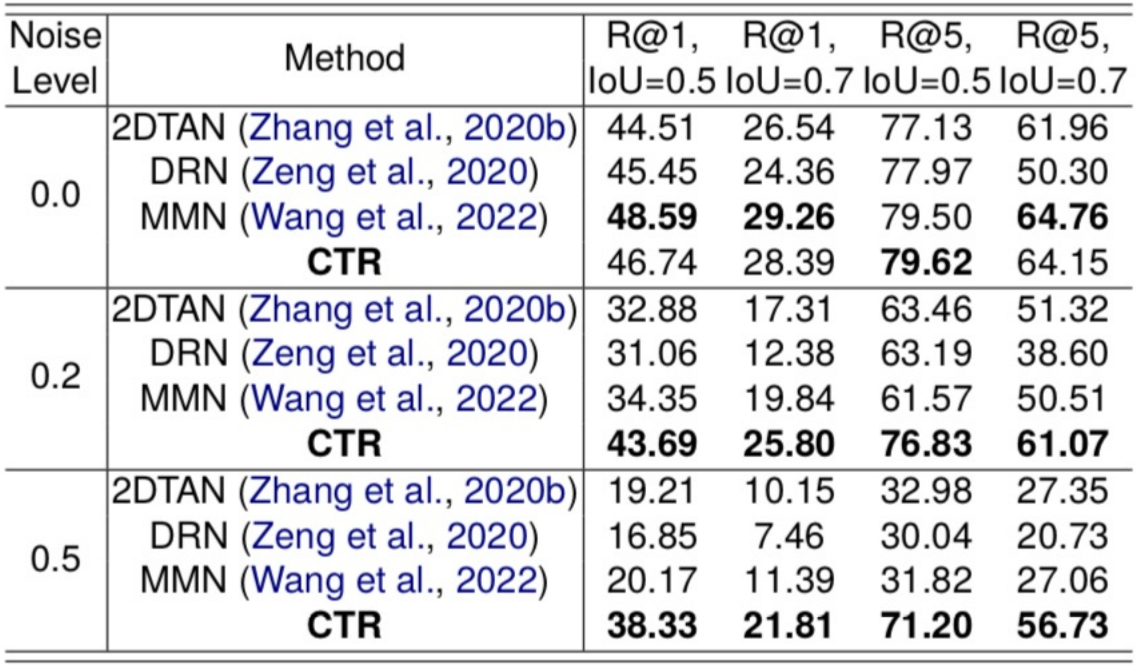
# Quantitative Comparison



| Model | Co-Teaching Paradigm | | | Curriculum Learning | | R@1, | R@1, | R@5, | R@5, |
|---|---|---|---|---|---|---|---|---|---|
| | Divide data | Refine label | Warm-up | Balanced weights | Controllers | IoU=0.5 | IoU=0.7 | IoU=0.5 | IoU=0.7 |
| Backbone | × | × | × | × | × | 23.82 | 15.25 | 39.14 | 32.37 |
| ① | ✓ | × | ✓ | × | × | 31.47 | 18.71 | 58.93 | 46.59 |
| ② | × | ✓ | ✓ | × | × | 25.63 | 15.19 | 42.38 | 33.43 |
| ③ | ✓ | ✓ | × | × | × | 3.94 | 0.85 | 16.46 | 11.71 |
| ④ | ✓ | ✓ | ✓ | × | × | 35.88 | 21.02 | 65.60 | 51.92 |
| ⑤ | ✓ | ✓ | ✓ | ✓ | × | 38.96 | 22.74 | 71.23 | 56.25 |
| ⑥ | ✓ | ✓ | ✓ | ✓ | ✓ | **40.92** | **23.86** | **74.37** | **59.17** |

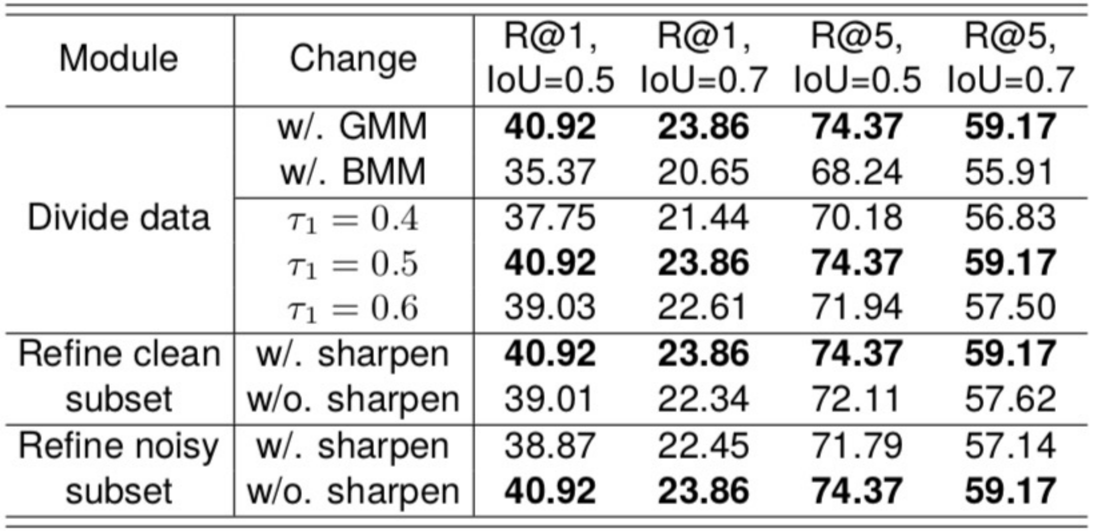
Table 2: Main ablation study on the ActivityNet Caption dataset with 50% noise ratio.

# Quantitative Comparison



| Noise Level | Method | R@1, IoU=0.5 | R@1, IoU=0.7 | R@5, IoU=0.5 | R@5, IoU=0.7 |
|---|---|---|---|---|---|
| 0.0 | 2DTAN (Zhang et al., 2020b) | 44.51 | 26.54 | 77.13 | 61.96 |
| | DRN (Zeng et al., 2020) | 45.45 | 24.36 | 77.97 | 50.30 |
| | MMN (Wang et al., 2022) | **48.59** | **29.26** | 79.50 | **64.76** |
| | **CTR** | 46.74 | 28.39 | **79.62** | 64.15 |
| 0.2 | 2DTAN (Zhang et al., 2020b) | 32.88 | 17.31 | 63.46 | 51.32 |
| | DRN (Zeng et al., 2020) | 31.06 | 12.38 | 63.19 | 38.60 |
| | MMN (Wang et al., 2022) | 34.35 | 19.84 | 61.57 | 50.51 |
| | **CTR** | **43.69** | **25.80** | **76.83** | **61.07** |
| 0.5 | 2DTAN (Zhang et al., 2020b) | 19.21 | 10.15 | 32.98 | 27.35 |
| | DRN (Zeng et al., 2020) | 16.85 | 7.46 | 30.04 | 20.73 |
| | MMN (Wang et al., 2022) | 20.17 | 11.39 | 31.82 | 27.06 |
| | **CTR** | **38.33** | **21.81** | **71.20** | **56.73** |

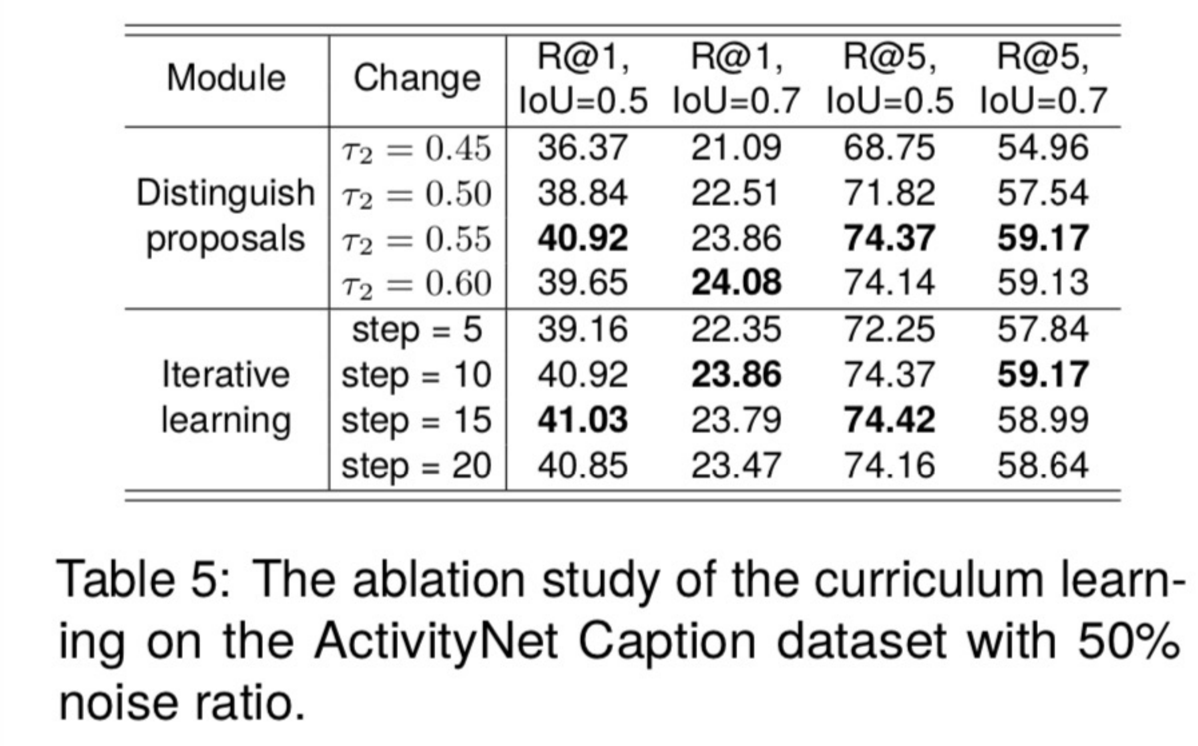
Table 3: Performance comparison on the ActivityNet dataset with different noise level.



| Module | Change | R@1, IoU=0.5 | R@1, IoU=0.7 | R@5, IoU=0.5 | R@5, IoU=0.7 |
|---|---|---|---|---|---|
| Divide data | w/. GMM | **40.92** | **23.86** | **74.37** | **59.17** |
| | w/. BMM | 35.37 | 20.65 | 68.24 | 55.91 |
| | $\tau_1 = 0.4$ | 37.75 | 21.44 | 70.18 | 56.83 |
| | $\tau_1 = 0.5$ | **40.92** | **23.86** | **74.37** | **59.17** |
| | $\tau_1 = 0.6$ | 39.03 | 22.61 | 71.94 | 57.50 |
| Refine clean subset | w/. sharpen | **40.92** | **23.86** | **74.37** | **59.17** |
| | w/o. sharpen | 39.01 | 22.34 | 72.11 | 57.62 |
| Refine noisy subset | w/. sharpen | 38.87 | 22.45 | 71.79 | 57.14 |
| | w/o. sharpen | **40.92** | **23.86** | **74.37** | **59.17** |

Table 4: The ablation study of the co-teaching paradigm on the ActivityNet Caption dataset with 50% noise ratio.

# Ablation Study



| Module | Change | R@1, IoU=0.5 | R@1, IoU=0.7 | R@5, IoU=0.5 | R@5, IoU=0.7 |
|---|---|---|---|---|---|
| Distinguish proposals | $\tau_2 = 0.45$ | 36.37 | 21.09 | 68.75 | 54.96 |
| | $\tau_2 = 0.50$ | 38.84 | 22.51 | 71.82 | 57.54 |
| | $\tau_2 = 0.55$ | **40.92** | 23.86 | **74.37** | **59.17** |
| | $\tau_2 = 0.60$ | 39.65 | **24.08** | 74.14 | 59.13 |
| Iterative learning | step = 5 | 39.16 | 22.35 | 72.25 | 57.84 |
| | step = 10 | 40.92 | **23.86** | 74.37 | **59.17** |
| | step = 15 | **41.03** | 23.79 | **74.42** | 58.99 |
| | step = 20 | 40.85 | 23.47 | 74.16 | 58.64 |

Table 5: The ablation study of the curriculum learning on the ActivityNet Caption dataset with 50% noise ratio.

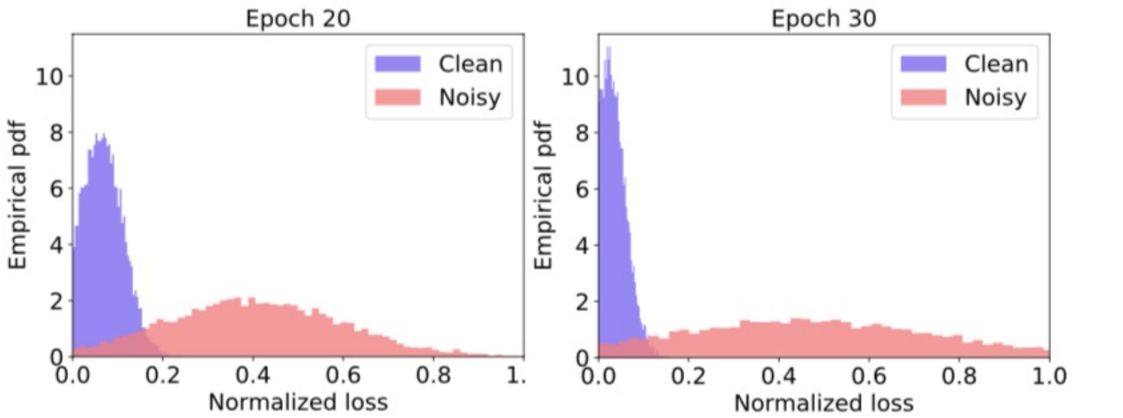
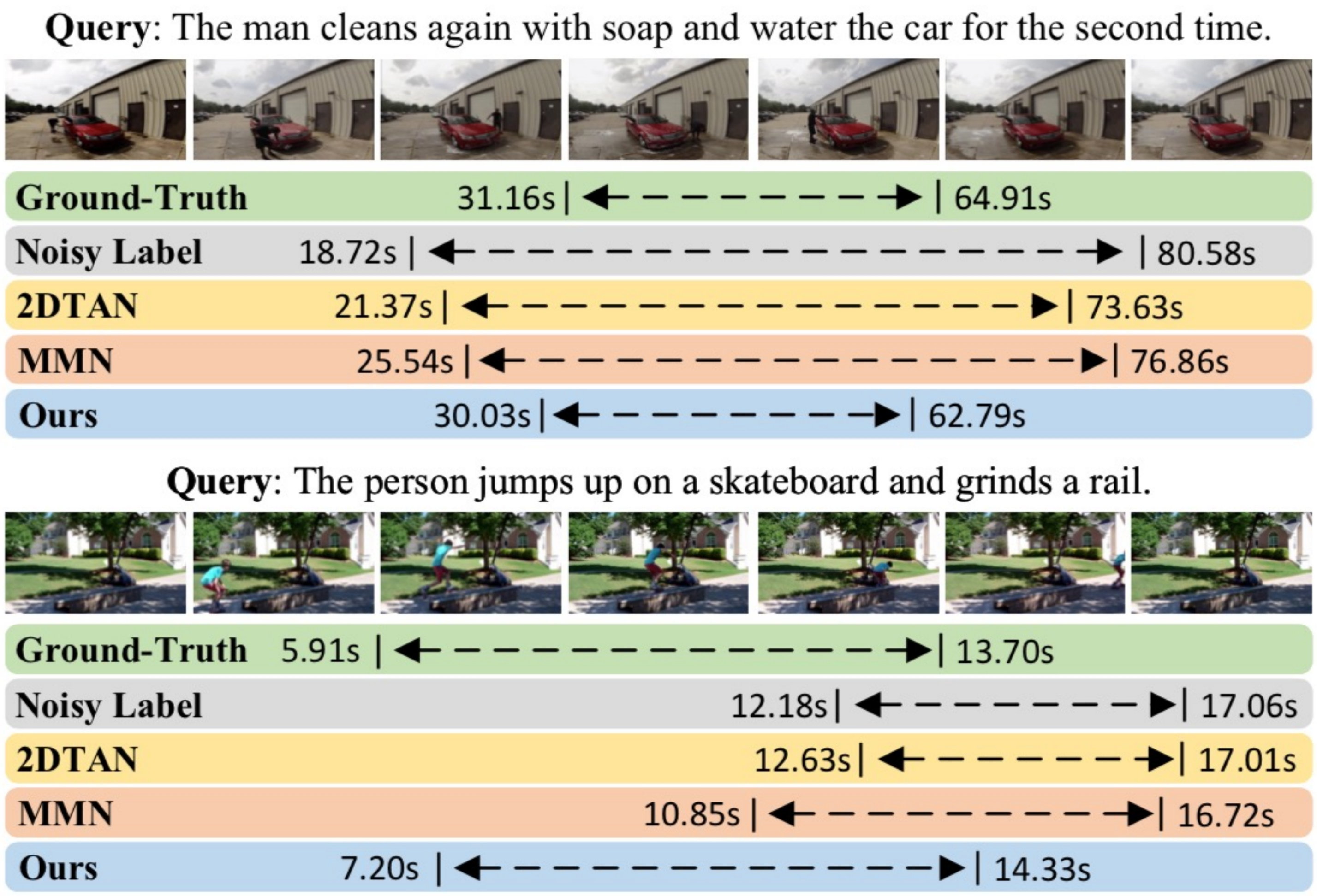


Figure 5: *Left:* The probability density function (PDF) on the clean and noisy sample when we re-train the model with 20 epochs. *Right:* The PDF when we re-train the model with 30 epochs.

# Qualitative Results



**Query**: The man cleans again with soap and water the car for the second time.

**Ground-Truth** 31.16s |← - - - - - - - - →| 64.91s

**Noisy Label** 18.72s |← - - - - - - - - - - - - - - - - →| 80.58s

**2DTAN** 21.37s |← - - - - - - - - - - - - - - ▶| 73.63s

**MMN** 25.54s |← - - - - - - - - - - - - - - - ▶| 76.86s

**Ours** 30.03s |← - - - - - - - - ▶| 62.79s

**Query**: The person jumps up on a skateboard and grinds a rail.

**Ground-Truth** 5.91s |← - - - - - - - - - - - - ▶| 13.70s

**Noisy Label** 12.18s |← - - - - - - ▶| 17.06s

**2DTAN** 12.63s |← - - - - - - ▶| 17.01s

**MMN** 10.85s |← - - - - - - - - →| 16.72s

**Ours** 7.20s |← - - - - - - - - - - - - ▶| 14.33s

# Thanks!

Email: dzliu@stu.pku.edu.cn