TP^2-DETR: Unlocking Deformable DETR for Zero-Shot Temporal Action Proposal Generation with Temporal Feature Pyramids

Supplementary Material

6. Additional Experimental Results

We present the detailed experimental results below to provide additional information.

6.1. Visualization of Prediction Quality

Figure 5 shows the visualization of prediction quality. We plot scatter charts of inference results from a random 50%/50% split of the THUMOS14 dataset to compare our TP^2-DETR method with the baseline Deformable DETR. Each point represents a predicted proposal, where its horizontal position indicates the confidence score (i.e., actionness score) and the vertical position represents its best temporal IoU with the ground truth. Since high-quality proposals typically appear in the top-right region (indicating both high confidence and accurate localization), TP^2-DETR demonstrates a significantly denser distribution in that area compared to the baseline.

6.2. Additional Qualitative Comparisons

Figure 6 shows two additional sets of qualitative comparisons in the same format as Figure 4 but under different numbers of action instances. Figure 7 shows the precision-recall chart of the dense video (`video_test_0000464`) from which we calculate the mAP value.

6.3. Visualization of Salient Logits

As mentioned in Section 5, our salient head is trained independently as an auxiliary component to facilitate early supervision, but we observe that its predictions (i.e., salient logits) often overlap with those of the primary head (i.e., bounding box head). Figure 8 shows several examples.

6.4. Comparison of THUMOS14 and ActivityNet1.3

As discussed in Section 4, the two datasets exhibit significant differences. Figure 9 shows the distributions of action durations. THUMOS14's action durations are significantly shorter than those of ActivityNet1.3. Figure 10 shows the distributions of relative action durations of the two datasets. The THUMOS14 distribution highly concentrates in the first bin (0-5%), while the ActivityNet1.3 distribution has a U shape.

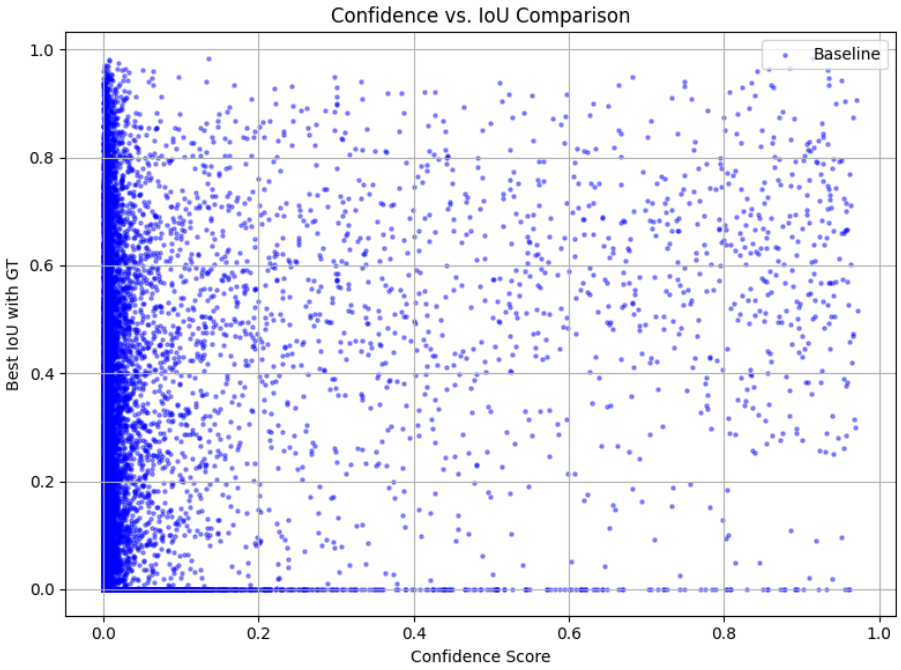
6.5. Class-Agnostic mAP Computation

Algorithm 2 outlines the detailed steps for computing the class-agnostic mAP reported in our manuscript.

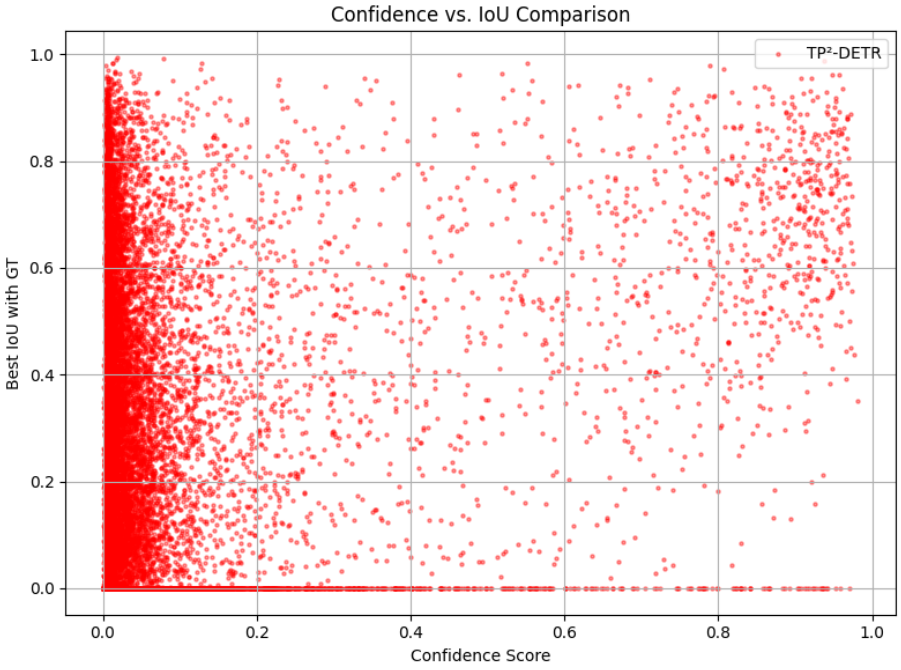
Algorithm 2: Class-Agnostic mAP Computation for Proposal Generation

Input:
G: Ground-truth proposals across all videos
P: Predicted proposals with actionness scores (i.e., confidence scores)
T: Set of tIoU thresholds
Output:
mAP: Mean Average Precision across all tIoU thresholds

foreach threshold t_{IoU} in T **do**
  Sort P by actionness scores in descending order;
  Initialize an empty set $matched\_gt$ to store matched ground truths;
  foreach prediction p in P **do**
    if exists $g \in G$ in same video s.t. $g \notin matched\_gt$ and IoU$(p, g) \geq t_{IoU}$ **then**
      Mark p as True Positive (TP);
      Add g to $matched\_gt$;
    end
    else
      Mark p as False Positive (FP);
    end
  end
  Compute precision and recall from TP/FP labels;
  Compute $\text{AP}_{t_{IoU}}$ as area under Precision-Recall curve;
end
Compute $mAP = \frac{1}{|T|} \sum_{t_{IoU} \in T} AP_{t_{IoU}}$;



(a) Baseline (Deformable DETR)



(b) TP^2-DETR

Figure 5. Visualizations of prediction quality (confidence vs. IoU with ground truth). Each point represents a predicted proposal. TP^2-DETR shows a denser concentration in the top-right region, indicating higher-quality proposals in terms of both localization and confidence. Results are reported on THUMOS14 using the 50%/50% split (split_id = 0).

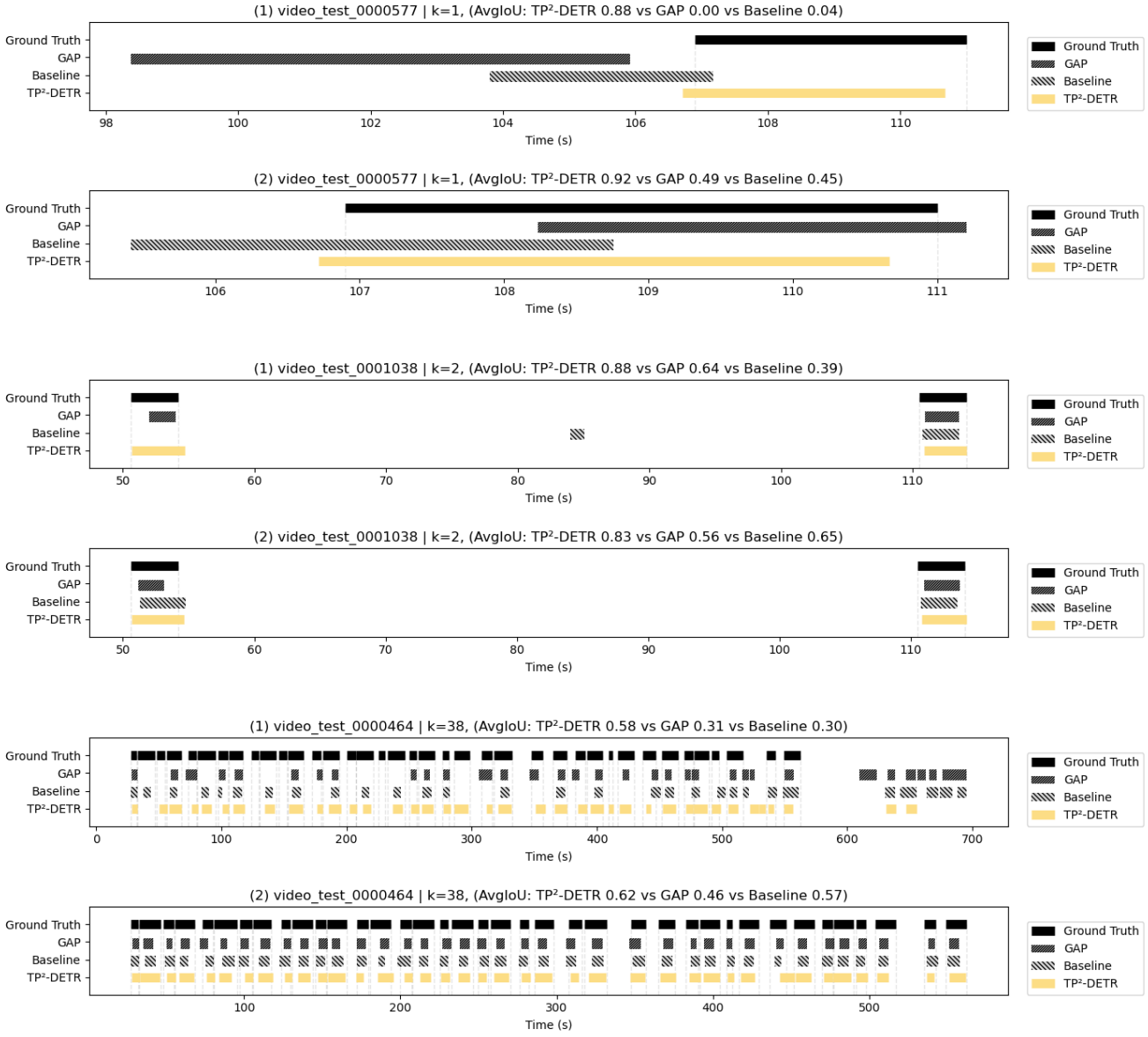


Figure 6. Visualizations of class-agnostic proposals as qualitative results. We present examples from sparse cases ($k = 1$, $k = 2$) to dense cases ($k = 38$), visualizing the top-k proposals selected in two ways: (1) by actionness scores (following the GAP approach) and (2) by bipartite matching using the same cost as in training. In addition to the intuitive metric AvgIoU for each case, we also provide a precision-recall curve for the dense video (`video_test_0000464`) to illustrate the mAP behavior. All results are from the 50/50 split (split_id = 0) on THUMOS'14.

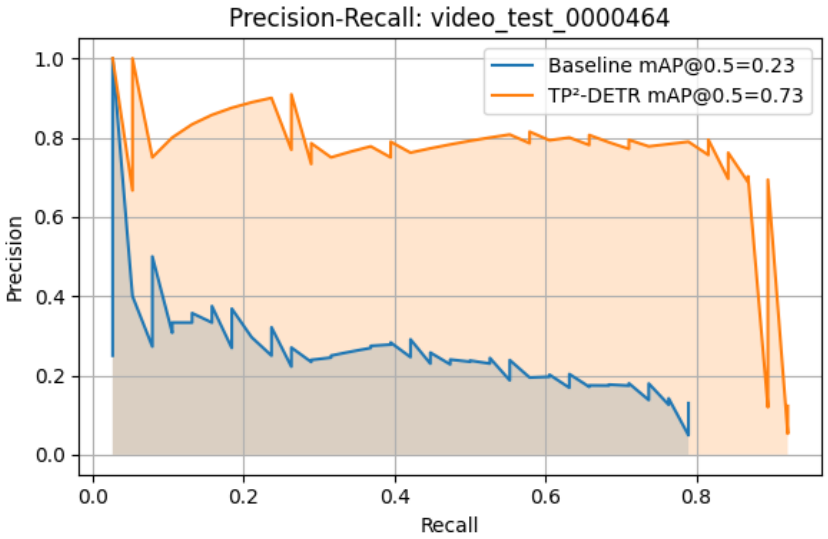


Figure 7. Precision-recall curve of the `0000464` video, whose proposal visualization results are shown in Figure 6. Here we show its precision-recall curve of the 38 actionness to illustrate the mAP behavior.

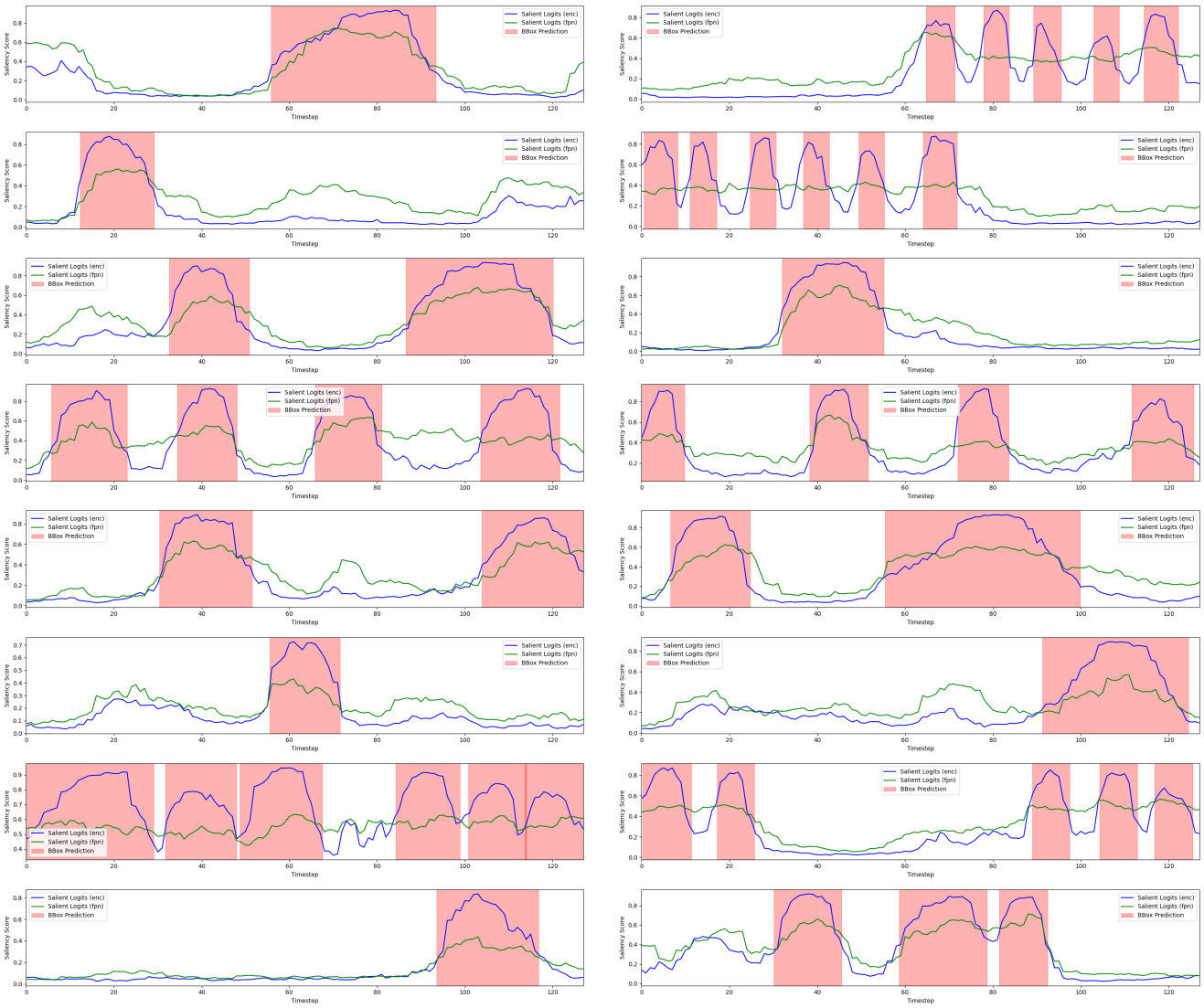


Figure 8. Visualization of salient logits (blue: from encoder, green: from temporal FPN) and predicted proposals (red) from matched queries for a randomly sampled training batch of videos on THUMOS'14 using the 50/50 split. Each subfigure shows the predicted saliency scores (y-axis) across timesteps (x-axis). The visualizations show that salient peaks—particularly those from the encoder output—often align with the predicted segments, despite the salient head being trained independently. This supports our hypothesis of potential cross-head consistency, which may be leveraged for future improvements.

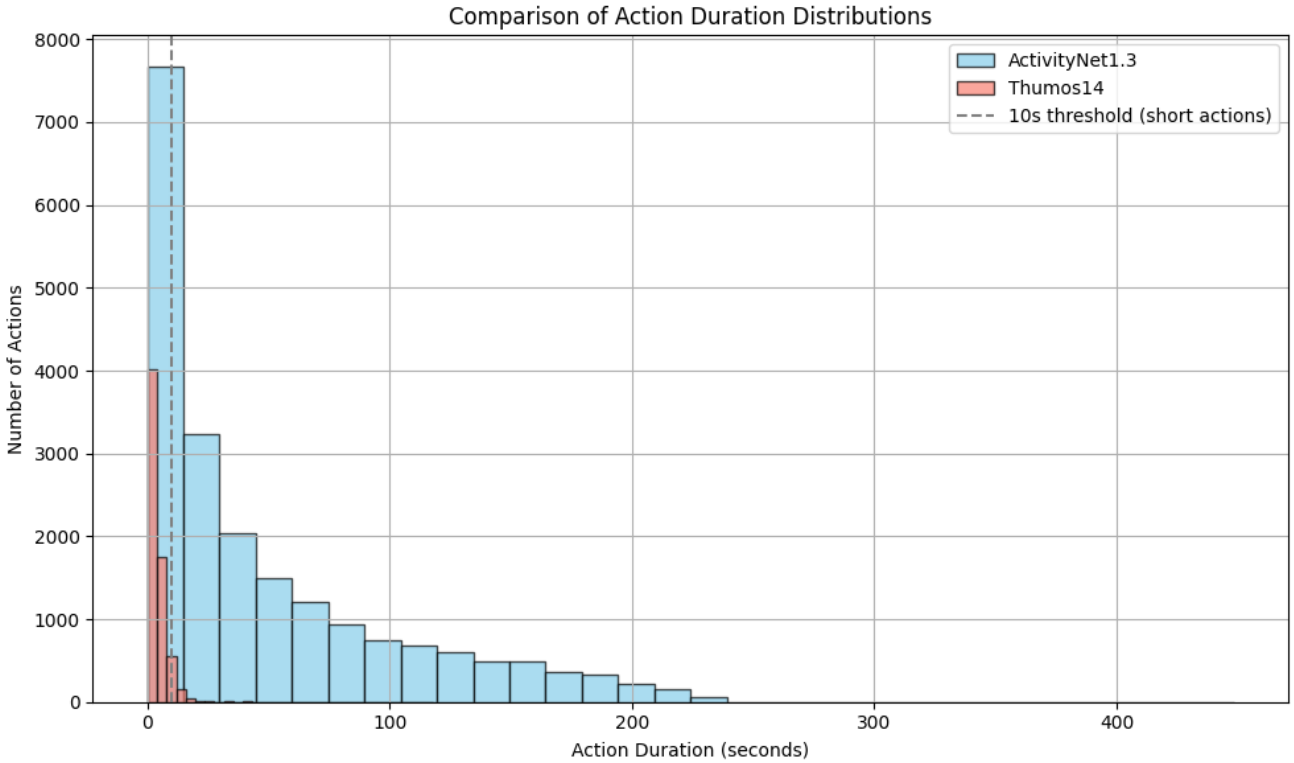


Figure 9. Histogram of action durations (seconds).

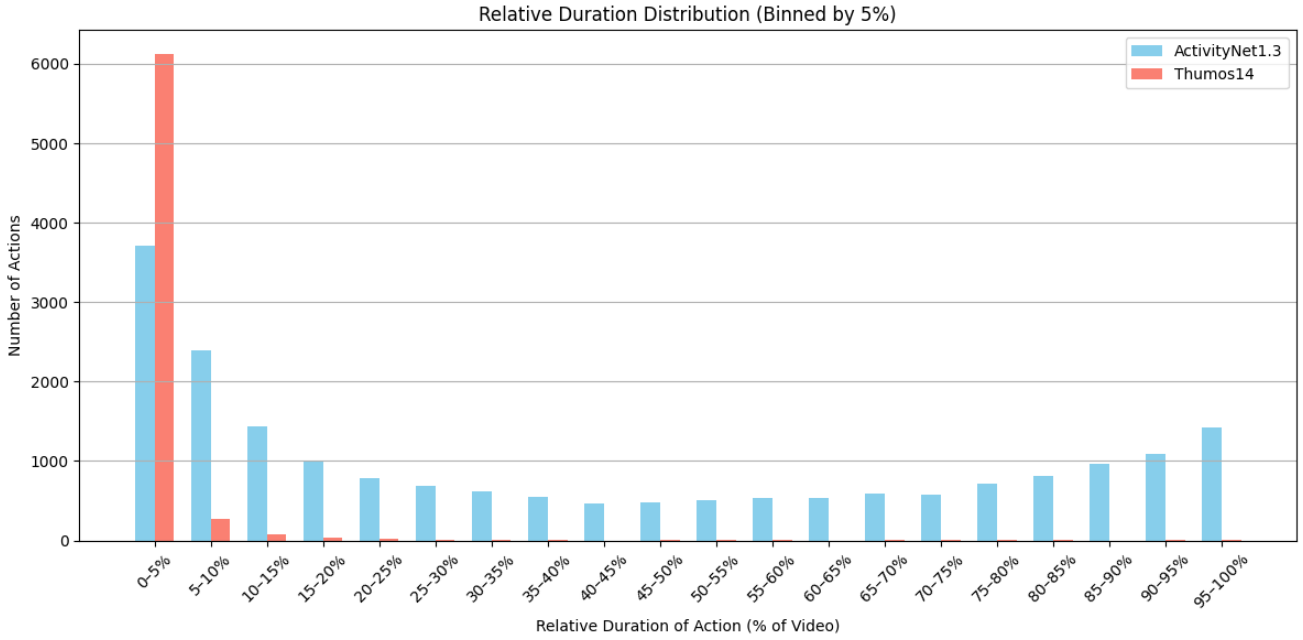


Figure 10. Relative view of action duration distributions on THUMOS14 and ActivityNet1.3. The x-axis indicates the relative duration of each action instance as a percentage of the corresponding video length. THUMOS14 exhibits a high concentration of short actions, with the majority lasting under 5% of video length. This observation aligns with our motivation to enhance short-action localization.