Appendix

For total transparency and a better understanding of our work, in this appendix, we supplement discussions about limitations and failure cases, differences with a pioneering work, as well as potential avenues for further explorations.

A. Limitations and Failure Cases

Though our adaptive token division is generally helpful, it is still possible that some tokens from distractors are mistakenly selected to interact with the template. This is typically when our tracker might fail. Fig. 1 shows a failure case. When the target (red box) is partially occluded by the distractor (blue box), the token selection becomes confusing and our tracker drifts for a while in the subsequent frames. A promising solution is using model update to exploit useful temporal cues, which we leave for future work.

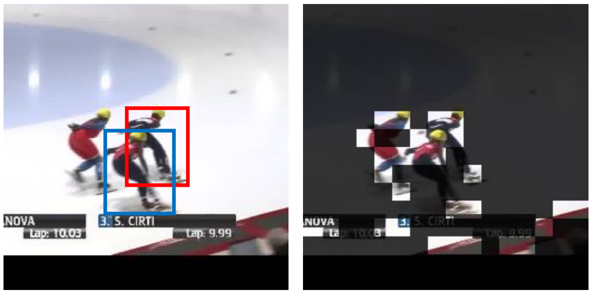

Figure 1. Visualization of a typical failure case.

B. Differences with DynamicViT

One may wonder about the differences between DynamicViT [1] and our GRM since the former method also incorporates an attention masking strategy and the Gumbel-Softmax technique. It is worth noting that the proposed method is driven by a completely different motivation. In particular, contrary to DynamicViT, which is designed to accelerate the inference speed with a definite accuracy drop, our method aims to prevent undesired cross-relations and thus improves the tracking performance. With this totally different objective, our method differs from DynamicViT in three major aspects: First, besides the task-specific (image classification) loss, DynamicViT needs three extra losses to constrain the token sparsification, whereas our method can implicitly learn the adaptive token division solely by the task-specific (target localization) loss. Second, in DynamicViT, once a token is pruned in a certain layer, it will never be used in the subsequent layers. Differently, our token division is independently determined by each layer, making it more robust to possible improper token selection in the earlier layers as it can still be recovered in the latter layers. Third, during inference, DynamicViT abandons a fixed ratio of tokens regardless of the prediction scores, whereas our method classifies an adaptive portion of tokens into each category based on the predictions.

C. Possible Explorations

C.1. Supervision of Token Division

Honestly, the outset of this work mainly stems from IA-SSD [3], which selects the sparse yet important foreground points for efficient 3D object detection. As a prototype in our early explorations, the prediction modules in each encoder layer are directly supervised by the ground-truth bounding boxes and do not affect the global cross-relation modeling during training. Essentially, they are learning a rough foreground classification based on each feature vector. During inference, we sort all the search tokens according to their foreground classification scores. Then, a portion of search tokens with higher scores is selected to model cross-relations with the template tokens. The remaining search tokens thus form another token group and model self-relations with themselves. Inspired by BOAT [2], we also maintain some shared search tokens around the selection boundary to allow the information flow between these two search token groups. Although the actual foreground classification results are far from precise, we surprisingly find a performance gain. We thus argue that selecting partial search tokens could be helpful as long as we refer to a relatively reasonable token ranking (*e.g.*, foreground classification scores). Nonetheless, the performance of this approach seems to be unstable on different benchmarks and the ratios of token selection in every encoder layer need to be tuned with lots of manual efforts.

Taking these explorations as foundations, the proposed GRM resorts to the Gumbel-Softmax technique to enable end-to-end optimized token division and achieves satisfactory results on multiple benchmarks. Nevertheless, the supervision from the ground-truth bounding boxes can be simultaneously applied to our token division modules as an auxiliary guidance, which is not investigated in this work. We conjecture that the integration of both implicit and explicit supervisions could lead to more interpretable token division results and contribute to better tracking performance.

C.2. Form of Relation Modeling

From the converged model, we notice that the relation modeling in most encoder layers permanently degenerates to two-stream or one-stream forms. For those degenerated cases, the learning of token division actually becomes a neural architecture search process. The calculation of the corresponding token division modules can be skipped during the inference to further reduce the computation overhead and reach a faster running speed than we reported. It is also noteworthy that there appears to be no obvious pattern in the form of relational modeling from the earlier layers to the latter layers. The irregular pattern may depend upon the initialization weights from the pretrained models.

We also attempt to use some regularizations (*e.g.*, the entropy of the token division ratio) to encourage all layers behave as the intermediate form. However, the eventual results become sensitive and we cannot witness on-par performance across all benchmarks. The underlying reason might be that for most encoder layers, the two degenerated forms can generalize better since each learned attention block only needs to handle a stationary situation, namely either global cross-relation modeling or no cross-relation modeling.

C.3. Alternative to Discrete Categorization

Prompted by one of the reviewers, we realize that we can also use the continuous estimations of the prediction modules to scale the raw attention weights. This flexible alternative can naturally bypass the non-differentiable obstacle caused by the strict constraints in our method and eliminate the need of Gumbel-Softmax. In contrast to the token-level attention weights, the scaling weights here can be devised at region-level to serve as a better complement. Such hierarchical attention weight design may hold more promises to improve the relation modeling for Transformer trackers, which is worth investigating in future work.

References

[1] Yongming Rao, Wenliang Zhao, Benlin Liu, Jiwen Lu, Jie Zhou, and Cho-Jui Hsieh. Dynamicvit: Efficient vision transformers with dynamic token sparsification. *Advances in neural information processing systems*, 34:13937–13949, 2021. 1

[2] Tan Yu, Gangming Zhao, Ping Li, and Yizhou Yu. Boat: Bilateral local attention vision transformer. *arXiv preprint arXiv:2201.13027*, 2022. 1

[3] Yifan Zhang, Qingyong Hu, Guoquan Xu, Yanxin Ma, Jianwei Wan, and Yulan Guo. Not all points are equal: Learning highly efficient point-based detectors for 3d lidar point clouds. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18953–18962, 2022. 1