A. Appendix

A.1. Additional Details

A.1.1 Non-Uniform Sparsity Loss

For each token, we determine whether or not it lies within a ground-truth bounding box. If the i-th token falls inside a bounding box, the corresponding heatmap's value is defined as

$$m_i = \exp\left(-\frac{(x_i - l_x)^2 + (y_i - l_y)^2}{2\sigma^2}\right), \tag{18}$$

where x_i and y_i are the xy-coordindates of the token, l_x and l_y are the xy-coordinates of the box's center, and σ is a hyper-parameter that controls the smoothness of the heatmap. For tokens that do not lie within a bounding box, the heatmap's value is set to zero. We create the heatmap for each object class, and we take the maximize over all heatmaps to obtain the final class-agnostic heatmap used by our non-uniform sparsity loss. The non-uniform sparsity loss is similar to focal loss [95], and it is defined as follows:

$$\mathcal{L}_s = -\sum_{l=1}^{L}\sum_{i\in\mathcal{K}}\frac{1}{|\mathcal{K}|}\left[(1-s_{l,i})^{\alpha}\log(s_{l,i})\mathbb{I}_{m_i\geq 1-\epsilon} + (1-m_i)^{\gamma}s_i^{\alpha}\log(1-s_{l,i})\mathbb{I}_{m_i<1-\epsilon}\right], \tag{19}$$

where $\mathcal{K} = \{i : k_{0:l-1,i} = 1\}$ is the set of tokens that have not been halted before the l-th layer, $s_{l,i}$ is the score for the i-th token at the l-th layer, $\alpha = 2$ and $\gamma = 4$ are hyper-parameters, and $\epsilon = 10^{-4}$ is used to improve numerical stability.

A.1.2 Analyzing the Pseudo-Gradient

In Section 5, we claim the following:

$$\Delta_i \approx \frac{\partial\mathcal{L}(q_1, q_2)}{\partial s_i} + O(u), \tag{20}$$

where

$$\Delta_i := \mathcal{L}(\tilde{q}_1, \tilde{q}_2) - \mathcal{L}(q_1, q_2) \tag{21}$$

is the difference in the detection loss when the i-th token is halted instead of being forwarded in a single layer network. In other words, we claim that the (pseudo)-gradient of $\mathcal{L}(\tilde{q}_1, \tilde{q}_2)$ with respect to s_i provided by our proposed EDF and the STE is a reasonable proxy of Δ_i.

To prove this claim, we begin by computing

$$\frac{\partial\mathcal{L}(q_1, q_2)}{\partial s_i} = \left\langle\frac{\partial\mathcal{L}(q_1, q_2)}{\partial q_1}, \frac{\partial q_1}{\partial s_i}\right\rangle + \left\langle\frac{\partial\mathcal{L}(q_1, q_2)}{\partial q_2}, \frac{\partial q_2}{\partial s_i}\right\rangle. \tag{22}$$

Recall that $q_1 = (\mathbf{1} - k) \circ f$ and $q_2 = k \circ \phi_2(\mathrm{WSA}(\phi_1(f), s \circ k), f)$. Using the definition of the STE,

$$\frac{\partial q_1}{\partial s_i} = -\frac{\partial(k \circ f)}{\partial s_i} = -\mathbf{1}_i \circ f \tag{23}$$

and

$$\begin{aligned}\frac{\partial q_2}{\partial s_i} &= k \circ \frac{\partial\phi_2(\mathrm{WSA}(\phi_1(f), s\circ k), f)}{\partial s_i} + \frac{\partial k}{\partial s_i} \circ \phi_2(\mathrm{WSA}(\phi_1(f), s \circ k), f) &(24)\\ &= k \circ \frac{\partial\phi_2(\mathrm{WSA}(\phi_1(f), s\circ k), f)}{\partial s_i} + \mathbf{1}_i \circ \phi_2(\mathrm{WSA}(\phi_1(f), s \circ k), f). &(25)\end{aligned}$$

Next, let us compute Δ_i. Using Taylor series approximation,

$$\Delta_i \approx \left\langle\frac{\partial\mathcal{L}(q_1, q_2)}{\partial q_1}, \tilde{q}_1 - q_1\right\rangle + \left\langle\frac{\partial\mathcal{L}(q_1, q_2)}{\partial q_2}, \tilde{q}_2 - q_2\right\rangle. \tag{26}$$

Recall that $\tilde{q}_1 = (\mathbf{1} - k - \mathbf{1}_i) \circ f$ and $\tilde{q}_2 = (k + \mathbf{1}_i) \circ \phi_2(\mathrm{WSA}(\phi_1(f), s \circ (k + \mathbf{1}_i)), f)$; therefore,

$$\tilde{q}_1 - q_1 = -\mathbf{1}_i \circ f \tag{27}$$

and

$$\begin{aligned}
\tilde{q}_2 - q_2 &= (k + \mathbf{1}_i) \circ \phi_2(\mathrm{WSA}(\phi_1(f), s \circ (k + \mathbf{1}_i)), f) - k \circ \phi_2(\mathrm{WSA}(\phi_1(f), s \circ k), f) && (28) \\
&= k \circ \phi_2(\mathrm{WSA}(\phi_1(f), s \circ (k + \mathbf{1}_i)), f) - k \circ \phi_2(\mathrm{WSA}(\phi_1(f), s \circ k), f) && (29) \\
&+ \mathbf{1}_i \circ \phi_2(\mathrm{WSA}(\phi_1(f), s \circ (k + \mathbf{1}_i)), f). && (30)
\end{aligned}$$

Again, using Taylor series approximation,

$$k \circ \phi_2(\mathrm{WSA}(\phi_1(f), s \circ (k + \mathbf{1}_i)), f) - k \circ \phi_2(\mathrm{WSA}(\phi_1(f), s \circ k), f) \approx k \circ \frac{\partial \phi_2(\mathrm{WSA}(\phi_1(f), s \circ k), f)}{\partial s_i} \tag{31}$$

and

$$\mathbf{1}_i \circ \phi_2(\mathrm{WSA}(\phi_1(f), s \circ (k + \mathbf{1}_i)), f) \approx \mathbf{1}_i \circ \phi_2(\mathrm{WSA}(\phi_1(f), s \circ k), f) + \mathbf{1}_i \circ \frac{\partial \phi_2(\mathrm{WSA}(\phi_1(f), s \circ k), f)}{\partial s_i} \tag{32}$$

As a result,

$$\tilde{q}_2 - q_2 \approx k \circ \frac{\partial \phi_2(\mathrm{WSA}(\phi_1(f), s \circ k), f)}{\partial s_i} + \mathbf{1}_i \circ \phi_2(\mathrm{WSA}(\phi_1(f), s \circ k), f) + \mathbf{1}_i \circ \frac{\partial \phi_2(\mathrm{WSA}(\phi_1(f), s \circ k), f)}{\partial s_i}. \tag{33}$$

Comparing Eq. (23) to Eq. (27) and Eq. (25) to Eq. (33), we see that

$$\begin{aligned}
\Delta_i - \frac{\partial \mathcal{L}(q_1, q_2)}{\partial s_i} &\approx \left\langle \frac{\partial \mathcal{L}(q_1, q_2)}{\partial q_2}, \mathbf{1}_i \circ \frac{\partial \phi_2(\mathrm{WSA}(\phi_1(f), s \circ k), f)}{\partial s_i} \right\rangle && (34) \\
&= \frac{\partial \mathcal{L}(q_1, q_2)}{\partial q_{2,i}} \frac{\partial \phi_2(\mathrm{WSA}(\phi_1(f), s \circ k), f)_i}{\partial s_i} && (35) \\
&= \frac{\partial \mathcal{L}(q_1, q_2)}{\partial q_{2,i}} \frac{\partial \phi_2(\mathrm{WSA}(\phi_1(f), s \circ k), f)_i}{\partial \mathrm{WSA}(\phi_1(f), s \circ k)_i} \frac{\partial \mathrm{WSA}(\phi_1(f), s \circ k)_i}{\partial (s \circ k)_i} s_i, && (36)
\end{aligned}$$

where

$$\frac{\partial \phi_2(\mathrm{WSA}(\phi_1(f), s \circ k), f)_i}{\partial s_i} = \frac{\partial \phi_2(\mathrm{WSA}(\phi_1(f), s \circ k), f)_i}{\partial \mathrm{WSA}(\phi_1(f), s \circ k)_i} \frac{\partial \mathrm{WSA}(\phi_1(f), s \circ k)_i}{\partial (s \circ k)_i} \frac{\partial (s \circ k)_i}{\partial s_i} \tag{37}$$

and

$$\frac{\partial (s \circ k)_i}{\partial s_i} = \frac{\partial s_i k_i}{\partial s_i} = k_i + s_i = s_i \tag{38}$$

using the definition of the STE and the fact that $k_i = 0$. In general, the derivative of $\mathcal{L}$ and ϕ_2 is bounded since the parameter space of the network is bounded (due to the weight decay) and the operators inside the network are Lipschitz continuous. However, $\partial \mathrm{WSA}(\phi_1(f), s \circ k)_i / \partial (s \circ k)_i$ can be singular when all the element of $s \circ k$ are zero. We argue that in our analysis, we can still treat this term as bounded for two reasons. Firstly, in practice, we add an ϵ to the denominator of the WSA to prevent numeric instability, which makes the gradient bounded even if all the elements of $s \circ k$ are zero. Secondly, we employ gradient clipping to enforce a bound on the gradient. All of this combine, we have

$$\Delta_i \approx \frac{\partial \mathcal{L}(q_1, q_2)}{\partial s_i} + O(u). \tag{39}$$

That is, the approximation error of the pseudo-gradient is proportional to s_i thanks to the usage of weighted attention. Furthermore, since $k_i = 0$, we have $|s_i| < u$ where the threshold u is in general a very small value. This demonstrates that our pseudo-gradient provides useful information for updating the halting module.

Speed Up/Sparsity	Method	Vehicle		Pedestrian		Cyclist	
		AP/APH L1	AP/APH L2	AP/APH L1	AP/APH L2	AP/APH L1	AP/APH L2
1.00/0.00	Original	76.2/75.7	67.7/67.2	79.9/71.4	72.7/64.8	67.7/66.3	65.2/63.8
1.07/0.00	Width scale	75.4/74.9	66.9/66.5	79.5/70.7	72.2/64.0	65.6/64.1	63.0/61.6
1.16/0.00	Num head scale	75.5/75.0	67.0/66.6	79.4/70.8	72.0/64.1	65.5/63.9	63.0/61.5
1.28/0.48	AViT$_{\text{adapted}}$ [87]	71.9/71.3	63.4/62.9	76.8/67.9	69.1/60.9	63.1/61.6	60.7/59.3
1.21/0.75	Ours	76.1/75.6	67.8/67.3	79.4/70.7	72.1/64.0	67.0/65.6	64.4/63.1
1.13/0.00	Width scale	73.8/73.3	65.3/64.8	78.2/69.0	70.6/62.1	61.7/60.2	59.3/57.9
1.45/0.00	Num head scale	73.8/73.3	65.4/64.9	78.2/68.8	70.7/62.1	62.0/60.3	59.6/58.0
1.39/0.57	AViT$_{\text{adapted}}$ [87]	70.3/69.7	61.9/61.4	76.2/67.2	68.4/60.1	60.8/59.3	58.5/57.0
1.37/0.82	Ours	76.1/75.6	67.7/67.2	79.9/71.5	72.6/64.7	67.4/66.1	64.8/63.6
1.18/0.00	Width scale	70.4/69.8	62.0/61.5	74.5/64.3	66.7/57.4	54.9/52.9	52.8/50.8
1.45/0.00	Num head scale	73.8/73.3	65.4/64.9	78.2/68.8	70.7/62.1	62.0/60.3	59.6/58.0
1.55/0.72	AViT$_{\text{adapted}}$ [87]	70.1/69.6	61.7/61.3	76.3/67.5	68.5/60.4	61.6/60.1	59.2/57.8
1.52/0.89	Ours	75.4/74.9	67.0/66.5	79.7/71.5	72.4/64.7	67.1/65.7	64.5/63.2

Table 5. Efficiency and accuracy trade-off. We report the relative backbone speed-up and the average sparsity across all the attention layers.

A.2. Additional Experiment Details and Results

A.2.1 Setup and Implementation Details

We use a mixed strategy to decide the halting threshold. We specific an upper and lower token score quantile (denoted as α_u and α_l) and enforce that the sparsity of each layer varies within $[\alpha_l, \alpha_u]$. To achieve this, the final threshold is given by clamping the pre-specific threshold u within $[Q(\alpha_l), Q(\alpha_u)]$ where $Q(\alpha_l)$ and $Q(\alpha_u)$ denote the score corresponding to the α_l and α_u quantile, respectively. We enforce such a constrain because we observe that the distribution of scores can vary considerably for different scenes during the early stages of training and selecting the threshold in this way helps to stabilize training. In Table 1, the sparsity is bounded between 80% and 90% for the first halting module, 90% and 99% for the second halting module, and the default value of u is 0.01. The following technique can be used to identify u for a new dataset/model: train a model for a short period, then select u such that it is higher than the score of most foreground voxels and less than the score of most background voxels.

A.2.2 Efficiency and Accuracy Trade-off

For the baselines, we vary the latent dimension of the attention mechanism by $\{16, 12, 8, 4\}$, and we vary the number of attention head by $\{8, 6, 5, 4\}$. We adapt AViT from [87], but we apply our token recycling to improve the performance. Also, we adjust the number of input token features for the halting module from 1 to 32 as this improves performance while having a negligible impact on latency. Table 5 summarizes the results. Overall, we observe that our method significantly improves over other model scaling approaches as well as AViT.

A.2.3 The SST$^{++}_{\text{halt}}$ Architecture

For our SST$^{++}_{\text{halt}}$ architecture, we use a U-Net [61] and a single layer MLP as the first and second halting module. We find that the latent features of the U-Net contain useful semantic information. To reuse those features, we fuse the token features with the U-Net's features by applying a linear transformation and sums the features. Furthermore, we add the U-Net's feature map to the BEV feature map. To leverage the latency savings provided by halting tokens, SST$^{++}_{\text{halt}}$ uses an extra convolutional block in the detection head for a total of two convolutional blocks. The first convolutional block contains four convolutional layers. The second convolutional block contains four convolutional layer where the first layer has a stride of 2. All the layers use a kernel size of 3. Afterwards, we use the Feature Pyramid Network [34] employed by SECOND [81] to fuse the two scales of the BEV feature map and make predictions.