A Reproducibility

The experimental results in this paper are reproducible. We explain the details of model training and configuration in the main text and supplement it in the appendix. Our codes and models of SpikeYOLO will be available on GitHub after review.

B Details of the SpikeYOLO architecture

The structural specifics of SpikeYOLO are outlined in Table 6. This architecture is implemented across four scales, distinguished by varying width and depth scaling factors (see Table 7). The width scaling factor regulates the number of channels, whereas the depth scaling factor establishes the number of repetitions.

C Firing Rate of Different D on the Gen1 Dataset

In our method, T is the timestep, and D signifies the maximum integer value for activation during training. d denotes The firing rate of different quantified values. Inference timestep is reported as $T \times D$, e.g., 2×4 denotes $T = 2, D = 4$. Table 4 shows that the increase of D will bring less energy cost on the Gen1 dataset. To delve deeper into this phenomenon, we evaluate the firing rate of the model (referenced in Table 4) on different D, and report results in 8. When D increases, there's a notable decrease in the firing rate for each d, leading to a reduction in power consumption. For instance, at $d = 1$, the firing rate of 1×1 vs. 1×4: 0.1942 vs. 0.1626. Our analysis also highlights the significance of outliers in influencing the model's performance. Take 2×4 as an example, there are only 0.0029 on $d = 3$ and 0.0001 on $d = 4$, but contribute to a 0.9% improvement in mAP@50.

D Object Detection Results on the Gen1 Dataset

Here, we present the object detection outcomes on the Gen1 Dataset, illustrating the effects of varying T and D(see Fig. 5). We observe a significant enhancement in the model's feature extraction capabilities with an increase in either T or D.

Table 6: The details of SpikeYOLO architecture. Input Layer is the input of the layer(-1 mains the previous layer). Depth refers to the frequency of repetition for each layer(repeat at least once).

Layer	Input Layer	Layer Specification		Kernel	Channel	Stride	Ratio	Depth
1	-1	downsampling	Conv	7	128	4		
2	-1	SNN-Block-1	SepConv	7	128	1	2	3
			Channel Conv	3	128	1	4	
3	-1	downsampling	Conv	3	256	2		
4	-1	SNN-Block-1	SepConv	7	256	1	2	6
			Channel Conv	3	256	1	4	
5	-1	downsampling	Conv	3	512	2		
6	-1	SNN-Block-2	SepConv	7	512	1	2	9
			Channel Conv	3	512	1	3	
7	-1	downsampling	Conv	3	512	2		
8	-1	SNN-Block-2	SepConv	7	1024	1	2	3
			Channel Conv	3	1024	1	2	
9	-1	SPPF						
10	-1	SNNConv	Conv	1	512	1		
11	-1	upsampling						
12	-1	SNN-Block-2	SepConv	7	512	1	2	1
			Channel Conv	3	512	1	3	
13	-1,6	concat						
14	-1	SNNConv	Conv	1	256	1		
15	-1	upsampling						
16	-1	SNN-Block-1	SepConv	7	256	1	2	1
			Channel Conv	3	256	1	4	
17	-1,4	concat						
18	-1	SNNConv	Conv	1	256	1		
19	-1	SNN-Block-1	SepConv	7	256	1	2	1
			Channel Conv	3	256	1	4	
20	-1	SNNConv	Conv	3	256	1		
21	-1,14	concat						
22	-1	SNN-Block-2	SepConv	7	512	1	2	1
			Channel Conv	3	512	1	3	
23	-1	SNNConv	Conv	3	512	1		
24	-1,10	concat						
25	-1	SNN-Block-2	SepConv	7	1024	1	2	1
			Channel Conv	3	1024	1	1	
26	19,22,25	Detect						

Table 7: Width and depth scaling factors of SpikeYOLO. Taking Table 6 as a benchmark, the model's channel will be multiplied by the Width Scale, and depth will be multiplied by the Depth Scale.

Model	13.2M	23.1M	48.1M	68.8M
Width Scale	×0.375	×0.5	×0.625	×0.75
Depth Scale	×0.333	×0.333	×0.5	×0.5

Table 8: The firing rate of different d on the Gen1 dataset. d denotes the firing rate of different quantified values, sum indicates the aggregate firing rates ranging from $d = 1$ to $d = 4$. All the models are in line with those in Table 4.

$T \times D$	$d = 1$	$d = 2$	$d = 3$	$d = 4$	sum	mAP@50(%)
1×1	0.1942	0	0	0	0.1942	59.3
1×4	0.1626	0.0198	0.0035	0.0001	0.1869(-0.0073)	65.1(+5.8)
2×1	0.1938	0	0	0	0.1938	63.6
2×2	0.1692	0.0185	0	0	0.1877(-0.0061)	66.1(+2.5)
2×4	0.1566	0.0165	0.0029	0.0001	0.1767(-0.0171)	67.0(+3.4)
4×1	0.1804	0	0	0	0.1804	66.0
4×2	0.1459	0.0098	0	0	0.1595(-0.0209)	67.2(+1.2)
8×1	0.1942	0	0	0	0.1942	66.7

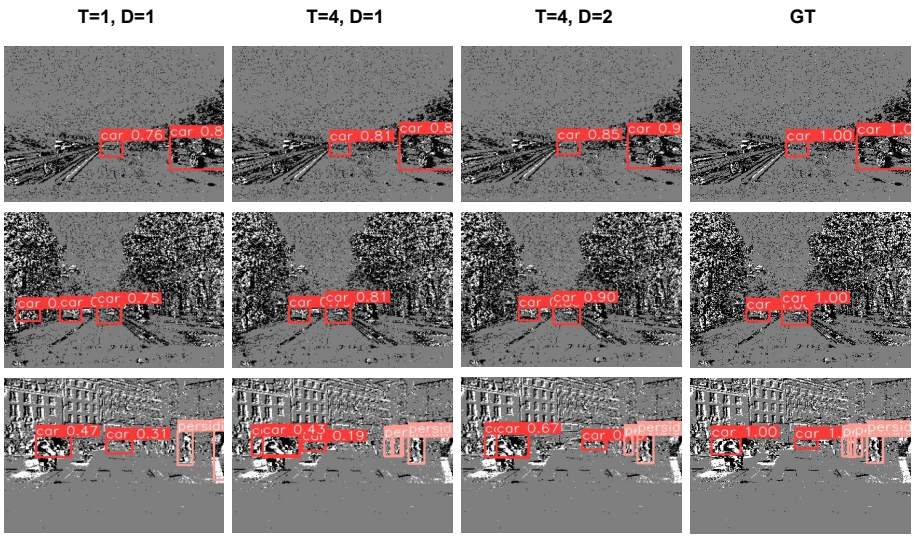

Fig. 5: The object detection results on the Gen1 dataset. The initial two columns analyze the impact of timestep T on performance, while the subsequent columns evaluate the influence of maximum integer value D on performance. Each model under consideration maintains a consistent parameter count of 23.1M.