DyRoNet: Dynamic Routing and Low-Rank Adapters for Autonomous Driving Streaming Perception (Supplementary Material)

The appendix completes the main paper by providing in-depth research details and extended experimental results. The structure of the appendix is organized as follows:

1. Analysis of Environmental Factors Affecting Streaming Perception: Sec. A
    - Impact of Weather Conditions: Sec. A.1
    - Quantitative Analysis of Objects: Sec. A.2
    - Proportion of Small Objects: Sec. A.3
    - Environmental Speed Dynamics: Sec. A.4
2. Expanded Experimental Results: Sec. B
    - Inference Time: Analysis Sec. B.1
    - Statistic of model selection: Sec. B.2
    - The comparison between Speed Router and $\mathbb{E}[\Delta I_t]$: Sec. B.3
3. Detailed Description of *DyRoNet*: Sec. C
    - Selection of Pre-trained Model: Sec. C.1
    - Hyperparameter Settings: Sec. C.2
4. Detailed Description of Experiments on nuScenes-H Dataset: Sec. D

A. Factor Analysis in Streaming Perception

In development of *DyRoNet*, we undertook an extensive survey and analysis to identify key influencing factors in autonomous driving scenarios that could potentially impact streaming perception. This analysis utilized the Argoverse-HD dataset [5], a benchmark in the field of streaming perception. The primary goal of this factor analysis was to isolate the most critical factor affecting streaming perception performance. As elaborated in the main text, our comprehensive analysis led to the identification of the speed of the environment as the predominant factor. Consequently, *DyRoNet* is tailored to address this specific aspect. Our analysis focuses on four primary elements: *weather conditions*, *object quantity*, *small object proportion*, and *environmental speed*. We methodically examined each of these factors to evaluate their respective impacts on streaming perception within autonomous driving.

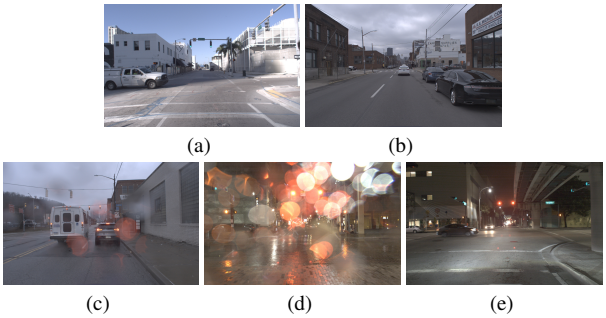


Figure 1. Illustrative Examples of Varied Weather Conditions and Times of Day: (a) Sunny during Daytime, (b) Cloudy during Daytime, (c) Rainy during Daytime, (d) Rainy during Nighttime, (e) Sunny during Nighttime.

A.1. Impact of Weather Conditions

The Argoverse-HD dataset, comprising testing, training, and validation sets, includes a diverse range of weather conditions. Specifically, the dataset contains 24, 65, and 24 video segments in the testing, training, and validation sets, respectively, with frame counts ranging from 400 to 900 per segment. Tab. 1 details the distribution of various weather types across these subsets. Fig. 1 provides visual examples of different weather conditions captured in the dataset. A clear variation in visual clarity and perception difficulty is observable under different conditions, with scenarios like Sunny + Day or Cloudy + Day appearing visually more challenging compared to Rainy + Night.

To evaluate the impact of weather conditions on streaming perception, we conducted tests using a range of pre-trained models from StreamYOLO [8], LongShortNet [4], and DAMO-StreamNet [3], employing various scales and settings. The results, presented in Tab. 2, indicate that performance is generally better during Day conditions compared to Night. This confirms that weather conditions indeed influence streaming perception.

However, it's noteworthy that even within the same weather conditions, model performance varies significantly, with accuracy ranging from below 10% to above 70%. Fig. 2 illustrates this point by comparing frames from two video segments (Clip ids: `00c561` and `395560`) under identical weather conditions, where the performance difference of the same model on these segments is as high as 32.1%. This observation suggests the presence of other crit-

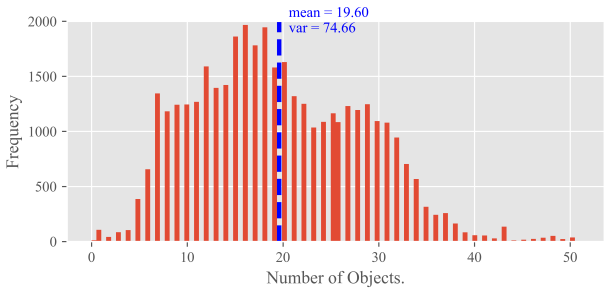
ical environmental factors that affect streaming perception, indicating that weather, while influential, is not the sole determinant of model performance.

	test	train	val
Sunny + Day	8	34	8
Cloudy + Day	13	27	15
Rainy + Day	1	1	0
Rainy + Night	1	0	0
Sunny + Night	1	3	1

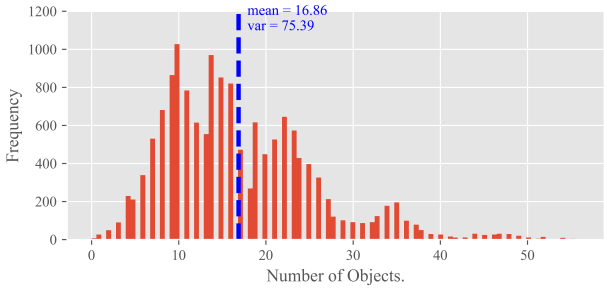
Table 1. Distribution of Weather Conditions in Testing, Training, and Validation Sets: This figure illustrates the frequency of different weather conditions in the testing, training, and validation sets of the Argoverse-HD dataset, providing an overview of the environmental variability within each dataset subset.

(a) (b)

Figure 2. Rapid Fluctuations in Performance Under Identical Weather Conditions: (a) Clip id: 00c561 shows a Streaming Average Precision (sAP) of 16.2% using the StreamYOLO-s model, (b) Clip id: 395560 demonstrates a significantly higher sAP of 48.3% under the same model and weather condition, illustrating the variability in model performance even under consistent environmental factors.



(a)



(b)

Figure 3. Histograms Depicting Object Quantity in the Argoverse-HD Dataset: This figure presents two histograms, (a) representing the distribution of the number of objects per frame in the training set of Argoverse-HD, and (b) showing the same distribution in the validation set. These histograms provide a visual analysis of object frequency and variability within different sets of the dataset.

A.2. Analysis of Object Quantity Impact

To assess the impact of the number of objects on streaming perception, we conducted a statistical analysis of object counts per frame in the Argoverse-HD dataset, encompassing both training and validation sets. The results of this analysis are depicted in Fig 3, which showcases a histogram representing the distribution of the number of objects in individual frames. The variance in the distribution is notable, with values of 74.66 for the training set and 75.39 for the validation set, indicating significant fluctuation in the number of objects across frames. Additionally, as shown in Tab. 2, there is considerable variability in object counts across different video segments. This observation led us to further investigate the potential correlation between object quantity and model performance fluctuations.

To explore this correlation, we calculated the average number of objects per frame for each segment within the Argoverse-HD validation set. The findings, detailed in Tab. 3, include the average object counts alongside Spearman correlation coefficients, which measure the relationship between object quantity and model performance. The absolute values of these coefficients range from 1e-1 to 1e-2. This range of correlation coefficients suggests that the number of objects present in the environment does not exhibit a strong or significant correlation with the performance of streaming perception models. In other words, our analysis indicates that the sheer quantity of objects within the environment is not a predominant factor influencing the efficacy of streaming perception.

A.3. Analysis of the Proportion of Small Objects

The influence of small objects on perception models, particularly in autonomous driving scenarios, has been underscored in studies like [6] and [8]. In such scenarios, even minor shifts in viewing angles can cause notable relative displacement of small objects, posing a challenge for perception models in processing streaming data effectively. This observation prompted us to closely examine the proportion of small objects in the environment.

To begin, we analyzed the area ratios of objects in both the training and validation sets of the Argoverse-HD dataset. This involved calculating the ratio of the pixel area covered by an object's bounding box to the total pixel area of the frame. We visualized these ratios in histograms

		StreamYOLO					LongShortNet				DAMO-StreamNet			
Clip ID	Weather	s 1x	m 1x	l 1x	l 2x	l still	s 1x	m 1x	l 1x	l high	s 1x	m 1x	l 1x	l high
1d6767	Cloudy + Day	**20.9**	**22.8**	**24.9**	7.0	**26.7**	**20.9**	**23.4**	**25.0**	**36.4**	**21.3**	**24.6**	**26.0**	**34.2**
5ab269	Cloudy + Day	25.6	30.0	31.6	**6.9**	33.3	25.2	29.5	31.4	40.1	26.9	29.0	31.7	41.2
70d2ae	Cloudy + Day	26.3	31.4	37.9	9.4	41.0	25.2	31.0	37.5	44.7	27.7	34.8	34.3	44.9
337375	Cloudy + Day	24.8	24.8	33.4	17.1	35.3	27.2	27.9	34.7	38.0	26.4	37.5	28.8	39.1
7d37fc	Cloudy + Day	32.5	36.4	41.5	15.5	42.1	33.6	37.7	40.8	45.8	35.2	40.1	39.4	45.7
f1008c	Cloudy + Day	38.6	42.0	44.4	11.3	46.2	40.0	40.4	45.3	50.3	39.1	42.4	45.8	54.1
f9fa39	Cloudy + Day	35.7	39.5	41.8	9.9	48.1	33.2	39.8	42.9	50.1	38.8	44.1	44.3	51.4
cd6473	Cloudy + Day	40.0	45.7	44.0	11.3	52.7	36.6	47.3	47.3	54.0	40.2	44.6	47.9	54.7
cb762b	Cloudy + Day	36.4	41.3	44.3	10.8	44.8	36.9	41.4	44.4	57.7	40.9	44.8	43.7	57.6
aeb73d	Cloudy + Day	39.6	44.6	45.2	12.5	46.7	39.2	46.7	45.9	52.3	42.6	46.4	47.5	51.3
cb0cba	Cloudy + Day	48.3	47.5	52.1	13.8	50.9	46.0	47.5	50.4	55.5	47.1	47.7	51.5	59.4
e9a962	Cloudy + Day	45.6	53.8	55.4	15.8	58.8	44.0	52.8	55.6	60.7	45.1	50.2	52.9	56.2
2d12da	Cloudy + Day	50.8	56.5	56.2	11.9	58.8	48.5	54.6	56.6	59.1	53.1	54.8	57.5	63.8
85bc13	Cloudy + Day	**56.2**	**56.8**	**60.1**	**19.5**	**62.1**	**55.3**	**58.2**	**59.2**	**63.5**	**54.9**	**58.3**	**59.6**	**67.3**
00c561	Sunny + Day	**16.2**	**19.0**	**20.5**	**5.1**	**22.2**	**17.6**	**20.1**	**20.2**	**26.4**	**17.9**	**19.3**	**21.5**	**25.2**
c9d6eb	Sunny + Day	22.5	28.9	32.5	07.5	35.3	22.6	28.8	32.9	39.1	24.5	26.0	28.4	38.6
cd5bb9	Sunny + Day	23.3	24.9	25.8	6.2	27.2	23.4	25.2	25.8	30.4	23.4	25.7	26.2	31.5
6db21f	Sunny + Day	24.1	26.4	27.0	6.7	28.9	23.3	27.0	27.0	34.7	25.1	28.0	28.7	37.0
647240	Sunny + Day	27.1	29.3	31.2	07.8	34.1	26.5	30.1	31.5	38.8	26.9	32.0	32.0	38.4
da734d	Sunny + Day	30.2	33.4	37.0	8.8	39.9	29.2	34.4	37.5	42.6	34.2	35.7	38.2	43.1
5f317f	Sunny + Day	31.9	42.3	45.9	8.9	50.1	32.8	42.0	46.1	51.2	40.0	44.6	47.0	54.0
395560	Sunny + Day	49.3	61.2	60.6	11.3	**72.1**	51.7	60.7	58.5	65.4	58.9	63.4	57.8	59.6
b1ca08	Sunny + Day	**60.0**	**62.1**	**68.4**	**22.4**	67.9	**61.7**	**61.4**	**67.7**	**70.6**	**59.6**	**65.0**	**67.7**	**68.6**
033669	Sunny + Night	18.0	23.5	25.7	6.6	27.4	18.5	23.6	25.1	27.6	21.8	22.7	23.8	27.5
Overall	–	29.8	33.7	36.9	34.6	39.4	29.8	34.1	37.1	42.7	31.8	35.5	37.8	43.3

Table 2. Offline Evaluation Results on the Argoverse-HD Validation Dataset: It records the sAP scores across the 0.50 to 0.95 range for each clip. The optimal and worst results are highlighted in **green** and **red** font under the same weather conditions. The notation "l high" is used as an abbreviation for the resolution 1200×1920, providing a concise representation of the data.

shown in Fig. 4. The analysis revealed that the mean object area ratio is below 1e-2, indicating a substantial presence of small objects in the dataset. For simplicity in subsequent discussions, we define objects with an area ratio less than 1% as 'small objects'.

Tab. 4 presents our findings on the proportion of small objects within the Argoverse-HD validation set. Despite some variability in the overall number of objects and small objects, the proportion of small objects remains relatively stable, as reflected in the variance of their proportion. This stability suggests that small objects are a consistent and prominent feature across various video segments, representing a persistent challenge of streaming perception.

A.4. Impact of Environmental Speed

In Sec. A.3, we highlighted how motion within the observer's viewpoint can affect the perception of small objects. This observation leads us to consider that the speed of the environment could interact with the proportion of small objects.

To investigate the relationship between the environmental speed and the performance variability of streaming perception models, we categorized the validation dataset into three distinct environmental states: *stop*, *straight*, and *turning*. We then manually divided the dataset based on these states. In this reorganized dataset, the clips with an ID's first digit as 0 exclusively represent the *stop* state, while the digits 1 and 2 correspond to *straight* and *turning* states, respectively.

Fig. 5 showcases the performance of StreamYOLO, LongShortNet, and DAMO-StreamNet across each of these segments. Additionally, the mean performance under each motion state is calculated and presented. The data reveals a consistent pattern across all three models: the performance ranking in different environmental motion states follows the order of *stop* being better than *straight*, which in turn is better than *turning*. This trend indicates an association between the state of environmental motion and fluctuations.

Consequently, based on this analysis, we infer that the speed of the environment, particularly when considering the substantial proportion of small objects and their sensitivity to environmental dynamics, emerges as the most influential environmental factor in the context of streaming perception.

Clip ID	Mean Obj ↑	sYOLO	LSN	DAMO
1d6767	35.30	20.9	20.9	21.3
7d37fc	30.89	32.5	33.6	35.2
da734d	25.16	30.2	29.2	34.2
cd6473	23.75	40.0	36.6	40.2
5ab269	23.37	25.6	25.2	26.9
cb762b	23.31	36.4	36.9	40.9
f1008c	23.08	38.6	40.0	39.1
e9a962	21.58	45.6	44.0	45.1
70d2ae	21.38	26.3	25.2	27.7
2d12da	19.33	50.8	48.5	53.1
337375	18.19	24.8	27.2	26.4
f9fa39	17.46	35.7	33.2	38.8
aeb73d	16.82	39.6	39.2	42.6
6db21f	16.30	24.1	23.3	25.1
647240	14.18	27.1	26.5	26.9
b1ca08	14.08	60.0	61.7	59.6
85bc13	12.06	56.2	55.3	54.9
033669	11.89	18.0	18.5	21.8
00c561	10.06	16.2	17.6	17.9
cb0cba	10.04	48.3	46.0	47.1
395560	10.00	49.3	51.7	58.9
cd5bb9	8.95	23.3	23.4	23.4
c9d6eb	7.88	22.5	22.6	24.5
5f317f	6.92	31.9	32.8	40.0
Coefficient	–	0.052	0.035	-0.020

Table 3. Table 3 shows the analysis of the average number of objects per frame for each segment in the Argoverse-HD validation set, along with the Spearman correlation coefficients. These coefficients determine the relationship between the quantity of objects and the performance of streaming perception models. The coefficients range from 1e-1 to 1e-2, indicating a weak correlation. This data suggests that the total number of objects in the environment does not significantly affect the performance of streaming perception models, indicating that object quantity is not a primary factor that affects the efficacy of streaming perception tasks.

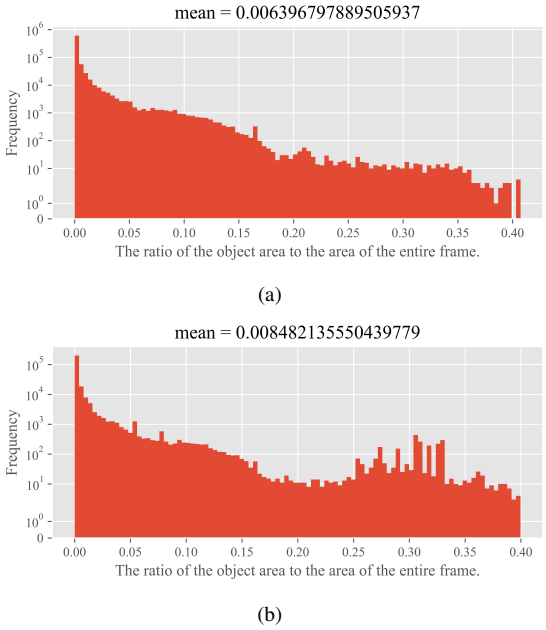


Figure 4. Histograms of Object Area Proportions in Argoverse-HD Dataset: This figure showcases two histograms depicting the proportion of area occupied by objects relative to the entire frame, for (a) the training set and (b) the validation set of the Argoverse-HD dataset. These histograms provide insights into the spatial distribution and size variation of objects within the frames of the dataset.

B. More Experiment Results

B.1. Inference Time Analysis

This subsection supplements Section 4.4 of the main paper, where we previously discussed the performance of *DyRoNet* but did not extensively delve into its inference time characteristics. To address this, Tab. 5 presents a detailed comparison of the inference times for each independent branch used in our model. It is important to note that the inference times reported here may show variations when compared to those published by the original authors of the models. This discrepancy is primarily due to differences in the hardware platforms used and the specific configurations of the corresponding models in our experiments.

An interesting observation from the results is that there are instances where *DyRoNet* exhibits a slower inference time compared to either the *random* selection method or *branch 1*. This slowdown is attributed to the incorporation of the speed router in our sample routing mechanism. Despite this, it is evident from the overall results that *DyRoNet*, employing the router strategy, still retains real-time processing capabilities across the various branches in the model bank. Moreover, in certain scenarios, *DyRoNet* demonstrates even faster inference speeds than when using individual branches independently. This detailed analysis underlines the dynamic and adaptive nature of *DyRoNet* in balancing between inference speed and accuracy, highlighting its capability to optimize streaming perception tasks in real-time scenarios.

B.2. Statistic of model selection

We also provide statistics on *DyRoNet*'s selection of different models during both training and inference time in Tab.6. From the results, it can be observed that DAMO-StreamNet (M+L) exhibits a bias to select the second model during inference time, leading to a similar performance as DAMO-StreamNet L. However, under normal circumstances, *DyRoNet* can still dynamically choose the appropriate model based on input conditions.

B.3. The comparison between Speed Router and $\mathbb{E}[\Delta I_t]$

We also consider a special case id Tab.7, where the model selection only base on the mean of ΔI_t without using

sid	# obj ↑	# small obj	proportion
12	27829	24033	86%
3	16557	15937	96%
14	15058	14260	95%
15	12685	10229	81%
9	12618	11216	89%
5	12189	9509	78%
21	11801	10259	87%
18	11073	9856	89%
20	11068	10203	92%
7	10962	9707	89%
23	10961	9839	90%
2	10717	9700	91%
10	10706	9001	84%
22	10122	8846	87%
11	9965	8976	90%
4	9180	7989	87%
1	9068	8153	90%
24	8293	7830	94%
19	8068	6552	81%
17	4709	4230	90%
6	4420	3708	84%
16	7001	6508	93%
13	5654	5251	93%
8	3237	2449	76%
mean	10580	9343	87.96%
var	–	–	0.0026

Table 4. Distribution of Small Objects in the Argoverse-HD Validation Set: This figure illustrates the count of objects in each video segment of the Argoverse-HD validation set, specifically focusing on objects with an area proportion less than 1%. The chart provides a detailed view of the prevalence and distribution of smaller-sized objects across different video segments in the dataset.

Branches	branch 0	branch 1	random	*DyRoNet*
$\text{DAMO}_{\text{S+M}}$	29.26	33.65	36.61	**33.22**
$\text{DAMO}_{\text{S+L}}$	29.26	36.63	**35.12**	39.60
$\text{DAMO}_{\text{M+L}}$	33.65	36.63	**37.30**	37.61
$\text{LSN}_{\text{S+M}}$	22.08	25.88	24.79	**21.47**
$\text{LSN}_{\text{S+L}}$	22.08	31.24	**21.49**	30.48
$\text{LSN}_{\text{M+L}}$	25.88	31.24	**24.75**	29.05
$\text{sYOLO}_{\text{S+M}}$	18.76	23.01	39.16	**26.25**
$\text{sYOLO}_{\text{S+L}}$	18.76	27.85	**24.04**	29.35
$\text{sYOLO}_{\text{M+L}}$	23.01	27.85	24.69	**23.51**

Table 5. In-Depth Analysis of *DyRoNet*'s Inference Time: This table presents a detailed comparison of inference times between the *random* selection method and *DyRoNet*. For ease of analysis, the optimal values in each comparison are highlighted in **green** font. This highlighting assists in quickly identifying which method—*random* or *DyRoNet* —achieves superior performance in terms of inference speed under various conditions.

Speed Router, which is denoted as $\mathbb{E}[\Delta I_t]$. To be specific, the larger model is selected when $\mathbb{E}[\Delta I_t] > 0$ and minor model is selected otherwise. Unlike Tab.5 in the main text, both methods here are trained for 5 epoch using LoRA finetuning. From the results in Tab.7, it can be seen that our proposed Speed Router has significant advantages compared to directly using $\mathbb{E}[\Delta I_t]$ to select branches.

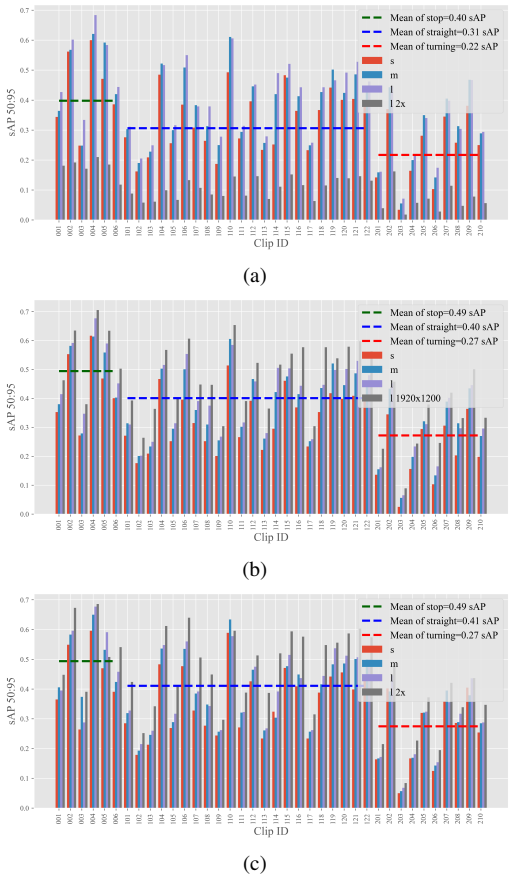


Figure 5. Performance Analysis by Environmental Speed in Validation Segments: This figure displays the performance outcomes of three different models—(a) StreamYOLO, (b) LongShortNet, and (c) DAMO-StreamNet—across various segments of the Argoverse-HD validation set, categorized by environmental speed. The charts provide a comparative view of how each model responds to different speeds in the environment, highlighting their effectiveness in varying dynamic conditions.

Model Combination	training time		inference time	
	Model 1	Model 2	Model 1	Model 2
SYOLO S+M	14.24%	85.76%	5.95%	94.05%
SYOLO S+L	10.98%	89.02%	4.83%	95.17%
SYOLO M+L	37.53%	62.47%	94.67%	5.33%
LSN S+M	13.05%	86.95%	81.65%	18.35%
LSN S+L	7.28%	92.72%	17.26%	82.74%
LSN M+L	30.86%	69.14%	19.87%	80.13%
DAMO S+M	6.26%	93.74%	0.00%	100.00%
DAMO S+L	35.29%	64.71%	3.69%	96.31%
DAMO M+L	84.61%	15.39%	0.02%	99.98%

Table 6. The statistics of model selection by *DyRoNet* under different model choices during both training and inference time.

Model Bank	$\mathbb{E}[\Delta I_t]$ (sAP)	Speed Router (sAP)
StreamYOLO$_{S+M}$	31.5	32.6 (+1.1)
StreamYOLO$_{S+L}$	32.9	35.0 (+2.1)
StreamYOLO$_{M+L}$	34.2	34.6 (+0.4)

Table 7. Comparison of Speed Router and $\mathbb{E}[\Delta I_t]$. Where $\mathbb{E}[\Delta I_t]$ means directly select model by the sign of $\mathbb{E}[\Delta I_t]$ without using Speed Router.

	$=0$	>0	<0
train	0.24%	48.22%	51.55%
test	0.30%	49.85%	49.85%
val	0.17%	49.18%	50.66%

Table 8. Statistics of the sign of $\mathbb{E}(\Delta I_t)$ over Argoverse-HD.

Furthermore, in Tab.8, we also conducted the statistic the sign of $\mathbb{E}[\Delta I_t]$ on the Argoverse-HD. Results with absolute values less than 1e-6 were considered equal to 0. The results reveal that evenly distributing training across models did not effectively adapt them to varying speeds as our Speed Router did.

C. More Details of *DyRoNet*

Model	Scale	# of params
StreamYOLO	S	9,137,319
	M	25,717,863
	L	54,914,343
LongShortNet	S	9,282,103
	M	25,847,783
	L	55,376,515
DAMO-StreamNet	S	18,656,357
	M	50,129,333
	L	94,156,945

Table 9. Parameter Count of Selected Pre-trained Models: This table lists the number of parameters for each pre-trained model chosen for our analysis. It provides a quantitative overview of the complexity and size of the models, facilitating a comparison of their computational requirements.

C.1. Pre-trained Model Selection

As outlined in the main paper, our implementation of *DyRoNet* incorporates three existing models as branches within the Model Bank $\mathcal{P}$: StreamYOLO [8], LongShortNet [4], and DAMO-StreamNet [3]. These models were selected due to their specialized features and proven effectiveness in streaming perception tasks. StreamYOLO is unique for its two additional pre-trained weight variants, each tailored for different streaming processing speeds. This feature allows for adaptable performance depending on the speed requirements of the streaming task. In contrast, LongShortNet and DAMO-StreamNet are equipped with pre-trained weights optimized for high-resolution image processing, making them suitable for scenarios where image clarity is paramount.

To ensure a diverse and versatile range of options within the Model Bank, our implementation of *DyRoNet* selectively utilizes the Small (S), Medium (M), and Large (L) variants of the pre-trained weights from each model. This choice enables a balanced mix of processing speeds and resolution handling capabilities, catering to a wide range of streaming perception scenarios. The specific details regarding the number of parameters for these pre-trained models can be found in Tab.9, which provides a comparative overview to help in understanding the computational complexity for different tasks.

C.2. Setting of Hyperparameters

For all our experiments, we maintained consistent training hyperparameters to ensure comparability and reproducibility of results. The experiments were executed on four RTX 3090 GPUs. Considering the need for selecting the optimal branch model for each sample during the routing process, we established a batch size of 4, effectively allocating one sample to each GPU for parallel computation.

In alignment with the configuration used in StreamYOLO, we employed Stochastic Gradient Descent (SGD) as our optimization technique. The learning rate was set to $0.001 \times \text{BatchSize}/64$, adapting to the batch size proportionally. Additionally, we incorporated a cosine annealing schedule for the learning rate, integrated with a warm-up phase lasting one epoch to stabilize the initial training process.

Regarding data preprocessing, we ensured uniformity by resizing all input frames to 600×960 pixels. This standardization was crucial for maintaining consistency across different datasets and ensuring that our model could generalize well across various input dimensions.

D. Details of experiment on NuScenes-H dataset

To meet the requirements of streaming perception tasks, nuScenes-H [7] enhances the commonly used autonomous driving perception dataset nuScenes [1] by increasing the annotation frequency from 2Hz to 12Hz. While nuScenes encompasses data from three modalities—Camera, LiDAR, and Radar—nuScenes-H provides dense 3D object annotations exclusively for the 6 sensors of Camera modality.

As mentioned in the main text, we trained and evaluated *DyRoNet* on the nuScenes-H dataset. To accommodate the requirements for 2D object detection, the 3D object annotations in nuScenes-H are converted to 2D using publicly available conversion scripts. All experiments were conducted exclusively using the `CAM_FRONT` viewpoint. The

	train set	test set
# of video clips	120	30
# of frames	26705	6697
# of anno	225346	71819
adult	32200	13920
child	22	142
wheelchair	0	0
stroller	0	174
personal_mobility	0	2
police_officer	0	0
construction_worker	1573	362
animal	22	0
car	100487	25356
motorcycle	4958	330
bicycle	1844	1248
bus.bendy	531	283
bus.rigid	4854	1161
truck	21801	4934
construction	2154	1200
emergency.ambulance	61	0
emergency.police	112	0
trailer	6799	805
barrier	33058	10568
trafficcone	8654	10096
pushable_pullable	5191	649
debris	666	348
bicycle_rack	359	241

Table 10. Dataset partition of nuScenes-H 2D, includes the number of video clips (# of video clip), video frames (# of video clip), and the instance counts for each object category (# of anno) within the subsets.

dataset partition details are summarized in Tab. 10, which includes the number of video clips, video frames, and the instance counts for each object category within the subsets. As it shows in Tab. 10, limited or even absent annotation for some categories resulted in lower overall test performance. For clarity, this dataset is referred to as nuScenes-H 2D as follows.

Before training *DyRoNet*, YOLOX [2] was trained for 80 epochs on nuScenes-H 2D to obtain pretrained weights. These weights were then used to initialize the branch models within *DyRoNet*. During the training of *DyRoNet*, each individual branch was trained for 10 epochs, followed by 5 epochs of training for the router. All other training settings were consistent with those described in the main text. As indicated by the experimental results presented, *DyRoNet* maintains strong selection capabilities across different branches on other datasets, demonstrating its adaptability under practical application conditions.

References

[1] Holger Caesar, Varun Bankiti, Alex H. Lang, Sourabh Vora, Venice Erin Liong, Qiang Xu, Anush Krishnan, Yu Pan, Giancarlo Baldan, and Oscar Beijbom. nuscenes: A multimodal dataset for autonomous driving. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, June 2020. 6

[2] Zheng Ge, Songtao Liu, Feng Wang, Zeming Li, and Jian Sun. Yolox: Exceeding yolo series in 2021. *arXiv preprint arXiv:2107.08430*, 2021. 7

[3] Jun-Yan He, Zhi-Qi Cheng, Chenyang Li, Wangmeng Xiang, Binghui Chen, Bin Luo, Yifeng Geng, and Xuansong Xie. Damo-streamnet: Optimizing streaming perception in autonomous driving. In *Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence, IJCAI-23*, pages 810–818. International Joint Conferences on Artificial Intelligence Organization, 8 2023. 1, 6

[4] Chenyang Li, Zhi-Qi Cheng, Jun-Yan He, Pengyu Li, Bin Luo, Hanyuan Chen, Yifeng Geng, Jin-Peng Lan, and Xuansong Xie. Longshortnet: Exploring temporal and semantic features fusion in streaming perception. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 1–5. IEEE, 2023. 1, 6

[5] Mengtian Li, Yu-Xiong Wang, and Deva Ramanan. Towards streaming perception. In *Proceedings of the European Conference on Computer Vision*, pages 473–488. Springer, 2020. 1

[6] Bharat Mahaur and KK Mishra. Small-object detection based on yolov5 in autonomous driving systems. *Pattern Recognition Letters*, 168:115–122, 2023. 2

[7] Xiaofeng Wang, Zheng Zhu, Yunpeng Zhang, Guan Huang, Yun Ye, Wenbo Xu, Ziwei Chen, and Xingang Wang. Are we ready for vision-centric driving streaming perception? the asap benchmark. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9600–9610, 2023. 6

[8] Jinrong Yang, Songtao Liu, Zeming Li, Xiaoping Li, and Jian Sun. Real-time object detection for streaming perception. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5385–5395, June 2022. 1, 2, 6