CCDiff: Causal Composition Diffusion Model for Closed-loop Traffic Generation (Supplementary Materials)

A. Additional Related Works

Table 4. Key features of related works in scenario generation for autonomous vehicles.

Paper	Controllability	Realism	Closed-loop	Safety-Critical	Compositionality
TrafficSim [26]	✓	✓	✓	✗	✗
BITS [29]	✓	✓	✓	✗	✗
SimNet [27]	✓	✓	✓	✗	✗
STRIVE [28]	✓	✓	✓	✓	✗
CTG [3]	✓	✓	✓	✗	STL
SceneGen [10]	✓	✓	✗	✗	✗
RealGen [5]	✓	✓	✗	✓	✗
CTG++ [4]	✓	✓	✓	✓	LLM
LCTGen [13]	✓	✓	✗	✗	LLM
CausalAF [11]	✓	✓	✗	✓	CG
Ours	✓	✓	✓	✓	CG

B. Additional Algorithm Details

Algorithm 2 presents the training of *CCDiff* similar to DDPM and outputs denoising scene encoder $\pi_{\phi,\psi}(\cdot|\boldsymbol{s}, \boldsymbol{c}; G)$. Algorithm 3 presents the causal discovery and ranking algorithm

Algorithm 2 Training of *CCDiff*

Require: Dropout p_{uncond}, threshold C_ρ, C_{ttc}
Require: Guidance loss $\{\mathcal{J}_i\}_{i=1}^N$, trajectories $\boldsymbol{\tau}$, map $\boldsymbol{c}$.
  while $M \leq M_{\text{max}}$ **do**
    $M \leftarrow M + 1$
    $(\boldsymbol{\tau}(0), \boldsymbol{c}) \sim \mathcal{D}$
    $G \leftarrow G(\boldsymbol{\tau}(0))$ with probability $1 - p_{\text{uncond}}$
    $G \leftarrow \boldsymbol{I}_N$ with probability p_{uncond}
    $k \sim \text{Unif}[K]$
    $\boldsymbol{\tau}(k) = \sqrt{\overline{\alpha}_k}\boldsymbol{\tau}(0) + \sqrt{1 - \overline{\alpha}_k}\boldsymbol{\epsilon}$
    Update $\pi_{\phi,\psi}$ with $\nabla_{\phi,\psi} \left\| \pi_{\phi,\psi}(\boldsymbol{\tau}(k), \boldsymbol{c}, k; G) - \boldsymbol{a}(0) \right\|^2$
  end while
  return Denoising scene encoder $\pi_{\phi,\psi}(\cdot|\boldsymbol{s}, \boldsymbol{c}; G)$

Algorithm 3 Causal discovery and Ranking for *CCDiff*

Require: History trajectories $\boldsymbol{\tau}$, TTC Graph M, attention matrix $\boldsymbol{\alpha}$, Top-K agents k
  $G = M \cdot \boldsymbol{\alpha} \triangleq (V, E, w)$
  for all $v_i \in G$ **do**
    $C_i \leftarrow \{v_i\}, w_i = 0$
    for all $v_j \in V \backslash C_i$ **do**
      if $(v_j, v) \in E, \forall v \in C_i$ **then**
        $w_i \leftarrow w_i + \sum_{v \in C_i} w(v_j, v)$
        $C_i \leftarrow C_i \cup \{v_j\}$
      end if
    end for
  end for
  $\boldsymbol{\rho} \leftarrow \text{argsort}(C, w)[:\text{k}]$
  return Importance ranking $\boldsymbol{\rho}$

C. Additional Experiment Details

C.1. Additional Quantitative Results

Table 5. Evaluation of Controllability and Realism across different scales of editable agents (N) and planning horizons (T). For each metric, we report the **best** and **second best** performance among all the methods. CCDiff has the best overall performance presented in the main text.

Methods	Metrics	K=2	3	4	5	10	Full	T=1s	2s	3s	4s	5s
SimNet	SCR (↑)	0.31	0.32	0.33	0.36	0.42	0.47	0.35	0.37	0.37	0.37	0.40
	ORR (↓)	1.76	2.19	2.62	2.67	2.90	3.17	**2.09**	**3.87**	**6.16**	**8.36**	**9.93**
	FDE (↓)	**3.76**	**4.34**	**4.98**	**5.26**	**6.63**	8.03	**4.11**	**3.78**	**4.90**	**4.83**	**3.83**
	CFD (↓)	2.56	2.95	2.86	3.16	5.00	7.00	4.02	5.03	5.51	5.57	8.04
TrafficSim	SCR (↑)	0.38	0.36	0.44	0.41	0.41	0.47	**0.53**	**0.53**	**0.58**	**0.55**	**0.53**
	ORR (↓)	2.09	2.25	2.45	2.48	2.66	2.73	3.56	6.36	8.98	10.96	12.21
	FDE (↓)	4.25	5.06	5.77	6.23	6.79	**7.13**	8.32	6.48	8.61	8.66	7.32
	CFD (↓)	7.76	9.53	10.64	10.99	10.96	11.57	5.00	10.06	7.89	10.38	9.90
STRIVE	SCR (↑)	**0.49**	**0.49**	**0.53**	**0.53**	**0.56**	**0.54**	0.49	0.41	0.41	0.39	0.38
	ORR (↓)	5.70	6.45	7.13	7.50	8.04	8.53	5.75	4.98	6.64	8.40	10.02
	FDE (↓)	9.01	10.79	12.13	13.00	13.76	14.52	11.48	11.20	14.56	15.00	12.41
	CFD (↓)	7.72	8.93	9.91	10.67	10.72	11.21	5.60	10.21	11.59	11.32	9.11
BITS	SCR (↑)	0.38	0.38	0.37	0.39	0.37	0.41	0.37	0.34	0.44	0.39	0.41
	ORR (↓)	**0.53**	**0.51**	**0.56**	**0.63**	**0.56**	**0.60**	**1.44**	**3.68**	**5.63**	**7.56**	**9.39**
	FDE (↓)	**3.20**	**3.95**	**4.42**	**4.67**	**5.05**	**5.35**	**4.68**	**4.69**	**6.10**	**6.36**	**5.44**
	CFD (↓)	7.43	8.32	9.15	9.42	9.46	10.23	8.79	10.35	10.75	11.30	11.65
CTG	SCR (↑)	**0.43**	0.42	**0.46**	0.42	0.44	0.46	**0.41**	**0.44**	**0.49**	0.52	0.49
	ORR (↓)	1.00	1.04	1.10	1.09	**1.12**	**1.23**	1.91	4.58	7.13	9.04	10.71
	FDE (↓)	5.32	6.18	6.83	7.40	8.10	9.19	7.58	7.91	10.26	10.30	8.27
	CFD (↓)	**2.37**	**2.31**	**2.68**	**2.59**	**2.57**	**3.13**	2.68	**4.06**	**2.43**	**2.80**	**3.00**
Ours	SCR (↑)	0.40	**0.44**	0.43	**0.46**	**0.49**	**0.51**	0.40	**0.44**	**0.49**	**0.55**	**0.52**
	ORR (↓)	**0.61**	**0.72**	**0.99**	**1.02**	1.80	2.05	2.92	4.52	7.10	9.35	10.51
	FDE (↓)	4.17	5.22	5.99	6.59	7.84	8.26	7.06	5.54	6.86	7.00	5.71
	CFD (↓)	**1.88**	**1.92**	**1.93**	**2.25**	**2.83**	**3.47**	**2.37**	**4.08**	**4.25**	**4.97**	**6.33**

Table 6. Ablation study on CCDiff's variants. Evaluation of Controllability (CO, OR) and Realism (FDE and CFD) over different agent scales. For each metric, we highlight the **best** and the **second best** results. Causal ranking has the greatest impact to the final performance.

Enc.	Guide	Rank	Metrics	K=2	3	4	5	10	Full	T=1s	2s	3s	4s	5s
	✓	✓	SCR (↑)	**0.43**	**0.44**	**0.43**	**0.42**	**0.42**	**0.48**	**0.41**	**0.48**	**0.46**	**0.50**	0.44
			ORR (↓)	1.10	0.98	**0.91**	**0.91**	**1.39**	**1.43**	**2.45**	**4.54**	**6.85**	**9.46**	**10.38**
			FDE (↓)	**4.00**	5.41	**5.87**	**5.79**	**7.65**	**8.22**	**6.33**	5.96	7.17	**7.01**	5.73
			CFD (↓)	**1.00**	**1.14**	**1.22**	**1.22**	**1.78**	**1.73**	2.47	3.86	**4.11**	**4.77**	**5.41**
✓		✓	SCR (↑)	0.38	0.45	0.40	0.40	0.39	0.40	**0.41**	**0.44**	**0.46**	0.48	**0.48**
			ORR (↓)	**0.81**	**0.76**	1.00	1.06	**1.47**	**1.60**	2.78	4.83	7.39	9.40	10.44
			FDE (↓)	4.33	5.28	6.13	6.82	8.65	9.20	7.03	6.14	8.02	6.99	**5.55**
			CFD (↓)	1.81	**1.60**	**1.84**	**1.94**	2.92	**2.62**	2.87	**3.64**	4.27	5.37	6.46
✓	✓	Dist	SCR (↑)	0.33	0.34	0.36	0.37	0.39	0.36	0.34	0.35	0.41	0.41	0.39
			ORR (↓)	1.38	1.50	1.59	**1.49**	1.56	1.74	3.06	5.21	7.41	10.14	**10.43**
			FDE (↓)	**4.15**	**5.15**	5.79	5.96	8.01	9.69	**6.51**	**5.73**	**6.82**	**7.01**	**5.38**
			CFD (↓)	**1.79**	2.44	2.03	2.34	3.09	3.30	**1.94**	**2.92**	**3.88**	**4.44**	**5.95**
✓	✓	Human	SCR (↑)	0.34	0.35	0.33	0.31	0.33	0.33	0.33	0.34	0.40	0.37	0.40
			ORR (↓)	1.66	1.65	1.73	1.93	1.66	1.75	3.10	5.25	7.44	10.37	10.51
			FDE (↓)	5.80	6.74	7.40	7.84	8.63	8.99	8.12	7.25	8.70	9.16	7.01
			CFD (↓)	2.21	2.51	2.83	3.14	**2.60**	2.96	3.39	5.20	6.17	6.65	8.43
✓	✓	✓	SCR (↑)	**0.40**	**0.44**	**0.43**	**0.46**	**0.49**	**0.51**	0.40	**0.44**	**0.49**	**0.55**	**0.52**
			ORR (↓)	**0.61**	**0.72**	**0.99**	**1.02**	1.80	2.05	2.92	**4.52**	**7.10**	**9.35**	10.51
			FDE (↓)	4.17	**5.22**	**5.99**	6.59	**7.84**	**8.26**	7.06	5.54	**6.86**	**7.00**	5.71
			CFD (↓)	1.88	1.92	1.93	2.25	2.83	3.47	**2.37**	4.08	4.25	4.97	6.33

We also extend our experiments to over-speed scenarios by incorporating an over-speed guidance function. We compare the Scene Overspeed Rate (**SOR**) with CTG in Table 7 (upper). CCDiff demonstrates better realism (ORR, CFD) with comparable controllability (SOR, SCR). This confirms that CCDiff is extensible to diverse safety-critical events under corresponding controllability guidance objectives.

We then analyze gradient conflicts in CTG and CCDiff, focusing on two aspects: (i) negative average cosine similarity among conflicted gradients and (ii) the percentage of agents with gradient conflicts (inner product < 0). Table 7 (lower) shows CCDiff reduces conflicting agents from $\sim$9.1% (CTG) to $\sim$4.8% and lowers negative average cosine similarity, demonstrating its effectiveness in mitigating gradient conflicts.

Table 7. **Upper**: additional controllability experiments with over-speed guidance. **Lower:** Gradient conflict statistics. In both cases, CCDiff outperforms CTG in both metrics by a clear margin.

Metric	CTG	Ours	Metric	CTG	Ours
SOR (↑)	0.68	**0.73**	SCR (↑)	**0.35**	0.33
ORR (↓)	4.23	**0.89**	CFD (↓)	15.81	**9.65**
Neg. grad. cosine similarity (1e-2, ↓)	1.85	**1.29**	The % of agents w/ grad conflict (%, ↓)	9.12	**4.79**

We further illustrate Decision Causal Graph (DCG) computation using attention and time-to-collision (TTC) masks in Figure 5. As is shown in Figure 5(a), Agent 7 tends to change lanes and interact with Agent 5. resulting in non-diagonal elements in the DCG matrix between Agents 5 and 7 in Figure 5(d). This is computed by the TTC mask in Figure 5(b) and attention map in Figure 5(c). We've included more qualitative results in our qualitative examples in the following subsection.

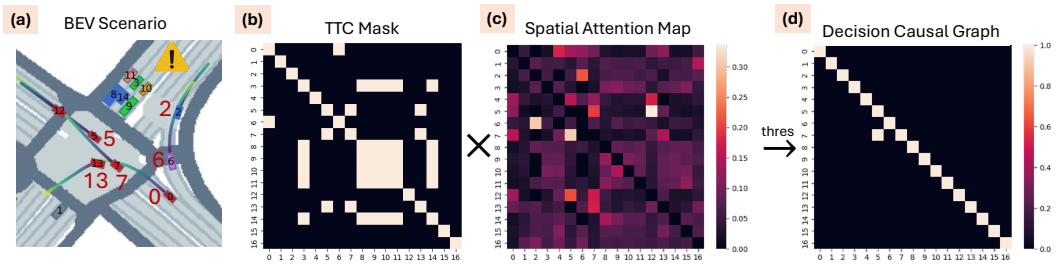

Figure 5. **(a)** Lane-changing at an intersection; **(b, c, d)** Interpretable computation of DCG from TTC mask and attention map.

The CCDiff model has 15.4M parameters, including a CNN-based map encoder and a transformer-based trajectory encoder. Its inference speed is comparable to CTG at $\sim$20 ms per frame per agent on an NVIDIA V100. Figure 6 illustrates full-scene generation time across agent scales.

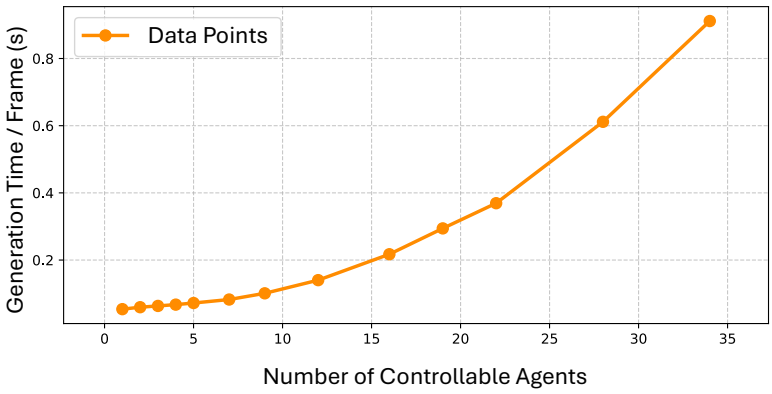

Figure 6. Inference speed with respect to the number of agents.

C.2. Additional Qualitative Results

C.2.1 Long-horizon Generation

We evaluate the long-horizon generation with different planning cycle for the scenarios with same length between *CCDiff* and all the baselines. We illustrate the qualitative examples below. The results demonstrate that *CCDiff* can consistently generate realistic cross-traffic violation scenarios for $1s \leq T \leq 5s$. In contrast, CTG baseline can only generate an opposite-lane collision when $T = 1s$.

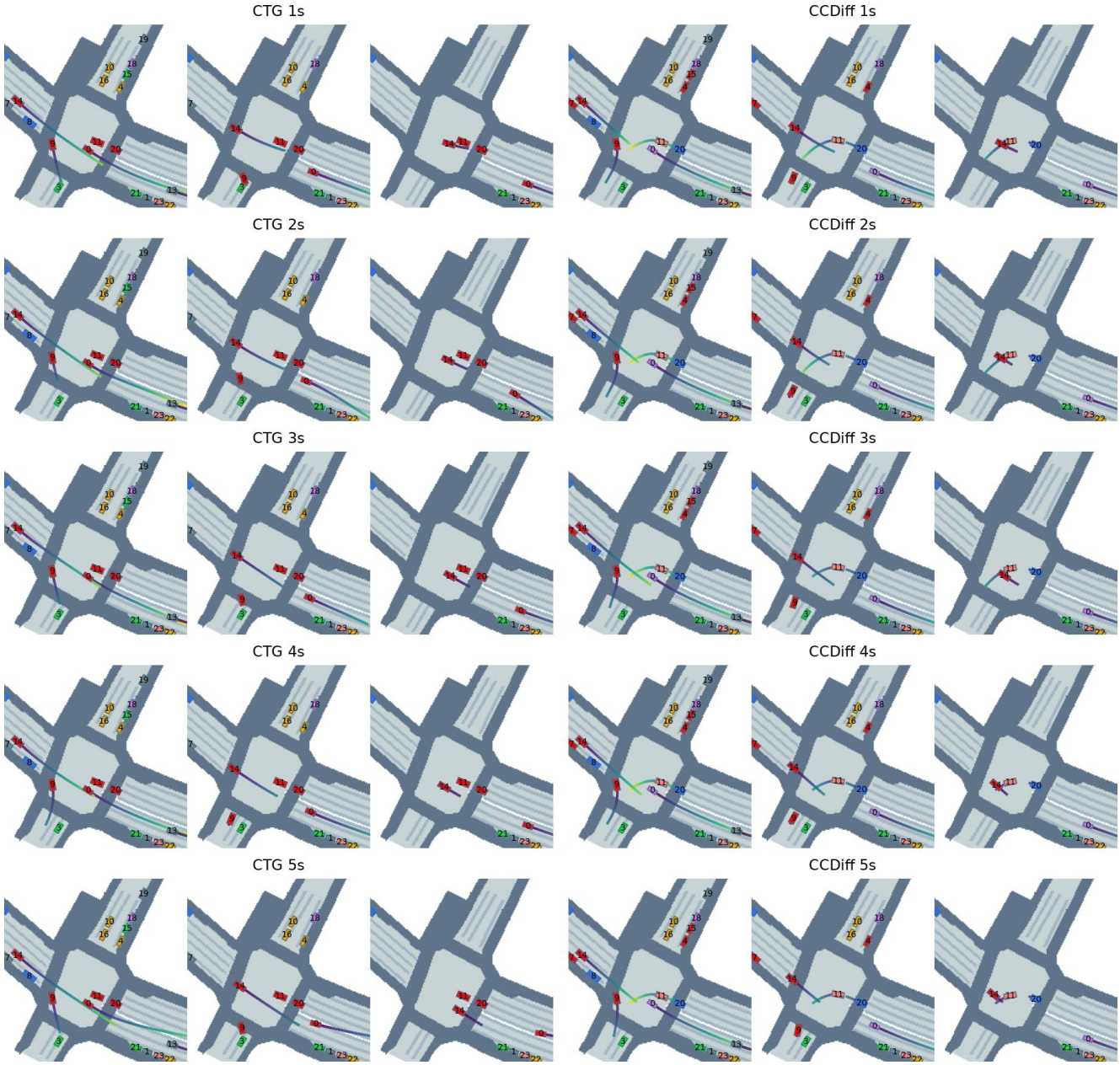

Figure 7. Comparison of *CCDiff* and CTG on the controllability and realism under different sizes of controllable agents. We can see that *CCDiff* can consistently generate realistic cross-traffic violation scenarios, yet CTG can only generate one with shorter planning cycle in 1s.

C.2.2 Multi-agent Generation

We evaluate the multi-agent generation with different sizes of controllable agents K. We illustrate the qualitative examples of unprotected left turn scenarios below. The results demonstrate that with abundant controllable access to the agents at the scene ($K \geq 2$ in this case), *CCDiff* can consistently generate realistic unprotected left-turn scenarios compared to the CTG baseline.

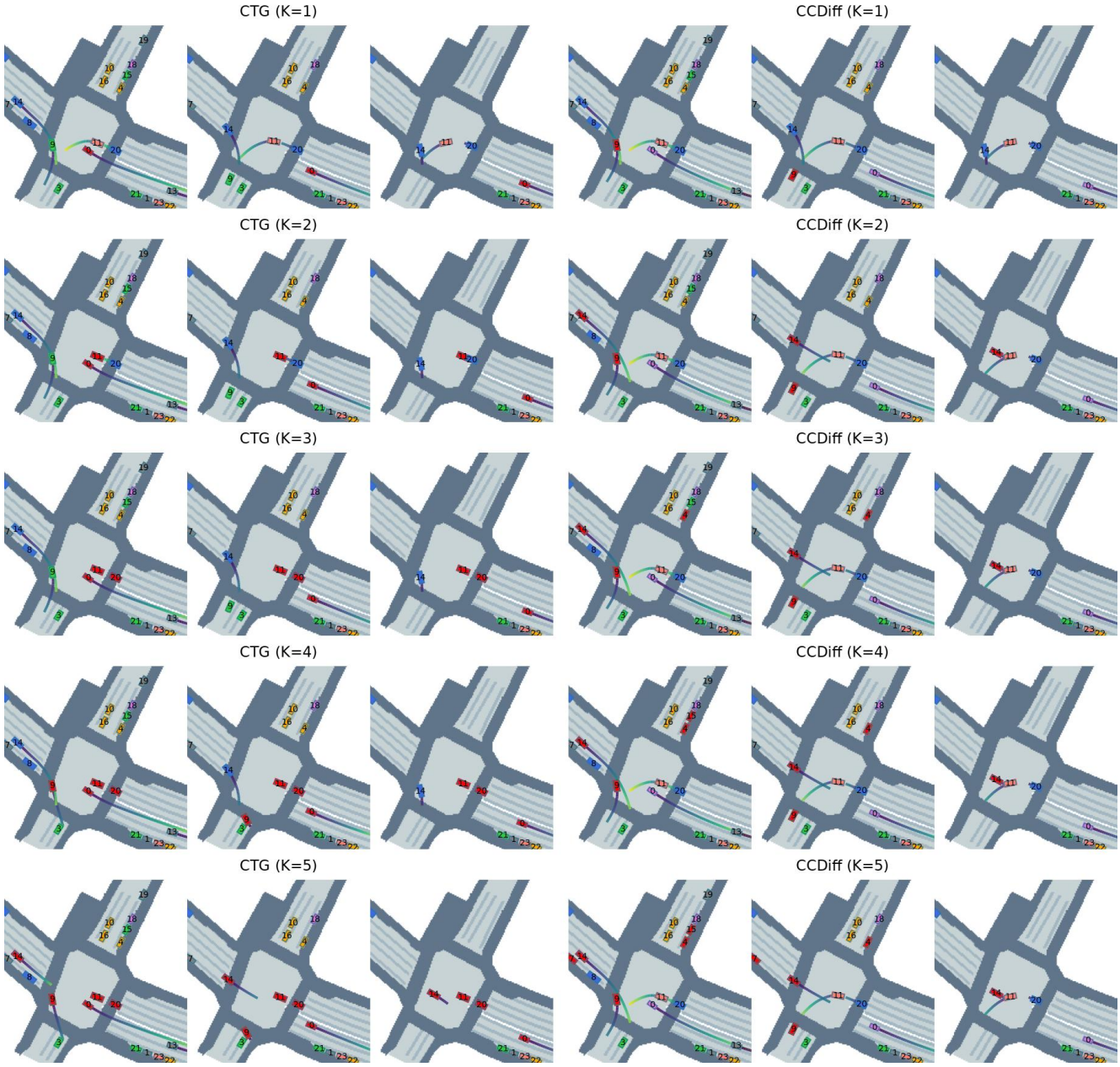

Figure 8. Comparison of *CCDiff* and CTG on the controllability and realism under different sizes of controllable agents. We can see that when the number of controllable agents is greater than 1, *CCDiff* can consistently generate realistic unprotected left-turn violations, yet CTG can only generate one unrealistic right turn collision with 5 controllable agents.

C.3. Detailed description of baselines

SimNet [27]: SimNet frames the problem as a Markov Process, and models state distributions and transitions directly from raw observational data, eliminating the need for handcrafted models. Trained on 1,000 hours of driving logs, it dynamically generates novel and adaptive scenarios that enable closed-loop evaluations. The system reveals subtle issues, such as causal confusion, in state-of-the-art planning models that traditional non-reactive simulations fail to detect.

TrafficSim [26]: TrafficSim uses an implicit latent variable model like conditional variational autoencoder (CVAE). The system parameterizes a joint actor policy that simultaneously generates plans for the agents in a scene. The model is jointly trained with (i) ELBO objective inspired by CVAE and (ii) common-sense following with agents' pair-wise collision loss. TrafficSim generates diverse, realistic traffic scenarios and can serve as effective data augmentation for improving autonomous motion planners.

STRIVE [28]: STRIVE employs a graph-based conditional variational autoencoder (VAE) to model realistic traffic motions and formulates scenario generation as an optimization problem in the latent space of this model. By perturbing real-world traffic data, STRIVE generates scenarios that stress-test planners. A subsequent optimization step ensures that the scenarios are useful for improving planner performance by being solvable and challenging. STRIVE has been successfully applied to attack two planners, showing its ability to produce diverse, accident-prone scenarios and improve planner robustness through hyperparameter tuning.

BITS [29]: BITS (Bi-level Imitation for Traffic Simulation) framework leverages the hierarchical structure of driving behaviors by decoupling the simulation into two levels: high-level intent inference and low-level driving behavior imitation. This structure enhances sample efficiency, behavior diversity, and long-horizon stability. BITS also integrates a planning module to ensure consistency over extended scenarios.

CTG [3]: CTG is a novel framework combining controllability and realism in traffic simulation by leveraging conditional diffusion models and Signal Temporal Logic (STL). The approach allows fine-grained control over trajectory properties, such as speed and goal-reaching, while maintaining realism and physical feasibility through enforced dynamics. Extending to multi-agent settings, the model incorporates interaction-based rules, such as collision avoidance, to simulate realistic agent interactions in traffic.

We list implementation details of all the methods are listed below with important hyperparameters and model structures information in Table 8.

Table 8. Hyper-parameters of models used in experiments of CCDiff and baselines

Parameter Name	Value	Parameter Name	Value
Step length	0.1s	Map Encoder	ResNet-18
History steps	31	Map feature dim.	256
Generation steps	52	Trajectory Encoder	MLP
Learning rate	0.0001	Trajectory feature dim.	128
Optimizer	Adam	Transformer decoder head	16
Batch size	100	Transformer decoder layers	2
Trajectory prediction loss weight	1.0	Guidance gradient Steps	30
Yaw regularization weight	0.1	Guidance constraint norm	100
EMA step	1	Guidance learning rate	0.001
EMA decay	0.995	Guidance optimizer	Adam
Denoising Steps	100	Guidance weight: off-road	1.0
Guidance discount factor	0.99	Guidance weight: collision	-50.0
Planning steps	10, 20, 30, 40, 50	TTC threshold	3.0 s
Controllable Agents	1, 2, 3, 4, 5, 10, Full	Distance threshold	50 m

Training and Inference Resources We conduct training and inference of all the models on 4x NVIDIA Tesla V100 with 16GB GPU memory each, and 48-core CPU Intel(R) Xeon(R) CPU @ 2.30GHz. The training of one model takes 3 hours per epoch on nuScenes training split, and we train 10 epochs for each baseline model and CCDiff. At inference time, the parallel evaluation takes an average of 3 minutes on each closed-loop testing scenario for all the methods under the same configuration (controllable agents and generation frequencies).

C.4. Detailed description of evaluation metrics

- **Controllability Score (CS)**: The computation of CS standardizes the scenario-wise collision rate (SCR) used in [13, 26]:

$$\text{CS} = \frac{\text{SCR} - \min(\text{SCR})}{\max(\text{SCR}) - \min(\text{SCR})}$$

  We then standardize SCR among all the methods to get the **CS**, a higher-the-better score between 0 and 1.
- **Realism Score (RS)**: We average over three widely-used quantitative metrics to evaluate the realism of the scenarios: (i) scenario off-road rate (ORR) used in [5, 13], (ii) final displacement error (FDE, m) and (iii) comfort distance (CFD) in [3, 29] to quantify the realism of the similarity in the smoothness of agents' trajectories in the generated scenarios. We standardize all the metrics among all the methods respectively and average them to get the **RS**, a higher-the-better score between 0 and 1:

$$\text{RS} = 1.0 - \frac{1}{3}\left(\frac{\text{ORR} - \min(\text{ORR})}{\max(\text{ORR}) - \min(\text{ORR})} + \frac{\text{FDE} - \min(\text{FDE})}{\max(\text{FDE}) - \min(\text{FDE})} + \frac{\text{CFD} - \min(\text{CFD})}{\max(\text{CFD}) - \min(\text{CFD})}\right)$$

  Specifically, FDE describes the trajectory closeness between the synthetic one and the original one, ORR describes how frequently the generated trajectories go off-road, while CFD measures the **smoothness** of the generated trajectories with their acceleration and jerk. All these raw metrics are lower the better, so after we revert it above, the resulting RS is a higher-the-better metric.
- **Multi-objective optimization metrics**: with the **RS** and **CS**, we further quantify the optimality of the solution based on generational distance (**GD**) and inverted generational distance (**IGD**), the average minimum distance between the methods and Pareto frontier [44, 53]:

$$\text{GD} = \left(\frac{1}{|\mathcal{D}|}\sum_{\mathbf{d}\in\mathcal{D}}\min_{\mathbf{p}\in\mathcal{P}}\|\mathbf{a}-\mathbf{p}\|^q\right)^{\frac{1}{q}},$$

  where $\|\cdot\|$ denotes the Euclidean distance, and q is typically set to 2. Conversely, IGD measures the average distance from each solution in the Pareto frontier $\mathcal{P}$ to its nearest solution in the obtained set $\mathcal{D}$, and is defined as

$$\text{IGD} = \left(\frac{1}{|\mathcal{P}|}\sum_{\mathbf{p}\in\mathcal{P}}\min_{\mathbf{d}\in\mathcal{D}}\|\mathbf{p}-\mathbf{d}\|^q\right)^{\frac{1}{q}}.$$

  Both metrics provide insights into the convergence and diversity of the obtained solution set: lower values of GD indicate better convergence to the Pareto frontier. On the other hand, lower values of IGD suggest better coverage over the Pareto frontier. We visualize an example for GD and IGD in Figure 9.

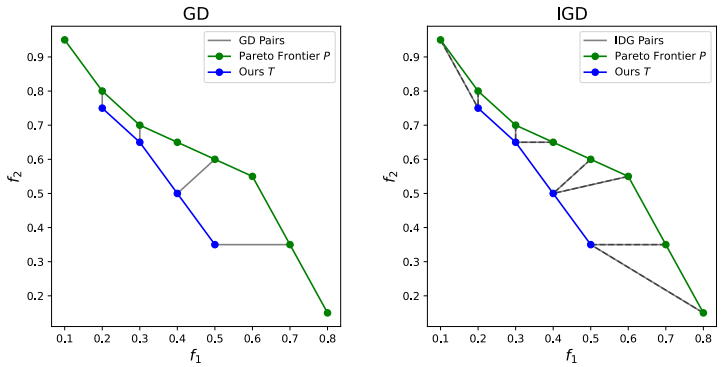

Figure 9. Examples of GD and IGD used to evaluate the multi-objective optimization. Two axes f_1, f_2 represent two objectives.

Quantitative Analysis on the design Causal Masking Design We also analyze the importance of different features w.r.t. the collision samples in the generated scenarios. The results show that TTC feature has the highest statistical correlation with the controllability score (i.e. the collision rate) in our setting.

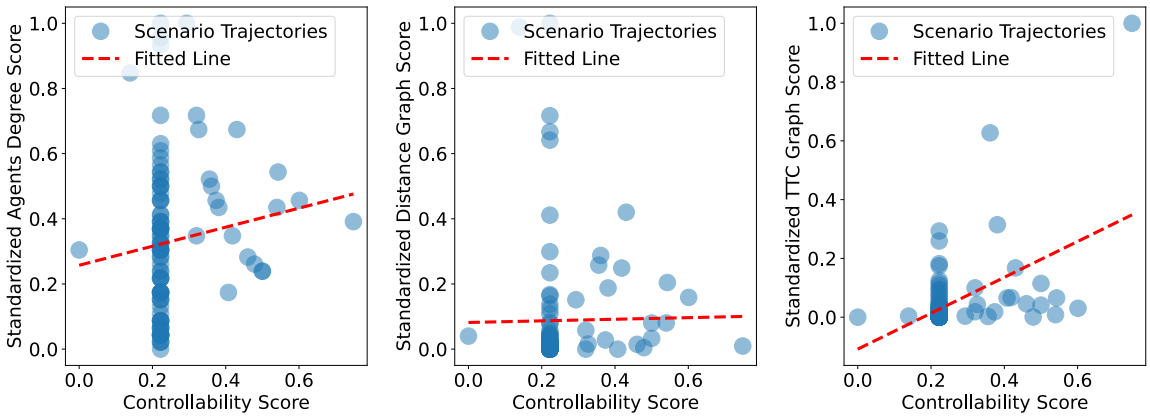

Figure 10. The number of cliques in the TTC graph is more informative causal features of safety-critical incidents (higher Pearson correlation) compared to Relative Distance and number of agents.

Table 9. Correlation analysis between the collision accidents and different causal structure features: standardized clique score for TTC graph, standardized clique score for distance graph, and the standardized number of agents at the scene. We list the Pearson correlation R^2 between the standardized controllability score for each scenario, as well as the significance level of each feature (p-value)

Causal Structure Feature	$R^2(\uparrow)$	p-value ($\downarrow$)
#Cliques in Dist. graph	0.01	0.89
#Agents	0.13	0.20
#Cliques in TTC graph (Ours)	**0.49**	$\mathbf{2.2 \times 10^{-7}}$

C.5. Additional Qualitative Analysis over Scenarios

In the following subsection, we present seven representative interactive scenarios that are safety-critical in urban traffic. We begin by analyzing the comparisons with baseline methods and highlighting the differences between distance-based graphs and TTC-based graphs. The results demonstrate that TTC-based graphs are generally sparser yet more informative, particularly for capturing safety-critical maneuvers.

Additionally, we provide examples of multi-agent, long-horizon trajectory generation for individual scenarios, showcasing the model's ability to handle complex interactions over extended time frames.

C.5.1 Unprotected Left Turn

Baseline Comparison Below, we present the unprotected left-turn scenarios. The relational reasoning of the distance-based graph fails to capture the interaction between the two involved vehicles (11 and 14). We omit the multi-agent and long-horizon generation examples for this scenario, as these have already been analyzed in previous comparisons.

Among all the baselines, CTG, SimNet, and BITS closely follow the ground-truth trajectories, successfully generating a left-lane right turn without producing collision samples. In contrast, STRIVE generates unrealistic collisions with parked vehicles in the side lane. Notably, only CCDiff manages to produce realistic unprotected left-turn behaviors. Only the TTC mask captures the interaction between agents 11 and 14.

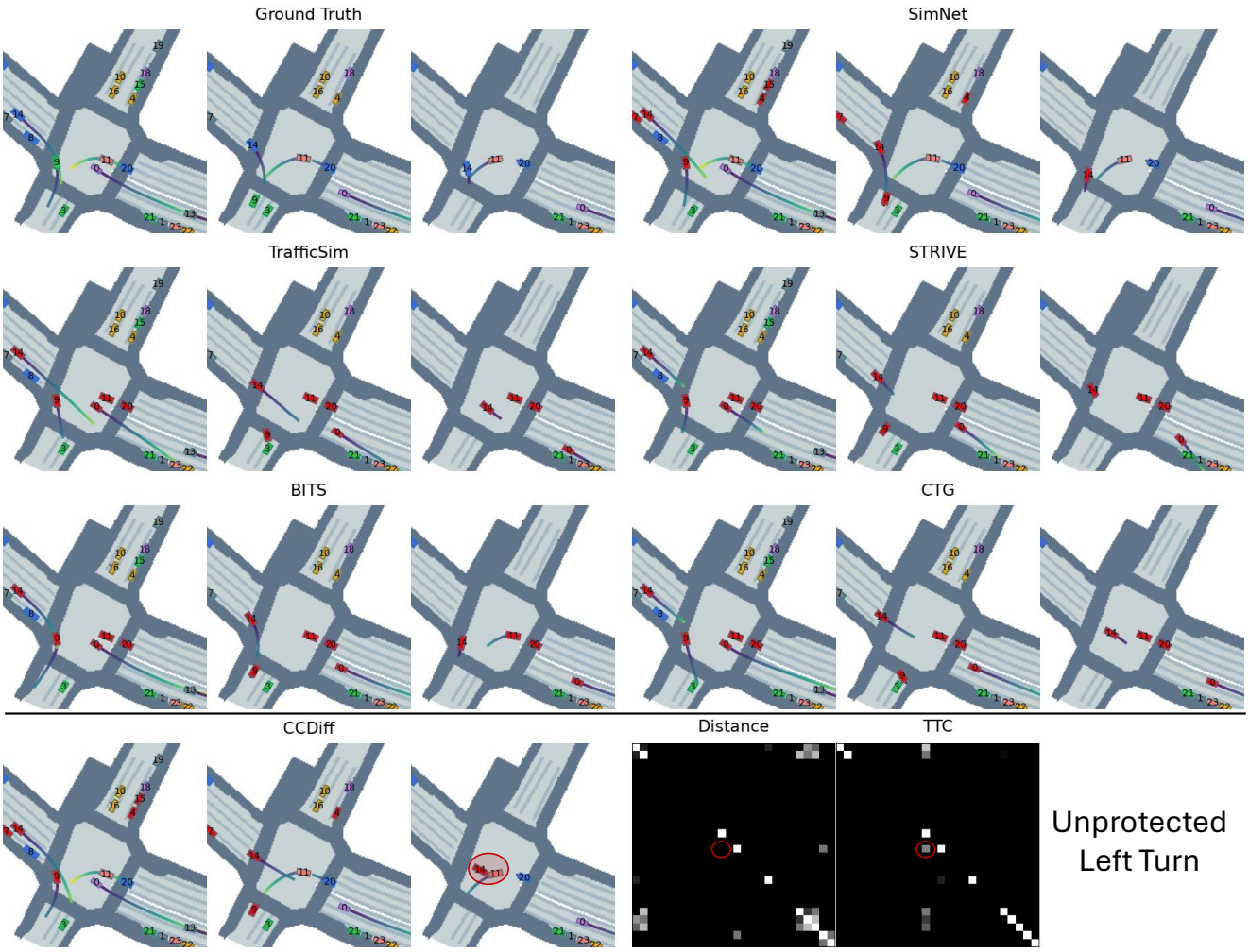

Figure 11. Qualitative of *CCDiff* and baselines in unprotected left turn scenarios.

C.5.2 Cross Traffic Violation

Baseline Comparison A cross-traffic violation occurs when a vehicle at a T-intersection fails to yield the right of way to a vehicle approaching from a perpendicular direction. Such violations often result in side-impact collisions, particularly when the violating driver misjudges the speed or distance of the cross-traffic vehicle. In *CCDiff*, agent 0 collides with agent 6, illustrating this scenario.

Among the baselines, BITS, TrafficSim, and CTG successfully avoid generating collision samples. However, SimNet also generates a collision between agent 0 and agent 6, failing to model the scenario accurately. Both TTC and distance mask manage to capture the interaction between agents 0 and 6.

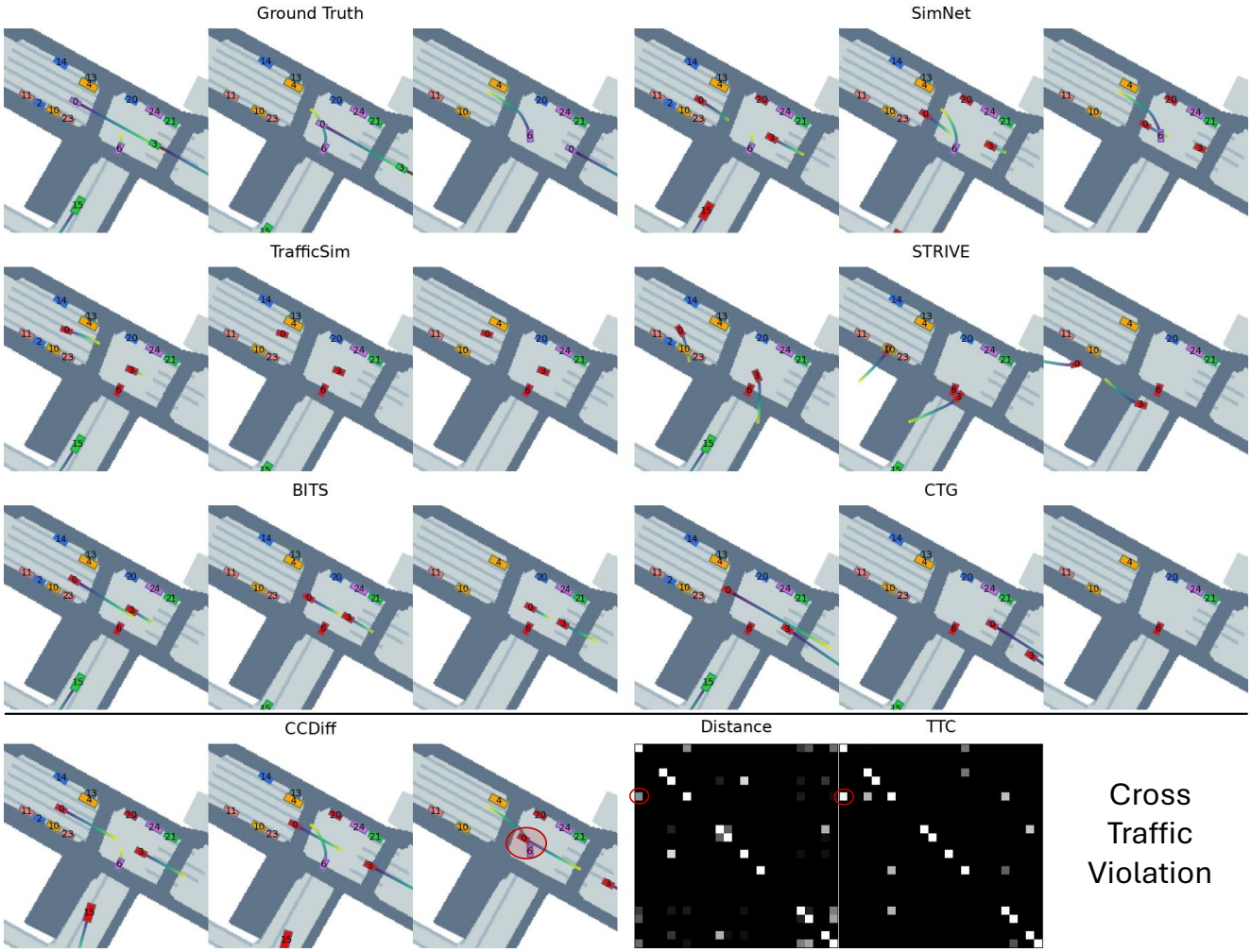

Figure 12. Qualitative of *CCDiff* and baselines in cross traffic violation scenarios.

Multi-agent Generation We compare the multi-agent generation results of *CCDiff* with CTG. *CCDiff* can consistently generate the cross traffic violation when the controllable agents $K \geq 2$.

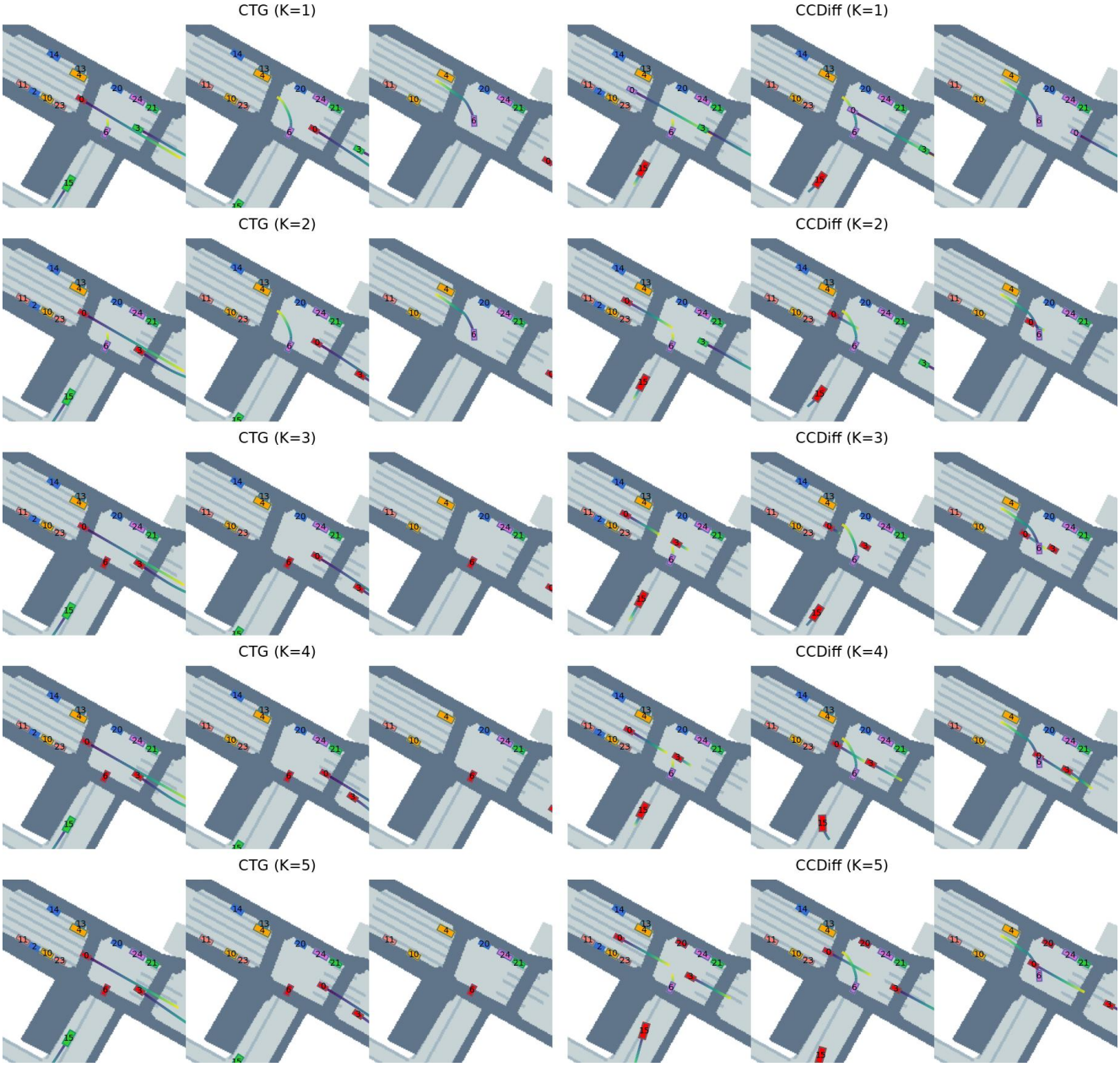

Figure 13. Qualitative comparison of *CCDiff* and CTG under cross traffic violation generation under different sizes of controllable agents.

Long-horizon Generation We compare the long-horizon generation results of *CCDiff* with CTG. *CCDiff* can consistently generate the cross traffic violation even with a generation horizon $T > 2s$, yet CTG generated scenarios are more conservative.

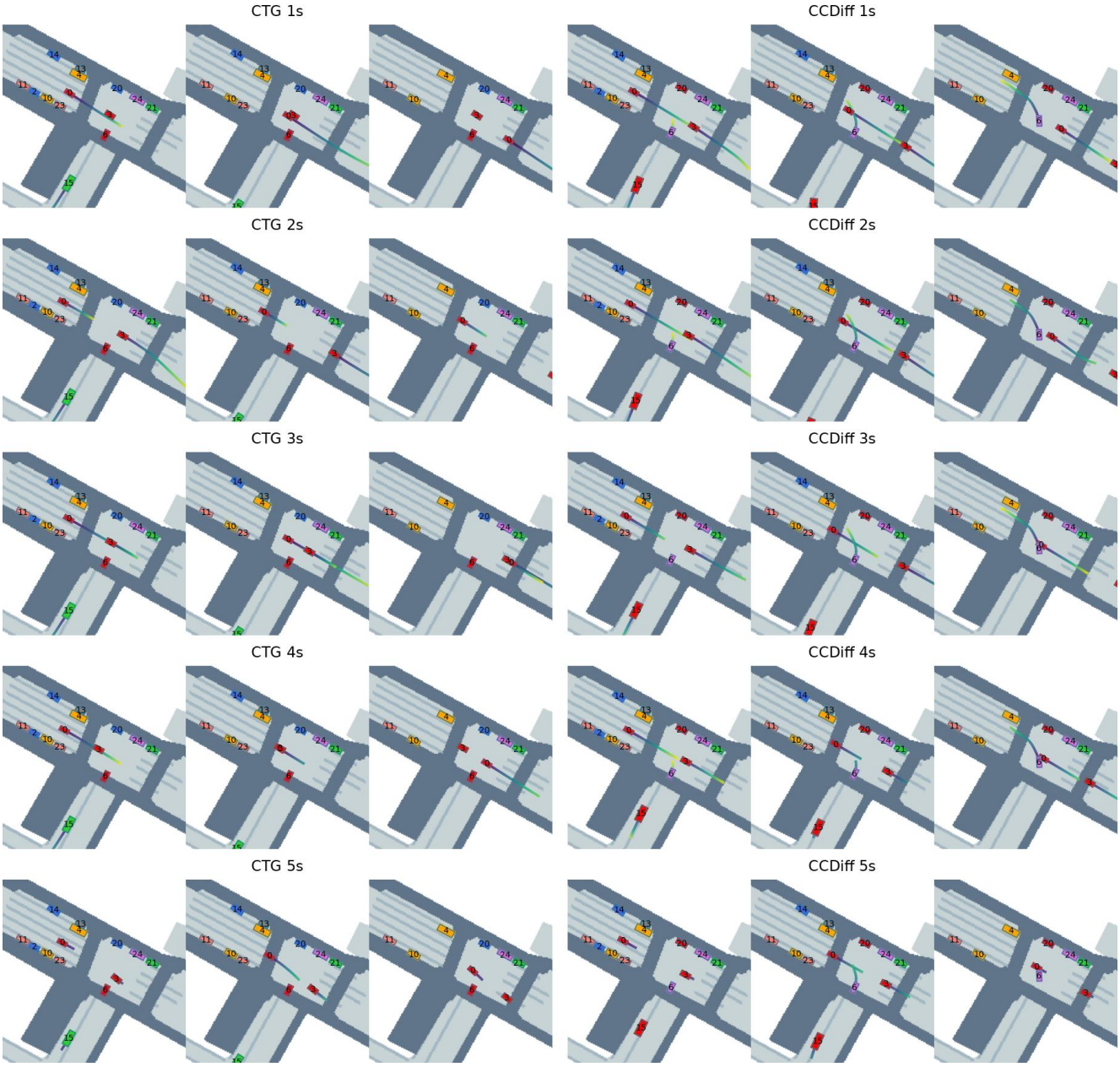

Figure 14. Qualitative comparison of *CCDiff* and CTG under cross traffic violation generation under different generation horizons.

C.5.3 Lane Cut-in

Baseline Comparison A lane cut-in at an intersection occurs when a vehicle abruptly changes lanes or merges into another lane while navigating through or approaching an intersection, often without sufficient clearance or signaling. This maneuver typically forces other vehicles in the affected lane to brake suddenly or adjust their trajectory, increasing the risk of collisions or near-misses. In our case, agent 3 will suddenly cut in from the left lane to the right lane and collide with agent 0.

Among all the baselines, CTG and SimNet generate some irregular behaviors and drive some of the controllable agents off-road. STRIVE generates relatively unrealistic right turn collision, and TrafficSim generates a wild unprotected left turn that is more unrealistic under this context. The TTC mask manages to capture the interaction between agents 0 and 3, while the distance mask misses it.

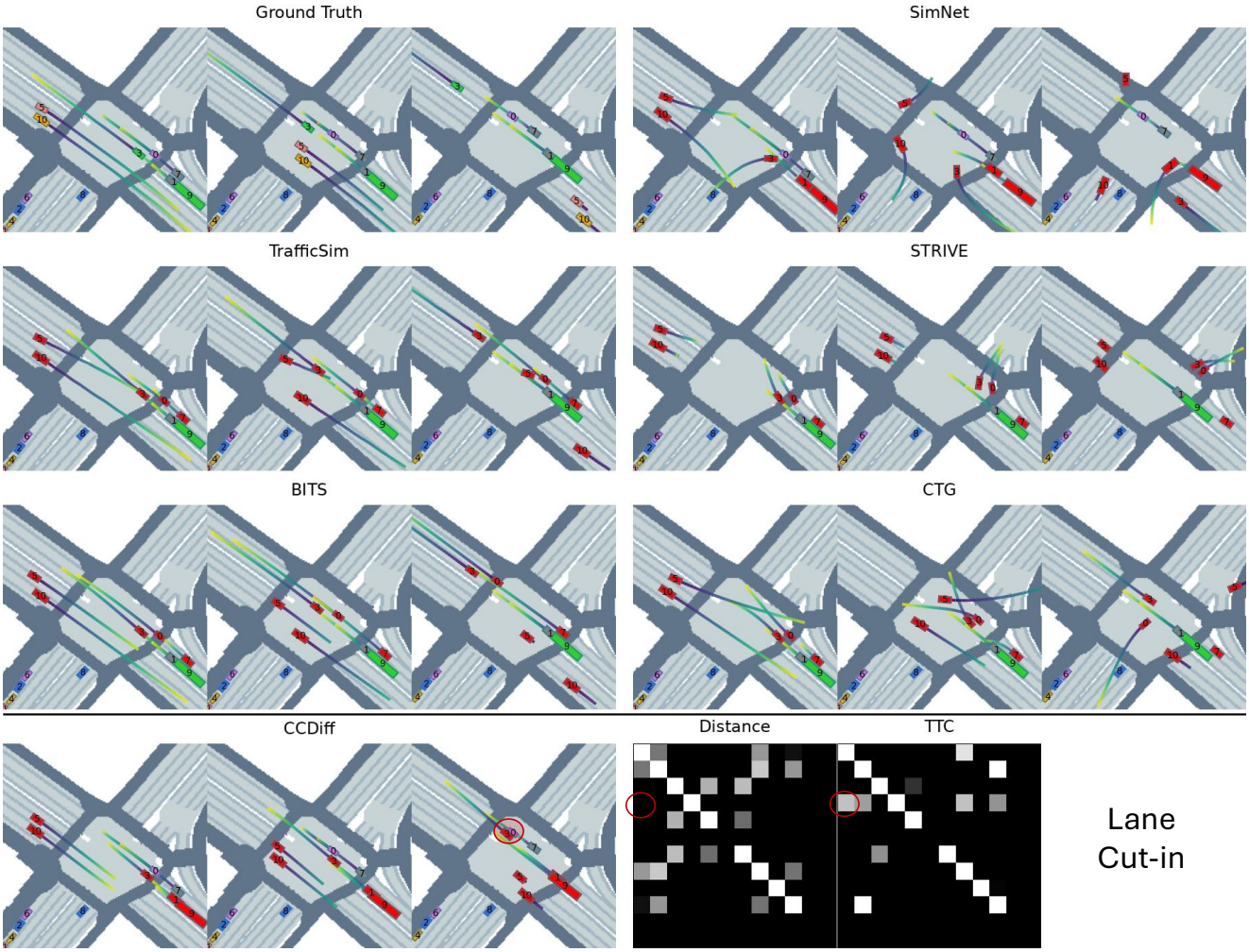

Figure 15. Qualitative of *CCDiff* and baselines in lane cut-in scenarios.

Multi-agent Generation We compare the multi-agent generation results of *CCDiff* with CTG. *CCDiff* can generate collision samples when $K = 5$, yet the CTG generates very wild behaviors that are unrealistic from the ground-truth trajectories.

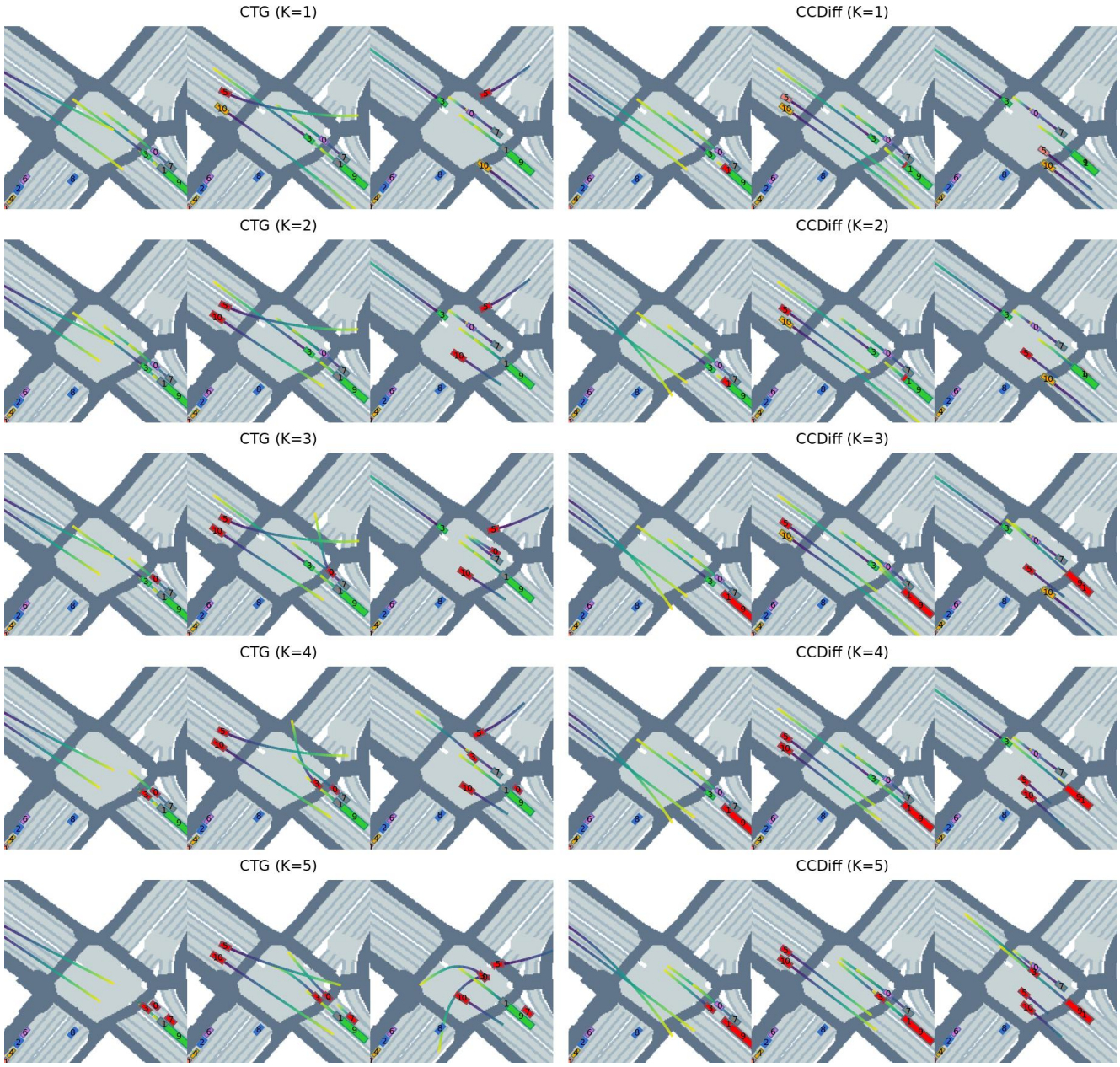

Figure 16. Qualitative comparison of *CCDiff* and CTG under cross traffic violation generation under different sizes of controllable agents.

Long-horizon Generation We compare the long-horizon generation results of *CCDiff* with CTG. *CCDiff* can consistently generate the cut-in violation scenarios with the generation horizon $1s \leq T \leq 4s$. In contrast, CTG attempts to generate some unprotected left turn in this context but fails.

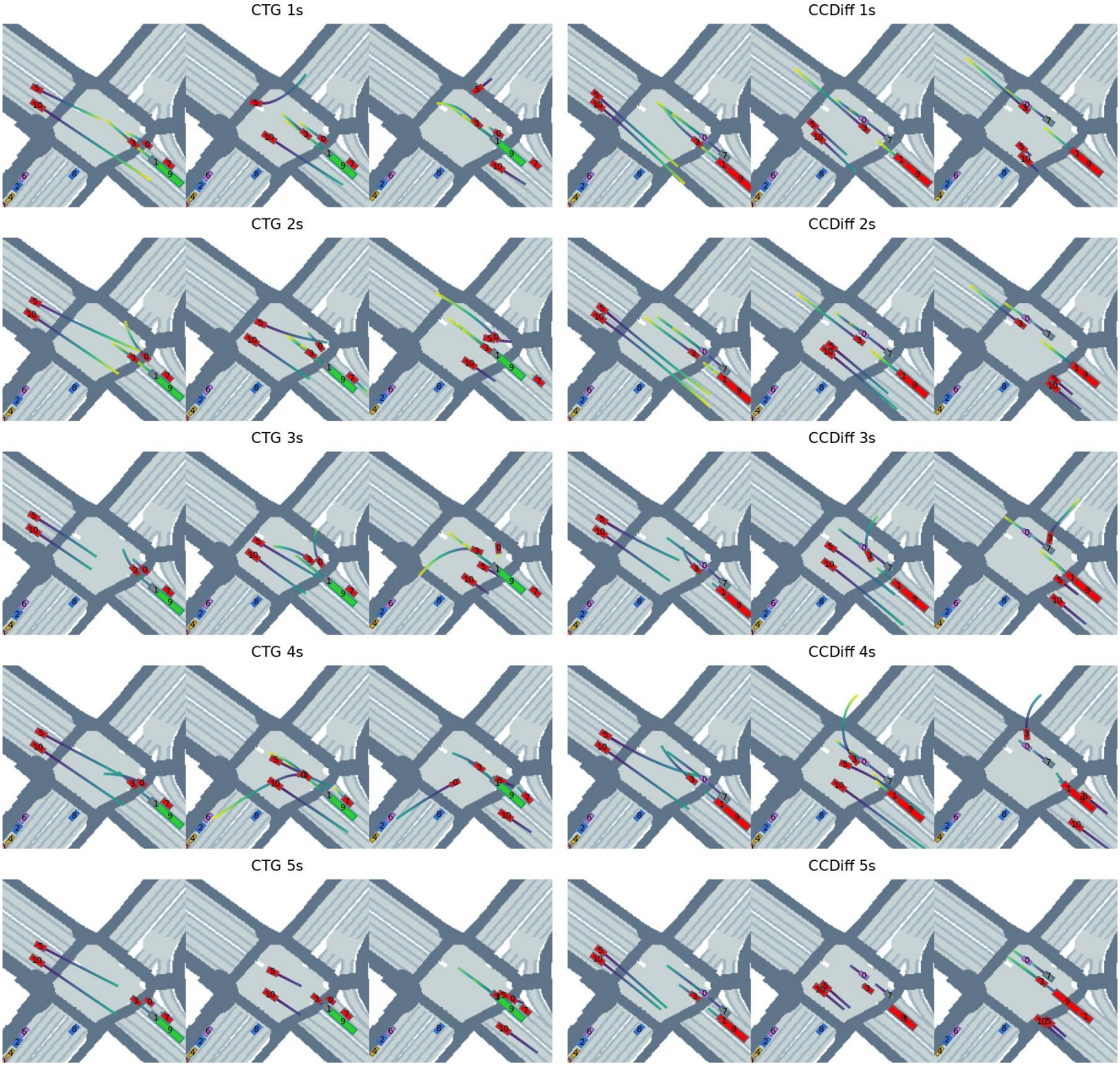

Figure 17. Qualitative comparison of *CCDiff* and CTG under cross traffic violation generation under different generation horizons.

C.5.4 Emergency Break

Baseline Comparison The emergency break occurs when the middle vehicle (agent 0) brakes to keep distance from the forward vehicle (agent 9) suddenly, causing the trailing vehicle (agent 8) to collide with it due to insufficient stopping distance.

Among all the baselines, STRIVE generates some irregular behaviors, which drive some of the controllable agents off-road. TrafficSim, BITS, and SimNet fail to generate safety-critical samples. Notably, although CTG also generates some collision samples, it accelerates the trailing vehicle 8 to collide with the middle vehicle 0, which does not break yet. In comparison, in our case, the middle vehicle 0 breaks and causes a collision with trailing vehicle 8 at normal speed, which is more realistic. Both the TTC mask and distance mask capture the interaction among agents 0, 8, and 9 in this scenario.

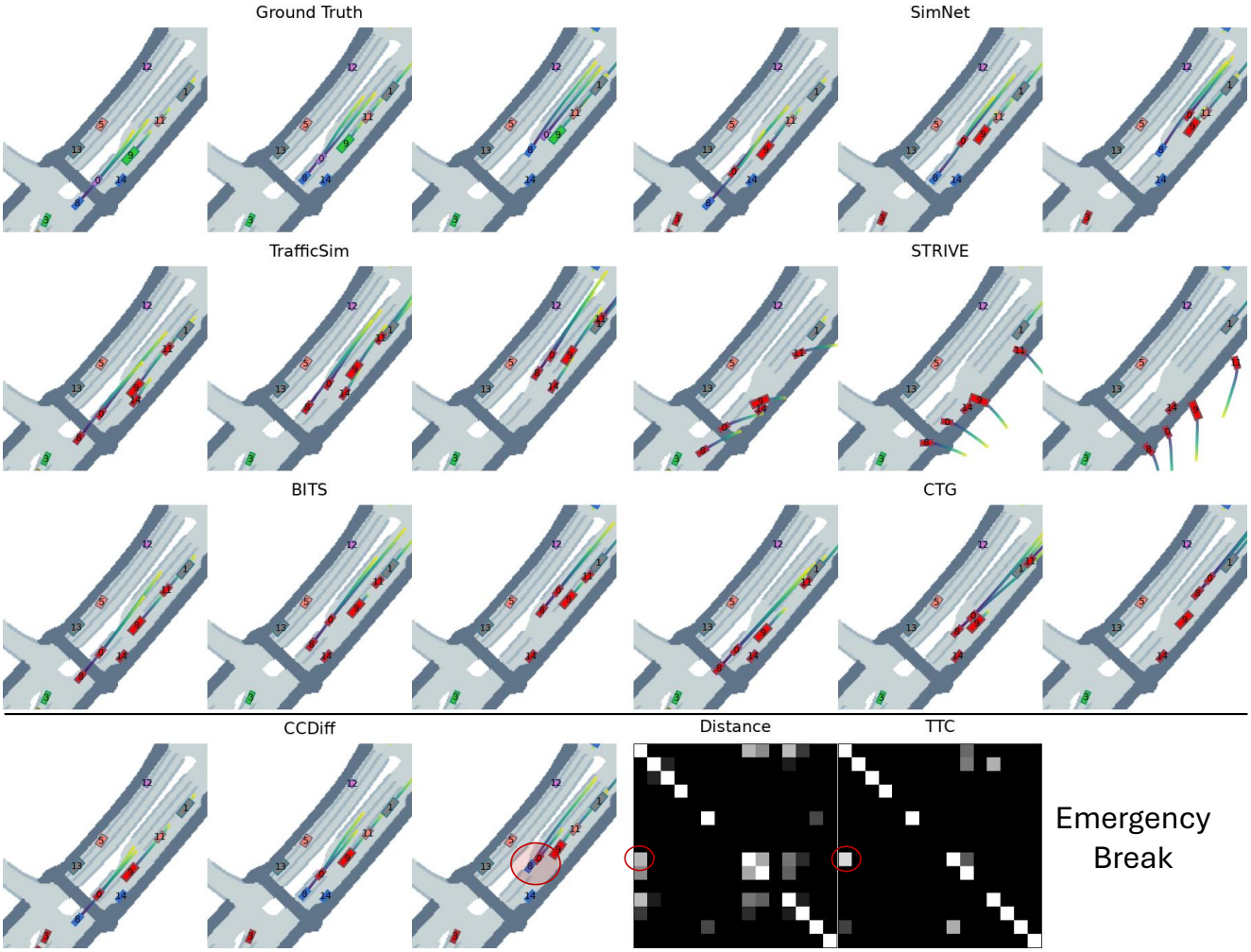

Figure 18. Qualitative of *CCDiff* and baselines in the emergency break scenarios.

Multi-agent Generation We compare the multi-agent generation results of *CCDiff* with CTG. *CCDiff* can consistently generate safety-critical emergency breaking samples when $K \geq 2$, with a control of the most important vehicle 8 in this context. In contrast, CTG keeps accelerating the rear vehicle 8 instead of slowing down the middle vehicle 0.

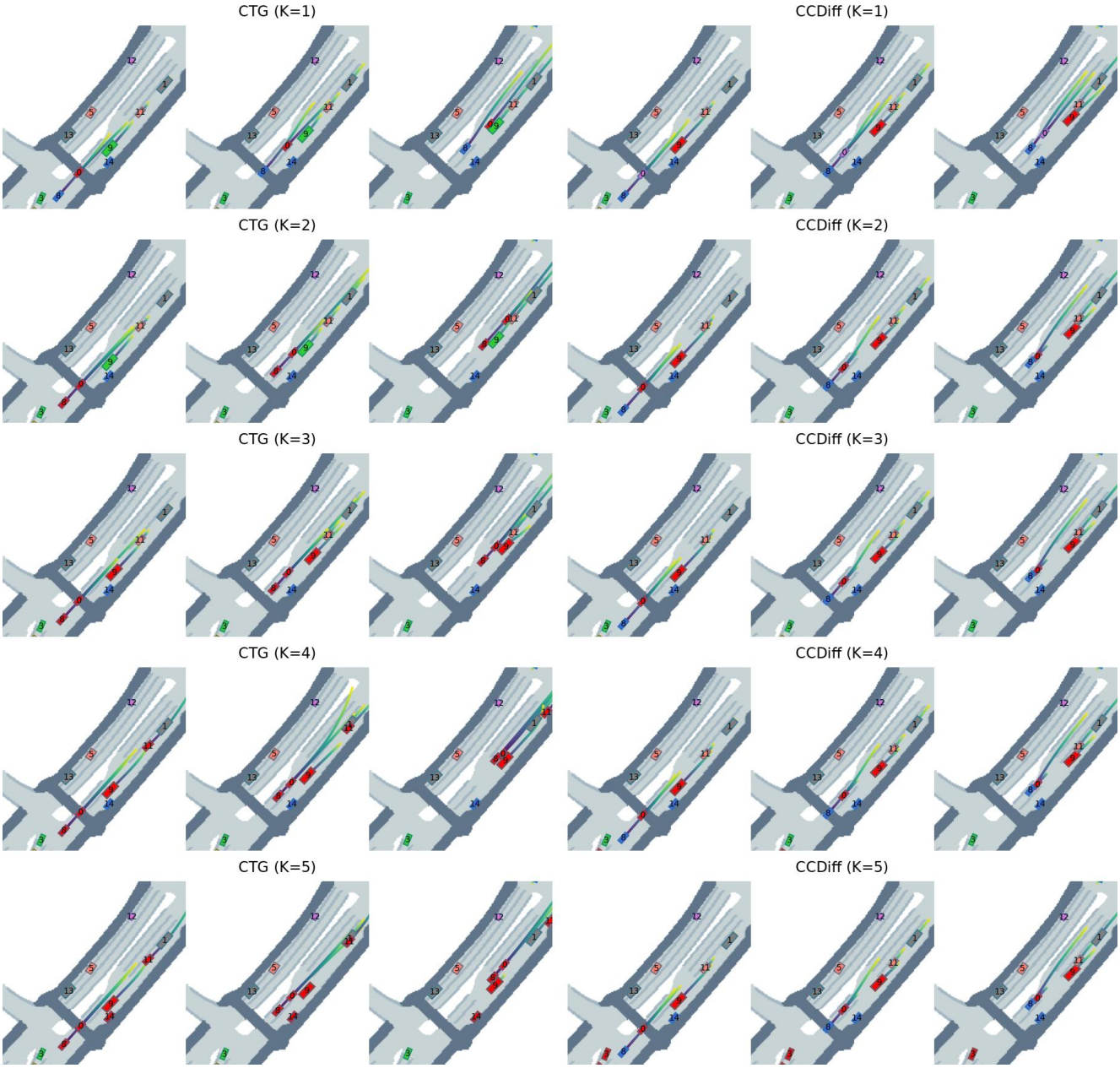

Figure 19. Qualitative comparison of *CCDiff* and CTG under cross traffic violation generation under different sizes of controllable agents.

Long-horizon Generation We compare the long-horizon generation results of *CCDiff* with CTG. *CCDiff* can consistently generate the cut-in violation scenarios with all different lengths of the generation horizon $1s \leq T \leq 5s$. In contrast, CTG attempts to accelerate the vehicle in the middle and cannot generate any near-miss samples with longer generation horizons.

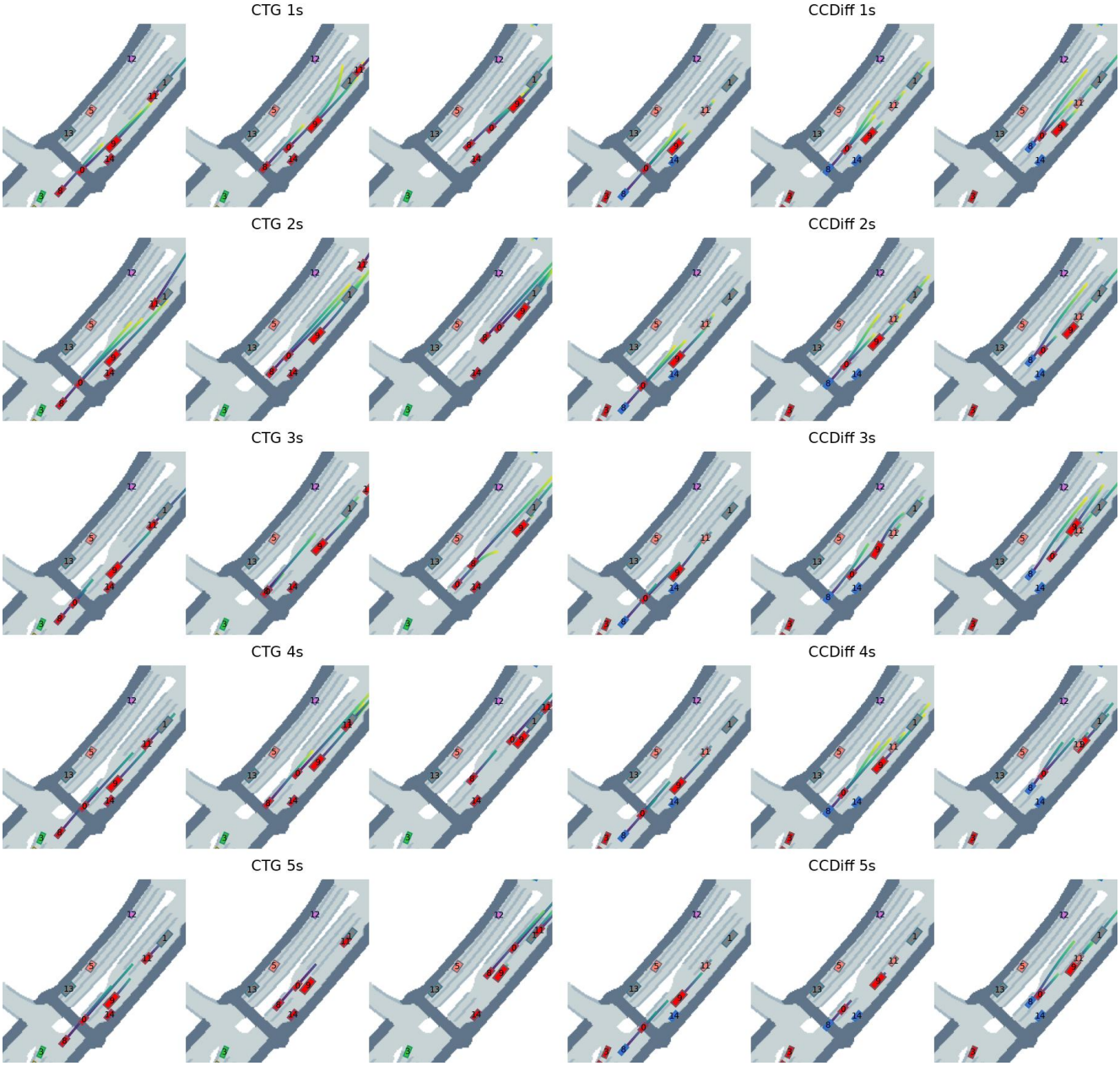

Figure 20. Qualitative comparison of *CCDiff* and CTG under cross traffic violation generation under different generation horizons.

C.5.5 Chain-reaction Crash

Baseline Comparison A chain-reaction crash involving five vehicles (agents 1, 2, 5, 7, 8) occurs when a sudden stop or collision causes a cascade of impacts among closely spaced vehicles in the same lane. This happens before an intersection when vehicles fail to maintain a safe following distance, leading to multiple rear-end collisions.

Among all the baselines, SimNet and BITS fail to generate safety-critical scenarios. TrafficSim, STRIVE, and CTG generate collisions between agent 0 on the side lane with agent 2 with a very unrealistic cut-in behavior. In comparison, CCDiff generates realistic collisions where the trailing vehicles 1, 7, and 8 fail to break timely and collide with static front vehicle 5, waiting for the right turn of 2. Both TTC graph and distance graph captures the interaction of 5 and 7, 8. Yet distance-based graphs fail to capture the indirect interaction between 2 and 7, 8.

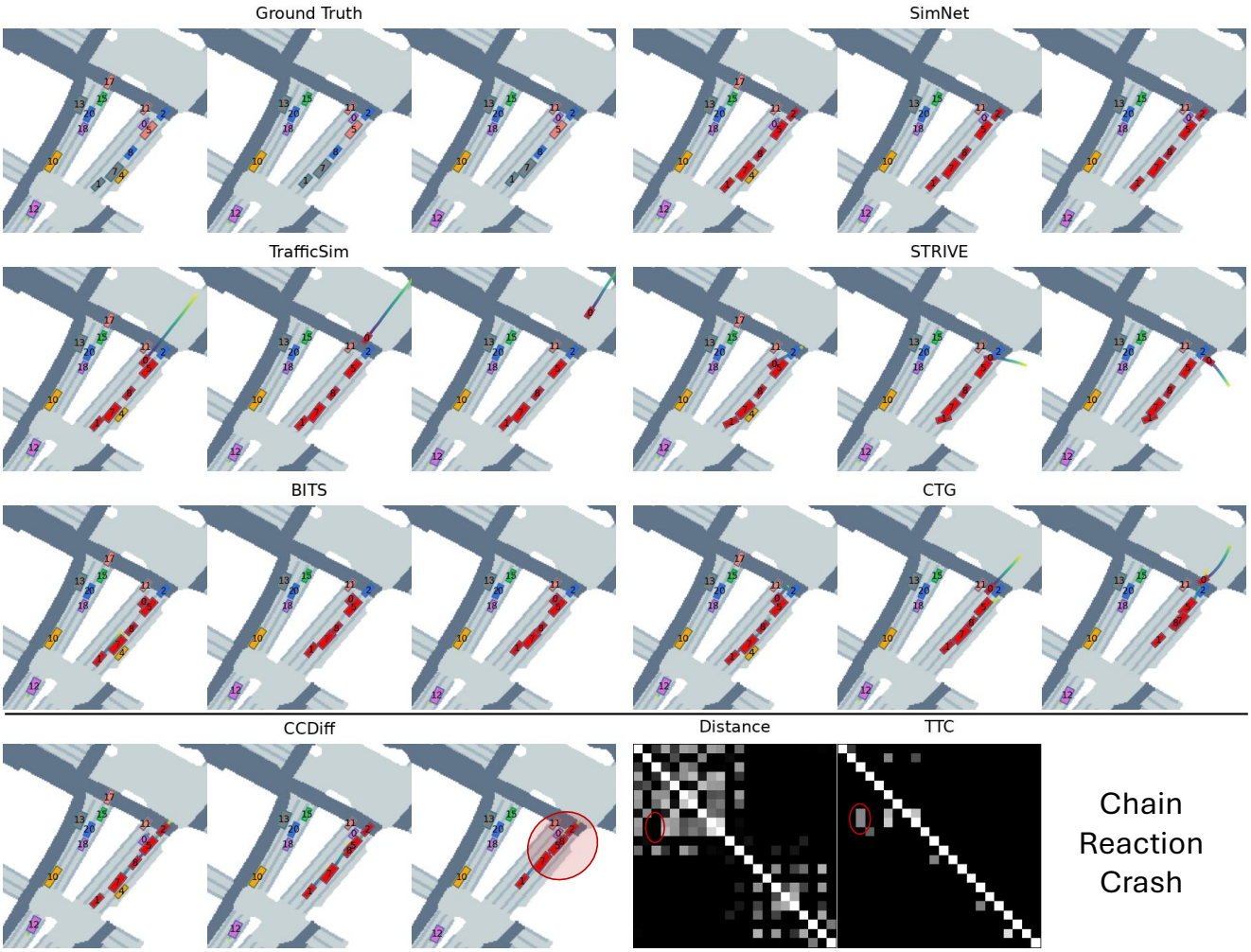

Figure 21. Qualitative of *CCDiff* and baselines in the chain-reaction crash scenarios.

Multi-agent Generation We compare the multi-agent generation results of *CCDiff* with CTG. *CCDiff* can consistently generate safety-critical emergency breaking samples when $K \geq 3$, with a control of the most important vehicle 7, 8 in this context. In contrast, CTG keep accelerating the side-lane vehicle 0 or rear vehicle 1 in a very unrealistic way.

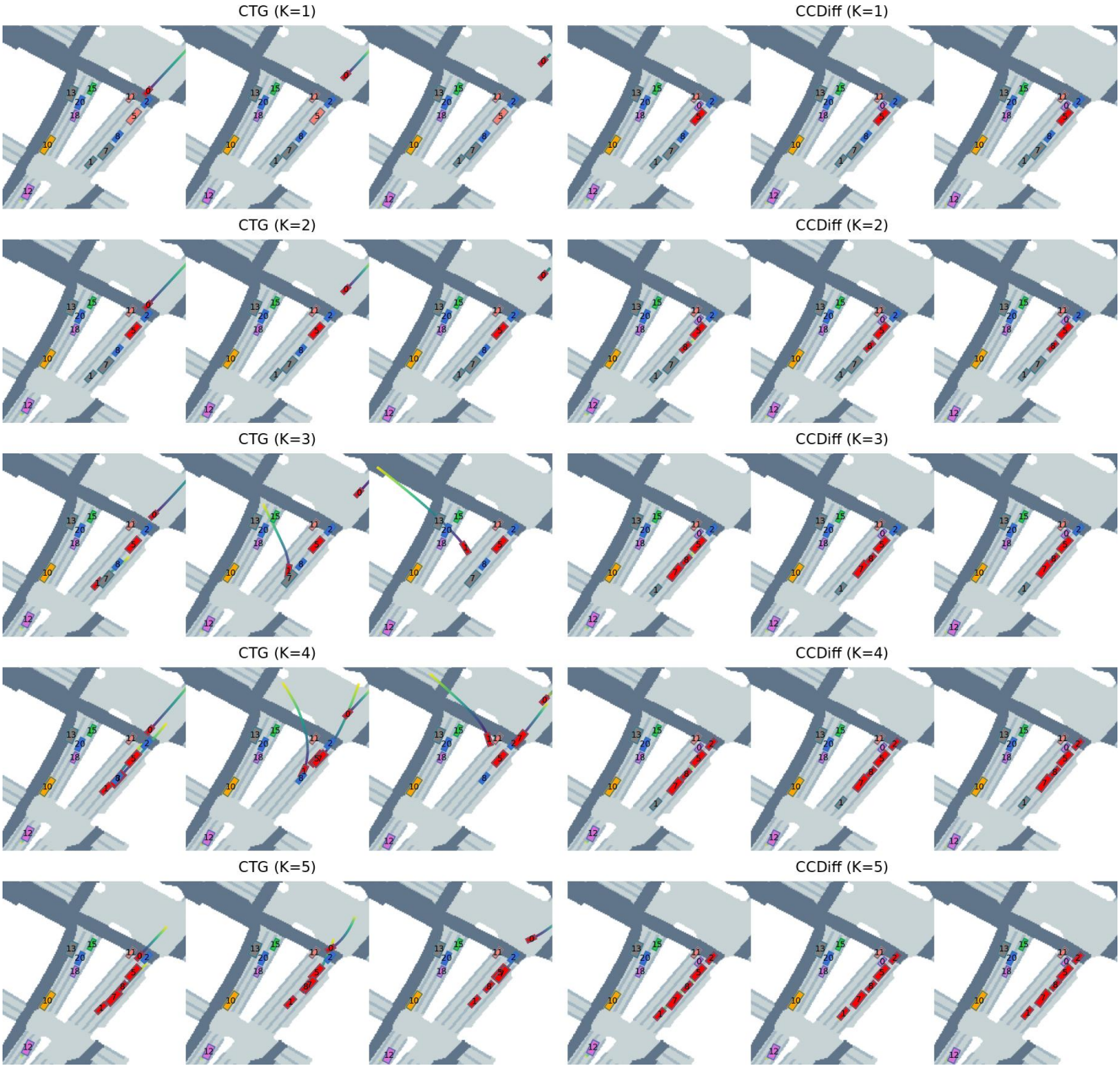

Figure 22. Qualitative comparison of *CCDiff* and CTG under cross traffic violation generation under different sizes of controllable agents.

Long-horizon Generation We compare the long-horizon generation results of *CCDiff* with CTG. *CCDiff* can consistently generate the cut-in violation scenarios with all different lengths of the generation horizon $1s \leq T \leq 5s$. In contrast, the trajectories generated by CTG seem to diverge by a great deal when $T \geq 2s$.

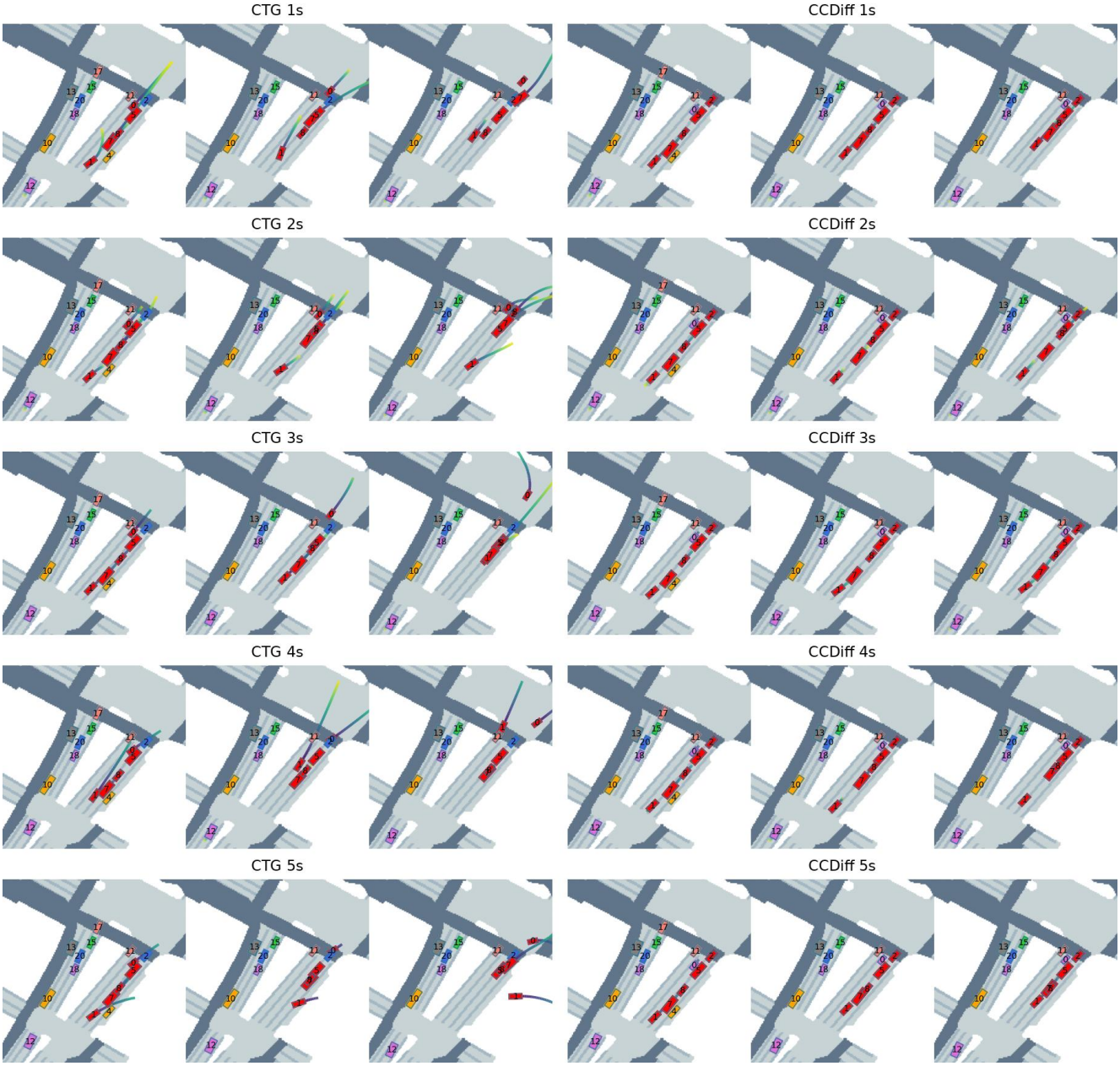

Figure 23. Qualitative comparison of *CCDiff* and CTG under cross traffic violation generation under different generation horizons.

C.5.6 Adjacent Left-turn Side-wipe

Baseline Comparison An adjacent left-turn sideswipe occurs when two vehicles (agent 1, 11) in neighboring left-turn lanes collide as Agent 1 veers into Agent 11's path.

Among all the baselines, STRIVE and CTG generate the motions of 1 and 11 to the straight lane reverse lane. TrafficSim generates the motions of 1 and 11 to the straight lane. BITS generally follows the original history scenarios with a rear collision between agents 18 and 11. CCDiff drifts 1 a little bit and let it veer into the agent 11's path.

Both the Distance graph and the TTC graph could detect the close interaction between agents 1 and 11 in this case.

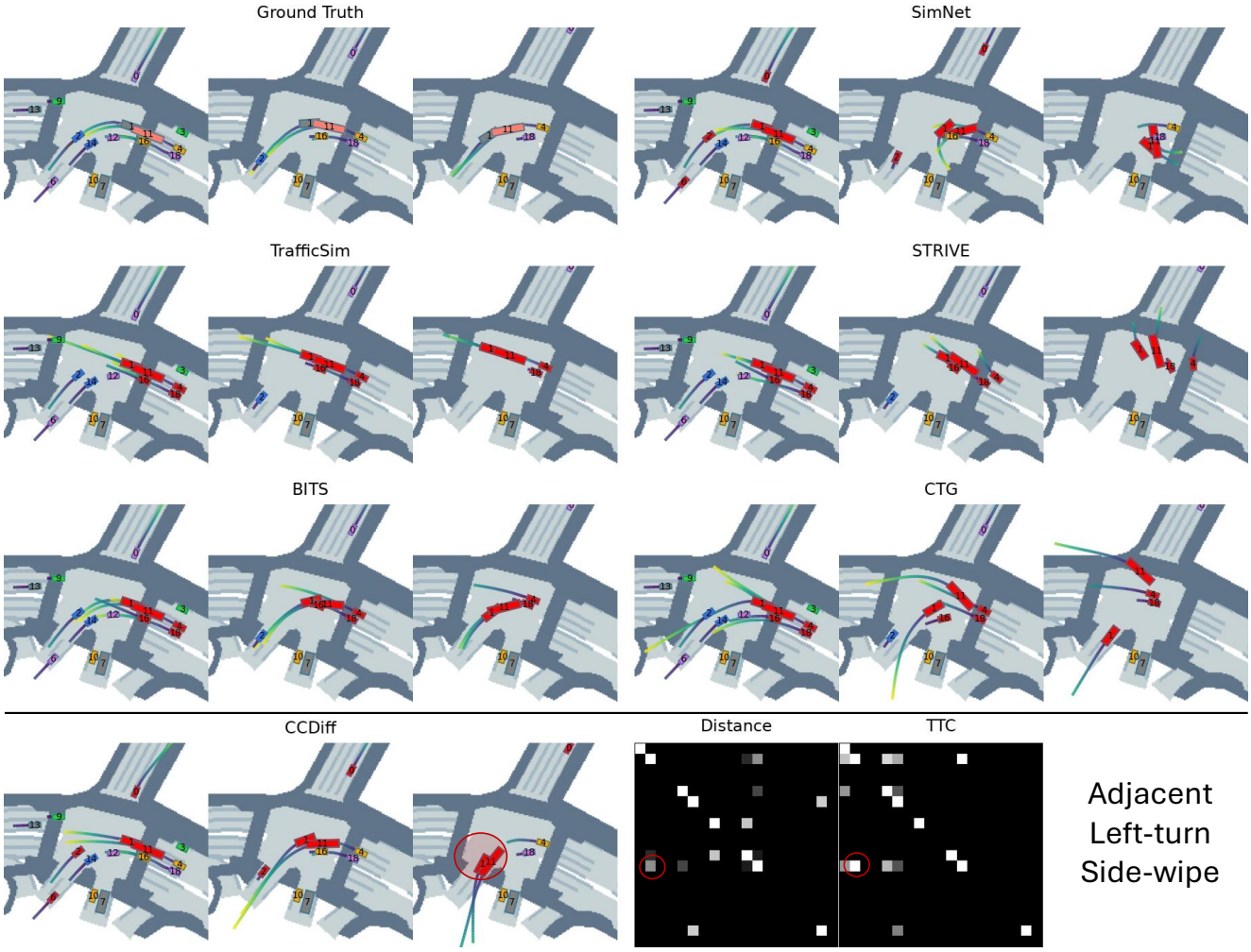

Figure 24. Qualitative of *CCDiff* and baselines in the adjacent left-turn side-wipe scenario.

Multi-agent Generation We compare the multi-agent generation results of *CCDiff* with CTG. *CCDiff* consistently generates safety-critical emergency braking scenarios when $K \geq 3$, effectively controlling the behavior of the most critical vehicle, agent 1, in this context. In contrast, CTG fails to accurately model the scenario, allowing agent 11 to continue in the wrong direction and being unable to generate collision samples, even when more agents are controllable.

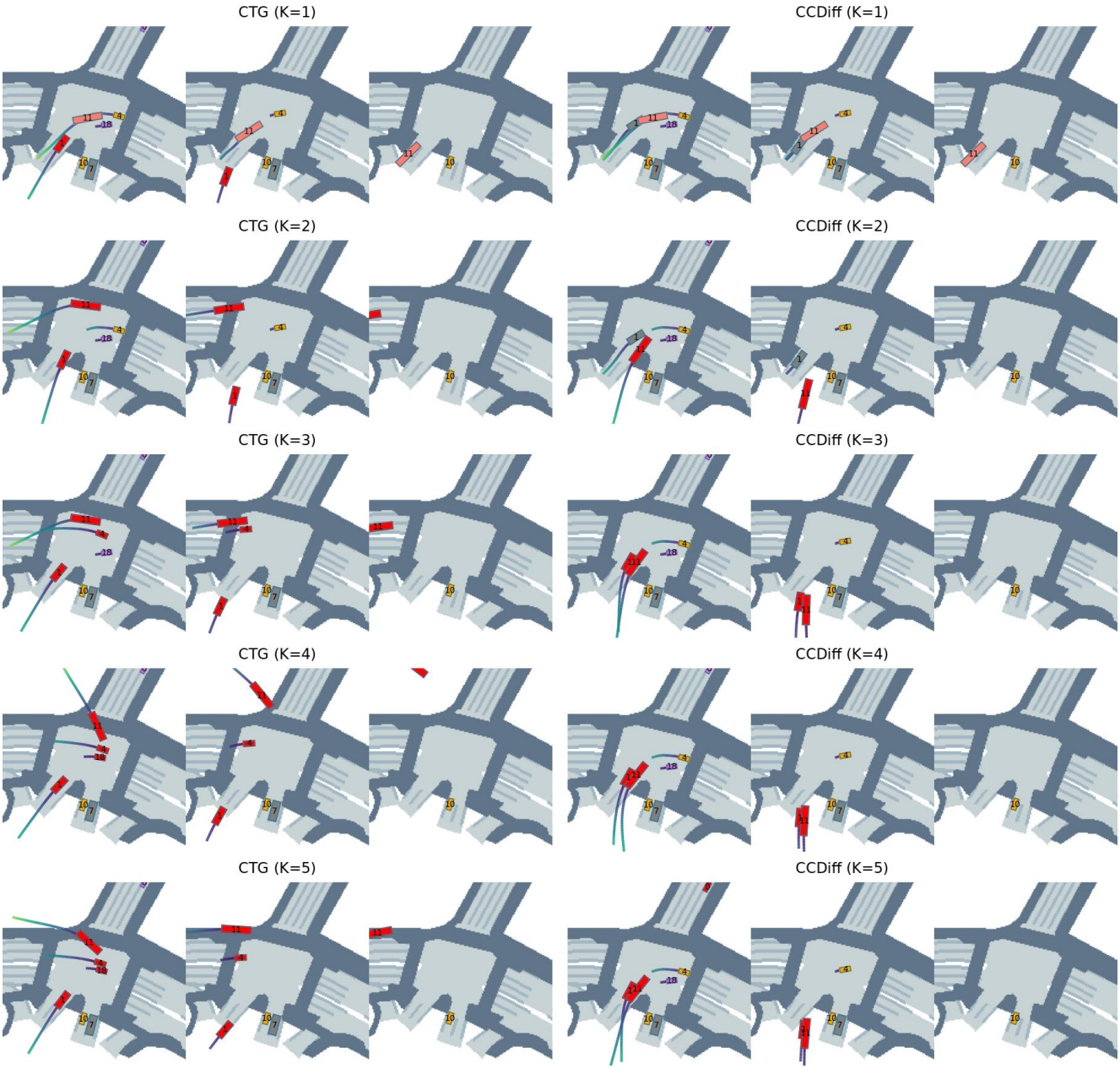

Figure 25. Qualitative comparison of *CCDiff* and CTG under cross traffic violation generation under different sizes of controllable agents.

Long-horizon Generation We compare the long-horizon generation results of *CCDiff* with CTG. *CCDiff* can consistently generate the left-turn side-wipe scenarios, while CTG diverges and fails to generate collision samples at $T = 3s, 4s$.

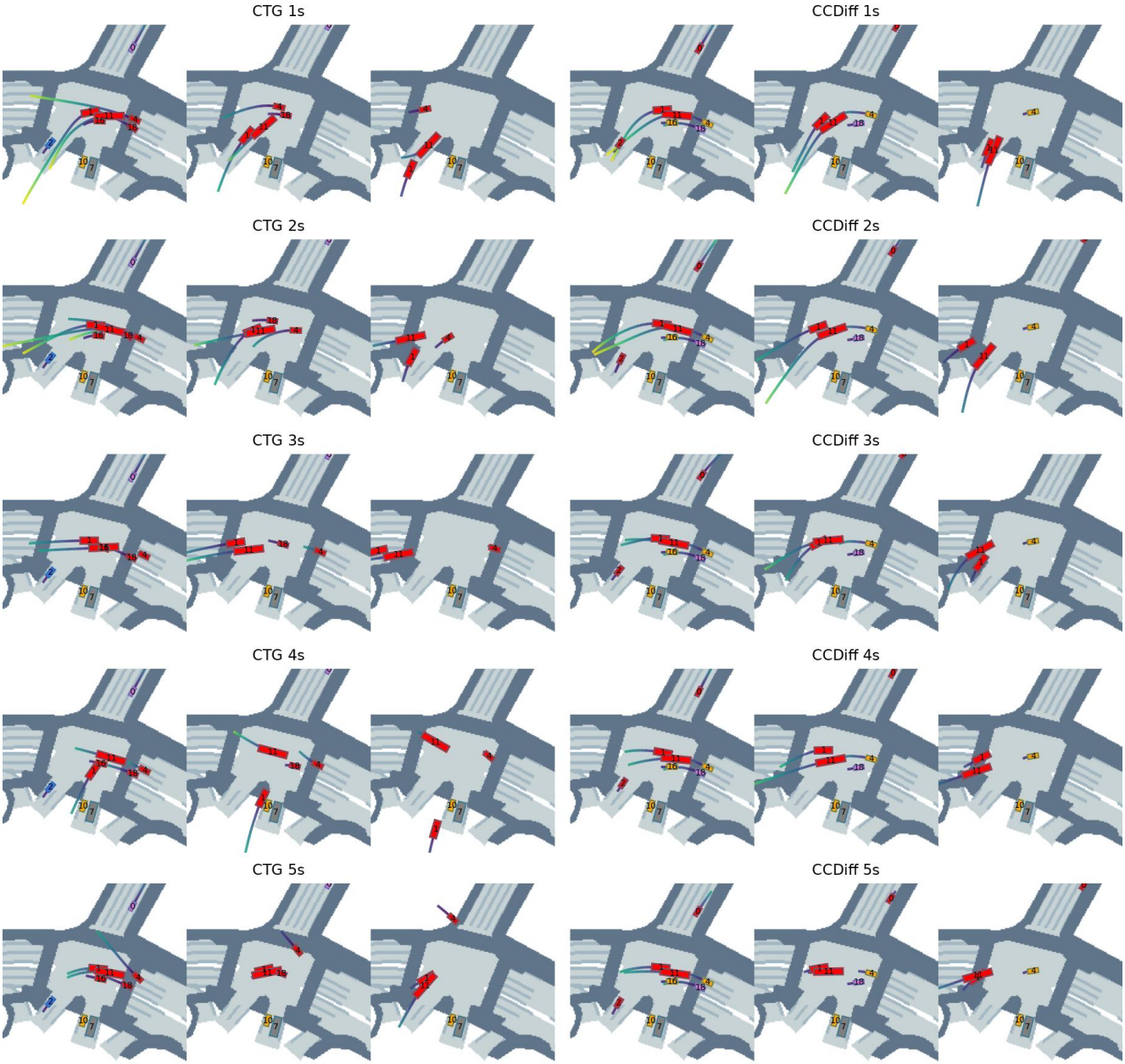

Figure 26. Qualitative comparison of *CCDiff* and CTG under cross traffic violation generation under different generation horizons.

C.5.7 Multi-vehicle Merge-in

Baseline Comparison Multi-vehicle merge-in occurs when a vehicle from a side lane (agent 13) attempts to merge into a single-lane traffic flow (agents 6, 2, 29), causing disruptions or collisions involving three vehicles 2 and 29.

Among all the baselines, SimNet does not generate collision samples, TrafficSim and CTG generate collision between 13 and 2 and manipulates the trajectory of 13 in an abrupt way. Our scenario just slows down agents 6 and 2 with an expectation of merge-in from agent 13, which causes the trailing agent 29 collides to agent 2. The generated final scenario of CCDiff have the closest layout with the ground-truth trajectories compared to other baselines.

TTC mask in this case is more sparse with necessary information (agent 2 and 29) compared to the distance mask.

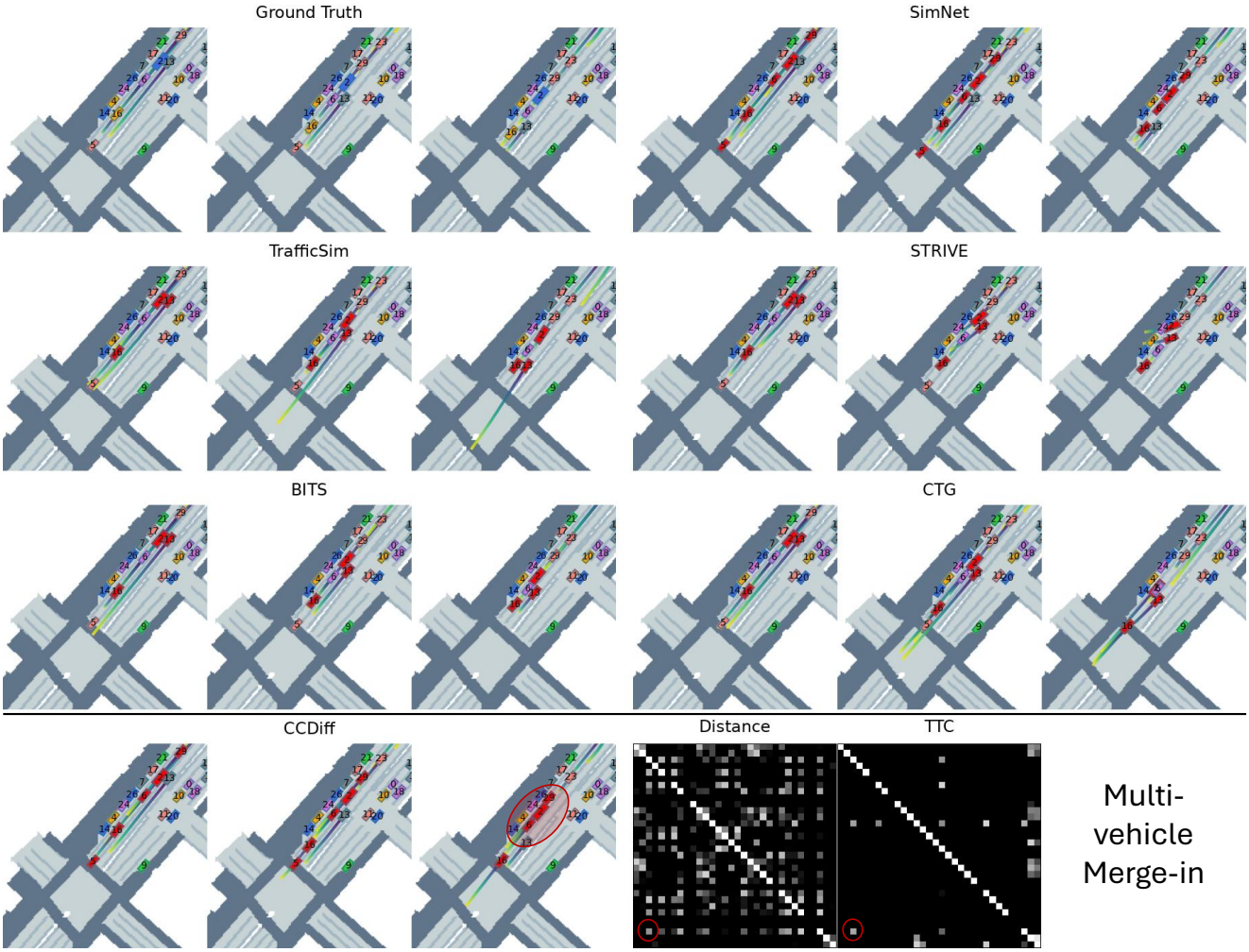

Figure 27. Qualitative of *CCDiff* and baselines in the multi-vehicle lane merge-in scenarios.

Multi-agent Generation We compare the multi-agent generation results of *CCDiff* with CTG. *CCDiff* can consistently generate safety-critical emergency breaking samples when $K \geq 4$, with a control of the most important vehicle 2, 6 in this context. In contrast, CTG keeps accelerating the side-lane vehicle 13 without generating any meaningful near-miss samples.

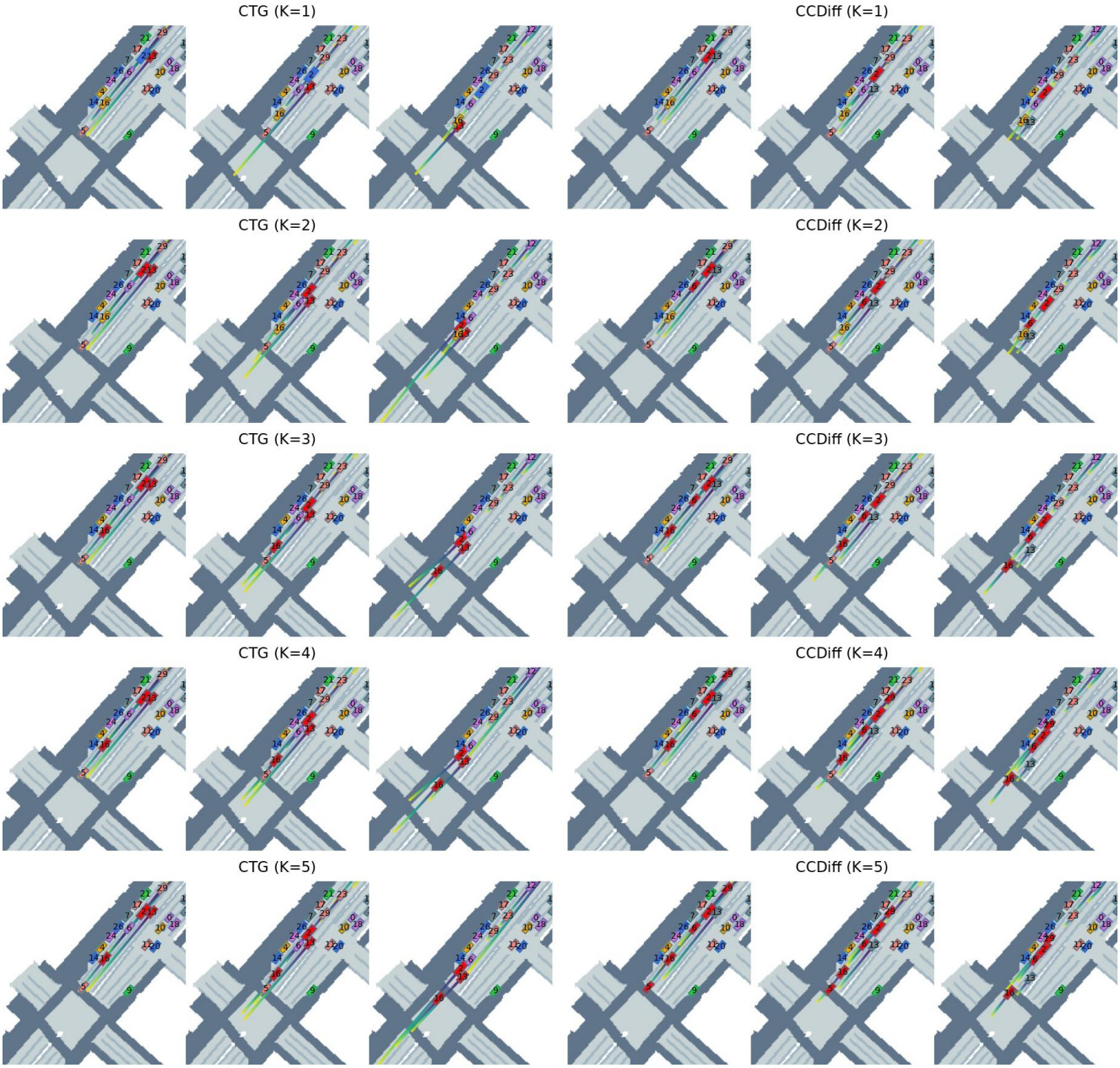

Figure 28. Qualitative comparison of *CCDiff* and CTG under cross traffic violation generation under different sizes of controllable agents.

Long-horizon Generation We compare the long-horizon generation results of *CCDiff* with CTG. *CCDiff* can consistently generate the multi-vehicle merge-in collision scenarios with all different lengths of the generation horizon $1s \leq T \leq 5s$. In contrast, CTG generates some cut-in collisions between 13 and 6 when $T \geq 2$, which is more unrealistic given the ground-truth layouts.

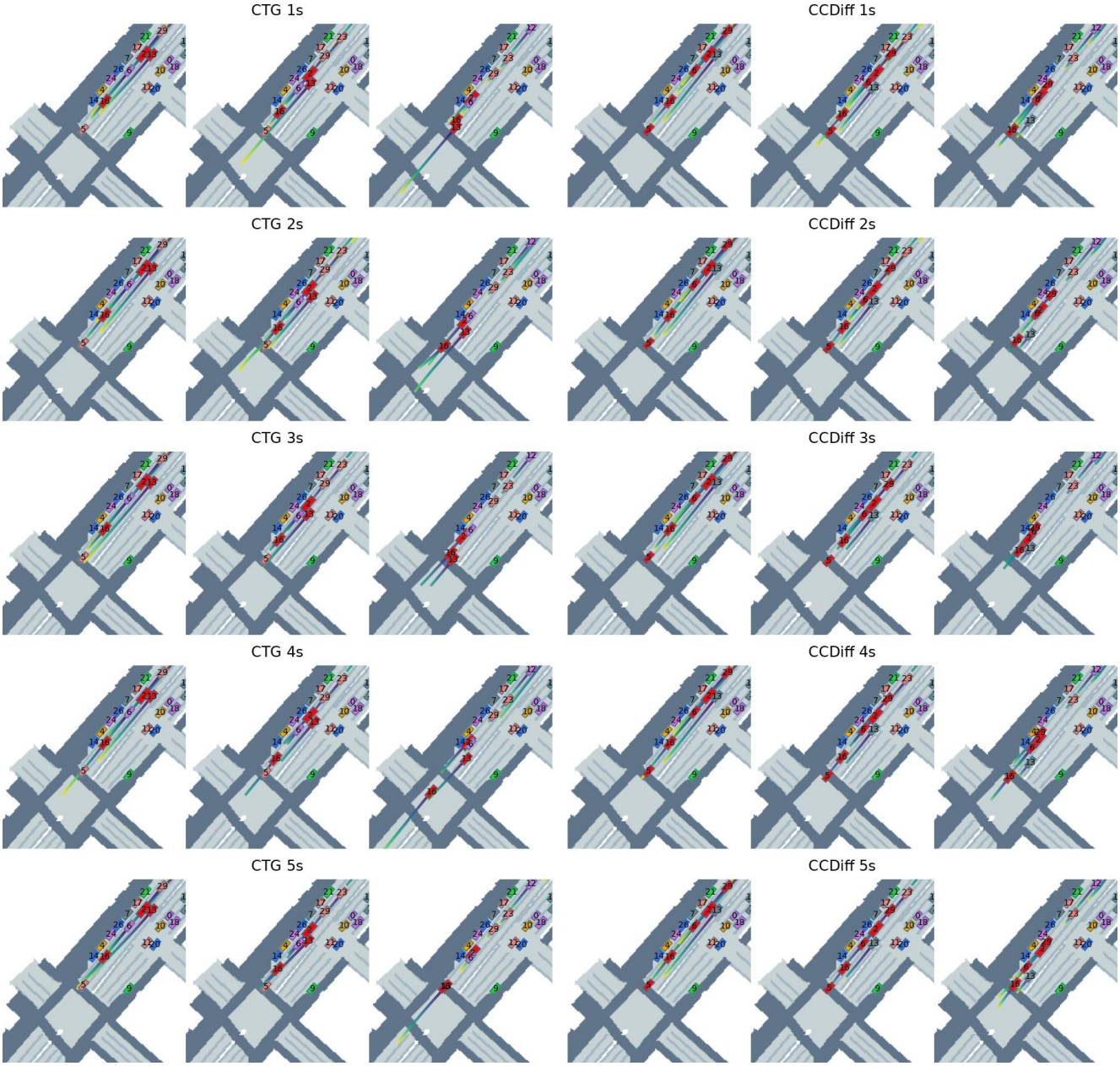

Figure 29. Qualitative comparison of *CCDiff* and CTG under cross traffic violation generation under different generation horizons.