

UNICST: Next-scale Latent Prediction for Continuous Spatio-Temporal World Modeling

Supplementary Material

865 This appendix is organized as follows. In Sec. 6, we provide
866 results of some additional tasks including novel view
867 synthesis (Sec. 6.1) and motion planning (Sec. 6.2). In
868 Sec. 7, extra ablation studies further show the explain the
869 roles of each module in our method. Sec. 8 gives more
870 qualitative results of synthetic driving scenarios. Finally,
871 we discuss the limitations of our work in Sec. 9.

872 6. Additional Tasks

873 6.1. Novel View Synthesis

874 We follow [68] to evaluate the performance of UNICST in
875 novel view synthesis task on nuScenes validation set. Different
876 shifts $\{1m, 2m, 4m\}$ along the x -axis and y -axis of
877 ego-vehicle are applied to the camera viewpoints. We measure
878 the FID and FVD metrics between the synthetic shifted
879 multiview videos and the original real videos. In Tab. 6,
880 our FID and FVD metrics only drop slightly after view-
881 point shifting. Compared to baselines, UNICST exhibits
882 a significant performance gain even if it does not include
883 any 3D inductive bias such as radiance field. These results
884 demonstrate the great capability of UNICST in novel view
885 synthesis in a data-driven manner.

886 6.2. Motion Planning

887 UNICST can also output future trajectories of ego-vehicle
888 in addition to multiview video generation. In Fig. 5,
889 we visualize some qualitative motion planning results on
890 nuScenes validation set, where the model takes ground-
891 truth multiview images as inputs. As a motion planning
892 model, UNICST can handle various scenarios including
893 turning and circumventing a stationary vehicle.

894 7. Additional Ablation Studies

895 In this part, we provide additional ablation studies to further
896 analyze the role of each module in UNICST. In Sec. 7.1, the
897 effect of classifier-free guidance is discussed. In Sec. 7.2,
898 we compare the image quality for single-frame or multi-
899 frame generation. In Sec. 7.3, we compare our proposed de-
900 coupled spatio-temporal condition module with vanilla full
901 attention design. In Sec. 7.2 and 7.1, we directly conduct
902 experiments on our base model (as in Tab. 2 in the main
903 paper) to explicitly reflect the influence on our final perfor-
904 mance. In Sec. 7.3, we follow the setting in Sec. 4.4 of our
905 main paper to apply a smaller model and less training data
906 to speed up the training process.

7.1. Classifier-Free Guidance 907

908 Similar with diffusion models [29], UNICST can also be
909 enhanced through classifier-free guidance (CFG) for the
910 output logits. In Tab. 2 of our main paper, we adopt CFG
911 weight 3 as the default setting. However, in Tab. 7, im-
912 proved performance is witnessed with higher CFG weights,
913 which can reflect the input conditions more significantly. In
914 the same time, we also notice that the CFG weight can get
915 saturated when the weight is high like (*e.g.* 7).

7.2. Single-Frame Generation 916

917 UNICST can also handle independent single-frame gener-
918 ation task, where the model only generates the first frame
919 without any actual cross-frame temporal reliance. In Tab. 8,
920 we report the metrics for both single-frame and multi-frame
921 generation pipelines with our base model. Single-frame
922 generation can achieve comparable FID with multi-frame
923 generation which reflects similar image quality. However,
924 this strategy sacrifices the coherence across frames, so it
925 cannot generate smooth videos.

7.3. Decoupled Spatio-Temporal Modules 926

927 In Sec. 3.4 of the main paper, we elaborate the decoupled
928 modules scale, spatial, and temporal conditions. Another
929 intuitive strategy is to apply full attention in a single self-
930 attention module over all camera views and all historical
931 frames. In this case, for each frame, the per-scale multi-
932 view features attend to all the all scale-wise features in past
933 frames from all camera views. In Tab. 9, we find the vanilla
934 full attention brings significantly worse performance, and
935 intuitively has higher time and memory costs. It cannot
936 effectively benefit from the pretraining since the scale-
937 wise causality no longer exists. The self-attention module
938 should learn more spatio-temporal correlation rather than
939 pure next-scale prediction.

8. Additional Qualitative Results 940

941 In attached three mp4 files, we provide synthetic 10Hz
942 videos with eight camera views with resolution $384 \times$
943 672 . Ideally, UNICST can generate infinite length videos
944 through sliding window in an autoregressive manner. Given
945 the 100MB limitation of supplementary materials, we pro-
946 vide 3 videos with 10s length of each. The synthetic high-
947 resolution videos show high visual quality and fidelity to
948 object conditions. They demonstrate realistic street views
949 and can also simulate the motion of ego vehicle and other

Table 6. Novel view synthesis performance with camera view shifts

Method	Shift 1m		Shift 2m		Shift 4m	
	FID	FVD	FID	FVD	FID	FVD
PVG [14]	48.15	246.74	60.44	356.23	84.50	501.16
EmerNeRF [82]	37.57	171.47	52.03	294.55	76.11	497.85
StreetGaussian [80]	32.12	153.45	43.24	256.91	67.44	429.98
OmniRe [16]	31.48	152.01	43.31	254.52	67.36	428.20
FreeVS [68]	51.26	431.99	62.04	497.37	77.14	556.14
UNICST _{x-shift}	14.75	139.60	15.04	139.94	15.34	147.36
UNICST _{y-shift}	14.54	138.47	14.63	140.83	15.90	151.60

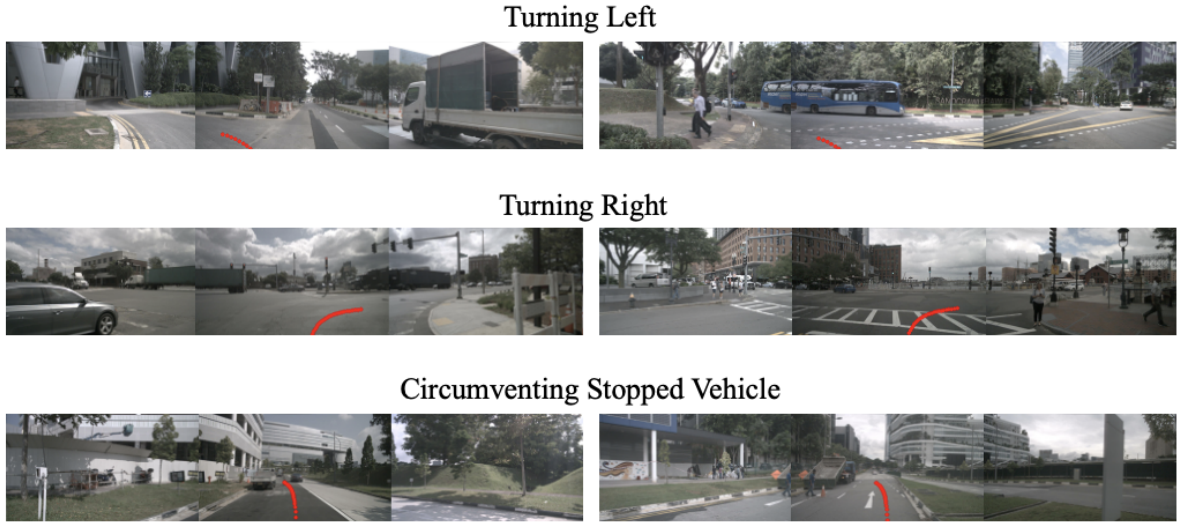


Figure 5. Motion planning qualitative results on nuScenes validation set. All the images are ground-truths from the front three camera views.

Table 7. Ablation on classifier-free guidance

CFG	FID	FVD
CFG=1	20.6	165
CFG=3	14.5	134
CFG=5	13.0	114
CFG=7	12.8	102
CFG=9	13.2	101

Table 8. Ablation on single-frame generation

Methods	FID	FVD
Multi-frame	14.5	134
Single-frame	14.3	-

Table 9. Ablation on decoupled spatio-temporal modules

Attention	FID	FVD
Full	70.2	618
Decouple	21.7	241

larger scale of training data.

9. Limitations

Currently, UNICST is mainly limited by the amount and diversity of training data. In this paper, the model is only trained with the combination of two open datasets: nuScenes [10] and nuPlan [12]. The total length is only about 60 hours, which is significantly less than thousands of hours in many previous works [20, 32, 83]. Besides, our training data are limited to driving scenarios captured from the ego-vehicle. Since UNICST can take heteroge-

agents. However, we also notice some slight shakes and object-wise inconsistency, which may be solved with a

962 neous training data with various sensor setups, we can po-
963 tentially include indoor scenes from navigation robots and
964 outdoor off-road scenarios from delivery robots to further
965 improve the generalization ability of UNICST.