MaskGWM: A Generalizable Driving World Model with Video Mask Reconstruction

Supplementary Material

6. More Ablation Experiments

6.1. Additional Results of Mask Ratio

Table 5 illustrates the impact of the mask ratio r on mask reconstruction across various branches and temporal attention strategies. Our findings reveal several key insights: (1) The temporal branch equipped with masked temporal self-attention is more sensitive to mask ratio and necessitates a substantially lower mask ratio compared to the spatial branch. (2) The influence of the mask ratio on the proposed shifted temporal self-attention is more consistent with that observed on the spatial branch. As depicted in Fig.4, the main difference in the DiT Encoder with the spatial branch is the positional shift, which can be effectively handled by positional encoding. Consequently, this allows for the attainment of an well-performed mask ratio (e.g. 0.25) in both spatial MR and temporal MR.

r	$\mathcal{M} = \mathcal{M}_{spatial}$	$\mathcal{M} = \mathcal{M}_{time}$	$\mathcal{M} = \hat{\mathcal{M}}_{time}$
0	136.5	136.5	136.5
0.1	125.8	133.2	123.7
0.25	116.7	142.6	121.3
0.4	155.9	179.1	149.8

Table 5. FVD comparisons on mask ratio.

6.2. Additional Results of Mask Reconstruction on NuScene Dataset

As described in GenAD [46], the training and validation sets of OpenDV-2K are sourced from different YouTube videos with significant scene changes. Therefore, the model's performance on this dataset can be used for the evaluation of its generalization ability. We also conduct ablation studies in the in-domain setting by evaluating metrics on the nuScenes validation dataset. As shown in Table 6, the proposed mask reconstruction method achieves significant improvements on both metrics.

row&shift att.	r	**FVD**↓	**FID**↓
	0	107.2	7.5
√	0.25	92.5	5.6

Table 6. Ablations on nuScene dataset.

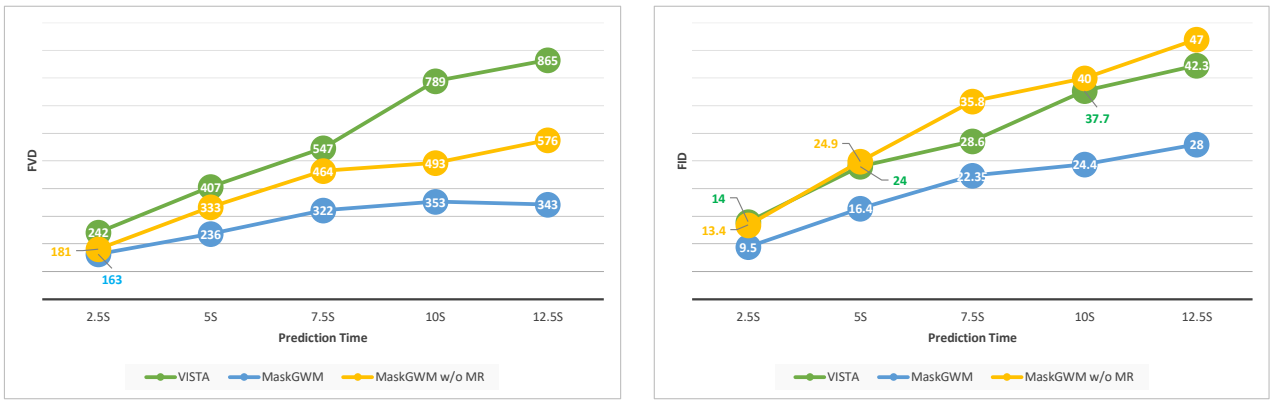
6.3. Additional Results of Mask Reconstruction on Long-Horizon Prediction

To further analyse the influence of MR on auto-regressive generation, we extend the video duration to approximately 12 seconds and documented the metrics in Fig.7. The results indicate that MR is also effective in enhancing performance in long-sequence prediction. Although our baseline without MR still outperforms Vista, the quality of generation begins to deteriorate notably from about 8 seconds, and the FID score increases to 37.7 at the 10 second, making it also incapable to predict the distant future. Consequently, we conclude that this baseline's improvements cannot translate to significant advancements in long-sequence. In contrast, when MR is integrated into our method, the fundamental enhancements in single-step generation lead to significantly alleviate quality degradation over time. As a result, MaskGWM is capable of generating 10 Hz videos with discernible scene elements for a long time, and even 60 second examples, which far exceeds both Vista and our non-MR baseline. Therefore, we regard MR as a pivotal design that enables the model to make generalized predictions over extended durations. Note that this evaluation is conducted on 300 videos of OpenDV-2K validation set only, due to their longer video sequence. Thus, the single-step (2.5 seconds) FID and FVD are higher than results in Tab.6, which is computed on 1800 videos.

7. Implementation Details

7.1. Concrete DiT Structure

We adopt the framework of SD3 and start our model with 2B parameters. We make several modifications to the original spatial transformer block to facilitate temporal and cross-view context modeling. First, Due to the limited availability of high-quality text data in our training datasets, e.g. only scene-level descriptions on nuScene, we skip the update of text feature by new-initialized temporal and cross-view transformer blocks. Then, for temporal transformer block, we make another modification to accommodate condition frames. To streamline the explanation, we represent the transformation within a transformer block as $z'_{out} = z'_{in} + f_b(z'_{in})$, where z'_{in} and z'_{out} are the input and out features, respectively, and f_b is the transformer block. Given the frame-level binary indicator m_c with value 0 on condition frames, the diffusion time-step τ, and time-step aware embeddings for scale f_{scale} and shift f_{shift}, we in-



(a) FVD (b) FID

Figure 7. **Comparison of Long-horizon FVD metric on OpenDV-2K validation set.** MR plays a crucial role in enhancing the capability to predict long video sequences, especially on FID.

troduce condition frames by:

$$z'_{out} = z'_{in} + f_b(f_{scale}(m_c\tau)z'_{in} + f_{shift}(m_c\tau)) \quad (5)$$

Here $m_c\tau$ is employed to reset the time-step for conditional frames to zero and time-step aware embeddings is applied for linear transform.

We append one temporal transformer block and one view transformer block after each spatial transformer block following the common practice of previous works [12, 40].

7.2. Detailed Training Parameters

We employ the Adam optimizer [28] for model training, using a learning rate of 5e-5. Throughout all training stages, we initiate the process with 1K warm-up steps and then maintain a constant learning rate. For condition frames, we randomly sample from zero to three frames following VISTA. We train Stage 1 for total 62K steps, Stage 2 for total 20K steps and Stage 3 for 6K steps. We select the training step based on numerical metrics from videos that are randomly sampled from the training set. Our training are conducted on 32 A800 GPUs with around 3 days on Stage 1.

7.3. Detailed Sampling Parameters

Our sampling strategy does not incorporate any special designs. We generate the video by sampling 30 steps and utilize a classifier-free guidance scale [17] of 4.0. Following Vista, we generate 25-frame videos containing one reference frame on full nuScene validation set with 5369 samples for our single-view model. All generated videos and corresponding frames are used for computing FVD and FID respectively. For our multi-view model, we generate 150 6-view videos for each nuScene scene, resulting in 900 single-view videos. Then, 10K frames are randomly sampled from these 900 videos for computing FID. This is align with the evaluation setting of DriveWM [40].

7.4. Details of Comparisons with Vista

For comparisons with Vista, we use the official sample script and checkpoint. For zero-infer on Waymo dataset, we infer both models without action and the number of condition frame is set to 1. For long-horizon rollout on OpenDV-2K dataset, we infer both models without action and the number of condition frame is set to 3 for better temporal continuity across auto-regressive steps. We find numeric improvement is similar for one condition frame but the qualitative continuity is reduced by one-frame auto-regression. For auto-regressive steps larger than 1, we randomly select 25 frames from the generated video sequences to calculate the FVD and FID metrics.

8. Qualitative Results

8.1. Long-horizon rollouts

(what is rollout) **Longer prediction** We provide more qualitative and longer visualizations with 42-seconds videos in Fig.8. We find MaskGWM can predict stable and consistent driving future, combined with unseen scene with initial scope.

Qualitative comparisons In Fig.11, we make qualitative comparisons with Vista, which is previous state-of-the-art method on generalizable driving world model. Our method can both make stable prediction and generate dynamic objects according to the future, e.g. unseen cars in initial visual scope.

Diverse scenes In Fig.9, we present the extended generation results across various scenes, demonstrating the robust generalization capability of our approach.

Action control In Figure 12, we illustrate the controllability of our method on the OpenDV-2K dataset, adhering to the action module in Vista.

Multi-view generation In Figure 10, we show the multi-

view generation ability coming from lifting our single-view model by extra view transformer blocks.

9. Discussions

9.1. Differences to Vista

Although both our method and Vista [12] aim to construct a generalizable world model using the large-scale OpenDV-2K dataset, we highlight several key distinctions here. First, our findings suggest that relying solely on the Diffusion loss may not be optimal for building a world model. We introduce a complementary MR task, which has demonstrated robust generalization capabilities in representation learning tasks. Second, our model enables multi-view video generation through an additional training stage. This also illustrates that multi-view generation can benefit from a well-trained single-view model trained on a dataset encompassing significantly longer durations—over 1,700 hours in the OpenDV-2K dataset. Third, our model achieves longer prediction durations than Vista. As indicated by the slope of the metric changes in Fig. 7, our method maintains stable video prediction results, up to 15 seconds by autoregressive generation, whereas the generation quality of Vista degrades notably at this point. Moreover, we have found that our model can sustain stable generation over longer time periods across diverse scenes. Regarding quantitative evaluation, our model exhibits superior generalization capabilities, as evidenced by results on both the OpenDV-2K and Waymo datasets. On the standard nuScene benchmark, our approach also yields better results, with a 19% improvement in FID and a 3% decrease in FVD.

9.2. Usage of Stable Diffusion 3

Our baseline, built upon the SD3 [7] model, yields superior results compared to GenAD (trained on SDXL [31]) and performance slightly lower than Vista (initialized with SVD [1]). Since both GenAD and our baseline are derived from image generation models, the improved performance of our baseline demonstrates the effectiveness of SD3. The superiority of SVD is attributed to its well-initialized temporal blocks, which have undergone multi-stage pre-training on extensive video datasets. Therefore, enhancing the data efficiency of SD3—as in our MR policy—and incorporating more video data present promising avenues to bridge this performance gap.

9.3. Future impact of MR.

In our method, MR acts as a complementary task to the diffusion loss, incorporating better video prediction abilities. Within the scope of representation learning, MR conducts context reasoning in a self-supervised way and can be generalized to various tasks. This aligns with our design: recovering the original MR at low noise levels using detailed local context. Our results show that diffusion models may excel in generating high-fidelity results but learn context reasoning slowly, which can be improved through the MR task. More generally, the effectiveness of MR shows that relying solely on diffusion may not be the optimal approach for driving world models. A similar inspiration can also be found in GAIA-1, where the prediction ability is decomposed into an auto-regressive model and a diffusion model. Exploring training targets for world models can be a promising direction.

9.4. Limitations.

Although better generalization ability and quality are achieved, there still exist some limitations that call for future works. (1) Controllability. Since we focused our main improvements on generalization ability and long-duration prediction, the action module follows the design of Vista. We have found several challenging cases in control, such as unreasonable commands. Similar to Vista, our method relies on resampling the nuScenes dataset to learn control ability. As a result, finding better feedback strategies and larger datasets for action learning is a promising direction. (2) Prediction of Uncertain Future. This phenomenon mainly arises when encountering complex traffic scenarios, especially when predicting the movement of each vehicle is difficult. (3) Generation of Non-Front View Images. Since multi-view capability is introduced only at the last training stage with a single nuScenes dataset, the images of non-front views lack exposure before this stage. Incorporating non-front view data at an earlier stage or adding more multi-view datasets (e.g., Waymo) may help address this problem.

Figure 8. Generalization ability of MaskGWM with longer time.

Figure 9. Generalization ability of MaskGWM in more scenarios.

Figure 10. Generalization ability of multi-view videos.

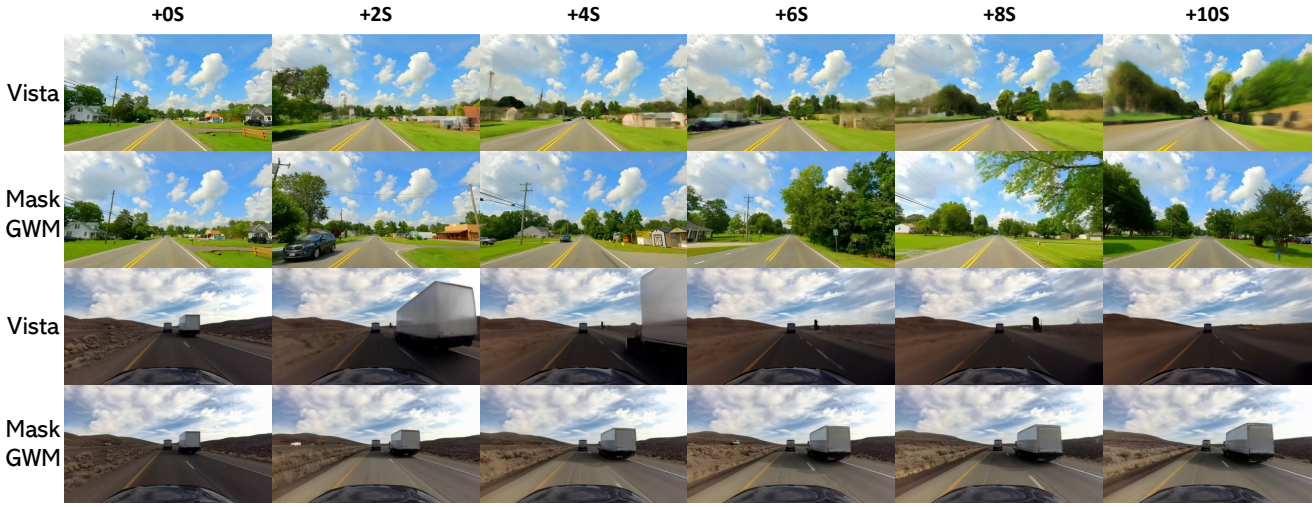


Figure 11. Qualitative comparison with Vista.

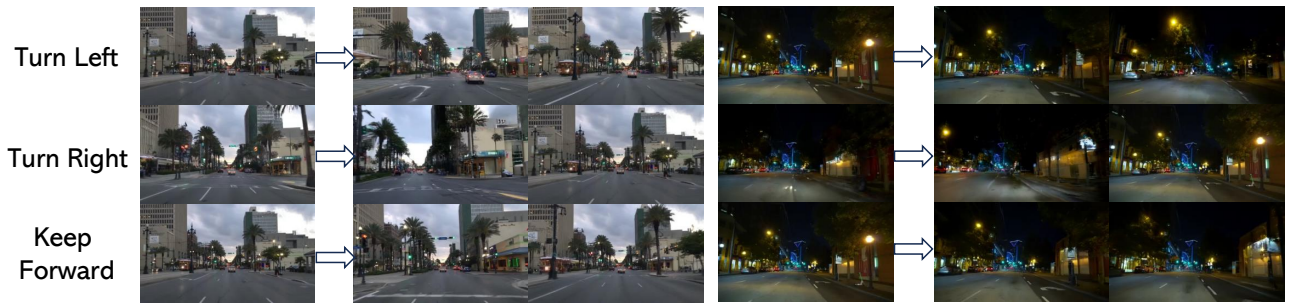


Figure 12. Action control ability of MaskGWM.