HMPDM: A Diffusion Model for Driving Video Prediction with Historical Motion Priors

Ke Li, *Graduate Student Member, IEEE*, Tianjia Yang, Kaidi Liang, Xianbiao Hu, Ruwen Qin*, *Member, IEEE*

***Abstract*—Video prediction is a useful function for autonomous driving, enabling intelligent vehicles to reliably anticipate how driving scenes will evolve and thereby supporting reasoning and safer planning. However, existing models are constrained by multi-stage training pipelines and remain insufficient in modeling the diverse motion patterns in real driving scenes, leading to degraded temporal consistency and visual quality. To address these challenges, this paper introduces the historical motion priors-informed diffusion model (HMPDM), a video prediction model that leverages historical motion priors to enhance motion understanding and temporal coherence. The proposed deep learning system introduces three key designs: *(i)* a Temporal-aware Latent Conditioning (TaLC) module for implicit historical motion injection; *(ii)* a Motion-aware Pyramid Encoder (MaPE) for multi-scale motion representation; *(iii)* a Self-Conditioning (SC) strategy for stable iterative denoising. Extensive experiments on the Cityscapes and KITTI benchmarks demonstrate that HMPDM outperforms state-of-the-art video prediction methods with efficiency, achieving a 28.2% improvement in FVD on Cityscapes under the same monocular RGB input configuration setting. The implementation codes are publicly available at https://github.com/KELISBU/HMPDM.**

***Index Terms*—Vehicle Intelligence, Driving Video Prediction, Diffusion Model, Historical Motion Priors**

I. Introduction

In the context of intelligent vehicles, modern autonomous driving systems need to not only perceive the present scenarios [1] but also anticipate their evolution over time, which is crucial for path planning, especially in safety-critical scenarios [2]. However, the complex dynamics and frequent occlusions in real-world driving scenes cause traditional object-centric predictors, which are based on low-dimensional states (e.g., trajectories), to discard essential appearance, structural, and contextual information. Emerging diffusion-based video prediction methods aim to forecast entire future scenes conditioned on past or present visual context. Crucially, scene-level video prediction offers a comprehensive and holistic understanding of the driving environment, capturing both the static background and the motion of dynamic traffic agents. Furthermore, by leveraging historical motion patterns, they naturally handle occlusions and can better represent the dynamic evolution of driving scene into the future.

Ke Li, Kaidi Liang, and Ruwen Qin are with the Department of Civil Engineering, Stony Brook University, Stony Brook, NY 11794, USA.

Tianjia Yang and Xianbiao Hu are with the Department of Civil and Environmental Engineering, Pennylvania State University, University Park, PA 16802, USA.

* Corresponding author: Ruwen Qin, email: ruwen.qin@stonybrook.edu

Despite notable advances in diffusion-based video prediction, several key challenges remain. Many models struggle with temporal consistency, since the visual quality degrades and objects appear distorted within the long-horizon prediction. Another limitation lies in the lack of robust historical motion modeling. Without properly leveraging historical motion priors, predicted agents often exhibit unrealistic trajectories or blurry motion patterns in driving scenarios. Accompanied with it is the concern of computational efficiency and adaptability. Despite promising performance, recent models often require substantial computation, complex training pipelines, and extra multimodal inputs, revealing the need for simpler yet effective solutions for driving video generation.

To bridge the gaps, the paper proposes a diffusion-based driving video prediction framework that is aware of historical motion, named **H**istorical **M**otion **P**riors-informed **D**iffusion **M**odel (**HMPDM**), making the following contributions:

- To implicitly inject historical motion context, **T**emporal-**a**ware **L**atent **C**onditioning (**TaLC**) is introduced to feed the model with latent representations of past frames as a learnable prior.
- Complementing this, a **M**otion-**a**ware **P**yramid **E**ncoder (**MaPE**) hierarchically encodes multi-scale motion features from the historical dynamic.
- To mitigate error accumulation in long-term generation, we implement **S**elf-**C**onditioning (**SC**), a strategy where the model conditions on its own intermediate predictions.

The remainder of this paper is organized as follows. Sec. II reviews the relevant literature. Sec. III presents the proposed HMPDM framework in detail. The experimental setting and results are reported in Sec. IV. Finally, Sec. V concludes the study and discusses future directions to pursue.

II. Related Work

This study is built upon the literature on diffusion-based video prediction, historical motion injection, and motion-enhanced video prediction.

A. Diffusion-based Video Prediction

A stream of studies primarily relies on LSTM, VAEs, and VRNNs [3]–[7] to learn a probabilistic latent space, model temporal dynamics, and capture spatial coherence. However, they meet the limitation on long-term video prediction with temporal consistency. Recently, diffusion-based paradigms

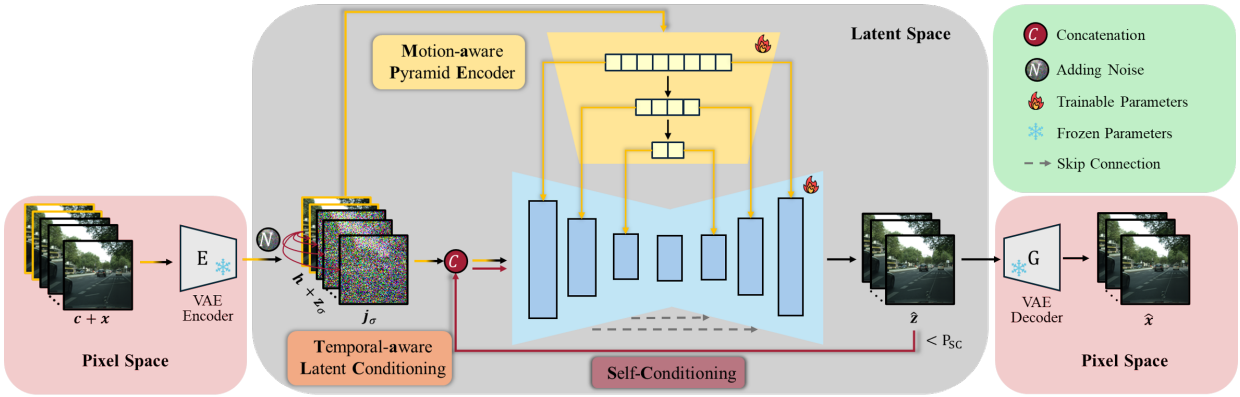

Fig. 1. Overview of the proposed HMPDM framework

have demonstrated remarkable performance in both video generation and prediction tasks. A diffusion model for video prediction is a generative framework that progressively transforms random Gaussian noise into coherent videos through an iterative denoising process [8]–[10]. Many studies build upon the U-Net denoising architecture, incorporating various conditioning strategies and spatio-temporal consistency techniques [11]–[17].

B. Historical Motion Injection

Historical frames naturally provide motion priors as the conditional context for future video prediction. Common ways for incorporating historical frames either concatenate them on the frame or channel dimension, or first encode them and subsequently inject the encoded representations into the cross-attention layers. NPVP [18] used a CNN autoencoder to extract appearance features from past frames, which are used as keys and values in the cross-attention layers to predict future frames. Similarly, STDiff [15] leveraged difference images from past frames as input to a specialized motion encoder, which disentangles motion and content features. While, VDT [19] compared three different historical motion injection schemes and illustrated the effectiveness of direct token concatenation. Integrating both injection methods, LGC-VD [20] introduced a two-stage training design, where recent local frames are concatenated on the channel dimension, while longer-range global history is encoded and injected via cross-attention mechanism. However, these methods struggle to effectively unify local and global motion priors while maintaining alignment across scales.

C. Motion-enhanced Multimodal Video Prediction

To enhance time coherence and motion consistency of video prediction, recent approaches have incorporated multimodal data, such as depth, optical flow, and contour as complementary information. ExtDM [21] and LFDM [22] designed motion autoencoders to extract flow-based motion information to guide the video diffusion process. Additionally, Syncvp [17] incorporated conditional depth information with RGB video within a synchronous denoising framework. Whereas, most of them rely on a two-stage training pipeline to handle multimodal inputs, thereby increasing computational cost and complicating cross-modal alignment.

III. Methodology

We propose a simple, yet effective, diffusion model for driving video prediction that uses RGB cameras and exploits historical motion priors to enhance the quality of generated data. The overall framework of the proposed method is illustrated in Fig. 1. HMPDM aims to generate and predict F future video frames, $\boldsymbol{x} = \{\boldsymbol{x}_t\}_{t=1}^{F}$, given a set of P past frames, $\boldsymbol{c} = \{\boldsymbol{c}_t\}_{t=1}^{P}$, where $\boldsymbol{x}_t$ and $\boldsymbol{c}_t \in \mathbb{R}^{C\times H\times W}$ are RGB images. Therefore, the objective is to learn the conditional distribution of future frames $p(\boldsymbol{x}|\boldsymbol{c})$. The proposed framework comprises three main components: temporal-aware latent conditioning, a motion-aware pyramid encoder, and a self-conditioning strategy.

A. Temporal-aware Latent Conditioning

HMPDM builds upon the latent diffusion modeling paradigm, specifically following the Elucidated Diffusion Model (EDM) framework originally proposed by Karras et al. [10] and widely adopted in Stable Video Diffusion (SVD) [23]. The latent autoencoder E encodes a sequence of past frames $\boldsymbol{c} \in \mathbb{R}^{B\times P\times C\times H\times W}$ and future frames $\boldsymbol{x} \in \mathbb{R}^{B\times F\times C\times H\times W}$ to a low-dimensional representations, $\boldsymbol{h}$ and $\boldsymbol{z}$, denoted as:

$$\boldsymbol{h} = E(\boldsymbol{c}), \quad \boldsymbol{z} = E(\boldsymbol{x}), \tag{1}$$

where $\boldsymbol{h} \in \mathbb{R}^{B\times P\times C'\times H'\times W'}$, $\boldsymbol{z} \in \mathbb{R}^{B\times F\times C'\times H'\times W'}$, and B denotes the batch size. As demonstrated in Vista [24], integrating the clean latent representations of past frames $\boldsymbol{h}$ with noisy latent representations of future frames $\boldsymbol{z}_\sigma$ does not degrade generation quality. The forward diffusion process of HMPDM progressively corrupts the clean latent $\boldsymbol{z}$ by adding noise with scale $\sigma \in [\sigma_{\min}, \sigma_{\max}]$:

$$\boldsymbol{z}_\sigma = \boldsymbol{z} + \sigma\boldsymbol{n}, \quad \boldsymbol{n} \sim \mathcal{N}(0, I), \tag{2}$$

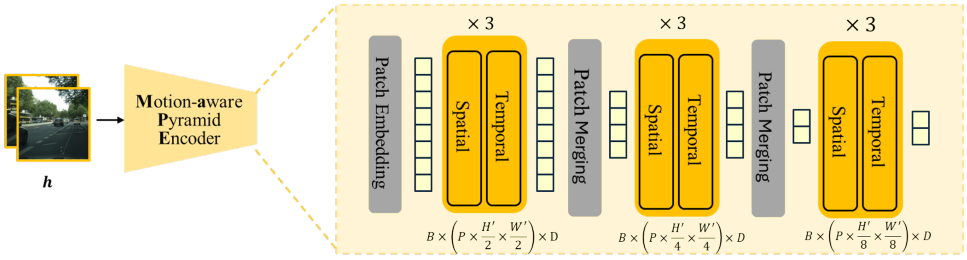

Fig. 2. MaPE architecture

where σ controls the corruption level. The resulting noisy latent is conditionally Gaussian: $\boldsymbol{z}_\sigma \mid \boldsymbol{z} \sim \mathcal{N}(\boldsymbol{z}, \sigma^2 I)$. A joint input, $\boldsymbol{j}_\sigma$, is defined as:

$$\boldsymbol{j}_\sigma = \mathrm{Concat}(\boldsymbol{h}, \boldsymbol{z}_\sigma), \tag{3}$$

which implicitly injects temporal context into the U-Net $\mathcal{E}$, allowing its internal spatio-temporal attention layers to jointly model and attend to the historical conditioning information.

To differentiate the deterministic nature of the observed domain from the stochasticity of the noisy domain, the time embedding for clean past frames and noisy future latent are defined as:

$$\begin{aligned} e_{\text{clean}}(\sigma) &= \mathrm{Embed}_{\text{clean}}(0.25 \log \sigma_{\min}), \\ e_{\text{noise}}(\sigma) &= \mathrm{Embed}_{\text{noise}}(0.25 \log \sigma), \end{aligned} \tag{4}$$

where $\mathrm{Embed}_{\text{clean}}(\cdot)$ and $\mathrm{Embed}_{\text{noise}}(\cdot)$ share the same architecture, a sinusoidal time embedding function, and are both initialized with the same pretrained weights. Then, we construct a binary mask $\boldsymbol{m} \in \{0,1\}^{(P+F)}$, where the first P positions (corresponding to historical frames) are set to 1, and the remaining F positions (future frames) are set to 0. This allows for defining the frame-wise time embedding as:

$$e(\sigma) = \boldsymbol{m} \cdot e_{\text{clean}}(\sigma) + (1 - \boldsymbol{m}) e_{\text{noise}}(\sigma). \tag{5}$$

Eqs. (4) and (5) means that $\boldsymbol{e}(\sigma)$ enforces a stationary, minimal noise level for historical frames, effectively distinguishing the clean and noisy domain to enhance the diffusion process's awareness of motion observed from past frames.

To predict the clean signal $\hat{\boldsymbol{x}}$ from the joint input $\boldsymbol{j}_\sigma$ conditioned on $\boldsymbol{h}$, the denoiser $\mathcal{D}$ is utilized:

$$\begin{aligned} \mathcal{D}(\boldsymbol{j}_\sigma; \sigma, \boldsymbol{h}) =& c_{\text{skip}}(\sigma) \boldsymbol{j}_\sigma \\ &+ c_{\text{out}}(\sigma) \mathcal{E}(c_{\text{in}}(\sigma) \boldsymbol{j}_\sigma; e(\sigma), \boldsymbol{h}), \end{aligned} \tag{6}$$

where $e(\cdot)$ is the time embedding defined in (5), and c_{skip}, c_{in}, and c_{out} are scale-dependent coefficients for normalization and conditioning:

$$\begin{aligned} c_{\text{skip}}(\sigma) &= 1/(\sigma^2 + 1), \\ c_{\text{in}}(\sigma) &= \sqrt{c_{\text{skip}}(\sigma)^2}, \; c_{\text{out}}(\sigma) = \sqrt{1 - c_{\text{in}}(\sigma)^2}. \end{aligned} \tag{7}$$

$\mathcal{E}$ in Eq. (6) is the diffusion model's encoder to be introduced in Sec. III-B, where the mechanism of explicitly using historical latent frames $\boldsymbol{h}$ to enhance video generation is introduced.

B. Motion-aware Pyramid Encoder

The Motion-aware Pyramid Encoder (MaPE), shown in Fig. 2, is a hierarchical spatio-temporal transformer encoder that convert the latent representation of historical video frames $\boldsymbol{h}$ into multi-scale token sequences. Following the design of SVD [23], $\mathcal{E}$ adopts a 3D U-Net architecture composed of down, mid, and up blocks of spatial and temporal layers:

$$\mathcal{E} = \mathrm{Up}(\mathrm{Mid}(\mathrm{Down}(\boldsymbol{j}_\sigma; \sigma, \boldsymbol{h}))). \tag{8}$$

$\boldsymbol{M}_s$, for $s = 1, 2, 3$, are token sequences produced by MaPE. Those tokens capture both local and global historical motion priors injected into $\mathcal{E}$ via cross-attention layers. $\boldsymbol{M}_s$ ($\in \mathbb{R}^{B \times N_s \times D}$) consists of N_s tokens, whose spatial token grids align with the corresponding grids used in the intermediate outputs of $\mathcal{E}$. As illustrated in Fig. 2, the framework of MaPE is composed of patch embedding module for tokenization, patch merging modules for pyramid downsampling, and three sequential blocks, each consisting of multiple spatio-temporal attention layers for capturing motion dynamics.

Concretely, patch embedding module divides the latent representation of historical frames, $\boldsymbol{h}$, into non-overlapping 2×2 patches and linearly projects them into a D-dimensional embedding space through a learnable patch embedding layer, where D is consistent with the hidden dimension of cross-attention layers in the U-Net. To better extract inner context and motion dynamics of historical condition cues, a stack of alternating spatio-temporal transformer blocks [25] is utilized. Driven by them, fine-grained local tokens $\boldsymbol{M}_1$ with short-term motion cues are denoted as:

$$\boldsymbol{M}_1 = (\mathrm{Spatial} \circ \mathrm{Temporal})^3(\mathrm{PatchEmbed}(\boldsymbol{h})), \tag{9}$$

where $\boldsymbol{M}_1 \in \mathbb{R}^{B \times N_1 \times D}$ is produced by 3 alternating spatio-temporal transformer blocks, and $N_1 = P \times \frac{H'}{4} \times \frac{W'}{4}$. Constructing a hierarchical feature pyramid by performing progressive 2×2 patch merging across stages facilitates the learning of structural and object-centric motion representation in $\boldsymbol{M}_2$,

$$\boldsymbol{M}_2 = (\mathrm{Spatial} \circ \mathrm{Temporal})^3(\mathrm{PatchMerg}(\boldsymbol{M}_1)), \tag{10}$$

where $\boldsymbol{M}_2 \in \mathbb{R}^{B \times N_2 \times D}$ and $N_2 = P \cdot \frac{H'}{4} \cdot \frac{W'}{4}$ tokens. Global semantic with long-term temporal dependencies are in $\boldsymbol{M}_3$,

$$\boldsymbol{M}_3 = (\mathrm{Spatial} \circ \mathrm{Temporal})^3(\mathrm{PatchMerg}(\boldsymbol{M}_2)), \tag{11}$$

where $\boldsymbol{M}_3 \in \mathbb{R}^{B\times N_3\times D}$ and $N_3 = P \cdot \frac{H'}{8} \cdot \frac{W'}{8}$.

These motion-aware tokens $[\boldsymbol{M}_1, \boldsymbol{M}_2, \boldsymbol{M}_3]$ are employed as conditional memory in U-Net through cross-attention layers at different depth. At each scale $s \in \{1, 2, 3\}$, the hidden state $\boldsymbol{Z}_s$ from each U-Net block serves as a sequence of query tokens, while the token sequences $\boldsymbol{M}_s$ act as the key-value memory. The attention operation at scale s is formulated as:

$$\begin{gathered} Q_s = \boldsymbol{W}_Q \boldsymbol{Z}_s, \quad K_s = \boldsymbol{W}_K \boldsymbol{M}_s, \quad V_s = \boldsymbol{W}_V \boldsymbol{M}_s, \\ \mathrm{Attn}_s = \mathrm{Softmax}(Q_s K_s^\top / \sqrt{d}), \end{gathered} \tag{12}$$

where $\boldsymbol{W}_Q$, $\boldsymbol{W}_K$, and $\boldsymbol{W}_V \in \mathbb{R}^{D\times d}$ are learnable matrices that project the query, key, and value token sequences into the attention subspace of dimension d.

The diffusion model then learns to reverse this process by training with the denoising score matching objective:

$$\mathcal{L}_{\text{diff}}(\theta) = \mathbb{E}_{\substack{\boldsymbol{z},\boldsymbol{h}\sim p_{\text{data}},\sigma\sim p(\sigma),\\ \boldsymbol{z}_\sigma|\boldsymbol{z}\sim\mathcal{N}(\boldsymbol{z},\sigma^2 I)}} \left[\lambda_\sigma \left\|\mathcal{D}_\theta(\boldsymbol{j}_\sigma;\sigma,\boldsymbol{h}) - \boldsymbol{z}\right\|_2^2\right], \tag{13}$$

where λ_σ balances noise levels and encourages prediction robustness across different σ values, p_{data} denotes the data distribution of clean video frames in latent space, and $p(\sigma)$ is defined as a discrete uniform distribution followed by EDM.

C. Self-conditioning

During inference, conditioning each denoising step on the model's previous generation, named self-conditioning, allows the model to review its historical trajectories and thus enhance motion-aware temporal consistency [27], [28]. Therefore, in the training stage, with probability p_{sc}, a forward pass is first performed without gradient updates. Then, the detached prediction is concatenated to the input along the channel dimension for a second forward pass on which gradients are computed. With probability $1 - p_{sc}$, we follow the default of SVD, the most recent historical frame is replicated across the temporal sequence and concatenated along the channel dimension. The empirical findings reported in W.A.L.T. [28] suggest setting p_{sc} as 0.9.

Specifically, following an n-step discrete noise schedule, the noise level σ_i decreases gradually, for $i = n, \ldots, 1$. At the beginning of the sampling process ($i = n$), a sequence of latent variables is initiated from a Gaussian distribution with variance $\sigma^2_{\max}$. Given conditioning frames $\boldsymbol{h}$, the denoiser then produces an estimate of the clean latent at each step. For the remaining steps ($i = n-1, \ldots, 1$), the latent variables are iteratively refined using the discrete EDM update, and the previous estimated variables are concatenated along channel dimension as the next input. After n denoising steps, the final clean future latent $\hat{\boldsymbol{z}}$ is decoded into RGB frames $\hat{\boldsymbol{x}}$ by a VAE decoder G, as follows:

$$\hat{\boldsymbol{x}} = G(\hat{\boldsymbol{z}}). \tag{14}$$

IV. Experiments and Results

A. Implementation Details

1) Setting: The proposed HMPDM was implemented using PyTorch 1.10.0 on a server equipped with an Nvidia L40S featuring 48 GB of memory. The diffusion model is trained for 10^5 steps using AdamW optimizer, with a batch size (B) of 4 and a learning rate 2×10^{-5}. The frozen VAE and the trainable U-Net are initialized with the pretrained weight from SVD [23].

2) Datasets: To evaluate the effectiveness of the proposed framework for driving video prediction, experiments are conducted on Cityscapes [29] and KITTI [30] datasets. Following the standard video prediction protocol, both datasets are resized to 128×128. Specifically, the Cityscapes dataset is partitioned into 2,975 training clips, 500 validation clips, and 1,525 test clips, each containing 30 consecutive frames. The KITTI dataset is divided into 759 training clips and 150 test clips, with each clip containing 9 frames.

3) Evaluation Metrics: Following the protocols established in prior work [14], [17], [21], four commonly used metrics are adopted to evaluate the performance of video prediction models: Structural Similarity Index Measure (SSIM) [31], Peak Signal-to-Noise Ratio (PSNR), Learned Perceptual Image Patch Similarity (LPIPS) [32], and Fréchet Video Distance (FVD) [33]. SSIM and PSNR evaluate pixel-wise reconstruction fidelity, while LPIPS measures perceptual similarity in deep feature space. FVD captures the spatio-temporal consistency by quantifying both motion dynamics and temporal coherence. For fair comparison, all metrics are reported on both a randomly sampled subset of 256 clips and the full test set across 10 random future denoising trajectories (#T), from which the best performing result is selected. To compare computational complexity, we additionally report the input modality, and the number of stages in the training pipeline.

B. Comparison with State-of-the-art Models

To evaluate the effectiveness of the proposed HMPDM on driving video prediction, we benchmark it against SOTA models on Cityscapes and KITTI datasets following the two standard protocols, with results summarized in Table III-B. Under a strictly RGB-only input setting, HMPDM demonstrates superior performance on Cityscapes. Specifically, it achieves an FVD of 151.2 on the 256-sample test protocol, representing a 17.8% relative reduction compared to MCVD [14], and 77.0 on the full-test protocol, corresponding to a 28.2% relative reduction to STDiff [15]. Although HMPDM is limited to a single RGB modality, it attains competitive results even against multimodal models (e.g., R+F, R+D), ranking the second on the 256-sample test set while achieving the best FVD on the full test set. Furthermore, HMPDM retains a one-stage training pipeline and requires fewer optimization steps, emphasizing its efficiency and compact architecture.

In contrast to FVD, HMPDM's performances on SSIM, PSNR, and LPIPS illustrate comparable yet slightly inferior results to the SOTA models. This disparity primarily stems from the inference mechanism. These evaluation metrics are calculated per frame, whereas HMPDM performs a one-time prediction of the entire future segment. Conversely, SOTA methods [14], [17], [20], [21] adopt autoregressive rollouts with shorter prediction horizons, which attenuate the drift,

TABLE I
QUANTITATIVE COMPARISON OF VIDEO PREDICTION MODELS ON CITYSCAPES AND KITTI DATASETS. ↓ MEANS LOWER IS BETTER, ↑ MEANS HIGHER IS BETTER

Methods	Year	Input	Pipeline	Cityscapes(128×128) 2 → 28				KITTI(128×128) 4 → 5		
				FVD ↓	SSIM ↑	PSNR ↑	LPIPS ↓	SSIM ↑	PSNR ↑	LPIPS ↓
256 Random Samples										
U-ViT [26]	CVPR23	R	1	1045.3	0.362	10.84	0.431	–	–	–
RaMViD [13]	TMLR24	R	1	812.6	0.454	13.14	0.395	–	–	–
VDIM [16]	AAAI23	R	2	724.7	0.539	18.49	0.252	–	–	–
RVD [12]	ArXiv22	R	1	465.0	0.489	17.21	0.226	–	–	–
MCVD-s* [14]	NeurIPS22	R	1	184.8	0.720	22.50	0.121	–	–	–
LFDM [22]	CVPR23	R+F	2	194.9	0.601	20.32	0.157	–	–	–
ExtDM-K4 [21]	CVPR24	R+F	2	**121.3**	**0.745**	**22.84**	**0.108**	–	–	–
HMPDM (Ours)*	IV26	R	1	151.2	0.633	21.42	0.142	–	–	–
Full Test Set										
NPVP [18]	CVPR23	R	2	768.0	**0.744**	–	0.183	**0.66**	–	0.279
LGC-VD [20]	IJCAI23	R	2	124.6	0.732	–	**0.069**	–	–	–
STDiff* [15]	AAAI24	R	1	107.3	0.658	–	0.136	0.54	–	**0.115**
SyncVP* [17]	CVPR25	R+D	2	84.0	0.649	–	0.160	–	–	–
HMPDM (Ours)*	IV26	R	1	**77.0**	0.626	**21.37**	0.145	0.54	**18.62**	0.149

* Results computed with #T=10; unmarked entries use #T=100 or the original paper's default is unspecified.
R = RGB; F = optical flow; D = depth.

TABLE II
QUANTITATIVE ABLATION RESULTS ON CITYSCAPES. ↓ MEANS LOWER IS BETTER, ↑ MEANS HIGHER IS BETTER

Models	Cityscapes(128×128) 2 → 28			
	FVD ↓	SSIM ↑	PSNR ↑	LPIPS ↓
Baseline	236.2	0.574	19.94	0.193
+ TaLC	197.5	0.627	21.42	0.153
+ TaLC + MP	188.7	0.632	**21.52**	0.154
+ TaLC + MaPE	155.7	**0.634**	21.44	0.146
+ TaLC + MaPE + SC	**151.2**	0.633	21.42	**0.142**

even in cases where the long-term motion consistency metric FVD is not superior.

To further assess the generalization of HMPDM in autonomous driving scenarios, we evaluate on KITTI with a short prediction horizon (4→5). Given the limited number of predicted frames, FVD is omitted for this dataset and we only report SSIM, PSNR, and LPIPS. The performance of high PSNR and competitive SSIM indicate strong pixel-level alignment and high structural similarity, while the LPIPS gap relative to the best method suggests remaining room in texture reconstruction. Notably, since HMPDM learns and samples in the VAE latent space at a low dimensional space (128 × 128), high-frequency details may be smoothed, contributing to its modest deficit on LPIPS.

C. Ablation Study

To assess the ability of the HMPDM framework in modeling historical motion priors and video prediction, experiments are conducted on the Cityscapes dataset. Sec. IV-C1 quantifies contribution of each component within HMPDM, while Sec. IV-C2 examines how the conditioning horizon influences prediction quality.

1) Effectiveness of Components: Table II summarizes the contribution of each component for predicting 28 future frames conditioned on 2 past frames. The architecture and pretrained weights of baseline model are identical to those of SVD [23] and then fine-tuned on the Cityscapes dataset. Compared to this baseline model, adding the TaLC module yields 16.4% lower FVD, 9.2% higher SSIM, 7.4% higher PSNR, and 21% lower LPIPS. These improvements indicate that the TaLC effectively leverages implicit historical context, leading to improved motion dynamics and stronger structural fidelity.

We further evaluate the MaPE module with respect to its pyramid design and patch merging mechanism. First, the variant (+TaLC+MP) improves over the variant (+TaLC) by 4.5% in FVD, highlighting the importance of multi-scale conditioning injected through cross-attention layers for capturing the semantically relevant features. +TaLC+MaPE, which integrates past frames within the spatio-temporal domain through stacked transformers, achieves a 17.5% reduction in FVD compared with the variant (+TaLC+MP), illustrating the effectiveness of the designed transformer blocks.

Finally, adding SC upon the variant (+TaLC+MaPE) constitutes the complete HMPDM framework, which delivers the best FVD and LPIPS and maintains competitive SSIM and PSNR. This strategy mitigates error accumulation and enhances temporal consistency.

The contribution of each designed component is further visualized in Fig. 3. Qualitative comparisons demonstrate that the TaLC enhances the spatio-temporal consistency, as evidenced by improved spatial structure. Specifically, in both samples, the front vehicle occupies lager spatial area in the frames generated by the baseline model than in the corresponding ground truth frames. Furthermore, the addition of MaPE allows HMPDM to better capture the historical motion priors. Compared with predicted frames generated

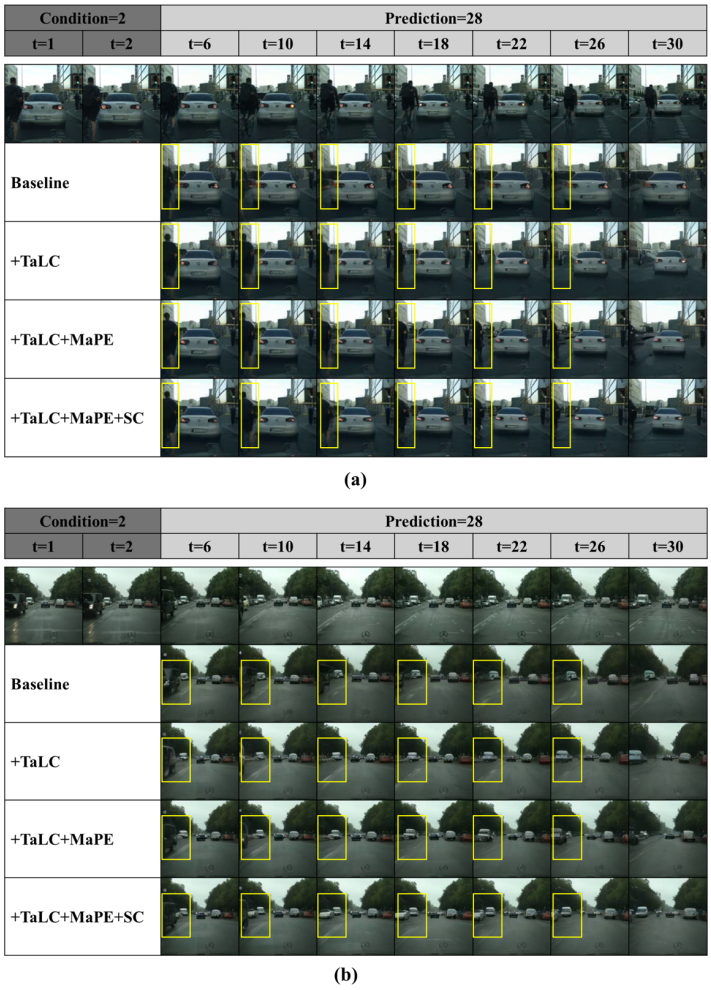

Fig. 3. Qualitative Ablation Results on Cityscapes (128× 128)

by (+TaLC) in sample (a), the (+TaLC+MaPE) variant produces clearer motion patterns for the cyclist, as evidenced by more visually coherent leg movements. Integrating the variant (+TaLC+MaPE) with SC, further enables HMPDM to handle occlusions and preserve the fidelity of traffic agents. As shown in sample (b), HMPDM successfully generates the stationary white sedan even though it never appears in the past frames. These improvements enhance the quality of generated traffic videos, thus providing more realistic and reliable information to support vehicles in predicting and reasoning about future scenes.

2) Effectiveness of Conditioning Horizon: Intuitively, providing sufficient historical motion information enhances the dynamic priors required for future frame generation. To investigate the impact of varying conditioning horizons on future driving video prediction, quantitative results on the Cityscapes dataset are reported in Table III. When increasing the number of historical frames from 2 to 6 while keeping the same prediction length, the FVD decreases from 117.4 to 104.9 (≈10.7%↓). Simultaneously, SSIM improves from 0.679 to 0.694, PSNR increases from 22.92 to 23.37 dB, and LPIPS concurrently decreases from 0.117 to 0.110. These improvements indicate that supplying richer motion context advances both temporal coherence and frame-wise fidelity, with diminishing but still positive returns as the conditioning length grows.

TABLE III
EFFECTIVENESS OF CONDITIONING HORIZON ON VIDEO PREDICTION PERFORMANCE. ↓ MEANS LOWER IS BETTER, ↑ MEANS HIGHER IS BETTER

Conditioning Horizon	**Cityscapes**(128×128)			
	FVD ↓	**SSIM** ↑	**PSNR** ↑	**LPIPS** ↓
$2 \rightarrow 14$	117.4	0.679	22.92	0.117
$4 \rightarrow 14$	108.9	0.691	23.24	0.112
$6 \rightarrow 14$	104.9	0.694	23.37	0.110

V. CONCLUSION

This paper proposes HMPDM, a diffusion-based video prediction model enhanced by mono-modal historical motion

priors and specifically tailored for driving scenarios. TaLC and MaPE modules effectively capture and utilize the historical motion priors from past driving patterns. In addition, SC improves the fidelity of traffic agents and long-term temporal consistency. The proposed HMPDM framework, owing to its efficiency and well-designed historical motion priors, achieves superior performance on the Cityscapes benchmark, outperforming existing methods by 17.8% and 28.2% in FVD under two standard evaluation protocols, respectively.

The HMPDM lays a foundation for the perception and prediction in emerging driving world models. Future work will explore efficient multimodal extensions and conduct deployment-oriented evaluations to improve robustness for real-world utilization.

ACKNOWLEDGMENT

Qin receives funding (69A3552348321) from the Rural Safe Efficient Advanced Transportation Center (R-SEAT), funded by the US Department of Transportation (USDOT). The contents of this paper reflect the views of the authors. USDOT assumes no liability for the contents or use thereof.

REFERENCES

[1] K. Li, C. Zhang, Y. Ding, X. Hu, and R. Qin, "Multi-label scene classification for autonomous vehicles: Acquiring and accumulating knowledge from diverse datasets," *arXiv e-prints*, pp. arXiv–2506, 2025.

[2] S. Teng, X. Hu, P. Deng, B. Li, Y. Li, Y. Ai, D. Yang, L. Li, Z. Xuanyuan, F. Zhu, and L. Chen, "Motion planning for autonomous driving: The state of the art and future perspectives," *IEEE Transactions on Intelligent Vehicles*, vol. 8, no. 6, pp. 3692–3711, 2023.

[3] R. Villegas, J. Yang, S. Hong, X. Lin, and H. Lee, "Decomposing motion and content for natural video sequence prediction," in *International Conference on Learning Representations*, 2017.

[4] W. Lotter, G. Kreiman, and D. Cox, "Deep predictive coding networks for video prediction and unsupervised learning," in *International Conference on Learning Representations*, 2017.

[5] L. Castrejon, N. Ballas, and A. Courville, "Improved conditional vrnns for video prediction," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2019.

[6] B. Wu, S. Nair, R. Martin-Martin, L. Fei-Fei, and C. Finn, "Greedy hierarchical variational autoencoders for large-scale video prediction," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2021, pp. 2318–2328.

[7] E. Denton and R. Fergus, "Stochastic video generation with a learned prior," in *International Conference on Machine Learning*. PMLR, 2018, pp. 1174–1183.

[8] J. Ho, A. Jain, and P. Abbeel, "Denoising diffusion probabilistic models," *Advances in Neural Information Processing Systems*, vol. 33, pp. 6840–6851, 2020.

[9] J. Song, C. Meng, and S. Ermon, "Denoising diffusion implicit models," in *International Conference on Learning Representations*, 2021.

[10] T. Karras, M. Aittala, T. Aila, and S. Laine, "Elucidating the design space of diffusion-based generative models," in *Advances in Neural Information Processing Systems*, S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, Eds., vol. 35. Curran Associates, Inc., 2022, pp. 26 565–26 577.

[11] J. Ho, T. Salimans, A. Gritsenko, W. Chan, M. Norouzi, and D. J. Fleet, "Video diffusion models," in *Advances in Neural Information Processing Systems*, S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, Eds., vol. 35. Curran Associates, Inc., 2022, pp. 8633–8646.

[12] R. Yang, P. Srivastava, and S. Mandt, "Diffusion probabilistic modeling for video generation," *Entropy*, vol. 25, no. 10, p. 1469, 2023.

[13] T. Höppe, A. Mehrjou, S. Bauer, D. Nielsen, and A. Dittadi, "Diffusion models for video prediction and infilling," *Transactions on Machine Learning Research*, vol. 2022-November, 2022.

[14] V. Voleti, A. Jolicoeur-Martineau, and C. Pal, "Mcvd - masked conditional video diffusion for prediction, generation, and interpolation," in *Advances in Neural Information Processing Systems*, S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, Eds., vol. 35. Curran Associates, Inc., 2022, pp. 23 371–23 385.

[15] X. Ye and G.-A. Bilodeau, "STDiff: Spatio-temporal diffusion for continuous stochastic video prediction," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 38, no. 7, 2024, pp. 6666–6674.

[16] K. Mei and V. Patel, "VIDM: Video implicit diffusion models," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 37, no. 8, 2023, pp. 9117–9125.

[17] E. Pallotta, S. M. Azar, S. Li, O. Zatsarynna, and J. Gall, "SyncVP: Joint diffusion for synchronous multi-modal video prediction," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2025, pp. 13 787–13 797.

[18] X. Ye and G.-A. Bilodeau, "A unified model for continuous conditional video prediction," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 3604–3613.

[19] H. Lu, G. Yang, N. Fei, Y. Huo, Z. Lu, P. Luo, and M. Ding, "VDT: General-purpose video diffusion transformers via mask modeling," in *The 12th International Conference on Learning Representations*, 2024.

[20] S. Yang, L. Zhang, Y. Liu, Z. Jiang, and Y. He, "Video diffusion models with local-global context guidance," in *Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence*, ser. IJCAI '23, 2023.

[21] Z. Zhang, J. Hu, W. Cheng, D. Paudel, and J. Yang, "Extdm: Distribution extrapolation diffusion model for video prediction," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2024, pp. 19 310–19 320.

[22] H. Ni, C. Shi, K. Li, S. X. Huang, and M. R. Min, "Conditional image-to-video generation with latent flow diffusion models," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2023, pp. 18 444–18 455.

[23] A. Blattmann, T. Dockhorn, S. Kulal, D. Mendelevitch, M. Kilian, D. Lorenz, Y. Levi, Z. English, V. Voleti, A. Letts, V. Jampani, and R. Rombach, "Stable video diffusion: Scaling latent video diffusion models to large datasets," *CoRR*, vol. abs/2311.15127, 2023.

[24] S. Gao, J. Yang, L. Chen, K. Chitta, Y. Qiu, A. Geiger, J. Zhang, and H. Li, "Vista: A generalizable driving world model with high fidelity and versatile controllability," *Advances in Neural Information Processing Systems*, vol. 37, pp. 91 560–91 596, 2024.

[25] X. Ma, Y. Wang, X. Chen, G. Jia, Z. Liu, Y.-F. Li, C. Chen, and Y. Qiao, "Latte: Latent diffusion transformer for video generation," *Transactions on Machine Learning Research*, 2025.

[26] F. Bao, S. Nie, K. Xue, Y. Cao, C. Li, H. Su, and J. Zhu, "All are worth words: A vit backbone for diffusion models," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 22 669–22 679.

[27] T. Chen, R. ZHANG, and G. Hinton, "Analog bits: Generating discrete data using diffusion models with self-conditioning," in *The Eleventh International Conference on Learning Representations*, 2023.

[28] A. Gupta, L. Yu, K. Sohn, X. Gu, M. Hahn, F.-F. Li, I. Essa, L. Jiang, and J. Lezama, "Photorealistic video generation with diffusion models," in *European Conference on Computer Vision*. Springer, 2024, pp. 393–411.

[29] M. Cordts, M. Omran, S. Ramos, T. Rehfeld, M. Enzweiler, R. Benenson, U. Franke, S. Roth, and B. Schiele, "The cityscapes dataset for semantic urban scene understanding," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 3213–3223.

[30] A. Geiger, P. Lenz, C. Stiller, and R. Urtasun, "Vision meets robotics: The kitti dataset," *The international journal of robotics research*, vol. 32, no. 11, pp. 1231–1237, 2013.

[31] Z. Wang, A. Bovik, H. Sheikh, and E. Simoncelli, "Image quality assessment: from error visibility to structural similarity," *IEEE Transactions on Image Processing*, vol. 13, no. 4, pp. 600–612, 2004.

[32] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang, "The unreasonable effectiveness of deep features as a perceptual metric," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

[33] T. Unterthiner, S. Van Steenkiste, K. Kurach, R. Marinier, M. Michalski, and S. Gelly, "Towards accurate generative models of video: A new metric & challenges," *arXiv preprint arXiv:1812.01717*, 2018.