LaMo: A Latent Motion World Model for Long-Horizon Prediction

Azwar Abdulsalam
azwarabdulsalam1729@gmail.com

Christopher Hoang
New York University
ch3451@nyu.edu

Mengye Ren
New York University
mengye@nyu.edu

Abstract

Visual world models learned from video are often trained to predict future observations from past frames, but pixel-space forecasting is expensive and forces the model to allocate capacity to high-frequency details. Latent-space video models reduce this cost, yet they typically predict a dense latent state at every step, repeatedly carrying largely static information and making long-horizon rollouts harder to sustain. We propose LaMo, probabilistic latent dynamics world model that instead predicts *motion tokens*—continuous latent variables that encode first-order temporal changes in a learned latent space. Given the current latent state and a short history of past motion tokens, our model autoregressively samples the next motion token and recurrently advances the latent state using a forward dynamics model. We instantiate the same contextual transformer backbone with two potential probabilistic parameterizations for next-token prediction using Gaussian mixtures and conditional flow matching We evaluate long-horizon rollouts on BDD100K, showing improved long-horizon modeling in complex driving scenes.

1 Introduction

Predictive world models enable embodied agents to *imagine* potential trajectories, enabling planning and control. For these agents to be useful over long horizons, world model rollouts must remain temporally coherent rather than drifting or hallucinating (Janner et al., 2019). Long-horizon prediction is therefore central to video-based world models for embodied decision-making (Ha & Schmidhuber, 2018; Hafner et al., 2019; 2020; Yang et al., 2023; Jang et al., 2025). Recent generative models operate in latent space to mitigate the cost of pixel-level forecasting (Blattmann et al., 2023a; Bar-Tal et al., 2024; Ma et al., 2025; Assran et al., 2025). However, these latent spaces are typically reconstructed via autoencoders that retain low-level details and result in dense token representations per frame. This can introduce significant redundancy as videos are typically dominated by static content, which can saturate model context and increase computational cost. Consequently, the temporal consistency and generation quality of these generative models can degrade over longer horizons (Xie et al., 2024; Cai et al., 2025; Yang et al., 2025).

To tackle the long-horizon prediction problem, we propose to capitalize on the efficiency advantages of modeling *latent dynamics* over latent states. Intuitively, dynamics is often of lower complexity than appearance; e.g., a rigid object undergoes coherent motion governed by a small number of shared degrees of freedom, so isolating dynamics can offer substantial compression benefits. Therefore, we aim to model videos by representing their first-order temporal changes as latent *motion tokens*. Latent motion tokens, also known as latent actions, stem from a recent line of work (Bruce et al., 2024; Schmidt & Jiang, 2024a; Ye et al., 2025; Cui et al., 2024) that seeks to learn a compressed representation of transition dynamics and models that can predict future observations conditioned on these latent representations. We specifically adopt Midway Network (Hoang & Ren, 2026), which employs a hierarchical formulation that refines latent motion tokens across multiple feature levels to handle complex scene dynamics in natural videos. However, Midway Network and other latent motion methods are restricted to inference between observed frames or only one-step prediction. To bridge this gap, we propose to learn a conditional generative model that predicts latent motion tokens over long horizons given a short history of motion tokens and the current latent state.

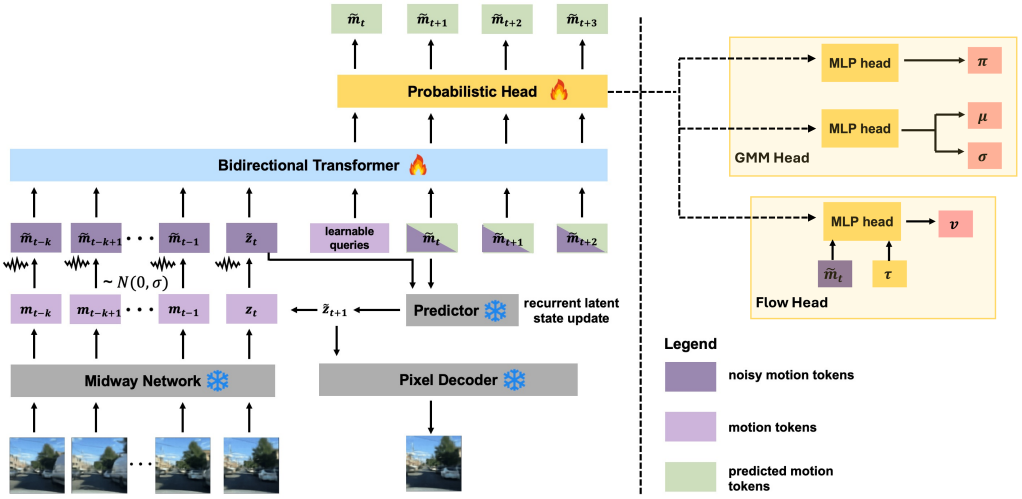

Figure 1: **LaMo overview: long-horizon rollout by predicting latent motion tokens.** A pretrained latent dynamics model (Midway Network) encodes frames into latent states z_t and extracts motion tokens m_t between consecutive frames. During training, we corrupt the conditioning motion token history and current latent state with Gaussian noise (dark purple) and feed them, together with learnable query tokens, into a bidirectional transformer and trained with teacher forcing. A lightweight probabilistic head then predicts the next motion token using either a *GMM* parameterization (global mixture weights π shared across tokens, and per-token (μ, σ)) or *conditional flow matching* (predicting a velocity field v for ODE-based sampling). At inference, predicted motion tokens $\hat{m}_t$ (green) are fed into the frozen forward dynamics predictor to update the latent state recurrently, and a separate decoder maps predicted latents back to pixel space.

We introduce LaMo, a probabilistic world model designed for long-horizon latent motion prediction. First, we use an autoregressive transformer backbone to process the context inputs and predict future latents. Then, to capture the inherent multi-modality of predicting future dynamics (e.g., a car turning left or right), we employ a probabilistic predictor head to generate the next motion token from the predicted latents. We investigate two formulations of the predictor head: **(a)** a Gaussian mixture model (GMM) trained to maximize conditional likelihood and **(b)** a flow matching head that learns a conditional velocity field. To improve robustness under autoregressive rollouts, we apply context-noise augmentation during training by injecting Gaussian noise into the conditioning motion token history and current latent state, encouraging the predictor to remain stable when conditioned on its own imperfect past predictions (Pasini et al., 2024; Valevski et al., 2024). We maintain a recurrent state by iteratively updating the current latent state features with the predicted motion tokens via a forward dynamics predictor. The resulting trajectory of latent states can be decoded for visual inspection or used towards downstream tasks.

In our experiments on BDD100K (Yu et al., 2020), a large-scale driving dataset, we demonstrate that LaMo produces more accurate long-horizon rollouts compared to an autoregressive transformer that conditions on and predicts dense latent states, demonstrating the benefit of modeling latent motion over latent states. We also observe that the two probabilistic head variants offer distinct benefits: the flow-matching head achieves lower FID across time horizons, while the GMM head attains better PSNR and SSIM in long-horizon settings.

To summarize, our contributions are:

- We propose LaMo, a probabilistic latent dynamics world model for long-horizon prediction that represents videos as sequences of compact *motion tokens* capturing first-order, multi-modal temporal changes in latent space.
- We show that conditioning on motion token history yields more accurate and stable long-horizon rollouts than a state-prediction baseline that conditions on dense latent state history.

2 Related Work

Future prediction. Neuroscience proposes that intelligent agents predict future sensory inputs to perceive and act, establishing predictive coding (Srinivasan et al., 1982; Rao & Ballard, 1999) and forward models (Wolpert et al., 1995; Miall & Wolpert, 1996) as central mechanisms in biological vision and motor control. In deep learning, this has motivated a long line of works in video prediction. Early attempts leveraged a variety of LSTM (Srivastava et al., 2015), CNN (Mathieu et al.), and motion-focused (Finn et al., 2016) architectures and showed that trained models could be useful towards video generation, action recognition, and robotic control tasks. These eventually evolved into today's modern video generators like Sora (Brooks et al., 2024), Veo (Google DeepMind, 2025), and Wan (Wan et al., 2025), which incorporated advances such as transformer architectures (Dosovitskiy et al., 2021; Peebles & Xie, 2022), diffusion modeling (Ho et al., 2020; Gupta et al., 2024), and massive data and parameter scaling to achieve highly realistic and physically plausible generations. However, consistent long-horizon prediction remains challenging due to computational cost and context limits. Prior works have turned to spatiotemporal tokenization (Yan et al., 2021; Villegas et al., 2022), and latent space diffusion (Blattmann et al., 2023b;a) for improved token efficiency. Yet, these approaches often still fail to exploit temporal redundancy, where the number of allocated latent tokens per frame still grows prohibitively high in long-horizon settings. In this work, we address this gap by explicitly disentangling videos as sparse latent states that are transformed by sequences of highly compressed latent motion representations.

Latent dynamics. Deep learning has aimed to acquire dynamics models that emulate how animals can predict their environment's future (Ha & Schmidhuber, 2018). While initial works pursued this goal through supervised learning with action-labeled data (Agrawal et al., 2016; Pathak et al., 2017), more recent efforts have shifted towards latent dynamics models, which can self-supervised learn from unlabeled videos and rely on either discretization or dimension-based information bottlenecks to prevent representation collapse (Garrido et al., 2026). These latent dynamics models can be used in a variety of ways, from data labeling (Schmidt & Jiang, 2024a; Jang et al., 2025; Ye et al., 2025) to environment simulation (Bruce et al., 2024; Gao et al., 2025). In particular, our work leverages Midway Network (Hoang & Ren, 2026), a recent latent dynamics architecture that uses hierarchical refinement and dense forward prediction to infer latent motion tokens that capture the dynamics of natural videos. While prior works either infer latent actions post-hoc or predict one step in the future, we aim to learn a model that can predict future latent motion tokens over long horizons and use them to evolve a sparse set of latent states over time. Closely related to this work is Moto (Chen et al., 2024b), which learns an autoregressive predictive model of latent motion tokens, though distinctly for the purpose of pretraining motion priors to enhance downstream robotic control finetuning.

3 Background

Motion tokens. Motion tokens are learned latent vectors designed to capture transition dynamics between source and target states z_t and z_{t+1}. They are typically used in latent dynamics modeling frameworks, where they are inferred by an inverse dynamics model, $m_t = f_{\text{inv}}(z_t, z_{t+1})$ and used by a forward dynamics model to predict future states, $\hat{z}_{t=1} = f_{\text{fwd}}(z_t, m_t)$. To prevent representation collapse, where m_t trivially compresses z_{t+1} rather than learning the underlying dynamics, prior methods apply information bottlenecks like VQ-VAE discretization (Schmidt & Jiang, 2024b; Bruce et al., 2024; Ye et al., 2025) or dimensionality reduction (Cui et al., 2024; Gao et al., 2025; Hoang & Ren, 2026).

Midway Network (Hoang & Ren, 2026) extends this paradigm to a hierarchical framework. Motivated by coarse-to-fine refinement in optical flow methods (Sun et al., 2018; Jonschkowski et al., 2020), Midway Network iteratively refines motion tokens in a top-down manner. For feature levels $\{l^1, l^2, ..., l^k\}$, motion tokens m^l are computed as an accumulation of higher level motion tokens plus a learned residual:

$$m^l = m^{l+1} + f_{\text{inv}}(\hat{z}^{l+1}_{t+1}, z^l_{t+1}). \tag{1}$$

Crucially, the inverse model is conditioned on the higher-level forward prediction $\hat{z}^{l+1}_{t+1}$ rather than source features, effectively forcing the network to learn residual dynamics to correct prediction errors from the previous scale.

Conditional flow matching. Flow-based generative models (Chen et al., 2018; Grathwohl et al., 2019) represent a target distribution by learning a time-dependent velocity field that transports samples from a simple base distribution to the data distribution via an ODE,

$$\frac{dx(\tau)}{d\tau} = v_\theta(x(\tau), \tau \mid c)\,, \tag{2}$$

where c denotes conditioning information (e.g., context from previous observations). A widely-used training approach is *flow matching* (Lipman et al., 2023), which regresses v_θ to the vector field induced by a chosen probability path between base and data samples. A particularly simple and effective choice is the straight-line ("rectified") path (Liu et al., 2022):

$$x_\tau = (1-\tau)x_0 + \tau x_1, \qquad x_0 \sim \mathcal{N}(0, I),\ x_1 \sim p(\cdot \mid c), \tag{3}$$

which yields the target velocity $v^*(x_\tau, \tau, c) = x_1 - x_0$. The conditional flow-matching objective is then

$$\mathcal{L}_{\text{CFM}} = \mathbb{E}_{x_0, x_1, \tau}\Big[\big\|v_\theta(x_\tau, \tau \mid c) - (x_1 - x_0)\big\|_2^2\Big]. \tag{4}$$

At inference time, one samples $x(0) \sim \mathcal{N}(0, I)$ and numerically integrates the ODE forward to obtain $x(1)$.

Continuous autoregressive modeling. Recent work has highlighted that sequence models need not rely on discrete codebooks: one can instead operate on *continuous-valued tokens* and attach a lightweight generative head that models the conditional distribution of each token given contextual features produced by the sequence backbone (Li et al., 2024; Chen et al., 2024a). This decouples *inter-token dependency modeling* (handled by the backbone) from *per-token distribution modeling* (handled by a small conditional generator) and motivates using conditional density models, including mixtures and flow-based objectives, for token prediction in continuous spaces.

4 LaMo: A Latent Motion World Model

LaMo performs long-horizon prediction by forecasting *latent motion tokens*—compact latent variables that describe the temporal dynamics between frames—instead of repeatedly predicting dense per-frame latent states. At each rollout step, LaMo (i) conditions on the current latent state and a short history of past motion tokens, (ii) samples the next motion token with a probabilistic predictor, and (iii) advances the latent state with a fixed forward predictor model; repeating this procedure yields long-horizon latent trajectories that can be decoded to pixels for visualization and evaluation. Figure 1 summarizes the full pipeline and the two interchangeable probabilistic parameterizations we study (GMM vs. conditional flow matching).

4.1 Setup: latent states and motion tokens

We factor long-horizon video prediction into (i) a pretrained latent dynamics model and (ii) a learnable stochastic sequence model over its motion token representation. In our experiments, the latent model is instantiated by the Midway Network (Hoang & Ren, 2026), which provides an encoder to extract a latent state representation z_t per frame, an inverse model that extracts motion tokens between consecutive states, and a forward predictor that advances the latent state given motion tokens. LaMo then learns a probabilistic motion token world model that predicts future motion tokens from a short history and the current state, enabling autoregressive rollouts in latent space.

Formally, let z_t denote the latent state (e.g., a set of latent tokens) at time t. The pretrained inverse dynamics model defines motion tokens as

$$m_t = f_{\text{inv}}(z_t, z_{t+1}), \tag{5}$$

and the corresponding pretrained forward transition updates the state via

$$z_{t+1} = f_{\text{fwd}}(z_t, m_t). \tag{6}$$

We keep f_{inv} and f_{fwd} fixed and learn a conditional distribution over the next motion token given the current state and a fixed-length motion history:

$$p_\theta(m_t \mid m_{t-K:t-1},\, z_t)\,. \tag{7}$$

4.2 AUTOREGRESSIVE BACKBONE

We use a bidirectional transformer encoder (Vaswani et al., 2017) over the fixed-length context window to produce *per-motion token features*—one feature vector for each motion token slot — which then parameterize a continuous conditional distribution over the next motion token.

We first embed the conditioning inputs into a common model dimension and add factorized embeddings indicating (i) time index within the context window, (ii) feature level, (iii) spatial patch index (when applicable), and (iv) token role (e.g., motion vs latent).

Learnable query tokens. To produce a context representation aligned with the motion token layout, we append a set of *learnable query tokens* arranged as a $(L \times N_m)$ grid, matching the structure of a motion token tensor. After self-attention over the full sequence, we take the transformer outputs at these query positions and reshape them into

$$h_t \in \mathbb{R}^{L \times N_m \times d}, \tag{8}$$

which provides one contextual feature vector per motion token position. These queries are learned parameters whose role is to extract per-motion token context features; they are trained end-to-end through the final prediction loss of the flow or GMM predictor head.

Global readout query (GMM only). For the GMM formulation, we additionally append a single learnable *global readout query*. The corresponding transformer output, $h_t^{\mathrm{g}} \in \mathbb{R}^d$, summarizes the context and is used to predict the mixture weights.

4.3 CONTEXT-NOISE AUGMENTATION

Autoregressive rollouts induce a train–test mismatch: at inference, the model conditions on its own previously predicted tokens and latents, which may contain small errors that can accumulate. To remedy this, we adopt context-noise augmentation during training by perturbing the *conditioning inputs* before they are embedded and passed to the transformer (Pasini et al., 2024; Valevski et al., 2024).

For the motion token history context, we corrupt both the motion token history and the current latent state features:

$$\tilde{m}_{t-K:t-1} = m_{t-K:t-1} + \sigma\, \epsilon_m, \qquad \tilde{z}_t = z_t + \sigma\, \epsilon_z, \tag{9}$$

with $\epsilon_m, \epsilon_z \sim \mathcal{N}(0, I)$. For the state-prediction baseline, we analogously corrupt the latent token history window $\tilde{z}_{t-K:t}$.

In our implementation, the noise scale $\sigma \sim \mathcal{U}(0, 1)$ is drawn per sample. Importantly, the model is *not provided* σ, i.e., it is not conditioned on the corruption level; the predictor must be robust to a range of plausible context errors.

4.4 TWO PROBABILISTIC HEADS FOR MOTION TOKEN PREDICTION

Using the same transformer-encoded context features, we study two conditional density parameterizations for m_t: a GMM head and a conditional flow-matching head.

4.4.1 CONDITIONAL FLOW MATCHING HEAD

We instantiate equation 7 using conditional flow matching. We treat the target motion token as $x_1 \equiv m_t$, draw $x_0 \sim \mathcal{N}(0, I)$ of matching shape, and sample $\tau \sim \mathcal{U}(0, 1)$. Using the straight path,

$$x_\tau = (1 - \tau)x_0 + \tau m_t, \tag{10}$$

we predict a conditional velocity field $v_\theta(x_\tau, \tau \mid c_t)$, where the context c_t is encoded by the transformer into motion-aligned features h_t. Concretely, we pass (x_τ, τ, h_t) to a lightweight per-token conditional head (AdaLN-modulated MLP) (Li et al., 2024) to obtain v_θ. The training loss is the conditional flow-matching objective:

$$\mathcal{L}_{\text{flow}} = \mathbb{E}\Big[\big\|v_\theta(x_\tau, \tau \mid c_t) - (m_t - x_0)\big\|_2^2\Big]. \tag{11}$$

At inference time, we sample $\hat{m}_t$ by integrating $\frac{dx}{d\tau} = v_\theta(x, \tau \mid c_t)$ forward from $x(0) \sim \mathcal{N}(0, I)$ using a fixed-step explicit integrator.

4.4.2 GAUSSIAN MIXTURE MODEL HEAD

Alternatively, we parameterize the conditional distribution of m_t as a Gaussian mixture with M components. Our GMM uses *global* mixture weights shared across all motion token positions and *position-wise* Gaussian parameters for each mixture component:

$$p_\theta(m_t \mid c_t) = \sum_{k=1}^{M} \pi_k(c_t) \prod_{l=1}^{L} \prod_{n=1}^{N_m} \mathcal{N}\big(m_t^{(l,n)};\, \mu_k^{(l,n)}(c_t),\, \mathrm{diag}(\sigma_k^{(l,n)}(c_t)^2)\big). \tag{12}$$

The transformer backbone provides (i) motion-aligned features h_t from the spatial prediction queries and (ii) a global context vector h_t^{g} from the global query token. A linear head on h_t^{g} predicts mixture logits and weights $\pi(c_t)$, while a per-position head on h_t predicts $\mu_k^{(l,n)}$ and $\log \sigma_k^{(l,n)}$ for all mixtures.

We train the backbone and GMM head by minimizing the negative log-likelihood, with an optional additional entropy regularizer on π):

$$\mathcal{L}_{\mathrm{gmm}} = -\mathbb{E}\big[\log p_\theta(m_t \mid c_t)\big] \;-\; \lambda\, \mathbb{E}\big[H(\pi(c_t))\big]. \tag{13}$$

At inference time, we sample a mixture index $k \sim \pi(c_t)$ (or take $\arg\max_k$) and then sample the motion tokens from the corresponding Gaussians.

4.5 STATE-PREDICTION BASELINE

LaMo predicts future dynamics by sampling the next motion token from a short motion token history and the current latent state, $p_\theta(m_t \mid m_{t-K:t-1}, z_t)$, and then advancing the latent state through a fixed transition model. To isolate the advantage of modeling motion tokens, we define a baseline that instead predicts the *next latent state directly* from a history of latent states:

$$p_\theta(z_{t+1} \mid z_{t-K:t}). \tag{14}$$

For a controlled comparison, we keep the sequence model capacity fixed by using the same transformer configuration and the same context length K; the only changes are (i) the signals provided in the context window (motion token history vs. latent state history) and (ii) the prediction target (m_t vs. z_{t+1}). Importantly, the latent state tokens z_t are substantially higher-dimensional and denser than motion tokens m_t (they retain more appearance/content information), so the state-prediction baseline must condition on and generate a much larger latent payload at each step.

4.6 AUTOREGRESSIVE ROLLOUT

At time t, we form a context c_t from a fixed-length window and use the transformer to parameterize a conditional distribution for the next motion token. We then sample $\hat{m}_t \sim p_\theta(\cdot \mid c_t)$ using either the flow head (§4.4.1) or the GMM head (§4.4.2) and advance the latent state via the frozen forward dynamics model:

$$\hat{z}_{t+1} = f_{\mathrm{fwd}}(\hat{z}_t, \hat{m}_t). \tag{15}$$

This state update acts as the model's *recurrent mechanism*: information from the past is carried forward through the evolving latent state $\hat{z}_t$, while stochasticity and multi-modality enter through the sampled motion tokens. This differs from standard RNN/LSTM (Hochreiter & Schmidhuber, 1997) formulations, where recurrence is implemented by learned hidden-state updates; here, recurrence is implemented by an explicit forward dynamics model f_{fwd}, and the learnable component focuses on modeling the conditional distribution of future dynamics.

We update the context using a sliding window and repeat this procedure to obtain long-horizon rollouts. In the motion token history setting, the window is over motion tokens and is advanced by appending $\hat{m}_t$; in the state-prediction baseline, the window is over latent states and is advanced by appending $\hat{z}_{t+1}$, with the baseline predicting $\hat{z}_{t+1}$ directly.

Decoding latents to pixels. To visualize rollouts in pixel space, we train a separate decoder that maps latent states back to pixels. Concretely, we fit a 6-layer transformer decoder d_ψ to reconstruct the next frame in pixel space from the latent representation via the mean squared error loss:

$$\mathcal{L}_{\mathrm{dec}} = \mathbb{E}\big[\|d_\psi(z_t) - x_t\|_2^2\big], \tag{16}$$

where x_t denotes the corresponding RGB frame. During evaluation, we decode predicted latent rollouts $\{\hat{z}_t\}$ using the frozen trained decoder d_ψ to obtain pixel-level predictions.

5 EXPERIMENTS

We evaluate LaMo on long-horizon autoregressive rollouts on BDD100K to test whether forecasting in terms of compact motion tokens yields more stable and accurate long-term predictions than forecasting dense latent states. Concretely, we compare motion token history conditioning against a matched state-prediction baseline (§4.5) under identical transformer capacity and context length, and report rollout quality on short (0.2s), medium (0.6s), and long-term horizons (1.2s) (Karypidis et al., 2025).

5.1 IMPLEMENTATION

Data and preprocessing. Following prior work on latent dynamics, we first train a Midway Network to (i) produce a frozen latent representation z_t for each frame and (ii) extract three levels of hierarchical motion tokens m_t with each level consisting of ten tokens describing transitions between consecutive latents. We then freeze the Midway Network and train separate probabilistic motion predictors to model the conditional distribution $p(m_t \mid m_{t-K:t-1}, z_t)$.

Models and training setup. All probabilistic heads share the same contextual transformer backbone described in §4.2. Throughout the paper, we use a motion token history of length $K{=}3$ together with the current latent token z_t. As a baseline, we train an architecture-matched variant that conditions on latent state history ($K{=}3$) rather than motion token history (see §4.5).

Both probabilistic heads are trained with context-noise augmentation (§4.3) to improve robustness under autoregressive rollouts. Training is performed on 4× H200 GPUs.

Flow head. For the conditional flow-matching model (§4.4.1), we use a 12-layer transformer backbone with model width $d = 1152$ and 8 attention heads and an AdaLN-modulated MLP flow head with 3 residual blocks. We train for 300 epochs with AdamW (Loshchilov & Hutter, 2019) at learning rate 1.5×10^{-4}. At inference time, motion tokens are sampled by fixed-step ODE integration; we evaluate different numbers of sampling steps and report their effect on rollout quality in Table 1.

GMM head. For the Gaussian mixture model variant (§4.4.2), we keep the same 12-layer, $d = 1152$ transformer backbone and predict a mixture of $K = 3$ Gaussians. The mixture weights are shared across motion token positions, while the Gaussian parameters are predicted per position. We train using the GMM negative log-likelihood objective with an entropy regularizer (§4.4.2), using AdamW at learning rate 1.5×10^{-4} for 300 epochs as well.

Evaluation protocol. We evaluate each model under autoregressive rollout using a fixed context window of length $K{=}3$ for each video evaluated. At each time step, the model predicts the next motion token $\hat{m}_t$ conditioned on the current latent token z_t and the history (motion token history for our method; latent state history for the baseline), then advances the latent state via a frozen forward dynamics model $\hat{z}_{t+1} = f_{\text{fwd}}(\hat{z}_t, \hat{m}_t)$ and decodes predictions to pixels with the trained decoder. We report image-space metrics, at three offsets: short-term ($t{=}0.2$s), medium-term ($t{=}0.6$s), and long-term ($t{=}1.6$s).

5.2 RESULTS

Motion token history conditioning improves long-horizon rollouts. Our main comparison is against an architecture-matched state-prediction baseline that predicts the next latent state $p_\theta(z_{t+1} \mid z_{t-K:t})$ and rolls out by feeding predicted latent states back into the context. In contrast, LaMo predicts the next motion token from motion token history and the current state, $\hat{m}_t \sim p_\theta(m_t \mid \hat{m}_{t-K:t-1}, \hat{z}_t)$, and advances the recurrent latent state via the frozen forward model $\hat{z}_{t+1} = f_{\text{fwd}}(\hat{z}_t, \hat{m}_t)$. Table 1 shows that replacing latent state history conditioning with motion token history conditioning yields markedly stronger long-horizon stability: at $t{=}1.6$s, the flow motion token history model reduces FID from 227 (state-prediction baseline) to 188 and improves

Figure 2: **Qualitative rollouts on BDD100K.** Yellow dashed region: conditioning context (observed frames). Remaining frames are autoregressive predictions decoded from predicted latents. Motion token–conditioned models better capture complex, scene-wide dynamics, including coordinated motion of multiple objects and structures (e.g., several cars, overhead signs, and lane markings). Top: GMM head conditioned on motion token history. Middle: Flow head with latent state history context (baseline). Bottom: Flow head conditioned on motion token history.

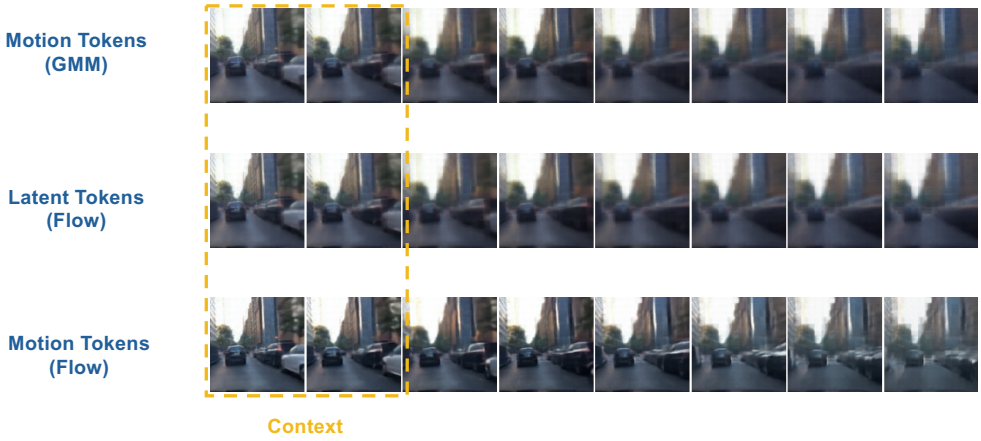

Figure 3: **Qualitative long-horizon rollouts on BDD100K.** Yellow dashed region: conditioning context (observed frames). Remaining frames are autoregressive predictions decoded from predicted latents. Top: GMM head conditioned on motion token history. Middle: Flow head with latent state history context (baseline). Bottom: Flow head conditioned on motion token history.

PSNR/SSIM from 14.9/0.4345 to 15.5/0.4653. These results support our central premise that explicitly representing dynamics via compact motion tokens yields a more predictable autoregressive context than conditioning on dense latent state history.

Qualitative rollouts preserve coherent scene evolution. Figure 3 illustrates the same trend visually. motion token history rollouts better preserve *coherent multi-object motion* over time—capturing the coupled motion of multiple cars together with background structure such as lane markings and signage—with fewer temporal artifacts and more stable long-range scene evolution than the state-prediction baseline.

Flow vs. GMM: a perception–distortion tradeoff across horizons. The flow head achieves lower FID across horizons, while the GMM head achieves higher long-horizon PSNR/SSIM (e.g.,

	Short-term (t=0.2s)			Med-term (t=0.6s)			Long-term (t=1.6s)		
Method	FID↓	PSNR↑	SSIM↑	FID↓	PSNR↑	SSIM↑	FID↓	PSNR↑	SSIM↑
Flow (state-prediction baseline)	58	23.3	0.7295	108	20.1	0.6385	227	14.9	0.4345
GMM	70	23.2	0.7277	113	**20.4**	**0.6474**	204	**16.9**	**0.5289**
Flow (5 steps)	61	23.3	0.7288	98	19.9	0.6165	238	15.2	0.4440
Flow (20 steps)	**52**	**23.7**	0.7426	74	20.2	0.6401	189	15.5	0.4665
Flow (50 steps)	**52**	**23.7**	**0.7427**	**72**	20.2	0.6395	**188**	15.5	0.4653

Table 1: We report short-term (t=0.2s), medium-term (t=0.6s), and long-term (t=1.6s). Lower FID is better; higher PSNR/SSIM are better.

16.9/0.5289 at t=1.6s), suggesting the head choice mainly affects which evaluation metrics are favored (Blau & Michaeli, 2018).

Effect of flow sampling steps. Increasing the number of ODE integration steps from 5 to 20 substantially improves quality, while gains saturate beyond 20 steps (Table 1). We therefore use 20 steps as a favorable compute–quality tradeoff unless otherwise stated.

6 Conclusion

We introduced **LaMo**, a probabilistic latent dynamics model that predicts *motion tokens*—compact latent variables representing the temporal dynamics between frames—instead of repeatedly regenerating dense per-frame latent states dominated by static content. By explicitly separating dynamics from content, LaMo reduces compounding drift and improves long-horizon rollouts.

On BDD100K, LaMo, by modeling latent dynamics, yields consistently more accurate rollouts than an architecture-matched *latent state history* baseline—a standard design in prior video world models that only predict in dense latent state space—improving perceptual quality (FID) and distortion metrics (PSNR/SSIM) at medium and long horizons. Looking ahead, LaMo could enable latent dynamics planning by sampling motion token rollouts and evaluating trajectories within an online model predictive control-style loop. Other future directions include designing a mechanism to update the latent state with ground-truth observations during rollout to account for new objects, and moving from two-stage training (latent dynamics model and latent motion predictor) to end-to-end training.

References

Pulkit Agrawal, Ashvin V Nair, Pieter Abbeel, Jitendra Malik, and Sergey Levine. Learning to poke by poking: Experiential learning of intuitive physics. In D. Lee, M. Sugiyama, U. Luxburg, I. Guyon, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 29. Curran Associates, Inc., 2016. URL https://proceedings.neurips.cc/paper_files/paper/2016/file/c203d8a151612acf12457e4d67635a95-Paper.pdf.

Mahmoud Assran, Adrien Bardes, David Fan, Quentin Garrido, Russell Howes, Mojtaba Komeili, Matthew Muckley, Ammar Rizvi, Claire Roberts, Koustuv Sinha, Artem Zholus, Sergio Arnaud, Abha Gejji, Ada Martin, Francois Robert Hogan, Daniel Dugas, Piotr Bojanowski, Vasil Khalidov, Patrick Labatut, Francisco Massa, Marc Szafraniec, Kapil Krishnakumar, Yong Li, Xiaodong Ma, Sarath Chandar, Franziska Meier, Yann LeCun, Michael Rabbat, and Nicolas Ballas. V-jepa 2: Self-supervised video models enable understanding, prediction and planning. *arXiv preprint arXiv:2506.09985*, 2025.

Omer Bar-Tal, Hila Chefer, Omer Tov, Charles Herrmann, Roni Paiss, Shiran Zada, Ariel Ephrat, Junhwa Hur, Guanghui Liu, Amit Raj, Yuanzhen Li, Michael Rubinstein, Tomer Michaeli, Oliver Wang, Deqing Sun, Tali Dekel, and Inbar Mosseri. Lumiere: A space-time diffusion model for video generation. In *SIGGRAPH Asia 2024 Conference Papers*, SA '24, New York, NY, USA, 2024. Association for Computing Machinery. ISBN 9798400711312. doi: 10.1145/3680528.3687614. URL https://doi.org/10.1145/3680528.3687614.

Andreas Blattmann, Tim Dockhorn, Sumith Kulal, Daniel Mendelevitch, Maciej Kilian, Dominik Lorenz, Yam Levi, Zion English, Vikram Voleti, Adam Letts, Varun Jampani, and Robin Rombach. Stable video diffusion: Scaling latent video diffusion models to large datasets. *CoRR*, abs/2311.15127, 2023a. URL https://doi.org/10.48550/arXiv.2311.15127.

Andreas Blattmann, Robin Rombach, Huan Ling, Tim Dockhorn, Seung Wook Kim, Sanja Fidler, and Karsten Kreis. Align your latents: High-resolution video synthesis with latent diffusion models. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023b.

Yochai Blau and Tomer Michaeli. The perception-distortion tradeoff. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6228–6237, 2018. doi: 10.1109/CVPR.2018.00652. URL https://openaccess.thecvf.com/content_cvpr_2018/papers/Blau_The_Perception-Distortion_Tradeoff_CVPR_2018_paper.pdf.

Tim Brooks, Bill Peebles, Connor Holmes, Will DePue, Yufei Guo, Li Jing, David Schnurr, Joe Taylor, Troy Luhman, Eric Luhman, Clarence Ng, Ricky Wang, and Aditya Ramesh. Video generation models as world simulators. 2024. URL https://openai.com/research/video-generation-models-as-world-simulators.

Jake Bruce, Michael D Dennis, Ashley Edwards, Jack Parker-Holder, Yuge Shi, Edward Hughes, Matthew Lai, Aditi Mavalankar, Richie Steigerwald, Chris Apps, Yusuf Aytar, Sarah Maria Elisabeth Bechtle, Feryal Behbahani, Stephanie C.Y. Chan, Nicolas Heess, Lucy Gonzalez, Simon Osindero, Sherjil Ozair, Scott Reed, Jingwei Zhang, Konrad Zolna, Jeff Clune, Nando De Freitas, Satinder Singh, and Tim Rocktäschel. Genie: Generative interactive environments. In *International Conference on Machine Learning*, 2024.

Shengqu Cai, Ceyuan Yang, Lvmin Zhang, Yuwei Guo, Junfei Xiao, Ziyan Yang, Yinghao Xu, Zhenheng Yang, Alan Yuille, Leonidas Guibas, Maneesh Agrawala, Lu Jiang, and Gordon Wetzstein. Mixture of contexts for long video generation. *arXiv preprint arXiv:2508.21058*, 2025. doi: 10.48550/arXiv.2508.21058. URL https://arxiv.org/abs/2508.21058.

Boyuan Chen, Diego Marti Monso, Yilun Du, Max Simchowitz, Russ Tedrake, and Vincent Sitzmann. Diffusion forcing: Next-token prediction meets full-sequence diffusion. *arXiv preprint arXiv:2407.01392*, 2024a. doi: 10.48550/arXiv.2407.01392. URL https://arxiv.org/abs/2407.01392.

Tian Qi Chen, Yulia Rubanova, Jesse Bettencourt, and David Duvenaud. Neural ordinary differential equations. In Samy Bengio, Hanna M. Wallach, Hugo Larochelle, Kristen Grauman, Nicolò Cesa-Bianchi, and Roman Garnett (eds.), *NeurIPS*, pp. 6572–6583, 2018. URL http://dblp.uni-trier.de/db/conf/nips/nips2018.html#ChenRBD18.

Yi Chen, Yuying Ge, Yizhuo Li, Yixiao Ge, Mingyu Ding, Ying Shan, and Xihui Liu. Moto: Latent motion token as the bridging language for robot manipulation. *arXiv preprint arXiv:2412.04445*, 2024b.

Zichen Jeff Cui, Hengkai Pan, Aadhithya Iyer, Siddhant Haldar, and Lerrel Pinto. Dynamo: In-domain dynamics pretraining for visuo-motor control. In *Advances in Neural Information Processing Systems*, 2024. URL https://openreview.net/forum?id=vUrOuc6NR3.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=YicbFdNTTy.

Chelsea Finn, Ian Goodfellow, and Sergey Levine. Unsupervised learning for physical interaction through video prediction. *Advances in Neural Information Processing Systems*, 2016.

Shenyuan Gao, Siyuan Zhou, Yilun Du, Jun Zhang, and Chuang Gan. Adaworld: Learning adaptable world models with latent actions. In *International Conference on Machine Learning (ICML)*, 2025.

Quentin Garrido, Tushar Nagarajan, Basile Terver, Nicolas Ballas, Yann LeCun, and Michael Rabbat. Learning latent action world models in the wild, 2026. URL https://arxiv.org/abs/2601.05230.

Google DeepMind. Veo 3 technical report. Technical report, Google DeepMind, may 2025. URL https://storage.googleapis.com/deepmind-media/veo/Veo-3-Tech-Report.pdf.

Will Grathwohl, Ricky T. Q. Chen, Jesse Bettencourt, Ilya Sutskever, and David Duvenaud. Ffjord: Free-form continuous dynamics for scalable reversible generative models. In *International Conference on Learning Representations (ICLR)*, 2019.

Agrim Gupta, Lijun Yu, Kihyuk Sohn, Xiuye Gu, Meera Hahn, Fei-Fei Li, Irfan Essa, Lu Jiang, and José Lezama. Photorealistic video generation withnbsp;diffusion models. In *Computer Vision – ECCV 2024: 18th European Conference, Milan, Italy, September 29–October 4, 2024, Proceedings, Part LXXIX*, pp. 393–411, Berlin, Heidelberg, 2024. Springer-Verlag. ISBN 978-3-031-72985-0. doi: 10.1007/978-3-031-72986-7_23. URL https://doi.org/10.1007/978-3-031-72986-7_23.

David Ha and Jürgen Schmidhuber. Recurrent world models facilitate policy evolution. In *Advances in Neural Information Processing Systems 31*, pp. 2451–2463. Curran Associates, Inc., 2018. URL https://papers.nips.cc/paper/7512-recurrent-world-models-facilitate-policy-evolution. https://worldmodels.github.io.

Danijar Hafner, Timothy P. Lillicrap, Ian Fischer, Ruben Villegas, David Ha, Honglak Lee, and James Davidson. Learning latent dynamics for planning from pixels. In *Proceedings of the 36th International Conference on Machine Learning (ICML)*, pp. 2555–2565, 2019. URL https://proceedings.mlr.press/v97/hafner19a.html.

Danijar Hafner, Timothy Lillicrap, Jimmy Ba, and Mohammad Norouzi. Dream to control: Learning behaviors by latent imagination. In *International Conference on Learning Representations (ICLR)*, 2020. URL https://openreview.net/forum?id=S1lOTC4tDS.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *arXiv preprint arxiv:2006.11239*, 2020.

Christopher Hoang and Mengye Ren. Midway network: Learning representations for recognition and motion from latent dynamics. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=ZenwrTwNcj.

Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural Computation*, 9(8): 1735–1780, 1997. doi: 10.1162/neco.1997.9.8.1735.

Joel Jang, Seonghyeon Ye, Zongyu Lin, Jiannan Xiang, Johan Bjorck, Yu Fang, Fengyuan Hu, Spencer Huang, Kaushil Kundalia, Yen-Chen Lin, et al. Dreamgen: Unlocking generalization in robot learning through video world models. *arXiv preprint arXiv:2505.12705v2*, 2025.

Michael Janner, Justin Fu, Marvin Zhang, and Sergey Levine. When to trust your model: Model-based policy optimization. In *Advances in Neural Information Processing Systems*, 2019.

Rico Jonschkowski, Austin Stone, Jonathan Barron, Ariel Gordon, Kurt Konolige, and Anelia Angelova. What matters in unsupervised optical flow. In *European Conference on Computer Vision*, 2020.

Efstathios Karypidis, Ioannis Kakogeorgiou, Spyros Gidaris, and Nikos Komodakis. Advancing semantic future prediction through multimodal visual sequence transformers. *arXiv preprint arXiv:2501.08303*, 2025. doi: 10.48550/arXiv.2501.08303. URL https://arxiv.org/abs/2501.08303.

Tianhong Li, Yonglong Tian, He Li, Mingyang Deng, and Kaiming He. Autoregressive image generation without vector quantization. *arXiv preprint arXiv:2406.11838*, 2024. URL https://arxiv.org/abs/2406.11838.

Yaron Lipman, Ricky T. Q. Chen, Heli Ben-Hamu, Maximilian Nickel, and Matthew Le. Flow matching for generative modeling. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2023. URL https://openreview.net/forum?id=PqvMRDCJT9t.

Xingchao Liu, Chengyue Gong, and Qiang Liu. Flow straight and fast: Learning to generate and transfer data with rectified flow. *arXiv preprint arXiv:2209.03003*, 2022. doi: 10.48550/arXiv.2209.03003. URL https://arxiv.org/abs/2209.03003.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=Bkg6RiCqY7.

Xin Ma, Yaohui Wang, Xinyuan Chen, Gengyun Jia, Ziwei Liu, Yuan-Fang Li, Cunjian Chen, and Yu Qiao. Latte: Latent diffusion transformer for video generation. *Transactions on Machine Learning Research*, 2025.

Michael Mathieu, Camille Couprie, and Yann LeCun. Deep multi-scale video prediction beyond mean square error. *International Conference on Learning Representations*.

R Chris Miall and Daniel M Wolpert. Forward models for physiological motor control. *Neural networks*, 9(8):1265–1279, 1996.

Marco Pasini, Javier Nistal, Stefan Lattner, and George Fazekas. Continuous autoregressive models with noise augmentation avoid error accumulation. *arXiv preprint arXiv:2411.18447*, 2024. URL https://arxiv.org/abs/2411.18447. Accepted to NeurIPS 2024 – Audio Imagination Workshop.

Deepak Pathak, Pulkit Agrawal, Alexei A. Efros, and Trevor Darrell. Curiosity-driven exploration by self-supervised prediction. In *ICML*, 2017.

William Peebles and Saining Xie. Scalable diffusion models with transformers. *arXiv preprint arXiv:2212.09748*, 2022.

Rajesh Rao and Dana Ballard. Predictive coding in the visual cortex: a functional interpretation of some extra-classical receptive-field effects. *Nature neuroscience*, 2:79–87, 02 1999. doi: 10.1038/4580.

Dominik Schmidt and Minqi Jiang. Learning to act without actions. In *The Twelfth International Conference on Learning Representations (ICLR)*, 2024a.

Dominik Schmidt and Minqi Jiang. Learning to act without actions. In *International Conference on Learning Representations*, 2024b.

Mandyam Veerambudi Srinivasan, Simon Barry Laughlin, and A. Dubs. Predictive coding: a fresh view of inhibition in the retina. *Proceedings of the Royal Society of London. B. Biological Sciences*, 216(1205):427–459, 11 1982. ISSN 0080-4649. doi: 10.1098/rspb.1982.0085. URL https://doi.org/10.1098/rspb.1982.0085.

Nitish Srivastava, Elman Mansimov, and Ruslan Salakhutdinov. Unsupervised learning of video representations using LSTMs. In *International Conference on Machine Learning*, 2015.

Deqing Sun, Xiaodong Yang, Ming-Yu Liu, and Jan Kautz. PWC-Net: CNNs for optical flow using pyramid, warping, and cost volume. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2018.

Dani Valevski, Yaniv Leviathan, Moab Arar, and Shlomi Fruchter. Diffusion models are real-time game engines. *arXiv preprint arXiv:2408.14837*, 2024. doi: 10.48550/arXiv.2408.14837. URL https://arxiv.org/abs/2408.14837. ICLR 2025 (accepted).

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. Attention is all you need. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. URL https://proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf.

Ruben Villegas, Mohammad Babaeizadeh, Pieter-Jan Kindermans, Hernan Moraldo, Han Zhang, Mohammad Taghi Saffar, Santiago Castro, Julius Kunze, and D. Erhan. Phenaki: Variable length video generation from open domain textual description. *ArXiv*, abs/2210.02399, 2022.

Team Wan, Ang Wang, Baole Ai, Bin Wen, Chaojie Mao, Chen-Wei Xie, Di Chen, Feiwu Yu, Haiming Zhao, Jianxiao Yang, Jianyuan Zeng, Jiayu Wang, Jingfeng Zhang, Jingren Zhou, Jinkai Wang, Jixuan Chen, Kai Zhu, Kang Zhao, Keyu Yan, Lianghua Huang, Mengyang Feng, Ningyi Zhang, Pandeng Li, Pingyu Wu, Ruihang Chu, Ruili Feng, Shiwei Zhang, Siyang Sun, Tao Fang, Tianxing Wang, Tianyi Gui, Tingyu Weng, Tong Shen, Wei Lin, Wei Wang, Wei Wang, Wenmeng Zhou, Wente Wang, Wenting Shen, Wenyuan Yu, Xianzhong Shi, Xiaoming Huang, Xin Xu, Yan Kou, Yangyu Lv, Yifei Li, Yijing Liu, Yiming Wang, Yingya Zhang, Yitong Huang, Yong Li, You Wu, Yu Liu, Yulin Pan, Yun Zheng, Yuntao Hong, Yupeng Shi, Yutong Feng, Zeyinzi Jiang, Zhen Han, Zhi-Fan Wu, and Ziyu Liu. Wan: Open and advanced large-scale video generative models. *arXiv preprint arXiv:2503.20314*, 2025.

Daniel M Wolpert, Zoubin Ghahramani, and Michael I Jordan. An internal model for sensorimotor integration. *Science*, 269(5232):1880–1882, 1995.

Desai Xie, Zhan Xu, Yicong Hong, Hao Tan, Difan Liu, Feng Liu, Arie Kaufman, and Yang Zhou. Progressive autoregressive video diffusion models. *arXiv preprint arXiv:2410.08151*, 2024. URL https://arxiv.org/abs/2410.08151.

Wilson Yan, Yunzhi Zhang, Pieter Abbeel, and Aravind Srinivas. Videogpt: Video generation using vq-vae and transformers, 2021.

Mengjiao Yang, Yilun Du, Kamyar Ghasemipour, Jonathan Tompson, Dale Schuurmans, and Pieter Abbeel. Learning interactive real-world simulators. *arXiv preprint arXiv:2310.06114*, 2023.

Shuo Yang, Haocheng Xi, Yilong Zhao, Muyang Li, Jintao Zhang, Han Cai, Yujun Lin, Xiuyu Li, Chenfeng Xu, Kelly Peng, Jianfei Chen, Song Han, Kurt Keutzer, and Ion Stoica. Sparse videogen2: Accelerate video generation with sparse attention via semantic-aware permutation. *arXiv preprint arXiv:2505.18875*, 2025. doi: 10.48550/arXiv.2505.18875. URL https://arxiv.org/abs/2505.18875.

Seonghyeon Ye, Joel Jang, Byeongguk Jeon, Se June Joo, Jianwei Yang, Baolin Peng, Ajay Mandlekar, Reuben Tan, Yu-Wei Chao, Bill Yuchen Lin, Lars Liden, Kimin Lee, Jianfeng Gao, Luke Zettlemoyer, Dieter Fox, and Minjoon Seo. Latent action pretraining from videos. In *International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=VYOe2eBQeh.

Fisher Yu, Haofeng Chen, Xin Wang, Wenqi Xian, Yingying Chen, Fangchen Liu, Vashisht Madhavan, and Trevor Darrell. Bdd100k: A diverse driving dataset for heterogeneous multitask learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020.