Enhancing Physical Consistency in Lightweight World Models

Dingrui Wang[1*], Zhexiao Sun[1*], Zhouheng Li[2], Cheng Wang[4], Youlun Peng[1], Hongyuan Ye[1], Baha Zarrouki[1], Wei Li[3], Mattia Piccinini[1], Lei Xie[2], Johannes Betz[1]

https://physics-wm.github.io/

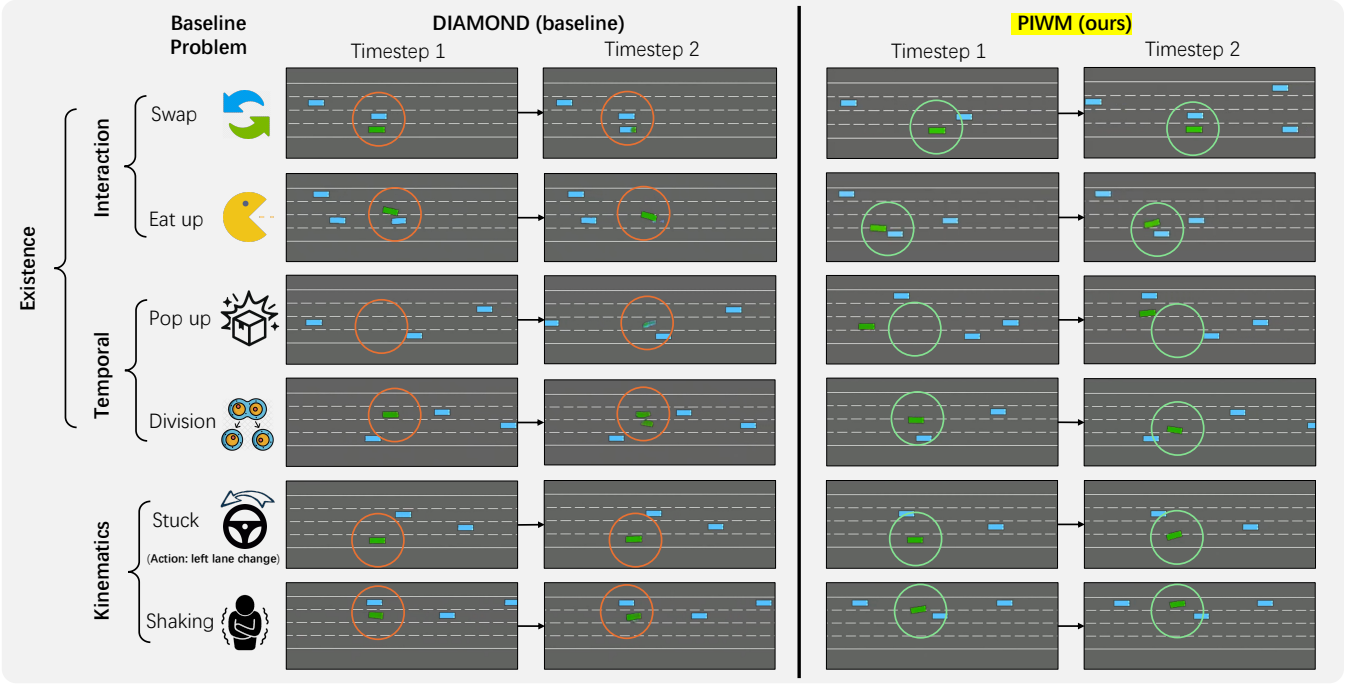


Fig. 1: **Performance comparison** between the DIAMOND [1] (baseline) and our Physics-Informed BEV World Model (PIWM), both trained on the dataset collected from HighwayEnv [2]. PIWM outperforms the baseline by generating more physically consistent results, with better existential consistency and kinematics response.

***Abstract*— A major challenge in deploying world models is the trade-off between size and performance. Large world models can capture rich physical dynamics but require massive computing resources, making them impractical for edge devices. Small world models are easier to deploy but often struggle to learn accurate physics, leading to poor predictions. We propose the Physics-Informed BEV World Model (PIWM), a compact model designed to efficiently capture physical interactions in bird's-eye-view (BEV) representations. PIWM uses *Soft Mask* during training to improve dynamic object modeling and future prediction. We also introduce a simple yet effective techniques—*Warm Start* for inference to enhance prediction quality with zero-shot model. Interactive experiments show that at the same parameter scale (400M), PIWM surpasses the baseline by 60.6% in weighted overall score. Moreover, even when compared with the largest baseline model (400M), the smallest PIWM (130M Soft Mask) achieves a 7.4% higher weighted overall score with a 28% faster inference speed.**

[1]Professorship of Autonomous Vehicle Systems, TUM School of Engineering and Design, Technical University of Munich, 85748 Garching, Germany; Munich Institute of Robotics and Machine Intelligence (MIRMI), {dingrui.wang, zhexiao.sun, mattia.piccinini, johannes.betz}@tum.de

[2]College of Control Science and Engineering, Zhejiang University, Hangzhou 310027, China. zh_li@zju.edu.cn

[3]Nanjing University, China.

[4]School of Mechatronics Engineering, Harbin Institute of Technology, Harbin 150001, China.

*These authors contributed equally to this work. Author order was determined at random.

I. Introduction

The Genie series of models [3]–[5] has significantly advanced research on world models. Beginning with the original Genie and Genie 2, these systems introduced the ability to generate interactive environments from data, enabling agents to learn and act within diverse virtual worlds. The most recent iteration, Genie 3, demonstrates emergent physical reasoning capabilities: it can simulate gravity, collisions, and object interactions without relying on an explicit physics engine. This capability indicates that scaling large world models can lead to a deeper, implicit understanding of the laws of physics, a critical step towards more general embodied intelligence. Among the world models, bird's-eye-view (BEV) world models [6], [7] are particularly promising for motion prediction and future trajectory modeling. By representing the environment in a top-down view [8], [9], BEV models can capture spatial relationships and object

interactions more effectively than first-person or ego-centric views, making them suitable for navigation tasks in robotics.

Running world models online is constrained by edge hardware performance. While devices such as the NVIDIA Jetson Orin Nano Super (67 TOPS) [10], AGX Orin (200–275 TOPS) [11], or Tesla HW4 (500-720 TOPS) [12] support moderate models, state-of-the-art video/world models like HunyuanVideo [13] and Veo 3 [14] with hundreds of billions of parameters remain infeasible outside GPU/TPU clusters. More compact models, e.g., DIAMOND (0.4B) runs at 21 FPS on an RTX 4080 Laptop GPU (542 TOPS) but only 9.5 FPS on 100-TOPS edge devices. Both rates fall below the human perceptual smoothness threshold ($\sim$24 FPS), which challenges real-time deployment.

While further shrinking the world model to run smoothly on edge devices is feasible, a fundamental challenge remains: can compact models capture the richness of physical dynamics without sacrificing reasoning ability? Reducing model size alone often leads to oversimplification and loss of fidelity. Addressing this gap requires shifting focus from scale to intelligent design—leveraging physics-informed data, inductive biases, and targeted training. Such models can move beyond pattern memorization, enabling robust generalization and effective reasoning under resource constraints. We address this challenge by extending the 400M-parameter world model DIAMOND [1] to better capture physical consistency (Fig. 1), by introducing a new Physics-Informed World Model (PIWM) with the following contributions:

- **Soft Mask.** We propose a method that extracts spatial semantic information, emphasizing the existence of dynamic objects while preserving action sensitivity. Our Soft Mask improves temporal and perceptual video consistency and achieves higher human-judged physics scores than the baseline [1]. Experiments further show that our method enables parameter reduction for edge deployment without compromising physical consistency.
- **Warm Starting.** We introduce a zero-shot Warm Start method, that injects contextual information at inference time to improve generation stability at small scales. It can be directly plugged into any pretrained diffusion-based world model, and yields FID gains over the baseline [1].
- **Open-source models and dataset.** We collect 2,000 episodes in HighwayEnv [2] with an MCTS agent, yielding 2 million BEV frames with aligned states and actions. Rewards and actions are carefully designed to evenly cover interaction and lane distributions, thereby supporting future exploration in physical consistency for the community.

II. Related Work

World Models have demonstrated the potential to learn complex world dynamics. Large-scale models such as the Genie series [3]–[5] and Veo 3 [14] show the benefits of scale for long-horizon prediction and richer environment simulation, while recent works [1], [15], [16] target lightweight models that can run with limited resources. DIAMOND [1] trains an RL agent fully inside a 0.4B-parameter diffusion world model. Inspired by these advances, researchers have applied world models to autonomous driving [17]–[20]. Within this domain, Bird's eye view (BEV) has emerged as a promising research direction, leading to attempts of interpreting BEV features as a world model. BEVDiffuser [21] designs a diffusion model to denoise the BEV feature maps, gaining a significant performance improvement, while Li et al. [7] introduce WoTE, a BEV world model for efficient and real-time future prediction and trajectory evaluation.

Edge computing. Large models are often unsuitable for edge deployment in autonomous driving due to high computational and memory demands. Although compression techniques offer efficiency gains, they pose critical limitations in safety-critical settings. Quantization [22]–[24] accelerates inference, but amplifies numerical errors under distribution changes. Pruning [25]–[27] risks removing weights vital for rare scenarios, harming robustness. Knowledge distillation [28]–[30] compresses models but often fails to preserve fine-grained spatiotemporal reasoning and precise decision boundaries, degrading performance in tasks. Thus, achieving compact, efficient, and reliable models for real-time edge deployment remains a key challenge. This motivates the design of small models yet capable of accurate and robust dynamic reasoning under resource constraints.

Encoding Physics into World Models. Because large models are hard to deploy on edge devices, distillation [31], [32] is widely used but often yields weaker or unstable performance. We instead train a compact world model tailored for edge constraints. Beyond efficiency, physics understanding is a key criterion for world models [33], [34], indicating whether they capture causal dynamics rather than merely reproducing appearance. DrivePhysica [35] improves physics awareness in driving world models by enforcing motion in a reference system, temporal consistency, and correct spatial relations. Using pretrained depth and semantic models, World4Drive [36] can gain a deep understanding of the spatial and semantic properties of the physical world. Yet maintaining strong physics reasoning in compact models remains open. Here, we show that PIWM effectively strengthens physical understanding under tight resource budgets.

III. Methodology

This section presents our data collection for world models' training and the design principles of our PIWM. The methods are split into two categories: training stage and inference stage. Section III-B introduces how we integrate masking into the training stage. In III-B.3, we elaborate the PIWM's training process, and III-C deals with our Warm Start strategy for the inference stage.

A. Data Collection

As illustrated in Fig. 3, we collect training data from the HighwayEnv simulator [2], a widely used platform for reinforcement learning with diverse highway scenarios and a BEV representation (Fig. 1). To promote diversity while maintaining a controlled collision rate, we designed a Monte Carlo Tree Search (MCTS) agent to autonomously control the ego vehicle. The agent is biased toward acceleration

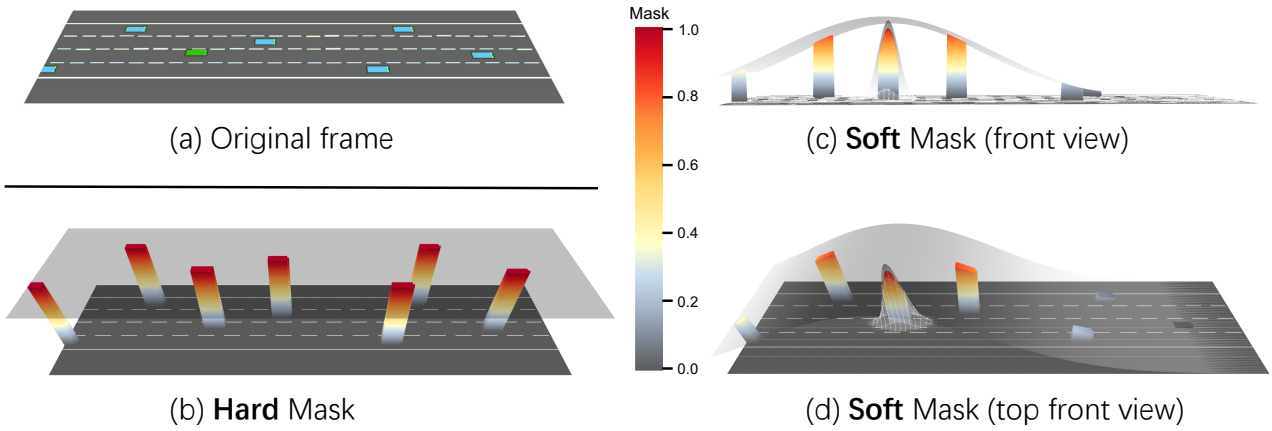


Fig. 2: **Illustration of soft masks and hard mask.** In each 3D cubic shape in the mask figures, the color gradient indicates the mask value according to the chosen mask weights. (a) shows the original frame, while (b) shows the hard mask. (c)-(d) show the Gaussian distributions for global scene softening and ego-centric softening.

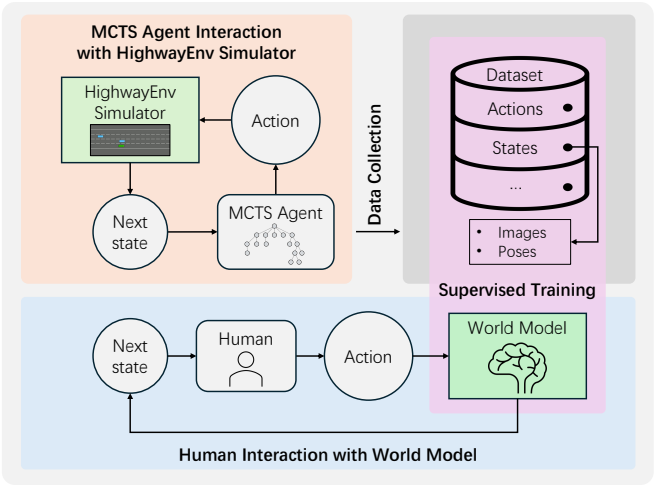


Fig. 3: The general framework: (Upper Left) Data collection using MCTS agent. (Right) Supervised World Model training. (Bottom) Online inference conditioned on human input.

maneuvers while maintaining collision avoidance, yielding a broad spectrum of realistic maneuvers. In total, we gather 2,000 episodes comprising 2,000,000 BEV frames, each paired with ground-truth physical states (poses and velocities) for the ego and surrounding vehicles.

B. Training with Hard & Soft Masks

We use masks to emphasize the existence of dynamic objects and thus aim to enhance physical consistency. Here, the mask can be interpreted as a matrix that shares the same shape as the image and selectively represents its characteristic regions. To design our masks, we need to first identify the ego vehicle (colored in green in the HighwayEnv simulator, Fig. 2a) and the surrounding vehicles (colored in blue) inside a BEV image. For this purpose, we adopt a color-checking module which detects the pixels belonging to the ego and the surrounding vehicles.

1) Hard Mask: Following state-of-the-art approaches [35], [37] that directly leverage geometric information as a condition, the hard mask employs a **binary** mask to distinguish between dynamic objects and background environment. Given a BEV image $\mathbf{I} \in \mathbb{R}^{H\times W\times C}$ where H denotes height, W width, and C the number of channels, The hard mask $\mathbf{m}_{\text{hard}} \in \mathbb{N}^{H\times W}$ is constructed as:

$$\mathbf{m}_{\text{hard}}(x,y) = \begin{cases} 1 & \text{if } (x,y) \text{ is green or blue} \\ 0 & \text{otherwise} \end{cases} \tag{1}$$

2) Soft Mask: Inspired by soft constraints in classical control, Soft Mask serves as an input-space soft constraint during both training and inference. It adds a conditioning channel with continuous spatial semantic weights that emphasize interaction-prone dynamic regions while preserving action sensitivity. Unlike the hard mask, our soft mask $\mathbf{m}_{\text{soft}} \in [0,1]^{H\times W}$ assigns continuous values between 0 and 1 to dynamic object regions, while maintaining zero values for static environmental areas. We design $\mathbf{m}_{\text{soft}}$ as:

$$\mathbf{m}_{\text{soft}} = (\underbrace{w_{\text{ego}} \cdot \mathbf{m}_{\text{ego}} \cdot \mathcal{N}_{\text{ego}}}_{\text{Weighted Gaussian Ego Mask}} + \overbrace{w_{\text{surr}} \cdot \mathbf{m}_{\text{surr}}}^{\text{Weighted Surrounding Mask}}) \cdot \mathcal{N}_{\text{global}} \tag{2}$$

In (2), $\mathbf{m}_{\text{ego}}$ and $\mathbf{m}_{\text{surr}}$ are hard masks for the ego vehicle (green channel) and surrounding vehicles (blue channel), weighted by the tunable factors w_{ego} and w_{surr}. In practice, we suggest w_{surr} slightly larger than w_{ego} during both training and inference to emphasize existential consistency while preserving action sensitivity of the ego vehicle:

$$\mathbf{m}_{\text{ego}}(x,y) = \begin{cases} 1 & \text{if } (x,y) \text{ is green} \\ 0 & \text{otherwise} \end{cases} \tag{3}$$

$$\mathbf{m}_{\text{surr}}(x,y) = \begin{cases} 1 & \text{if } (x,y) \text{ is blue} \\ 0 & \text{otherwise} \end{cases} \tag{4}$$

As shown in Fig. 2, we devise two softening dimensions: ego-centric and global scene softening. **Ego-centric Softening** uses a two-dimensional Gaussian distribution centered around the detected ego vehicle:

$$\mathcal{N}_{\text{ego}} = \mathcal{N}\left(\begin{bmatrix} x \\ y \end{bmatrix} \middle| \begin{bmatrix} x_{\text{ego}} \\ y_{\text{ego}} \end{bmatrix}, \begin{bmatrix} \sigma_x^2 & 0 \\ 0 & \sigma_y^2 \end{bmatrix}\right) \tag{5}$$

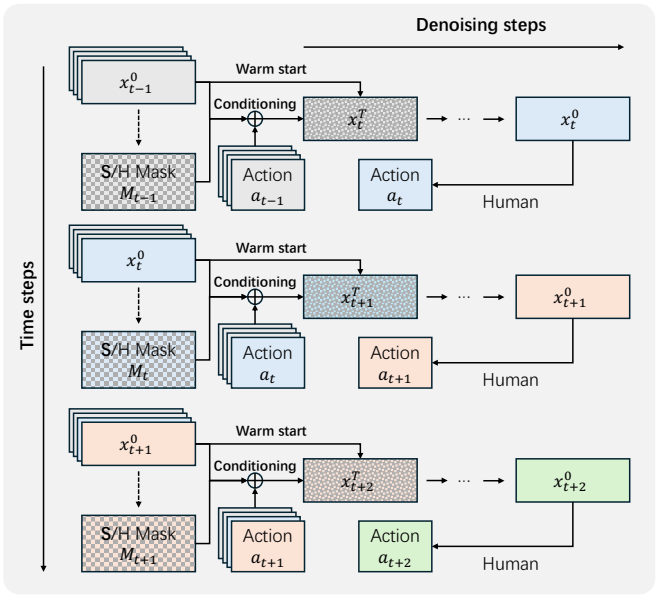


Fig. 4: Training pipeline.

where $(x_{\text{ego}}, y_{\text{ego}})$ denote the ego vehicle's centroid, while σ_x and σ_y are tunable Gaussian parameters. **Global Scene Softening** uses a global Gaussian distribution along the longitudinal direction of the BEV scene, centered at the ego vehicle's x-coordinate:

$$\mathcal{N}_{\text{global}} = \mathcal{N}\big(x \mid x_{\text{ego}},\ (\sigma_{\text{global}}\, W)^2\big) \tag{6}$$

where σ_{global} is a tunable parameter regulating the Gaussian's width, and W is the BEV image width.

Downsampling process uses bicubic interpolation to downsample the softened mask to match the input resolution of the denoising model: $\mathbf{m}_{\text{soft}}^{\text{down}} \leftarrow \text{bicubic}(\mathbf{m}_{\text{soft}})$.

3) World Model Training: World Models are generative models that learn to simulate how the environment evolves over time. Given a history of what the agent has seen (observations, actions), the model predicts future observations. As illustrated in Fig. 4, we follow a similar EDM [38] training setup as the baseline DIAMOND model [1]. Our soft mask is integrated into the EDM architecture as part of the input for the diffusion model $\mathbf{D}_\theta$, where θ are the trainable parameters. During training, we sample a sequence of length L containing past action–observation pairs, where each pair consists of an action a and an observation $\mathbf{x}$ (image). A sequence at timestamp t is represented as $(\mathbf{x}^0_{t-L+1}, a_{t-L+1}, \ldots, \mathbf{x}^0_t, a_t, \mathbf{x}^0_{t+1})$, drawn from the dataset $\mathcal{D}$. The denoising process is given as:

$$\hat{\mathbf{x}}^0_{t+1} = \mathbf{D}_\theta(\mathbf{x}^\tau_{t+1}, \tau, \mathbf{x}^0_{t-L+1:t}, a_{t-L+1:t}, \mathbf{m}_{\circ,t}) \tag{7}$$

where τ is noise level, $\mathbf{m}_{\circ,t}$ can be either $\mathbf{m}_{\text{soft},t}$ or $\mathbf{m}_{\text{hard},t}$ ($\mathbf{m}_{\text{soft},t}$ means Soft Mask at timestep t). The diffusion model $\mathbf{D}_\theta$ is trained to denoise a corrupted version of $\mathbf{x}^0_{t+1}$, conditioned on the history of observations. To guide the model's attention, we integrate our soft mask into the EDM architecture by concatenating it with the observations. The training loss is defined as:

$$\mathcal{L}(\theta) = \mathbb{E}\left[\|\hat{\mathbf{x}}^0_{t+1} - \mathbf{x}^0_{t+1}\|^2\right]. \tag{8}$$

Algorithm 1 Physics-Informed BEV World Model

Training

Require: $\mathcal{D}, P_{mean}, P_{std}$

while not converged **do**
  $(\mathbf{x}^0_{t-L+1}, a_{t-L+1}, \ldots, \mathbf{x}^0_t, a_t, \mathbf{x}^0_{t+1}) \sim \mathcal{D}$ ▷▷ sample
  $\log(\sigma) \sim \mathcal{N}(P_{mean}, P^2_{std})$ ▷▷ noise level
  $\tau := \sigma$ ▷▷ default identity schedule from EDM
  $\mathbf{x}^\tau_{t+1} \sim \mathcal{N}(\mathbf{x}^0_{t+1}, \sigma^2\mathbf{I})$ ▷▷ add gaussian noise
  $\hat{\mathbf{x}}^0_{t+1} \leftarrow \mathbf{D}_\theta(\mathbf{x}^\tau_{t+1}, \tau, \mathbf{x}^0_{t-L+1:t}, a_{t-L+1:t}, \mathbf{m}_{\circ,t})$
  $\mathcal{L} = \|\hat{\mathbf{x}}^0_{t+1} - \mathbf{x}^0_{t+1}\|^2$ ▷▷ loss
  $\theta \leftarrow \theta - \nabla_\theta \mathcal{L}$ ▷▷ gradient step
end while

Inference

Require: noise covariance: Σ_{off} and Σ_{ew}.

if $i == 0$ **then**
  $x_i \sim \mathcal{N}(0, \sigma_0^2 I)$ ▷▷ Initialize image
else if $i > 0$ **then**
  $x_i \sim \mathcal{N}(x_{i-1,0},\ \Sigma_{\text{off}} + \Sigma_{\text{ew}})$ ▷▷ Warm Start
end if
$\tau_0[0] \leftarrow s_0$ ▷▷ known initial state
for $i = 0$ to $N-1$ **do**
  $\tau_{i+1} \leftarrow S_\theta(\tau_i;\ \sigma_i, \sigma_{i+1})$ ▷▷ denoised prediction (2)
  $\tau_{i+1}[0] \leftarrow s_0$ ▷▷ known initial state
end for
return τ_N ▷▷ noise-free sample

C. Inference with Warm Starting

To improve temporal consistency in the generative process at inference time, we design a warm-start strategy. The method is model-agnostic (zero-shot) and can be applied to any trained world model without retraining. The intuition follows autoregressive generation: each frame is synthesized by perturbing the previously generated clean frame, which promotes spatial and temporal coherence across the sequence. At generation step i, instead of sampling from pure Gaussian noise, we initialize the reverse process by perturbing the clean image from step $i-1$, denoted $x_{i-1,0}$. The perturbed sample $\tilde{x}_{i,T}$ is drawn from

$$\begin{aligned} q(\tilde{x}_{i,T} \mid x_{i-1,0}) &= \mathcal{N}(x_{i-1,0},\ \Sigma_{\text{off}} + \Sigma_{\text{ew}}), \\ \Sigma_{\text{off}} &= \sigma^2_{\text{off}}\, \text{blkdiag}\big(\underbrace{\mathbf{K}_l, \ldots, \mathbf{K}_l}_{C \text{ times}}\big), \\ \Sigma_{\text{ew}} &= \sigma^2_{\text{ew}}\, \mathbf{J}_n. \end{aligned} \tag{9}$$

Here, $x_{i-1,0}$ and $\tilde{x}_{i,T}$ are flattened vectors in $\mathbb{R}^{n\times 1}$ with $n = H \times W \times C$; H, W, and C denote image height, width, and channel count. The matrix $\mathbf{K}_l = \mathbf{1}\mathbf{1}^\top \in \mathbb{R}^{l\times l}$ with $l = H \times W$ specifies a rank-1 covariance within each channel, inducing a channel-wise global offset (fully

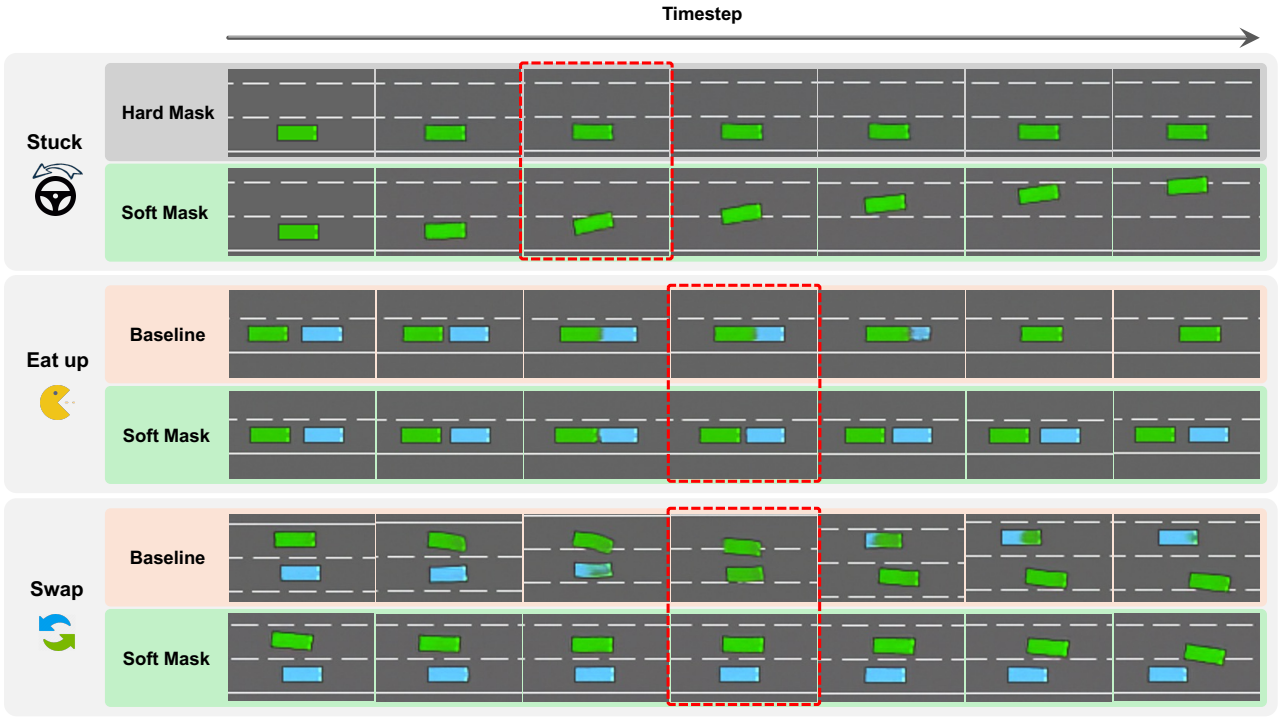


Fig. 5: **Qualitative comparison of different methods on common failure cases.** The first row ("Stuck") compares Hard Mask and Soft Mask for the "Stuck" issue, which concerns whether the agent is insensitive to actions. The second row ("Eat up") and third row ("Swap") compare the baseline [1] with Soft Mask for the "Eat up" and "Swap" problems, respectively.

correlated spatial positions). The operator $\mathrm{blkdiag}(\cdot)$ places C copies of $\mathbf{K}_l$ along the diagonal, ensuring no cross-channel correlation. The term $\sigma_{\mathrm{ew}}^2 \mathbf{J}_n$ adds element-wise independent noise, where $\mathbf{J}_n$ denotes the identity matrix. In practice, $\tilde{x}_{i,T}$ serves as the terminal-time initialization for the denoiser at step i, balancing coherence (via Σ_{off}) and flexibility for local changes (via Σ_{ew}).

IV. Results

A. Experimental Setup

We evaluate action-conditioned BEV highway video generation: given the last four frames and the current action, the model predicts the next frame and then rolls out autoregressively. We compare four variants using the same U-Net backbone under the EDM framework: (i) **Baseline** (DIAMOND [1]), (ii) **Hard Mask**, (iii) **Warm Start**, and (iv) **Soft Mask**. Method-specific mechanisms were detailed in III-B and III-C. We report three parameter budgets (130M, 170M, 400M) by changing only the denoiser channel width. Consistent with our evaluation scope, reconstruction quality was reported for baseline/Warm Start/Soft Mask; physical consistency was reported for all four methods; efficiency was reported for baseline/Soft Mask.

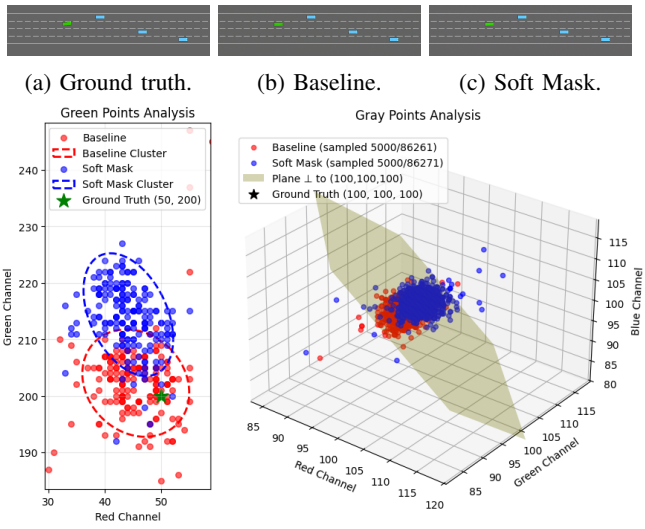
1) Reconstruction Quality: We sample 100 initial 16-frame segments from the test set as real observations and generate corresponding 16-frame rollouts with identical spawn points (Starting player position) and action sequences. FID is computed with `pytorch-fid` (Inception-V3 (`pool3`)) over all frames jointly. FVD is computed with `cd-fvd` (I3D backbone). LPIPS loss is computed with `lpips` (AlexNet backbone) between paired real/generated frames and averaged over spawn points.

Parameters	Method	Reconstruction metrics		
		FID ↓	FVD ↓	LPIPS ↓
130M	Baseline	52.9	304.1	0.021
	Warm Start	**50.9**	298.9	0.022
	Soft Mask	74.1	**269.4**	0.023
170M	Baseline	36.4	362.8	0.022
	Warm Start	**33.3**	367.8	0.022
	Soft Mask	79.7	**189.3**	0.026
400M	Baseline	23.3	204.2	0.013
	Warm Start	**22.2**	209.4	0.013
	Soft Mask	52.1	**156.8**	0.014

TABLE I: Results of FID, FVD and LPIPS.

2) Physics Consistency: As noted by the creators of Genie 3, evaluating visual world models can be subjective.[1] To obtain quantitative judgments, we adopt Mean Opinion Score (MOS) as our primary evaluation method, following ITU-R BT.500 [39], [40] on a five-point scale (10=Excellent, 8=Good, 6=Fair, 4=Poor, 2=Bad). We conduct a within-subject, double-blind user study with 24 non-expert raters, who evaluated all four methods under the same user interface and task prompts. We report Interactive Existential Consistency (IEC), Kinematics Response (KIR), and Temporal Existential Consistency (TEC), averaged and mapped to a

[1] *How Do You Measure the Quality of a World Model?*



(a) Ground truth. (b) Baseline. (c) Soft Mask.

(d) Green Points. (e) Gray Points.

Fig. 6: Color distribution difference concerning the RGB channel values for baseline and Soft Mask.

percentage scale. . We also report a weighted overall score (WO) emphasizing interactive existential stability:

$$\text{WO} = 0.5 \cdot \text{IEC} + 0.25 \cdot \text{KIR} + 0.25 \cdot \text{TEC}. \quad (10)$$

3) Efficiency: We benchmark on three platforms: RTX 4090, RTX 4080 (laptop), and RTX 3060 (laptop). For each parameter budget and configuration, we run 10 trials of 1,000 frames under identical resolution, sampling settings, initial frames, and action sequence; the first five frames of each run are discarded. We report p95 FPS, p95 inference-only latency measured with CUDA events, and peak GPU memory, averaged across trials.

B. Reconstruction Quality

At three model sizes (130M, 170M, 400M), Table I shows three consistent trends. **(i)** Soft Mask achieves the best FVD at all scales—269.4 (130M), 189.3 (170M), 156.8 (400M)—improving over baseline by 11.4%, 47.8%, and 23.2%, respectively, indicating stronger temporal coherence. **(ii)** Warm Start yields a small but consistent FID gain over baseline. **(iii)** LPIPS is low for all methods (all < 0.03), which reflecting high perceptual similarity to ground truth and differences between methods are minor.

Visual and distributional analysis (Fig. 6) suggests the higher FID for Soft Mask is driven by color-distribution shifts rather than perceptible degradations in image quality: compare to baseline, Soft Mask slightly increases green-channel intensity on green pixels and RGB values on gray background (Fig. 6d, 6e), which changes dataset-level statistics that FID captures despite minimal visual differences between Fig. 6b and Fig. 6c.

In summary, perceptual quality is comparable across methods (low LPIPS), while FVD—more sensitive to motion quality—consistently favors Soft Mask at all parameter sizes, indicating better temporal dynamics.

Param.	Method	IEC ↑	KIR ↑	TEC ↑	WO ↑
75M	Baseline†	22.50	45.00	47.50	34.38
	Warm Start†	22.50	50.00	52.50	**36.88**
130M	Baseline	28.12	53.59	56.46	41.57
	Hard Mask†	38.75	32.50	17.50	31.88
	Soft Mask	**43.75**	**64.17**	**56.87**	**52.14**
170M	Baseline	32.09	51.04	69.38	46.15
	Hard Mask†	42.50	46.25	27.50	39.69
	Soft Mask	**60.21**	**63.16**	**75.83**	**64.85**
400M	Baseline	30.63	**70.63**	62.29	48.55
	Hard Mask†	47.50	38.75	40.00	43.44
	Soft Mask	**82.08**	64.68	**82.90**	**77.94**

TABLE II: Results of Human evaluation scores. The metrics considered are Interactive Existential Consistency (IEC), Kinematics Response (KIR), and Temporal Existential Consistency (TEC), Weighted Overall (WO). Baseline is DIAMOND [1]. † indicate the experiments are evaluated by 4 humans. While the rest are evaluated by 24 humans.

C. Physical Consistency

Beyond generative quality, adherence to physical constraints more directly determines practical utility. We therefore conduct qualitative and quantitative evaluations. The first row ("Stuck") in Fig. 5 shows that the binary spatial guidance of Hard Mask can over-constrain behavior, suppressing actions and even preventing lane changes. Under the same scenario and action inputs, Soft Mask applies weighted, continuous spatial-semantic guidance that preserves object existence while remaining action-sensitive. The second ("Eat up") and third ("Swap") rows illustrate that the baseline frequently exhibits existential errors in interactive scenes, while Soft Mask is better in object continuity.

Quantitatively, we evaluated all methods using a double-blind, within-subject MOS protocol; results are reported in Table II. Hard Mask increases IEC relative to baseline (e.g., +37.8% @130M, +55.1% @400M) but, due to over-constraint, degrades KIR and TEC (e.g., -39.3% / -69.2% @130M), consistent with the "Stuck" failure mode in Fig. 5. Warm Start, as a training-free zero-shot approach, yields modest gains at small model scales. At 75M with four expert raters, it shows little change in IEC but improves kinematic stability (KIR +11.1%) and temporal stability (TEC +10.5%), producing a WO gain of +7.3% without retraining.

A detailed per-metric comparison between Soft Mask and baseline is shown below.

1) Interactive Existential Consistency (IEC): IEC quantifies identity/existence preservation under interactive, extreme maneuvers (e.g., collisions, squeezes). Soft Mask significantly outperforms baseline at all scales: **+55.6%** (130M), **+87.6%** (170M), **+168.1**% (400M). Baseline's IEC scores average around 30 with limited gains from scaling, whereas Soft Mask's continuous pixel-weighted guidance stabilizes identity and existence, achieving **82.08** at 400M.

Param.	Method	RTX 4090 (1321 TOPS)			RTX 4080 Laptop (542 TOPS)			RTX 3060 Laptop (105 TOPS)		
		FPS ↑	Latency ↓	GPU ↓	FPS ↑	Latency ↓	GPU ↓	FPS ↑	Latency ↓	GPU ↓
130M	Baseline	32.61	16.93	834.3	28.14	29.98	749.6	12.41	71.90	923.4
	Soft Mask	32.19	17.48	834.5	27.99	30.39	749.7	12.40	72.04	923.6
170M	Baseline	32.16	17.15	975.8	27.40	31.23	889.0	12.12	73.88	1064.8
	Soft Mask	30.99	17.34	974.7	27.53	30.89	889.8	12.04	74.42	1063.7
400M	Baseline	27.49	21.69	1795.7	21.83	40.17	1727.9	9.72	93.42	1884.8
	Soft Mask	28.32	22.08	1795.9	21.37	41.82	1727.0	9.67	93.99	1885.0

TABLE III: Results of FPS, latency, and peak GPU memory usage. GPU memory in MiB. FPS indicates 95th percentile (p95) Frames per second. Latency in ms. All measurements with compilation enabled.

2) Kinematics Response (KIR): KIR captures immediate controllability and the amplitude response to actions. Soft Mask maintains an advantage at small-to-medium scales: +19.7% (130M) and +23.8% (170M); at 400M it is slightly below baseline (-8.4%). Together with improved FVD at each scale, the results suggest comparable or better kinematic behavior via smoothed per-step action amplitudes.

3) Temporal Existential Consistency (TEC): TEC measures existence consistency over time. Soft Mask leads by +9.3% and +33.1% at 170M and 400M, respectively, and is on par with baseline at 130M, indicating that continuous spatial modulation improves cross-timestep stability.

Across metrics, our proposed Soft Mask dominates IEC and TEC while remaining competitive on KIR, yielding markedly higher weighted overall (WO) than baseline: **+25.4%** (130M), **+40.5%** (170M) , and **+60.6%** (400M). Notably, even the smallest Soft Mask model (130M) achieves higher IEC and WO than the largest baseline model (400M). Overall, continuous spatial-semantic guidance provided by Soft Mask mitigates the "stuck" issue of Hard Mask and significantly improves interactive and long-term physical consistency without sacrificing action flexibility.

D. Edge Computing Efficiency

As shown in Table III, we compare baseline and Soft Mask across parameter scales and hardware spanning a broad TOPS (Tera Operations Per Second) range representative of robotics and autonomous driving deployments. Soft Mask matches baseline's resource footprint across settings, adding no measurable computational overhead. On a representative 542 TOPS device, downsizing to 170M reduces p95 latency by $\sim$25% and increases p95 FPS from $\sim$21 to $\sim$27, surpassing the 24 FPS perceptual smoothness threshold. Further downsizing to 130M, and combined with Table II, the 130M Soft Mask attains a WO of **52.14** at **27.99** FPS, whereas the 400M baseline attains **48.55** at **21.83** FPS (below 24 FPS), supporting parameter reduction as a practical path for edge deployment without compromising physical consistency.

V. Discussion

We first establish a clear efficacy anchor at a higher parameter budget (400M) and then verify scale-insensitive improvements at smaller budgets (170M and 130M). Under edge computing, when p95 FPS meets real-time display (e.g., 24 FPS), smaller configurations still outperform the larger baseline on physics-oriented human scores. This indicates that the gains stem from the mechanisms themselves rather than from parameter count, and it supports parameter reduction as a practical path to edge deployment without sacrificing physics consistency. Trends on TEC further suggest that continuous spatial modulation improves cross-timestep stability, consistent with the intended role of Soft Mask in stabilizing temporal dynamics.

FID results reveal that Soft Mask may exhibit channel-level shifts during generation, likely caused by out-of-distribution deviations in generalization. Nevertheless, this does not compromise practical utility: considering perceptual similarity (LPIPS) and temporal consistency (FVD and TEC), Soft Mask consistently achieves superior performance.

Limitations remain. First, our evaluation is conducted solely in simulation using HighwayEnv, which may not capture the full complexity of real-world driving scenarios including adverse weather conditions, complex urban environments, etc. Second, the current mask construction relies on color-based detection and could be made more robust. Finally, the physics consistency evaluation relies primarily on subjective Mean Opinion Scores (MOS), which, despite following established standards, may miss subtle physical violations critical in safety-critical applications.

VI. Conclusion and Future Work

We present a Physics-Informed BEV World Model (PIWM) that improves physical consistency while avoiding action suppression and additional computational overhead. PIWM provides two mechanisms: Soft Mask, a training-time conditioning channel with continuous spatial semantic weights that highlight interaction-prone regions while preserving action sensitivity, and Warm Start, a training-free inference strategy that enhances generation stability. Across parameter scales, PIWM with Soft Mask substantially improves physics-oriented human scores and generative dynamics metric (FVD), and under edge-computing budgets that satisfy real-time display, even smaller models can maintain better physical consistency than larger baseline.

Future work will adapt PIWM to real-world datasets with richer conditions and rare events, develop a decoder that

directly predicts future states to guide generation and reduce drift, and explore general objective physics metrics for comprehensive evaluation of spatiotemporal consistency and dynamics. We also plan to evaluate our methods in closed-loop planning and control to assess safety and robustness under domain shifts.

REFERENCES

[1] E. Alonso, A. Jelley, V. Micheli, A. Kanervisto, A. J. Storkey, T. Pearce, and F. Fleuret, "Diffusion for world modeling: Visual details matter in atari," *Advances in Neural Information Processing Systems*, vol. 37, 2024.

[2] E. Leurent, "An environment for autonomous driving decision-making," https://github.com/eleurent/highway-env, 2018.

[3] J. Bruce, M. D. Dennis, A. Edwards, J. Parker-Holder, Y. Shi, E. Hughes, M. Lai, A. Mavalankar, R. Steigerwald, C. Apps *et al.*, "Genie: Generative interactive environments," in *Forty-first International Conference on Machine Learning*, 2024.

[4] J. Parker-Holder, P. Ball, J. Bruce, V. Dasagi, K. Holsheimer, C. Kaplanis, A. Moufarek, G. Scully, J. Shar, J. Shi, S. Spencer, J. Yung, M. Dennis *et al.*, "Genie 2: A large-scale foundation world model," 2024.

[5] P. J. Ball, J. Bauer, F. Belletti, B. Brownfield, A. Ephrat, S. Fruchter, A. Gupta, K. Holsheimer, othersAleksander Holynski, J. Hron, C. Kaplanis, M. Limont, M. McGill *et al.*, "Genie 3: A new frontier for world models," 2025.

[6] B. Liao, S. Chen, H. Yin, B. Jiang, C. Wang, S. Yan, X. Zhang, X. Li, Y. Zhang, Q. Zhang *et al.*, "Diffusiondrive: Truncated diffusion model for end-to-end autonomous driving," in *Proceedings of the Computer Vision and Pattern Recognition Conference*, 2025.

[7] Y. Li, Y. Wang, Y. Liu, J. He, L. Fan, and Z. Zhang, "End-to-end driving with online trajectory evaluation via bev world model," *arXiv preprint arXiv:2504.01941*, 2025.

[8] R. Krajewski, J. Bock, L. Kloeker, and L. Eckstein, "The highd dataset: A drone dataset of naturalistic vehicle trajectories on german highways for validation of highly automated driving systems," in *2018 21st international conference on intelligent transportation systems (ITSC)*. IEEE, 2018.

[9] H. Caesar, J. Kabzan, K. S. Tan, W. K. Fong, E. Wolff, A. Lang, L. Fletcher, O. Beijbom, and S. Omari, "nuplan: A closed-loop ml-based planning benchmark for autonomous vehicles," *arXiv preprint arXiv:2106.11810*, 2021.

[10] NVIDIA Corporation, *Jetson Orin Nano Super Developer Kit*, Dec. 2024, version 1.0. [Online]. Available: https://www.nvidia.com/en-us/autonomous-machines/embedded-systems/jetson-orin/nano-super-developer-kit/

[11] ——, *Jetson AGX Orin Module*, 2022, version 1.0. [Online]. Available: https://www.nvidia.com/en-us/autonomous-machines/embedded-systems/jetson-orin/

[12] ha1nk, "Tesla's latest advancement: Hardware 4.0," https://after1989.com/teslas-latest-advancement-hardware-4-0/, 2023.

[13] W. Kong, Q. Tian, Z. Zhang, R. Min, Z. Dai, J. Zhou, J. Xiong, X. Li, B. Wu, J. Zhang *et al.*, "Hunyuanvideo: A systematic framework for large video generative models," *arXiv preprint arXiv:2412.03603*, 2024.

[14] A. Sharma, A. Kuznetsova, A. Razavi, A. Holynski, A. Kuznetsova, A. Gupta, A. Waters, B. Poole, D. Tanis, D. Gasaway, D. Erhan, E. Corona, F. Belletti *et al.*, "Veo: a text-to-video generation system," 2025. [Online]. Available: https://deepmind.google/models/veo/

[15] A. Bar, G. Zhou, D. Tran, T. Darrell, and Y. LeCun, "Navigation world models," in *Proceedings of the Computer Vision and Pattern Recognition Conference*, 2025.

[16] G. Zhou, H. Pan, Y. LeCun, and L. Pinto, "Dino-wm: World models on pre-trained visual features enable zero-shot planning," *arXiv preprint arXiv:2411.04983*, 2024.

[17] A. Hu, L. Russell, H. Yeo, Z. Murez, G. Fedoseev, A. Kendall, J. Shotton, and G. Corrado, "Gaia-1: A generative world model for autonomous driving," *arXiv preprint arXiv:2309.17080*, 2023.

[18] X. Wang, Z. Zhu, G. Huang, X. Chen, J. Zhu, and J. Lu, "Drivedreamer: Towards real-world-drive world models for autonomous driving," in *European conference on computer vision*. Springer, 2024.

[19] Y. Li, L. Fan, J. He, Y. Wang, Y. Chen, Z. Zhang, and T. Tan, "Enhancing end-to-end autonomous driving with latent world model," *arXiv preprint arXiv:2406.08481*, 2024.

[20] C. Min, D. Zhao, L. Xiao, J. Zhao, X. Xu, Z. Zhu, L. Jin, J. Li, Y. Guo, J. Xing *et al.*, "Driveworld: 4d pre-trained scene understanding via world models for autonomous driving," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2024.

[21] X. Ye, B. Yaman, S. Cheng, F. Tao, A. Mallik, and L. Ren, "Bevdiffuser: Plug-and-play diffusion model for bev denoising with ground-truth guidance," in *Proceedings of the Computer Vision and Pattern Recognition Conference*, 2025.

[22] Y. Zhang, Z. Dong, H. Yang, M. Lu, C. C. Tseng, Y. Du, and S. Zhang, "Qd-bev: Quantization-aware view-guided distillation for multi-view 3d object detection," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023.

[23] X. Wang, Z. Lin, Z. Xia, and Y. Wang, "Ptqat: A hybrid parameter-efficient quantization algorithm for 3d perception tasks," *arXiv preprint*, 2025.

[24] P. Yu, Z. Kong, P. Zhao, P. Dong, H. Tang, F. Sun, and Y. Wang, "Q-tempfusion: Quantization-aware temporal multi-sensor fusion on bird's-eye view representation," in *2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*. IEEE, Feb. 2025.

[25] Y. Li, Y. Li, X. Yang, M. Yu, Z. Huang, X. Wu, and C. K. Yeo, "Learning content-aware multi-modal joint input pruning via birds'-eye-view representation," in *2024 IEEE 27th International Conference on Intelligent Transportation Systems (ITSC)*. IEEE, Sep. 2024.

[26] T. Castells, H. K. Song, B. K. Kim, and S. Choi, "Ld-pruner: Efficient pruning of latent diffusion models using task-agnostic insights," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024.

[27] R. Shirkavand, P. Yu, S. Gao, G. Somepalli, T. Goldstein, and H. Huang, "Efficient fine-tuning and concept suppression for pruned diffusion models," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025.

[28] T. Yin, M. Gharbi, R. Zhang, E. Shechtman, F. Durand, W. T. Freeman, and T. Park, "One-step diffusion with distribution matching distillation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024.

[29] A. Sauer, D. Lorenz, A. Blattmann, and R. Rombach, "Adversarial diffusion distillation," in *European Conference on Computer Vision (ECCV)*. Cham: Springer Nature Switzerland, Sep. 2024.

[30] T. H. Nguyen and A. Tran, "Swiftbrush: One-step text-to-image diffusion model with variational score distillation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024.

[31] G. Hinton, O. Vinyals, and J. Dean, "Distilling the knowledge in a neural network," *arXiv preprint arXiv:1503.02531*, 2015.

[32] A. Mishra and D. Marr, "Apprentice: Using knowledge distillation techniques to improve low-precision network accuracy," *arXiv preprint arXiv:1711.05852*, 2017.

[33] H. Duan, H.-X. Yu, S. Chen, L. Fei-Fei, and J. Wu, "Worldscore: A unified evaluation benchmark for world generation," *arXiv preprint arXiv:2504.00983*, 2025.

[34] D. Li, Y. Fang, Y. Chen, S. Yang, S. Cao, J. Wong, M. Luo, X. Wang, H. Yin, J. E. Gonzalez *et al.*, "Worldmodelbench: Judging video generation models as world models," *arXiv preprint arXiv:2502.20694*, 2025.

[35] Z. Yang, X. Guo, C. Ding, C. Wang, and W. Wu, "Physical informed driving world model," *arXiv preprint arXiv:2412.08410*, 2024.

[36] Y. Zheng, P. Yang, Z. Xing, Q. Zhang, Y. Zheng, Y. Gao, P. Li, T. Zhang, Z. Xia, P. Jia *et al.*, "World4drive: End-to-end autonomous driving via intention-aware physical latent world model," *arXiv preprint arXiv:2507.00603*, 2025.

[37] H. Wu, D. Wu, T. He, J. Guo, Y. Ye, Y. Duan, and J. Bian, "Geometry forcing: Marrying video diffusion and 3d representation for consistent world modeling," *arXiv preprint arXiv:2507.07982*, 2025.

[38] T. Karras, M. Aittala, T. Aila, and S. Laine, "Elucidating the design space of diffusion-based generative models," *Advances in neural information processing systems*, vol. 35, 2022.

[39] ?, "Recommendation 500-10: Methodology for the subjective assessment of the quality of television pictures," ITU-R Rec. BT.500, 2000.

[40] International Telecommunication Union - Telecommunication Standardization Sector (ITU-T), "Mean opinion score (mos) terminology," ITU, Geneva, Switzerland, Recommendation P.800.1, Jul. 2016, series P: Telephone Transmission Quality, Telephone Installations, Local Line Networks. [Online]. Available: https://www.itu.int/rec/T-REC-P.800.1-201607-I/en