arXiv:2607.15898v1 [cs.CV] 17 Jul 2026

Orbis 2: A Hierarchical World Model for Driving

Sudhanshu Mittal* **Arian Mousakhan*** **Silvio Galesso***

Karim Farid **Jonannes Dienert** **Rajat Sahay** **Thomas Brox**

University of Freiburg, Germany

Abstract

Current world models operate at a single level of abstraction, with most prioritizing perceptual fidelity while lacking the spatial reasoning and semantic understanding required for real-world downstream tasks. We present a hierarchical driving world model that factorizes future prediction across two levels operating at distinct temporal and abstraction scales: a high-level predictor that forecasts coarse scene structure over extended temporal horizons, and a low-level generator that produces detailed predictions conditioned on the high-level output. This decomposition yields high perceptual fidelity while also capturing strong spatial and semantic representations. We further show that pretraining with a diffusion forcing objective yields substantially richer internal representations than the standard teacher forcing objective, while teacher forcing — predicting only the next frame from clean context — produces more stable autoregressive rollouts. We therefore introduce a generic two-stage training paradigm that pretrains the model with diffusion forcing and fine-tunes with teacher forcing, combining the representational benefits of the former with the rollout stability of the latter. Our approach achieves state-of-the-art results across the standard suite of driving world model evaluations on established benchmarks, including long-horizon generation fidelity, steering responsiveness evaluated on counterfactual scenarios, and internal representation quality. Project page with code, demo, checkpoints and qualitative results: https://lmb-freiburg.github.io/orbis2.github.io

1 Introduction

World models [15, 17, 30] are learned dynamical systems that approximate how an environment evolves in response to an agent's actions, providing internal predictions of transition dynamics that support planning, simulation, and policy learning [45]. In high-dimensional, complex domains such as autonomous driving, world models are commonly formulated in a learned latent space where an encoder maps raw observations to compact latent representations and a predictor propagates these states forward, conditioned on the agent's actions. Current implementations of this paradigm are limited in two key respects. First, the predictor operates at a single temporal scale [3, 2]: rollouts are generated frame-by-frame, accumulating error over time and limiting the model's ability to explicitly reason about long-horizon scene evolution. Second, the latent space is typically shaped by a reconstruction objective that preserves all fine visual details such as texture and appearance, even though long-horizon dynamics are primarily governed by coarse, slowly varying semantic properties, making such representations suboptimal for long-horizon forecasting. This aligns with a well-established principle in cognitive science: humans plan over distant goals using simplified,

*Main Contributors. {mittal,mousakha,galessos}@cs.uni-freiburg.de

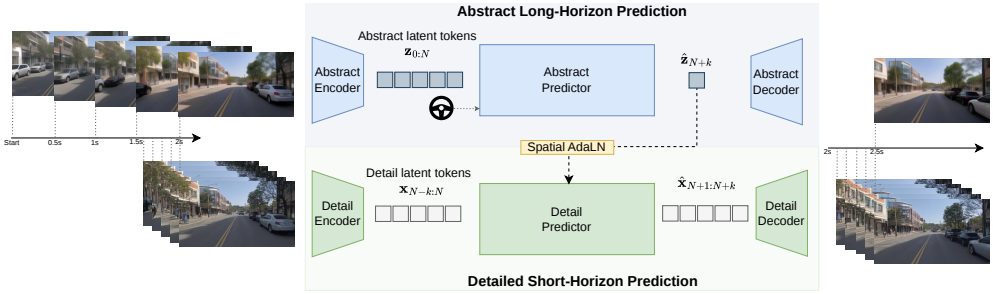


Figure 1: **Orbis 2: a hierarchical driving world model.** The *abstract predictor* operates over a long temporal context to forecast a future state in latent space, capturing abstract scene dynamics over long horizons, and enabling steering control. The *detail predictor* is conditioned on this abstract prediction and generates fine-grained short-horizon frames, enabling high-fidelity local prediction grounded in long-range temporal context.

abstract representations and refine details only as actions approach execution [47, 21]. Building on this principle, we propose a setup that decomposes future prediction across two levels of abstraction. A high-level predictor forecasts long-horizon dynamics in an abstract latent space, producing subgoals that condition a low-level predictor. The latter operates on a finer-grained latent space, specializing in detailed, short-horizon prediction. Figure 1 provides an overview of our hierarchical framework.

For the high-level predictor, we require a latent space that strongly captures spatial structure and semantic understanding, enabling long-horizon temporal estimation. We therefore utilize a compressed DINO [38] representation as the target for our high-level predictor. Since directly applying flow matching [33] in the high-dimensional space is challenging [56, 29], we introduce a lightweight projection that compresses DINOv2 features into a lower-dimensional latent space, which stabilizes training and improves long-horizon rollout quality, while retaining the representation properties of DINO. The goal of the low-level predictor is then to capture fine-grained details and precise motion dynamics, conditioned on the high-level prediction. This is naturally suited to pixel-reconstruction based latent spaces such as VAEs [28], which preserve the fine-grained visual detail needed for short-horizon, high-fidelity generation.

To implement the two predictors, we rely on flow matching [33], which was shown to be particularly effective for long-horizon prediction in driving world models [35]. As the major objective of world models is to generalize to unseen scenarios observed previously only in different contexts, an important indicator of the quality of a world model is the quality of its learned internal representations. Driving world models have typically been evaluated with FVD [12, 1, 35], which measures the visual fidelity of generated frames. The high-level predictor operating in an abstract latent space, however, is not expected to reproduce fine-grained appearance, even though it may capture the scene structure that matters for long-horizon prediction. Thus, we additionally evaluate the quality of the internal representations of both predictors via linear probing on semantic segmentation and depth estimation, which directly measures how well the learned latent space encodes scene structure and geometry. In the proposed hierarchical world model, both predictors can play to their advantages: the abstract predictor learns higher-quality structural representations and dynamics, whereas the low-level predictor can satisfy the requirements of high-fidelity, short-horizon roll-outs with strong FVD scores. By probing the internal representations of the model, we find that the standard teacher forcing recipe — conditioning on clean ground-truth context to predict the target frame — fails to learn strong internal representations, as the model never observes out-of-distribution context during training. Pretraining with a *diffusion forcing* objective [6], which independently corrupts and denoises all frames in the sequence, produces substantially richer internal representations.

We summarize our contributions as follows:

A hierarchical driving world model. We introduce a two-level world model that decomposes future prediction into a high-level predictor operating in an abstract latent space and a low-level predictor producing high-fidelity frames.

A two-stage training paradigm for predictive representation learning. We propose pretraining with a diffusion forcing objective followed by short fine-tuning with teacher forcing. This combination enables self-supervised representation learning directly from the temporal prediction loss and substantially improves representation quality over only next frame prediction objective.

State-of-the-art results on driving world modeling. As shown in Fig. 2, our model achieves state-of-the-art performance on video-based driving world modeling across complementary axes: long-horizon perceptual fidelity (FVD) and representation quality (linear probing for semantic segmentation and depth). Our model also outperforms prior methods in steering responsiveness on counterfactual scenarios.

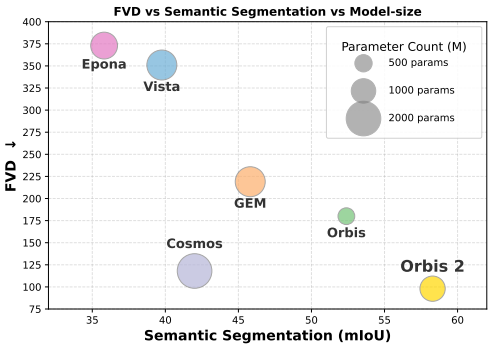


Figure 2: Comparison of driving world models over model scale (parameters), FVD (6s rollouts) and representation quality based on probing for semantic segmentation. FVD evaluated on Waymo and segmentation on Cityscapes.

2 Related Work

World models for autonomous driving aim to predict future scene evolution conditioned on past observations and ego actions. A first line of work models this prediction in *discrete token space*: GAIA-1 [22], DrivingWorld [23], and DrivingGPT [7] employ autoregressive transformers over quantized visual tokens, while Cosmos [1] scales this paradigm to large-scale video pretraining. A second line operates in *continuous latent space*, typically with diffusion [20, 25] or flow-matching [33] objectives: Vista [12], DriveDreamer [51], GenAD [52], MagicDrive-V2 [11], Epona [54], and Orbis [35] demonstrate high-fidelity multi-frame generation across diverse driving scenarios. Orbis directly compares these two paradigms under a hybrid tokenizer and finds that continuous representations are less brittle and more effective for long-horizon prediction than discrete tokens. Despite differing in their representation choices, all of these models share a single level of abstraction in their predictor.

2.1 Hierarchical planners

In reinforcement learning, hierarchical approaches have been a popular paradigm to learn both high-level skills and detailed control [46, 31]. Director [16] learns a two-level hierarchy where a high-level planner selects subgoals in latent space and a low-level policy executes them. Many recent vision-language-action models can also be regarded as hierarchical planners, with language-based planning covering the abstract, strategic level, whereas a fine-grained policy covers low-level control. For example, Hi Robot [43] interprets complex instructions with a VLM to determine the next step, which conditions a low-level action policy. HiLAM [26] discovers latent skills in an unsupervised fashion by modeling long-term temporal structure, then uses them to condition short-horizon prediction.

2.2 World models with abstract latent spaces

THICK [14] learns multi-scale temporal abstractions in the context of hierarchical model-based RL. This is closest to our work, as it includes multiple abstractions on the side of the predictor. DINO-WM [57] does not implement a hierarchy, but uses DINO features as latents for its predictor, which is similar to our abstract-level predictor. In concurrent work, Zhang et al. [55] use predictors at two levels for hierarchical planning, but both predictors share the same latent space. They demonstrate that planning with this hierarchy of predictors substantially improves the success rate compared to various flat models.

3 Method

3.1 Preliminary: Flow Matching

We adopt the flow matching objective [33] to model the distribution of future frames $\mathbf{x}_N$ in latent space. A noisy interpolant is constructed as $\mathbf{x}_N^\tau = (1-\tau)\mathbf{x}_N + \tau\epsilon$, where $\epsilon \sim \mathcal{N}(0, I)$ and $\tau \in [0, 1]$,

defining a linear path from data to noise. A velocity network $\mathbf{v}(\mathbf{x}_N^\tau;\, \mathbf{x}_{0:N-1})$, conditioned on context frames $\mathbf{x}_{0:N-1}$, is trained to regress the ground-truth velocity field:

$$\mathcal{L}_{\mathrm{FM}} = \mathbb{E}_{\tau,\,\epsilon}\big[\|\mathbf{v}(\mathbf{x}_N^\tau;\, \mathbf{x}_{0:N-1}) - (\epsilon - \mathbf{x}_N)\|^2\big]. \tag{1}$$

During inference, the predicted frame is generated by integrating the learned velocity field from $\tau{=}1$ to $\tau{=}0$, initialized from $\mathbf{x}_N^1 \sim \mathcal{N}(0, I)$.

Teacher forcing and diffusion forcing: A central design choice is how the velocity network conditions on context frames during training. Under **teacher forcing**, context frames are always provided as clean ground-truth observations and the velocity is predicted only for the last frame, as in eq. (1). In contrast, **diffusion forcing** [6] independently corrupts each frame $\mathbf{x}_i$ in the full sequence $\mathbf{x}_{0:N}$ with an independent noise level $\tau_i \sim [0, 1]$, and trains the model to predict the velocity for *all* frames jointly:

$$\mathcal{L}_{\mathrm{DF}} = \mathbb{E}_{\tau_{0:N},\,\epsilon_{0:N}}\big[\|\mathbf{v}(\mathbf{x}_{0:N}^{\tau_{0:N}}) - (\epsilon_{0:N} - \mathbf{x}_{0:N})\|^2\big]. \tag{2}$$

While teacher forcing has emerged as a common training objective for video world models [35, 1, 54, 49], we find that it yields weaker internal representations, as the model is never exposed to corrupted context during training. We demonstrate that diffusion forcing leads to stronger internal representations.

3.2 Hierarchical World Model

We propose a two-level hierarchical world model that decomposes future prediction along both abstraction and temporal axes. The high-level predictor $\mathcal{F}_H$ operates at a coarse temporal resolution in an abstract latent space, capturing long-horizon dynamics. The low-level predictor $\mathcal{F}_L$ operates at a finer temporal resolution in a pixel-aligned latent space, making detailed short-horizon prediction conditioned on the high-level prediction. An overview of the architecture is shown in Figure 1.

Tokenizer latent representations: Let o denote an observed driving frame. For each frame o, we extract two complementary representations. The fine-grained latent $\mathbf{x} = \mathcal{E}_L(o)$ is produced by a tokenizer encoder $\mathcal{E}_L$ trained with a reconstruction objective [42, 10]. This latent preserves pixel-aligned details and is used by the low-level predictor. In parallel, we learn an abstract encoder $\mathcal{E}_H$ that produces a high-level latent representation, $\mathbf{z} = \mathcal{E}_H(o)$. Unlike the detail tokenizer, $\mathbf{z}$ is optimized to capture semantic and structural information useful for long-horizon prediction. We train $\mathcal{E}_H$ by aligning its representation with a frozen DINO embedding [38] through a learned projection head $\mathcal{P}$. We add a weak reconstruction regularizer through a lightweight decoder $\mathcal{D}_H$, weighted by $\lambda_{\mathrm{rec}} = 0.1$, so that the abstract latent retains task-relevant visual information without degrading its semantic alignment. Formally, given an observation o and its frozen DINO embedding $\mathcal{E}_{\mathrm{DINO}}(o)$, we train $\mathcal{E}_H$, $\mathcal{P}$, and $\mathcal{D}_H$ with:

$$\mathcal{L}_{\mathrm{abs}} \;=\; \underbrace{\big\|\mathcal{P}\big(\mathcal{E}_H(o)\big) - \mathcal{E}_{\mathrm{DINO}}(o)\big\|^2}_{\text{DINO alignment}} + \lambda_{\mathrm{rec}} \underbrace{\big\|\mathcal{D}_H\big(\mathcal{E}_H(o)\big) - o\big\|^2}_{\text{reconstruction regularizer}}, \tag{3}$$

The resulting abstract latent $\mathbf{z}$ is only 16-dimensional latent space, providing a compact DINO-aligned representation that is well suited to the flow-matching objective of the high-level predictor.

Abstract long-horizon predictor: The high-level predictor $\mathcal{F}_H$ models future dynamics in the abstract latent space at a coarse temporal resolution. Let $\mathbf{z}_i$ denote the abstract latent of the frame at time step i, and let $\mathbf{Z}_N^k = \{\mathbf{z}_0, \mathbf{z}_k, \mathbf{z}_{2k}, \ldots, \mathbf{z}_N\}$ denote the strided context of past abstract latents sampled every k frames. Conditioned on this context and the current ego action a_N, $\mathcal{F}_H$ predicts the abstract latent k steps into the future:

$$\hat{\mathbf{z}}_{N+k} \;=\; \mathcal{F}_H\big(\mathbf{Z}_N^k,\, a_N\big). \tag{4}$$

Detailed short-horizon predictor: The low-level predictor $\mathcal{F}_L$ operates at fine temporal resolution and is responsible for generating the k intermediate fine-grained frames between $\mathbf{z}_N$ and $\hat{\mathbf{z}}_{N+k}$. Let $\mathbf{x}_i$ denote the fine-grained latent at time step i. Conditioned on its own short high-frequency context window of m frames and the abstract subgoal $\hat{\mathbf{z}}_{N+k}$, $\mathcal{F}_L$ predicts the next k frames:

$$\hat{\mathbf{x}}_{N+1:N+k} = \mathcal{F}_L\big(\mathbf{x}_{N-m:N},\, \hat{\mathbf{z}}_{N+k}\big). \tag{5}$$

During training, the abstract high-level predictor is pre-trained using the diffusion forcing loss eq. (2) and fine-tuned by teacher forcing loss eq. (1). The detailed low-level predictor is trained directly with a teacher-forcing objective conditioned on the high-level predictor output.

Action-conditioning: We first pre-train the hierarchical predictors unconditionally, without action information. We then fine-tune the world model to enable conditional generation on planar trajectories. The embedded steering vector is added to the τ embedding, and fed to the predictor via adaptive layer normalization layers [40, 39, 3]. We only condition the top-level predictor branch.

3.3 Diffusion-forcing pre-training for improved predictor representations

Diffusion forcing [6] was originally introduced to improve rollout length by mitigating exposure bias in sequential generation: models trained with teacher forcing observe clean ground-truth context during training, but must condition on their own prediction at inference time. This typically causes error to compound over long rollouts. We discover that diffusion forcing also substantially improves the internal representations learned by the predictor. Building on this finding, we propose using diffusion forcing as a pre-training stage followed by teacher-forcing fine-tuning for the predictor training. The fine-tuning stage aligns the predictor with the downstream future-prediction objective while retaining the representation quality gained during diffusion forcing.

We validate this effect through linear probes on semantic segmentation and depth estimation, and observe consistent gains across predictors operating in both detail and abstract tokenizer latent spaces, as well as across both probing tasks. This training strategy is independent of our hierarchical architecture and generalizes to any predictor level. In our model, we apply it to the abstract predictor $\mathcal{F}_H$, where it yields stronger high-level dynamics representations and better downstream world modeling performance.

4 Experiments

4.1 Datasets and Training Details

Table 1 provides an overview of the training data used in training both predictors. Both levels are trained for a single epoch over the full 5,890 hours of video. Both tokenizers are trained on a curated mix of 2.6 million frames sampled from BDD100K [53], OpenDV [52], Honda HAD [27], Honda HDD [41], ONCE [34], nuScenes [4], and nuPlan [5]. For steering-conditional generation, the model is fine-tuned on a 500h sub-set of the NVIDIA PhysicalAI AV dataset [37] using the dataset's IMU annotations to compute planar trajectory inputs.

The video dataset [36] used in this study was acquired in partnership with NATIX. The dataset is partially publicly available. It is a large-scale, crowd-sourced collection of multi-camera driving data spanning the United States, Europe, and Japan, covering varied road conditions, traffic environments, and regional driving behaviors. For this work, we used a 2,374-hour subset of front-camera driving video data. In future work, we plan to extend this approach using multi-camera data.

The high-level predictor operates at 2Hz with a window of 6 frames (3 seconds), taking 5 frames as input and predicting 1 frame. Its predicted abstract latent serves as the subgoal that conditions the low-level predictor. The low-level predictor operates at 10 Hz with a window of 10 frames (1second), taking 5 frames as context and generating the next 5 frames. During training, the low-level predictor is conditioned on ground-truth abstract latents. At inference, it uses the predicted latents from the high-level predictor. Rollouts are performed autoregressively: the high-level predictor advances with a sliding window of 1 frame and the low-level predictor with a sliding window of 5 frames. Both predictor models are trained at a resolution of 512×288.

Table 1: Overview of the training datasets used for the world model.*partially public

	OpenDV [52]	BDD100k [53]	PhysicalAI-AV [37]	NATIX [36]*	Total
Hours (h)	1124	659	1633	2374	5890

4.2 Model design

For the first stage, we use the hybrid tokenizer proposed in [35], consisting of a ViT-based [9] image encoder with a vector-quantized bottleneck for regularization, trained with a VQGAN [10] objective. The encoder contains 86M parameters, while the decoder is a CNN-based architecture with 91M parameters. In the hierarchical setup, we train two separate tokenizers for each predictor. The high-level tokenizer is trained primarily for DINOv2B embedding reconstruction, with a small pixel reconstruction weight of $\lambda_{\text{rec}} = 0.1$. The low-level tokenizer is trained with pixel-aligned reconstruction objectives following VQGAN [10].

Following [35], we extend the DiT block [39] into a Spatio-Temporal Transformer (ST-Transformer) backbone with 512M parameters. The low-level detail predictor adds 43M parameters through spatially aligned conditioning via adaLN, summing to 555M parameters. The full hierarchical predictor totals 1067M parameters. We maintain an exponential moving average (EMA) of the model weights with a decay of 0.9999 and use the EMA model at inference. For the model-internal design studies, we train at a lower resolution of 336×192 with a smaller 257M abstract predictor and 279M detail predictor.

4.3 Evaluation

Representation quality. We evaluate internal representations via linear probing on two downstream tasks: semantic segmentation on Cityscapes [8], reporting mIoU, and depth estimation on KITTI [13], reporting RMSE. Details about linear probing setup are included in Appendix A.7.

Predictive performance. We assess the visual quality and diversity of predicted rollouts using Fréchet Video Distance (FVD) [48], evaluated on three benchmarks: Waymo [44], nuPlan, and nuPlan-turns [35]. nuPlan-turns is a subset of nuPlan dataset focused on turning scenarios. 800 sequences are used for nuPlan benchmark, and 400 sequences for Waymo and nuPlan-turns benchmarks.

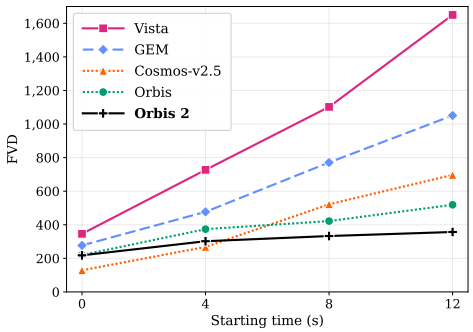


Figure 3: Performance (FVD) of long-horizon rollouts on nuPlan-turns over consecutive 4s windows. Models with good initial video quality can quickly degrade over time. FVD-slope captures this progressive fidelity loss.

Long-horizon stability. In a predictive setting, standard FVD can be confounded by shared conditioning frames at the beginning of each rollout. Methods that prioritize perceptual fidelity or are optimized for long-chunk generation can achieve strong short-horizon FVD without truly capturing semantics or physical structure. To address this, we adopt the chunked-FVD metric from [35], which measures FVD over consecutive 4-second windows. We additionally introduce **FVD-slope**, a scalar metric capturing the temporal degradation of video quality across consecutive chunks: FVD-slope $= \frac{\text{FVD}_T - \text{FVD}_1}{T-1}$, where FVD_t denotes the FVD of the t-th chunk and T is the total number of chunks. A lower FVD-slope indicates greater stability over long-horizon prediction.

Action controllability. Given the importance of action-conditional generation for world models, we evaluate their ability to follow given steering commands in the form of 2D trajectories. Furthermore, we evaluate the quality of the ego-trajectories generated by the model in the unconditional case, measuring how well the world model covers the real distribution of motion patterns.

5 Results

5.1 Comparison to State-of-the-art

Long-horizon rollout performance The ability to perform long-horizon rollouts reflects the quality of the learned transition function - how well the world model maps past states to future states. We first evaluate this by comparing driving world models using FVD. Table 2 (first three columns) reports FVD over the 6-second horizon window. Cosmos-v2.5 (version: Predict-2B, post-trained, Video2World) achieves the lowest FVD scores for nuPlan benchmarks. However, it is 2× larger

Table 2: FVD over 6s rollouts on nuPlan, Waymo and nuPlan-turns datasets. FVD-slope is shown for 4 chunks of 4 seconds each. Linear probing results are shown for semantic segmentation on Cityscapes and depth estimation on KITTI. *Epona is trained only on the nuPlan and nuScenes datasets. The training mix of Cosmos is not public.

Model	FVD↓			FVD-slope↓	Segmentation	Depth
	nuPlan	Waymo	nuPlan-turns	nuPlan-turns	mIoU↑	RMSE↓
Vista	289.95	351.42	353.27	434.67	39.76	4.884
GEM	348.36	218.61	318.73	258.00	45.81	4.373
Orbis	132.25	180.54	239.05	99.67	52.39	4.321
Epona*	251.81	373.03	123.60	93.89	35.81	5.720
Cosmos-v2.5*	**67.74**	116.12	**126.26**	189.34	43.20	6.384
Orbis 2	98.97	**98.01**	187.53	**46.45**	**58.29**	**4.157**

than our model and is trained on over 3× more driving data from an undisclosed data mix. Despite this, our model outperforms Cosmos-v2.5 on the Waymo benchmark. We further compare long-horizon rollout stability using chunked-FVD (see Figure 3) and FVD-slope (see Table 2), where our model achieves the best scores, demonstrating significantly greater stability than all baselines including Cosmos-v2.5. Epona, which is trained only on nuPlan and nuScenes datasets, performs well on nuPlan benchmarks but fails to generalize to the unseen Waymo. Orbis shows the next-best rollout stability. It also uses two abstraction levels, but in a single branch setting and at same temporal resolution. A direct comparison with the Orbis (v1) model is added in Appendix A.2. Please find qualitative examples of long-horizon prediction for different baselines here: https://lmb-freiburg.github.io/orbis2.github.io/model_comparison.html

Conditional trajectory evaluation To evaluate the responsiveness of our model to steering commands, we consider two settings. 1) **Original trajectories**: using the unaltered ground truth steering signal corresponding to each validation sample (as done in [12, 18, 35]) and 2) **Counterfactual trajectories**: obtained by altering the original odometry, to construct alternative steering events (e.g. under/oversteering). Specifically, we obtain four counterfactual trajectories per sample by scaling the original speeds and yaw rates by a factor 0.5 and 1.5. It is essential to evaluate steering in a counterfactual setting, because the original trajectories are partly leaked by the context frames, i.e. a model that cannot steer could still score well by relying on context only. We evaluate the models by estimating the trajectory of the generated videos using an external inverse dynamics model, as done in [12, 18, 35], specifically VGGT [50]. We then compute the average displacement error (ADE) between the estimated and ground truth trajectories.

Table 3: Complete steering evaluation on original and counterfactual trajectories.

Model	ADE↓ (m)				
	Original	Counterfactual			
		speed×0.5	speed×1.5	yaw rate×0.5	yaw rate×1.5
Orbis	2.23	3.49	1.42	3.02	1.31
Epona	1.36	**2.08**	1.48	4.01	0.91
Orbis 2	**1.20**	2.18	**0.53**	**1.65**	**0.73**

Because VGGT trajectories are scale-ambiguous, we align each generated trajectory $\mathbf{x}$ to its real counterpart $\bar{\mathbf{x}}$ before computing ADE using the least-squares scale $s^* = \left(\sum_t \mathbf{x}_t^\top \bar{\mathbf{x}}_t\right) / \left(\sum_t \mathbf{x}_t^\top \mathbf{x}_t\right)$. This removes global scale error while preserving penalties for shape, direction, and curvature. We evaluate one original and four counterfactual trajectory modes on 100 six-second nuPlan-turns samples (500 generations total). Table 3 shows that our model is best on the original trajectory and three counterfactual cases; Epona is slightly better for speed×0.5. Some qualitative examples are included in Appendix A.5.

5.2 Strong predictor representations lead to stable world models

We evaluate the internal representation quality of driving world models by linear probing on semantic segmentation and depth estimation (Table 2). We observe a clear trend: models with stronger semantic and geometric understanding also achieve more stable long-horizon prediction, as measured by FVD-slope. Our model achieves the best results on both probing tasks and the long-horizon stability metric. Orbis (v1), which also learns strong representations, achieves the second-best long-horizon stability. Several baselines operate at significantly higher resolutions to maximize perceptual fidelity, yet our model is more stable at 512×288 resolution. This suggests that representation quality plays a complementary, and arguably more important role in world modeling than perceptual fidelity alone.

5.3 Comparison between hierarchical variants

Table 4 compares the design choices for the top and bottom predictors (illustrated in fig. 1), varying the level of latent abstraction, frame rate, and training mode. The first two rows establish flat (single-branch) baselines at different abstraction levels. The detail-level model (Exp. 1) achieves strong but progressively degrading FVD and learns weak internal representations, whereas the abstract-level model (Exp. 2) yields substantially richer representations and remarkably stable rollout quality over long horizons, despite its high absolute FVD. The hierarchical configurations (Exps. 3–5) all fix the bottom predictor to operate on detailed latents at 10 fps, and vary the top predictor. Exp. 3 shows the best hierarchical design, where the top predictor operates on high abstraction latents at a longer horizon of 2 fps: it achieves both strong representations and the best FVD-slope with a competitive absolute FVD performance. Exp. 4 has high FVD-slope: the subpar long-horizon rollouts of the top predictor in the detailed latent space cause the model to drift out-of-distribution. In Exp. 5, the top predictor operating in abstract space in high-temporal resolution leads to stopping behavior in rollouts. We attribute this to the abstract latent space is not detailed enough to capture small motion at 10 fps, which leads to loss in motion signal. All bottom predictors do not learn strong representations due to their detailed predictive task with strong conditioning. The best model design is hierarchical with an abstract-latent top predictor at 2 fps (Exp. 3): this offers the best balance between overall (FVD) and long-horizon video fidelity (FVD-slope), simultaneously achieving the strongest representation quality. Figure 4 qualitatively confirms these trends: Exp. 1 drifts, Exp. 2 is stable but lacks detail, Exp. 4 diverges out-of-distribution, and Exp. 5 progressively loses motion, whereas Exp. 3 remains stable and perceptually detailed.

Table 4: Hierarchy design study across different abstractions and temporal resolutions. All experiments have comparable number of model parameters. Exp. 3 corresponds to our final hierarchical setting. Top/Bottom refers position of predictors in fig. 1. FVD and FVD-slope are reported on the nuPlan-turns benchmark.

Exp.	Model branch	Abstraction detail/abstract	Frame-rate fps	Sem. Seg. mIoU ↑	Depth RMSE ↓	FVD↓ 16s	FVD-slope↓ 4 x 4s
1.	Single-branch	detail	10	32.72	6.107	170.02	101.63
2.	Single-branch	abstract	10	47.88	4.624	490.04	48.38
3.	Top ↳ Bottom	abstract detail	2 10	49.39 31.97	4.498 6.396	209.15	47.87
4.	Top ↳ Bottom	detail detail	2 10	41.02 25.52	5.256 6.514	207.71	81.40
5.	Top ↳ Bottom	abstract detail	10 10	49.27 31.46	4.460 6.311	551.38	137.07

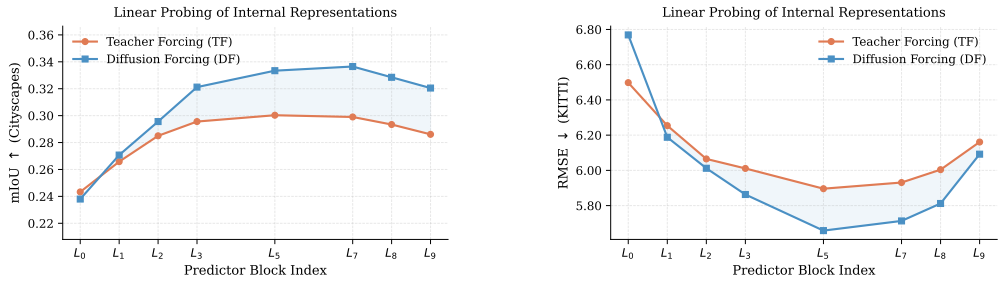
5.4 Training strategy: diffusion forcing + teacher forcing

We evaluate the effect of diffusion forcing pre-training followed by teacher forcing fine-tuning on the representation quality of the abstract predictor. Table 5 reports linear probing results for semantic segmentation and depth estimation. Diffusion forcing improves over teacher forcing alone on both tasks, and subsequent teacher forcing fine-tuning provides additional gains, yielding the best overall probing performance among our predictor variants. The final predictor outperforms the abstract

Figure 4: Qualitative examples showing behavior of different hierarchy variants discussed in Table 4. We observe that Exp-2 rollouts are stable but lacks details, Exp-4 drifts out-of-distribution and Exp-5 suffers from loss in motion.

tokenizer based on DINOv2-B on both probing tasks, improving mIoU from 57.19 to 58.29 and reducing RMSE from 4.574 to 4.157. Further, Figure 5 shows the same trend for the predictor trained on pixel reconstruction-based tokenizer latents, where diffusion forcing improves predictor representation across all blocks of the model.

DINOv2-Base shows stronger overall performance in segmentation probing, but our predictor achieves better performance on frequent driving classes such as road, car, sidewalk, and vegetation, while underperforming on rarer classes such as bus, train, and motorcycle (see Appendix A.1 for class-wise performance). Compressing DINOv2 features into the tokenizer reduces depth estimation performance, but our training strategy significantly narrows the gap to DINOv2.



(a) Semantic segmentation (mIoU ↑) on Cityscapes.

(b) Depth estimation (RMSE ↓) on KITTI.

Figure 5: Representation quality per predictor block via linear probing for the predictor trained on detail latents. Diffusion forcing consistently improves intermediate representations on both semantic segmentation and depth estimation.

5.5 More analysis

Model efficiency. Our final 1.07B-parameter model is trained in under 6k H100-GPU-hours (6×10^{21} FLOPs at 25% MFU), substantially less compute than comparable driving world models.

Table 5: Ablation of the proposed training strategy. All metrics are reported for the abstract branch including tokenizer and predictor. mIoU and val loss are for semantic segmentation, RMSE for depth estimation.

Model	Segmentation mIoU ↑ (val loss↓)	Depth RMSE ↓
Abstract Tokenizer	57.19 (0.265)	4.574
Teacher forcing (TF) only	55.01 (0.270)	4.355
Diffusion forcing (DF) only	58.03 (0.254)	4.274
DF + TF (Ours final)	58.29 (0.250)	4.157
DINOv2-Base	60.01 (0.284)	3.890

Table 6: Inference speed and VRAM requirements across models. VRAM shown in GB.

Method	FPS↑	VRAM ↓
Cosmos v2.5	0.66	22
Vista	0.58	86
GEM	0.44	45
Epona	0.4	17
Orbis	0.70	24
Orbis 2	3.64	19

Figure 6: Prediction in full 768-dimensional DINO space results in poor generation quality. Compression enables stable, long-horizon generation even with small models. Rollouts shown at 2fps for our compressed (16-dim) and full DINO latents. More qualitative examples in Appendix.

GAIA-1's 6.5B world model required 23k A100-hours, Epona used 16k A100-hours for a similarly sized 2.5B model, and Orbis (v1) used 8k A100-hours to train a 469M-parameter model, less than half our size. Our model achieves the fastest inference at 3.64 FPS while requiring only 19 GB VRAM, at least $5\times$ faster than all baselines, shown in Table 6. Inference can be further accelerated with `torch.compile`, with minimal code changes.

Abstract latent design: compressed vs. non-compressed DINO. A natural alternative to our compressed DINO latent space is to predict directly in the high-dimensional DINO space, as proposed in Representation Autoencoders (RAE) [56]. RAE could in principle serve as the abstract predictor in our hierarchy by replacing our compressed latent with raw DINO features. However, we find that their dimensionality is prohibitive for autoregressive rollouts: as shown in fig. 6, frame quality degrades progressively over the prediction horizon when the predictor operates directly in full DINO space. Our compressed DINO latent retains the semantic structure of DINO at reduced dimensionality, which proves essential for stable long-horizon prediction with a flow matching objective. More qualitative results are included in Appendix A.3.

6 Conclusion

We demonstrated that the proposed hierarchical approach makes effective use of abstract and detailed latent spaces, yielding a world model that combines high-fidelity short-term predictions with stability over long horizons. The model is steerable through action conditioning. We further demonstrated a strong correlation between representation quality and rollout stability, and showed that a diffusion forcing pre-training strategy consistently strengthens the learned representations.

Limitations. Although diffusion forcing improves predictor representations, the reason for this improvement is not fully studied in this work. We hypothesize that independent noise levels across frames act as implicit data augmentation, while different noise patterns encourage the model to learn from related tasks such as denoising, interpolation, and forecasting. We do not study these factors separately in this work, and leave a more detailed analysis of noise schedules and corruption patterns to future work. Our hierarchy relies on separate abstract and fine-grained latent spaces, each optimized with a different objective. While this separation clarifies the role of each level, it limits

the interaction between abstraction levels. Improving the connection between the two spaces is an important direction for future work.

Societal Impact. At this stage, the work remains an early research contribution, and any practical societal effects will depend on substantial future progress in model design, scaling, validation, and system integration.

Acknowledgment

This work was funded by the German Federal Ministry for Economic Affairs and Energy within the project “NXT GEN AI METHODS" (19A23014R). The authors gratefully acknowledge the Gauss Centre for Supercomputing e.V. (www.gauss-centre.eu) for providing computing time through the John von Neumann Institute for Computing (NIC) on the GCS Supercomputer JUWELS [24] and JUPITER at Jülich Supercomputing Centre (JSC). This project was selected as a winner of Gauss AI Compute Competition, organized by the Gauss Centre for Supercomputing (GCS) and Jülich Supercomputing Centre (JSC). The project was accomplished under GCS compute grants - *multiscale-wm* and *nxtaim-1*. Further local compute resources were funded through the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under grant numbers 417962828, 539134284, through EFRE (FEIH_2698644), and the state of Baden-Württemberg.

We thank NATIX for providing access to their large-scale, crowd-sourced, multi-camera driving dataset [36] in support of this open-source research. We thank Marcel Aach for his technical support and assistance with the GCS HPC cluster. We also thank Muhammad Ali for improving the model latency for the Hugging Face demo.

References

[1] Niket Agarwal, Arslan Ali, Maciej Bala, Yogesh Balaji, Erik Barker, Tiffany Cai, Prithvijit Chattopadhyay, Yongxin Chen, Yin Cui, Yifan Ding, et al. Cosmos world foundation model platform for physical ai. *arXiv preprint arXiv:2501.03575*, 2025.

[2] Mido Assran, Adrien Bardes, David Fan, Quentin Garrido, Russell Howes, Matthew Muckley, Ammar Rizvi, Claire Roberts, Koustuv Sinha, Artem Zholus, et al. V-jepa 2: Self-supervised video models enable understanding, prediction and planning. *arXiv preprint arXiv:2506.09985*, 2025.

[3] Amir Bar, Gaoyue Zhou, Danny Tran, Trevor Darrell, and Yann LeCun. Navigation world models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 15791–15801, 2025.

[4] Holger Caesar, Varun Bankiti, Alex H. Lang, Sourabh Vora, Venice Erin Liong, Qiang Xu, Anush Krishnan, Yu Pan, Giancarlo Baldan, and Oscar Beijbom. nuscenes: A multimodal dataset for autonomous driving. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.

[5] Holger Caesar, Juraj Kabzan, Kok Seang Tan, Whye Kit Fong, Eric Wolff, Alex Lang, Luke Fletcher, Oscar Beijbom, and Sammy Omari. nuplan: A closed-loop ml-based planning benchmark for autonomous vehicles. *arXiv preprint arXiv:2106.11810*, 2021.

[6] Boyuan Chen, Diego Martí Monsó, Yilun Du, Max Simchowitz, Russ Tedrake, and Vincent Sitzmann. Diffusion forcing: Next-token prediction meets full-sequence diffusion. *Advances in Neural Information Processing Systems*, 37:24081–24125, 2024.

[7] Yuntao Chen, Yuqi Wang, and Zhaoxiang Zhang. Drivinggpt: Unifying driving world modeling and planning with multi-modal autoregressive transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 26890–26900, October 2025.

[8] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The cityscapes dataset for semantic urban scene understanding. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3213–3223, 2016.

[9] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

[10] Patrick Esser, Robin Rombach, and Bjorn Ommer. Taming transformers for high-resolution image synthesis. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12873–12883, 2021.

[11] Ruiyuan Gao, Kai Chen, Bo Xiao, Lanqing Hong, Zhenguo Li, and Qiang Xu. Magicdrive-v2: High-resolution long video generation for autonomous driving with adaptive control. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 28135–28144, October 2025.

[12] Shenyuan Gao, Jiazhi Yang, Li Chen, Kashyap Chitta, Yihang Qiu, Andreas Geiger, Jun Zhang, and Hongyang Li. Vista: A Generalizable Driving World Model with High Fidelity and Versatile Controllability. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang, editors, *Advances in Neural Information Processing Systems*, volume 37, pages 91560–91596. Curran Associates, Inc., 2024.

[13] Andreas Geiger, Philip Lenz, and Raquel Urtasun. Are we ready for autonomous driving? the kitti vision benchmark suite. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2012.

[14] Christian Gumbsch, Noor Sajid, Georg Martius, and Martin V. Butz. Learning hierarchical world models with adaptive temporal abstractions from discrete latent dynamics. In *The Twelfth International Conference on Learning Representations*, 2024.

[15] David Ha and Jürgen Schmidhuber. World models. *arXiv preprint arXiv:1803.10122*, 2(3):440, 2018.

[16] Danijar Hafner, Kuang-Huei Lee, Ian Fischer, and Pieter Abbeel. Deep hierarchical planning from pixels. *Advances in Neural Information Processing Systems*, 35:26091–26104, 2022.

[17] Danijar Hafner, Jurgis Pasukonis, Jimmy Ba, and Timothy Lillicrap. Mastering diverse domains through world models. *arXiv preprint arXiv:2301.04104*, 2023.

[18] Mariam Hassan, Sebastian Stapf, Ahmad Rahimi, Pedro Rezende, Yasaman Haghighi, David Brüggemann, Isinsu Katircioglu, Lin Zhang, Xiaoran Chen, Suman Saha, et al. Gem: A generalizable ego-vision multimodal world model for fine-grained ego-motion, object dynamics, and scene composition control. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22404–22415, 2025.

[19] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16000–16009, 2022.

[20] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020.

[21] Mark K Ho, David Abel, Carlos G Correa, Michael L Littman, Jonathan D Cohen, and Thomas L Griffiths. People construct simplified mental representations to plan. *Nature*, 606(7912):129–136, 2022.

[22] Anthony Hu, Lloyd Russell, Hudson Yeo, Zak Murez, George Fedoseev, Alex Kendall, Jamie Shotton, and Gianluca Corrado. Gaia-1: A generative world model for autonomous driving. *arXiv preprint arXiv:2309.17080*, 2023.

[23] Xiaotao Hu, Wei Yin, Mingkai Jia, Junyuan Deng, Xiaoyang Guo, Qian Zhang, Xiaoxiao Long, and Ping Tan. Drivingworld: Constructing world model for autonomous driving via video gpt. *arXiv preprint arXiv:2412.19505*, 2024.

[24] Jülich Supercomputing Centre. JUWELS Cluster and Booster: Exascale Pathfinder with Modular Supercomputing Architecture at Juelich Supercomputing Centre. *Journal of large-scale research facilities*, 7(A138), 2021.

[25] Tero Karras, Miika Aittala, Timo Aila, and Samuli Laine. Elucidating the design space of diffusion-based generative models. *Advances in neural information processing systems*, 35:26565–26577, 2022.

[26] Hanjung Kim, Lerrel Pinto, and Seon Joo Kim. Hierarchical latent action model. *arXiv preprint arXiv:2603.05815*, 2026.

[27] Jinkyu Kim, Teruhisa Misu, Yi-Ting Chen, Ashish Tawari, and John Canny. Grounding human-to-vehicle advice for self-driving vehicles. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

[28] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

[29] Amandeep Kumar and Vishal M Patel. Learning on the manifold: Unlocking standard diffusion transformers with representation encoders. *arXiv preprint arXiv:2602.10099*, 2026.

[30] Yann LeCun and Courant. A path towards autonomous machine intelligence version 0.9.2, 2022-06-27. 2022.

[31] Alexander Li, Carlos Florensa, Ignasi Clavera, and Pieter Abbeel. Sub-policy adaptation for hierarchical reinforcement learning. In *International Conference on Learning Representations*, 2020.

[32] Xiang Li, Kai Qiu, Hao Chen, Jason Kuen, Jiuxiang Gu, Bhiksha Raj, and Zhe Lin. Imagefolder: Autoregressive image generation with folded tokens. *arXiv preprint arXiv:2410.01756*, 2024.

[33] Yaron Lipman, Ricky TQ Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow matching for generative modeling. *arXiv preprint arXiv:2210.02747*, 2022.

[34] Jiageng Mao, Minzhe Niu, Chenhan Jiang, Hanxue Liang, Jingheng Chen, Xiaodan Liang, Yamin Li, Chaoqiang Ye, Wei Zhang, Zhenguo Li, et al. One million scenes for autonomous driving: Once dataset. *arXiv preprint arXiv:2106.11037*, 2021.

[35] Arian Mousakhan, Sudhanshu Mittal, Silvio Galesso, Karim Farid, and Thomas Brox. Overcoming challenges of long-horizon prediction in driving world models. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2026.

[36] NATIX. NATIX Multi-Camera Driving Dataset. Hugging Face Hub: https://huggingface.co/datasets/natix-network-org/natix-multi-camera-driving-dataset, 2026. Multi-camera driving dataset with telemetry metadata. NATIX website: https://www.natix.network/.

[37] NVIDIA. Physical ai autonomous vehicles dataset, 2025.

[38] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023.

[39] William Peebles and Saining Xie. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 4195–4205, 2023.

[40] Ethan Perez, Florian Strub, Harm De Vries, Vincent Dumoulin, and Aaron Courville. Film: Visual reasoning with a general conditioning layer. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32, 2018.

[41] Vasili Ramanishka, Yi-Ting Chen, Teruhisa Misu, and Kate Saenko. Toward driving scene understanding: A dataset for learning driver behavior and causal reasoning. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

[42] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022.

[43] Lucy Xiaoyang Shi, Brian Ichter, Michael Equi, Liyiming Ke, Karl Pertsch, Quan Vuong, James Tanner, Anna Walling, Haohuan Wang, Niccolo Fusai, et al. Hi robot: Open-ended instruction following with hierarchical vision-language-action models. *arXiv preprint arXiv:2502.19417*, 2025.

[44] Pei Sun, Henrik Kretzschmar, Xerxes Dotiwalla, Aurelien Chouard, Vijaysai Patnaik, Paul Tsui, James Guo, Yin Zhou, Yuning Chai, Benjamin Caine, et al. Scalability in perception for autonomous driving: Waymo open dataset. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2446–2454, 2020.

[45] Richard S Sutton. An adaptive network that constructs and uses and internal model of its world. *Cognition and Brain Theory*, 4(3):217–246, 1981.

[46] Richard S. Sutton, Doina Precup, and Satinder Singh. Between mdps and semi-mdps: A framework for temporal abstraction in reinforcement learning. *Artificial Intelligence*, 112(1):181–211, 1999.

[47] Momchil S. Tomov, Samyukta Yagati, Agni Kumar, Wanqian Yang, and Samuel J. Gershman. Discovery of hierarchical representations for efficient planning. *bioRxiv*, 2018.

[48] Thomas Unterthiner, Sjoerd van Steenkiste, Karol Kurach, Raphaël Marinier, Marcin Michalski, and Sylvain Gelly. FVD: A new metric for video generation, 2019.

[49] Dani Valevski, Yaniv Leviathan, Moab Arar, and Shlomi Fruchter. Diffusion models are real-time game engines. *arXiv preprint arXiv:2408.14837*, 2024.

[50] Jianyuan Wang, Minghao Chen, Nikita Karaev, Andrea Vedaldi, Christian Rupprecht, and David Novotny. Vggt: Visual geometry grounded transformer. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 5294–5306, 2025.

[51] Xiaofeng Wang, Zheng Zhu, Guan Huang, Xinze Chen, Jiagang Zhu, and Jiwen Lu. Drivedreamer: Towards real-world-drive world models for autonomous driving. In *European conference on computer vision*, pages 55–72. Springer, 2024.

[52] Jiazhi Yang, Shenyuan Gao, Yihang Qiu, Li Chen, Tianyu Li, Bo Dai, Kashyap Chitta, Penghao Wu, Jia Zeng, Ping Luo, et al. Genad: Generalized predictive model for autonomous driving. *arXiv preprint arXiv:2403.09630*, 2024.

[53] 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*, pages 2636–2645, 2020.

[54] Kaiwen Zhang, Zhenyu Tang, Xiaotao Hu, Xingang Pan, Xiaoyang Guo, Yuan Liu, Jingwei Huang, Li Yuan, Qian Zhang, Xiao-Xiao Long, et al. Epona: Autoregressive diffusion world model for autonomous driving. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 27220–27230, 2025.

[55] Wancong Zhang, Basile Terver, Artem Zholus, Soham Chitnis, Harsh Sutaria, Mido Assran, Randall Balestriero, Amir Bar, Adrien Bardes, Yann LeCun, et al. Hierarchical planning with latent world models. *arXiv preprint arXiv:2604.03208*, 2026.

[56] Boyang Zheng, Nanye Ma, Shengbang Tong, and Saining Xie. Diffusion transformers with representation autoencoders. *arXiv preprint arXiv:2510.11690*, 2025.

[57] Gaoyue Zhou, Hengkai Pan, Yann LeCun, and Lerrel Pinto. Dino-wm: World models on pre-trained visual features enable zero-shot planning. *arXiv preprint arXiv:2411.04983*, 2024.

A Appendix

A.1 Class-wise semantic segmentation probing results

In Table 7, we compare class-wise semantic segmentation probing performance for the tokenizer, predictor, and DINOv2-Base representations. DINOv2-Base achieves stronger overall segmentation performance, but our predictor performs better on frequent driving classes such as road, car, sidewalk, and vegetation, improving by 2–5 mIoU points. In contrast, it underperforms on rarer classes such as bus, train, rider, and motorcycle, with a gap of 5–20 mIoU points.

Table 7: Comparison of our tokenizer and abstract predictor model with DINOv2-base. Per-class IoU for semantic segmentation reported on Cityscapes.

Class	Dino v2-B	Ours tokenizer	Ours predictor	Δ (Predictor $-$ Dino v2-B)
Road	94.88	95.3	95.9	+1.02
Sidewalk	66.58	67.6	71.0	+4.42
Building	82.98	84.4	84.8	+1.82
Wall	48.28	47.7	47.9	−0.38
Fence	45.84	44.8	44.9	−0.94
Pole	17.54	16.7	18.5	+0.96
Traffic Light	33.97	30.1	31.5	−2.47
Traffic Sign	42.93	41.4	43.6	+0.67
Vegetation	80.87	83.3	83.6	+2.73
Terrain	53.05	53.5	53.4	+0.35
Sky	82.11	86.6	87.0	+4.89
Person	54.18	53.0	52.7	−1.48
Rider	34.04	28.1	28.2	−5.84
Car	84.27	85.0	86.4	+2.13
Truck	70.25	67.0	67.1	−3.15
Bus	77.80	63.4	69.5	−8.30
Train	70.57	48.5	50.0	−20.57
Motorcycle	45.48	37.4	38.1	−7.38
Bicycle	54.57	52.9	53.1	−1.47
mIoU	60.01	57.2	58.3	−1.71

A.2 Single-branch, mixed abstraction

Table 8 shows the comparison between our model and a single branch model with mixed abstraction levels, combined via token factorization [32], as in [35]. Despite good FVD and representation quality, video fidelity degrades faster over time, as shown by the higher FVD-slope.

Table 8: Comparison of single-branch and hierarchical predictors.

Model	Parameters M	Sem. Seg. mIoU ↑	Depth RMSE ↓	FVD↓ 16s	FVD-slope↓ 4 × 4s
Single-branch, mixed	512	47.16	4.680	184.69	96.55
Hierarchical (ours)	257+279	49.39	4.498	209.15	47.87

A.3 Qualitative rollouts with full DINO space vs compressed DINO space

Figure 7 shows qualitative results comparing 1-second rollouts generated using our compressed DINO latent targets versus the full 768-dim DINOv2 latent targets. Rollouts in the full-dimensional DINO space degrade rapidly, while the compressed DINO latent space is effectively modeled by the flow-matching objective, producing stable predictions throughout the horizon.

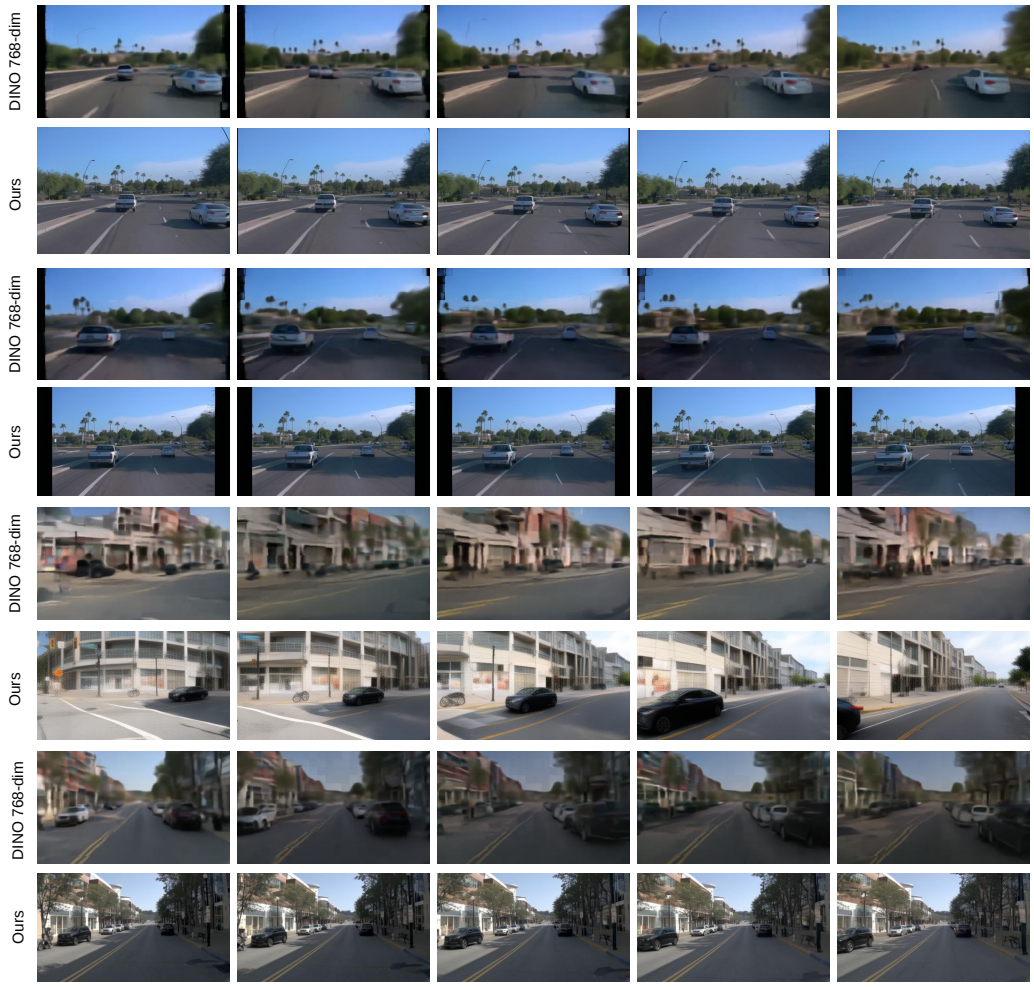


Figure 7: Comparison of raw DINO features versus our compressed DINO latent as the tokenizer for the abstract predictor. (Rollouts at 2hz)

A.4 Qualitative examples for different hierarchy variants

In Figure 4, we present qualitative examples illustrating the failure modes of the hierarchical design variants discussed in Section 5.3. Experiment-1, which operates in single-branch mode at the detailed abstraction level, produces rollouts that drift away from the starting distribution. Experiment-2, a single-branch model at the abstract level, yields stable rollouts but fails to capture fine-grained motion and perceptual detail. Experiment-3 corresponds to our proposed hierarchical design, achieving stable rollouts with high perceptual quality. Experiment-4, which uses detailed latents for the top branch, diverges out-of-distribution rapidly, resulting in poor rollouts. Experiment-5, which operates abstract latents at high temporal frequency, fails to capture detailed motion, leading to motion degradation despite retaining high perceptual quality.

A.5 Counterfactual steering

Evaluation details Evaluating with counterfactual steering inputs comes with additional challenges compared to the standard real-only steering evaluation done in [12, 18, 35], where the inverse dynamics model is applied to real and generated videos alike, compensating for scale ambiguities. Here, we compare the estimated generated trajectories with the actual input trajectories, but this

exposes us to scale mismatches due to the fact that the output of VGGT is not in metric space. As a solution, for each pair of real and generated trajectories $(\bar{\mathbf{x}}, \mathbf{x})$ we compute an optimal scale parameter for the latter:

$$s^* = \frac{\sum_{t=1}^{T} \mathbf{x}_t^\top \bar{\mathbf{x}}_t}{\sum_{t=1}^{T} \mathbf{x}_t^\top \mathbf{x}_t} \tag{6}$$

and rescale it accordingly before computing ADE. This still penalizes shape, direction, and curvature, but it largely removes global trajectory-scale error.

Figure 8 shows an example of steering the model on counterfactual trajectories. It is important for a world model to be able to follow commands that deviate from the original context-bound trajectories, even if they lead to out-of-distribution or illegal maneuvers.

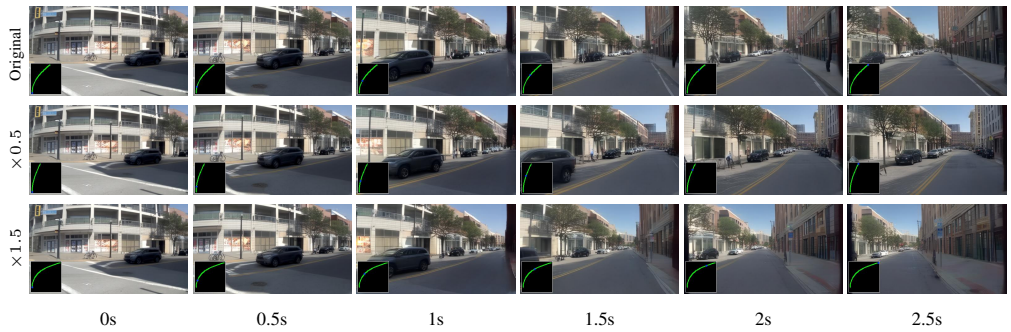


Figure 8: Example of counterfactual steering. Top row: the original steering trajectory, the model performs the turn as intended. Second row: the yaw rate was scaled by 0.5 (understeering), the model follows the command and executes a more gentle turn. Third row: the yaw rate was scaled by 0.5 (oversteering), the model follows the command and executes a sharper turn.

A.6 Model details

Tokenizer: We train two hybrid tokenizers with separate encoder and decoder branches for detail and abstract latents. The hybrid architecture follows the design proposed by Mousakhan et al. [35]. Both encoders are initialized from a MAE-pretrained [19] ViT with patch size 16×16, producing 768-dimensional embeddings that are quantized into 16-dimensional codes.

Abtract long-horizon predictor This predictor generates semantic latents at 2 Hz autoregressively with *causal* temporal attention (`causal_time_attn=True`). It takes 5 context frames and predicts 1 future frame per rollout step.

Detailed short-horizon predictor This predictor operates at 10 Hz and is conditioned on the abstract predictor's predicted latent. Temporal attention is *non-causal*. Conditioning is injected spatially per-frame via small non-zero-initialized linear projections and a learned frame-position MLP, ensuring the pretrained STDiT backbone is minimally perturbed at initialization. Each block contains an additional gated spatial cross-attention layer (small non-zero-initialized gate) for the abstract predictor's signal. The detail predictor is trained from scratch with abstract conditioning.

Inference Protocol. At inference, generation proceeds in a cascaded fashion. The abstract predictor takes 5 context frames and generates 1 frame at 2 Hz using 15 NFE. Its predicted abstract latent serves as the subgoal that conditions the detail predictor. The detail predictor then takes 5 context frames and generates the next 5 frames at 10 Hz using 30 NFE, conditioned on the predicted abstract latent. During training, the detail predictor is conditioned on ground-truth abstract latents, whereas at inference it operates on the predicted latents from the abstract predictor. Rollouts are performed auto-regressively. The abstract predictor advances with a sliding window of 1 frame, operating over a 6-frame window (3 seconds) at 2 Hz. The detail predictor advances with a sliding window of 5 frames, operating over a 10-frame window (1 second) at 10 Hz.

A.7 Linear probing details

Semantic Segmentation To evaluate the quality of representations learned by the Stage-2 predictor model, we train a linear probe for semantic segmentation on the Cityscapes fine-annotation benchmark (19 classes) using the model's frozen intermediate features. Input images are normalized to $[-1, 1]$ and encoded through the frozen backbone. During probing, context is dropped (set to null) and the input image is tiled across all predicted-frame slots, with features extracted from the last frame slot. Intermediate patch-token features are extracted from the 4th transformer block from the end for the main model and from the 2nd block from the end for ablation models, then reshaped from the flat patch sequence into a 2D spatial feature map. The linear probe head is a single 1×1 LazyConv2d that maps this feature map to 19-class logits, followed by bilinear upsampling to match the ground-truth mask resolution. Only the probe parameters are trained, the backbone is kept entirely frozen throughout. Training uses AdamW (lr=1e-3, wd=1e-4) with cross-entropy loss for 100 epochs. Performance is reported as mean Intersection-over-Union (mIoU) on the Cityscapes validation set.

Depth Estimation To probe the geometry-awareness of the Stage-2 predictor model, we train a linear depth probe on the KITTI Eigen split following the DINOv2 linear depth evaluation protocol. Input images are encoded through the frozen backbone; during probing, context is dropped (set to null) and the input image is passed as the sole target frame, with features extracted from the last frame slot. Intermediate patch-token features are extracted from the 4th transformer block from the end, then reshaped into a 2D spatial feature map. The linear probe head follows the DINOv2 Lin-1 design: a single 1×1 LazyConv2d producing logits over 256 uniformly-spaced depth bins covering [0, 80] m, preceded by a 4× bilinear upsample to recover spatial resolution, followed by a softmax and weighted sum over bin centers to yield a continuous depth prediction. At training time, ground-truth depth values are discretized into the same uniform bins and cross-entropy loss is minimized with AdamW (lr=1e-3, wd=1e-4) for 10 epochs; missing and out-of-range depth pixels are excluded from the loss. Only the probe parameters are trained; the backbone is kept entirely frozen. Performance is reported as RMSE on the KITTI Eigen validation split.