FOURIER-WORLD: TEMPORAL FOURIER STATES FOR VISUAL WORLD MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

World models aim to learn the dynamics of the world by predicting the future from visual observations. Recent approaches have shifted from directly generating RGB frames to predicting future features of vision foundation models (VFMs), which can support diverse downstream tasks. However, small feature errors accumulate during autoregressive prediction, degrading long-term rollout performance, while how existing predictors internally represent and use temporal changes remains unclear. We propose **Fourier-World**, which mitigates error accumulation by explicitly structuring temporal context as a frequency-domain state. Fourier-World applies a Fourier transform to the four most recent context features. A query initialized with learnable mask embedding reads the resulting Fourier coefficients to predict feature-space deltas, which are then accumulated to generate future dense VFM features. Pretrained on the large-scale *Kinetics-700* dataset, Fourier-World demonstrates strong transfer performance in segmentation, depth estimation, action anticipation, and causal future-point forecasting across five benchmarks: *Cityscapes*, *VSPW*, *KITTI*, *EK100*, and *TAP-Vid-RGB-Stacking*. It also maintains stable predictions over long autoregressive rollouts, achieving high computational efficiency and competitive performance.

1 INTRODUCTION

World models (Ha & Schmidhuber, 2018; Gkountouras et al., 2024; Tehenan et al., 2025; Team et al., 2026) aim to understand the current state of the world from visual observations and predict future states, with applications in robotics, autonomous driving, and embodied AI. Early visual prediction models (Lotter et al., 2016; Finn & Levine, 2017; Tong et al., 2022) aimed to predict RGB observations directly. However, pixel-level prediction has limitations as an approach to world modeling (Hafner et al., 2019b;a; Bardes et al., 2024; Assran et al., 2025; Pallotta et al., 2025): it requires reconstructing even fine-grained details that are unnecessary for scene understanding, and the predicted frames must subsequently be processed by a vision foundation model (VFM) to interpret the scene (Agarwal et al., 2025). This additional latency is a major drawback in world modeling tasks that require real-time operation, as the predicted RGB frames cannot be directly fed into downstream task heads.

Accordingly, recent studies have shifted toward predicting representations directly in the latent space of a VFM, defining the goal of world modeling as predicting the diverse features provided by the underlying VFM (Vondrick et al., 2016; Luc et al., 2017; Oquab et al., 2023; Zheng et al., 2025; Siméoni et al., 2025). However, several major challenges remain: (1) prediction speed, as fast prediction is necessary for applicability across diverse domains (Yin et al., 2026); (2) long-term prediction, which suffers from error accumulation during autoregressive rollouts (Huang et al., 2025); and (3) limited interpretability of the prediction process, which makes these models difficult to apply (Boduljak et al., 2025; Kerssies et al., 2026; Peper et al., 2025). Joint encoder-predictor methods also exist (Assran et al., 2023; Mur-Labadia et al., 2026), but questions remain regarding the accuracy of their predictors.

We begin with a simple question: can established transforms and analysis techniques for understanding motion benefit world models? We choose the **Fast Fourier Transform (FFT)** (Cochran et al., 1967) and propose **Fourier-World**, which directly uses the information obtained through the Fourier transform. Given the input context, Fourier-World incorporates the Fourier state into the

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

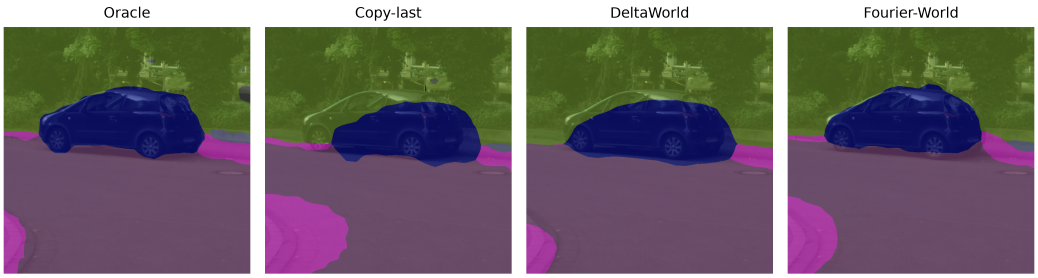


Figure 1: **Qualitative comparison of semantic segmentation forecasting.** Fourier-World predicts a future vehicle region that more closely matches the Oracle in position and shape than Copy-last and DeltaWorld in this example.

queries, keys, and values of a ViT-based predictor. It performs temporal Fourier cross-attention to update motion representations using the static, dynamic, and alternating modes derived from the Fourier state. This is followed by spatial self-attention and a feed-forward network (FFN), with the aim of enabling the predictor to directly understand object motion using the motion information extracted from the Fourier representation (Figure 2).

We evaluate Fourier-World from multiple perspectives. We use *DINOv3* (Siméoni et al., 2025) as the backbone and pretrain Fourier-World on the publicly available *Kinetics-700* (Smaira et al., 2020) dataset. To assess whether Fourier-World provides strong transfer capabilities as a world model, we evaluate it across four domains and five benchmarks: semantic segmentation on *Cityscapes* (Cordts et al., 2016) and *VSPW* (Miao et al., 2021), depth estimation on *KITTI* (Geiger et al., 2013), action anticipation on *EPIC-KITCHENS-100(EK100)* (Damen et al., 2022), and future-point forecasting on *TAP-Vid-RGB-Stacking* (Doersch et al., 2022). Beyond its strong performance, we demonstrate its advantages in autoregressive rollout generation and its high computational efficiency. Through comparisons with baselines and further analysis, we demonstrate that context decomposition using FFT can offer substantial benefits for visual world modeling.

The contributions of Fourier-World can be summarized as follows:

- We demonstrate that Fourier-World can serve effectively as a world model by achieving strong transfer capabilities after pretraining on a large-scale dataset.
- We demonstrate performance advantages over baselines and high computational efficiency through evaluations across diverse benchmarks and domains.
- We provide interpretability into how the Fourier state supports prediction through the predictor and its attention mechanism.

2 Related Work

World Models and Future Prediction World models (Ha & Schmidhuber, 2018; Hafner et al., 2019a; Rigter et al., 2024; Team et al., 2026; Agarwal et al., 2025) aim to learn predictive models of environment dynamics, enabling future states or outcomes to be anticipated from observations. This ability is central to many real-world visual intelligence problems, including autonomous driving, robotics, and embodied AI (Zhang et al., 2025; Wu et al., 2023; Maes et al., 2026). In visual future prediction, while intuitive, pixel-level prediction requires models to account for low-level visual variations that are often stochastic or only weakly related to downstream decisions, which has motivated residual or delta prediction approaches that model future changes relative to the observed context (Franceschi et al., 2020; Kerssies et al., 2026).

More recently, VFMs (Oquab et al., 2023; Siméoni et al., 2025) have made representation-space prediction an attractive alternative: predictors can operate on compact latent representations produced by strong pretrained encoders. Some methods (Luc et al., 2018; Karypidis et al., 2025a; Boduljak et al., 2025; Karypidis et al., 2025b; Kerssies et al., 2026; Porcher et al., 2026) reduce the prediction burden by extracting compressed or task-specific features from VFM embeddings, but such feature transformations can make the predictor harder to interpret, diagnose, and transfer

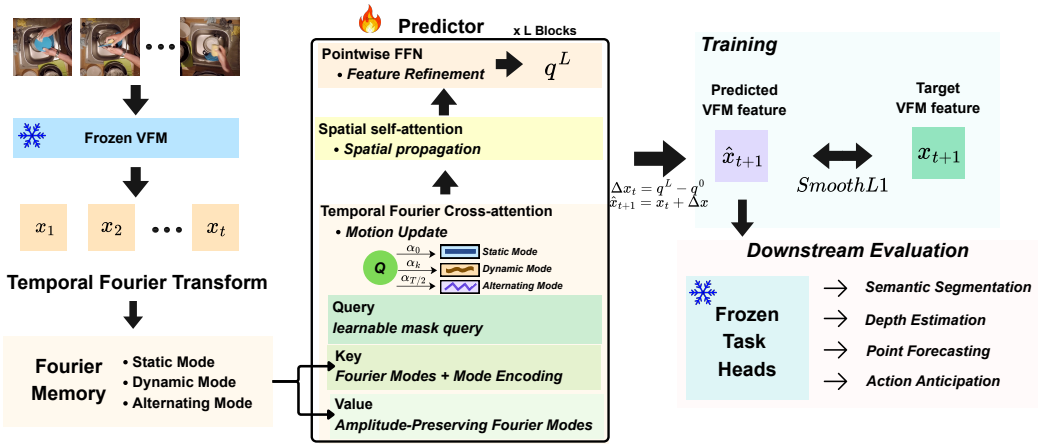


Figure 2: **Overview of Fourier-World.** Four observed VFM feature maps are transformed into temporal Fourier memory containing DC, complex AC, and Nyquist modes. Queries initialized from a shared learnable mask retrieve local dynamics through temporal cross-attention and propagate them through spatial attention and FFNs over L residual blocks. The resulting update is added to the latest feature. Training uses future-feature Smooth L1 loss; downstream evaluation uses frozen task heads or feature matching.

across different VFMs. Joint embedding predictive architectures, such as V-JEPA-style encoder-predictor systems (Assran et al., 2023; Mur-Labadia et al., 2026), provide a powerful framework for self-supervised predictive representation learning. However, because the encoder and predictor are optimized together (Assran et al., 2023; Chen et al., 2025), it can be difficult to isolate whether the predictor itself captures future dynamics or adapts to a jointly shaped representation space.

Deep Learning with Fourier Transforms Fourier analysis (Cooley, 2003; Rockmore, 2000; Zafar et al., 2022) has long been used to decompose complex signals into frequency components, and it has also become a useful tool in deep learning (Tancik et al., 2020; Canatar et al., 2023; Gao et al., 2025; Karkada et al., 2026). In visual recognition, frequency-domain methods such as Global Filter Networks (Rao et al., 2021) use discrete Fourier transforms and learnable frequency-domain filters to efficiently model long-range spatial interactions Xiao et al. (2026). Fourier-based designs have also been explored in video prediction and video understanding, often by modeling motion or temporal structure in the frequency domain (Mosbach & Behnke, 2021; Long et al., 2022). Recent studies have also explored wavelet (Graps, 1995) and Fourier representations for motion prediction (Zhu et al., 2026) and video generation (Agarwal et al., 2025). Fourier-World explicitly decomposes temporal trajectories of frozen VFM features into Fourier memory for attention-based residual prediction.

3 METHOD

We directly predict future states in the dense feature space of a frozen vision foundation model (VFM). Let I_t denote the observed RGB frame at time t and E the frozen VFM. We define the frame features and temporal context as

$$\mathbf{x}_t = E(I_t) \in \mathbb{R}^{S\times D}, \qquad \mathbf{X}_t = [\mathbf{x}_{t-T+1}, \ldots, \mathbf{x}_t] \in \mathbb{R}^{T\times S\times D}, \tag{1}$$

where S is the number of spatial tokens and D is the feature dimension. All VFM parameters remain frozen throughout training and inference.

Given $\mathbf{X}_t$, we predict $\hat{\mathbf{x}}_{t+1} = \mathbf{x}_t + \Delta\mathbf{x}_t$. The latest features preserve the scene's semantic appearance and spatial geometry, while the predictor estimates changes conditioned on Fourier memory. This formulation asks where and by how much the latest state should change. Fourier-World separates state from dynamics: the latest VFM features preserve the current scene state, Fourier memory represents how local features have evolved, and the query predictor selects and spatially propagates these dynamics to estimate the next feature update.

3.1 TEMPORAL FOURIER MEMORY

Each spatial token is a time-varying D-dimensional signal. At each location s, we apply a real-input discrete Fourier transform along the temporal axis:

$$\mathbf{F}_{t,s,k} = \frac{1}{\sqrt{T}} \sum_{j=0}^{T-1} \mathbf{x}_{t-T+1+j,s} \exp\left(-\mathrm{i}\frac{2\pi k j}{T}\right), \tag{2}$$

for $k = 0, \ldots, \lfloor T/2 \rfloor$, without mixing spatial tokens or feature channels. Before normalization and projection, this is an invertible change of temporal coordinates rather than a learned, lossy tokenizer. DC represents the temporal mean, while non-DC components capture deviations from it. Their magnitudes encode the strength of variation at each temporal frequency, and their phases encode alignment relative to the window origin. For even T, the real-valued Nyquist component captures alternation between successive frames. These coordinates explicitly expose temporal structure that the predictor would otherwise need to infer from the sequence.

We jointly normalize coefficients by a common window-wise RMS scale to preserve relative mode energies. Mode-specific projections map real-valued bins to $\mathbb{R}^H$ and complex bins to $\mathbb{C}^{H/2}$, where H is the predictor's hidden dimension. Let ρ_k be the identity for real-valued bins and $\rho_k(\mathbf{z}) = [\mathrm{Re}(\mathbf{z}); \mathrm{Im}(\mathbf{z})]$ for complex bins. The Fourier memory is

$$\mathbf{m}_{t,s,k} = \rho_k\big(\mathcal{L}_k(\bar{\mathbf{F}}_{t,s,k})\big), \qquad \mathcal{M}_t \in \mathbb{R}^{S\times K\times H}, \tag{3}$$

where $K = \lfloor T/2 \rfloor + 1$. Each frequency bin forms a single real-valued memory token, rather than separate real and imaginary tokens.

For complex bins, the real and imaginary components form a quadrature pair. A circular delay of Δ samples induces $\mathbf{F}'_k = e^{-\mathrm{i}2\pi k\Delta/T}\mathbf{F}_k$, omitting t and s for clarity. Discarding the imaginary component therefore loses information that distinguishes temporal alignments. We retain both components and, before packing, apply the complex-linear projection

$$\mathcal{L}(\mathbf{a} + \mathrm{i}\mathbf{b}) = (\mathbf{A}\mathbf{a} - \mathbf{B}\mathbf{b}) + \mathrm{i}(\mathbf{B}\mathbf{a} + \mathbf{A}\mathbf{b}). \tag{4}$$

It satisfies $\mathcal{L}(e^{\mathrm{i}\theta}\mathbf{F}) = e^{\mathrm{i}\theta}\mathcal{L}(\mathbf{F})$, preserving the phase-rotation structure induced by circular temporal shifts. An unconstrained linear map of concatenated real and imaginary components does not generally guarantee this property. The real and imaginary components jointly specify temporal alignment: trajectories with identical spectral magnitudes can still differ in phase. Retaining the quadrature pair therefore allows the predictor to distinguish temporal patterns that a magnitude-only representation would conflate.

3.2 FOURIER STATE-QUERY PREDICTOR

The predictor consists of L residual query blocks with hidden dimension H. We use $H = D$ in our implementation so that its final query displacement can be added directly to the VFM features. All spatial queries are initialized as $\mathbf{q}_s^0 = \mathbf{z}$ using a shared learnable mask embedding $\mathbf{z} \in \mathbb{R}^H$.

The latest feature is already preserved through the output residual connection. Initializing queries from it would introduce a second path for copying current semantic content into the update branch. Content-independent mask initialization removes this direct shortcut, making the query a neutral workspace that reads and accumulates changes from Fourier memory. Although initially identical, queries attend to distinct local memories and can become spatially varying from the first block. Subsequent blocks reuse the preceding output $\mathbf{q}^l$, applying temporal Fourier cross-attention, spatial propagation, and a pointwise FFN.

3.2.1 TEMPORAL FOURIER CROSS-ATTENTION

Each query attends to Fourier modes at the same spatial location:

$$\begin{aligned}
\mathbf{Q}_s^l &= \mathbf{W}_Q^{T,l}\,\mathrm{Norm}(\mathbf{q}_s^l),\\
\mathbf{K}_{s,k}^l &= \mathbf{W}_K^{T,l}\,\mathrm{CRMSNorm}(\mathbf{m}_{t,s,k}) + \mathbf{e}_k^l,\\
\mathbf{V}_{s,k}^l &= \eta_k \mathbf{W}_V^{T,l}\mathbf{m}_{t,s,k}.
\end{aligned} \tag{5}$$

Here, $\mathbf{e}_k^l$ encodes frequency-mode identity. $\mathrm{CRMSNorm}$ applies RMS normalization with a shared learnable gain for paired real and imaginary coordinates, preserving their phase coupling. Keys determine mode relevance, whereas values carry the strength of the selected temporal variations. Normalizing keys limits attention dominance driven solely by coefficient scale, while retaining relative spectral amplitudes in values preserves information used to scale the update.

DC provides temporal-mean context for mode selection but does not itself represent change. We therefore retain it in keys and exclude it from values by setting $\eta_0 = 0$ and $\eta_k = 1$ for $k > 0$.

Attention is entirely real-valued, with corresponding real and imaginary coordinates assigned to the same head. Tildes denote continuous temporal RoPE applied to queries and keys using prediction and memory timestamps, respectively. Before learned projection, we define $g_{t,s} = \mathbb{I}[\exists\, k > 0 : \bar{\mathbf{F}}_{t,s,k} \neq \mathbf{0}]$. This gate detects non-DC spectral activity, is shared across blocks, and is broadcast over hidden channels. The attention weights and gated local update are

$$\alpha^l_{s,h,k} = \mathrm{softmax}_k\left(\frac{\langle \tilde{\mathbf{Q}}^l_{s,h}, \tilde{\mathbf{K}}^l_{s,h,k}\rangle_{\mathbb{R}}}{\sqrt{d_h}}\right), \quad \mathbf{u}^l_s = g_{t,s}\mathbf{W}_O^{T,l}\,\mathrm{Concat}_h\Big(\textstyle\sum_k \alpha^l_{s,h,k}\mathbf{V}^l_{s,h,k}\Big). \tag{6}$$

Here, h indexes attention heads and d_h is the head dimension. Softmax operates only over Fourier modes, without mixing spatial locations. The gate suppresses local motion updates at locations with zero spectral activity before spatial propagation.

3.2.2 Spatial Propagation and Query Update

Spatial attention propagates the gated local updates across tokens, using the current query state to construct queries and keys and the Fourier updates to construct values:

$$\mathbf{p}^l = \mathrm{Attn}_S\left(\mathbf{W}_Q^{S,l}\mathbf{q}^l, \mathbf{W}_K^{S,l}\mathbf{q}^l, \mathbf{W}_V^{S,l}\mathbf{u}^l\right). \tag{7}$$

We apply 2D positional encoding to spatial queries and keys. Queries and keys determine where information is routed, while values determine what is transmitted. Thus, the evolving query state controls spatial routing, but only retrieved motion evidence is propagated, preventing direct mixing of original VFM content or DC components into spatial updates. Using $\mathbf{q}^l$ as values would instead recirculate the mask embedding and accumulated query state.

For $l = 0, \ldots, L-1$, we combine the temporal and spatial updates and refine them with a pointwise FFN. Since spatial attention can bring updates from other locations, we gate the entire residual again to prevent these updates from modifying locations with zero spectral activity. The two gates serve complementary roles: the first suppresses locally generated motion before spatial propagation, whereas the second also suppresses incoming spatial updates and FFN outputs. Their combination ensures that learned transformations cannot introduce a residual update at a location whose observed features contain no temporal variation:

$$\mathbf{q}^{l+1}_s = \mathbf{q}^l_s + g_{t,s}\left[\mathbf{u}^l_s + \mathbf{p}^l_s + \mathrm{FFN}^l\left(\mathbf{u}^l_s + \mathbf{p}^l_s\right)\right]. \tag{8}$$

The displacement from the initial mask query to the final query defines the predicted change and next state:

$$\Delta\mathbf{x}_t = \mathbf{q}^L - \mathbf{q}^0, \qquad \hat{\mathbf{x}}_{t+1} = \mathbf{x}_t + \Delta\mathbf{x}_t. \tag{9}$$

This residual endpoint preserves the latest VFM features and adds only the inferred changes, rather than replacing the semantic state with the mask query. For static contexts, all non-DC coefficients vanish, giving $g_{t,s} = 0$, $\mathbf{q}^L = \mathbf{q}^0$, and $\hat{\mathbf{x}}_{t+1} = \mathbf{x}_t$ exactly. Reinserting this prediction preserves the static context, so the property also holds under autoregressive rollout.

We train with Smooth L1 loss between predicted features and actual future features extracted by the frozen VFM. For multi-step forecasting, we insert predictions back into the context, recompute the Fourier memory and activity gate, and repeatedly apply the same predictor.

4 Experiments

We evaluate Fourier-World's predictive capability, use of Fourier state, autoregressive rollout stability, and computational efficiency.

Table 1: Datasets used for transfer evaluation. TAP-Vid RGB-Stacking includes 1,210 point queries across 50 videos.

Dataset	Domain	Task	Evaluation set	Metric
VSPW	In-the-wild videos	Semantic segmentation	758 sources	mIoU ↑
Cityscapes	Urban driving	Semantic segmentation	500 sources	mIoU ↑
KITTI	Driving	Depth estimation	603 sources	RMSE ↓
TAP-Vid RGB-Stacking	Robotic manipulation	Point forecasting	50 videos	δ_{avg} ↑
EK100	Egocentric actions	Action anticipation	8,293 clips	MCR@5 ↑

Data and implementation. We pretrain Fourier-World on *Kinetics-700* (Smaira et al., 2020) and evaluate it without further predictor training across four tasks and five benchmarks: *Cityscapes* (Cordts et al., 2016), *VSPW* (Miao et al., 2021), *KITTI* (Geiger et al., 2013), *EK100* (Damen et al., 2022), and *TAP-Vid-RGB-Stacking* (Doersch et al., 2022). Table 1 summarizes the datasets, tasks, and metrics. We use frozen DINOv3 (Siméoni et al., 2025) ViT-B/16 features with 512×512 inputs and four context frames.

Baselines. We compare against **DeltaWorld** (Kerssies et al., 2026), **COSMOS** (Agarwal et al., 2025), **DINO-Foresight** (Karypidis et al., 2025b), **DINO-World** (Baldassarre et al., 2025), and **V-JEPA** (Assran et al., 2025). We evaluate the public COSMOS 4B/12B models and encode their RGB predictions with DINOv3. DeltaWorld uses its compatible public DINOv3 checkpoint with `BoM-mean 20`. For DINO-Foresight, DINO-World, and V-JEPA, we use `-style` reimplementations with frozen DINOv3 and predictor parameter counts matched to Fourier-World. These reimplementations and corresponding Fourier-World variants are trained on the same benchmark-specific training subsets. For V-JEPA, only the reimplemented predictor is trained.

Evaluation. We evaluate one- and three-step forecasts (H_1 and H_3), generating H_3 autoregressively. Cityscapes, VSPW, and KITTI use DeltaWorld's pretrained task heads, each applied to a single predicted feature map. For EK100, a linear classifier combines four observed and three predicted frame representations under a custom action-stop-aligned protocol, evaluating nouns, verbs, and actions using mean class recall at 5 (MCR@5). For TAP-Vid-RGB-Stacking, we match query-point features to predicted feature maps using cosine similarity and report δ_{avg}. Further dataset, implementation, baseline, and evaluation details appear in Appendices A.1, A.2, A.3, and A.4, respectively.

5 Results

5.1 Main World-Modeling Results

Table 2 compares Fourier-World, pretrained on Kinetics-700, with the baselines across five benchmarks. Fourier-World achieves the best result among the evaluated predictors on 10 of the 11 reported metrics, trailing DeltaWorld only on Cityscapes at H_1 by 0.93 mIoU points. Both checkpoints were trained on the same Kinetics-700 dataset, making this comparison particularly informative.

On TAP-Vid-RGB-Stacking, Fourier-World improves δ_{avg} over the strongest baseline, DeltaWorld, by 0.32 and 1.76 points at H_1 and H_3, respectively. On EK100, it exceeds COSMOS 12B, the strongest baseline for all three targets, by 2.06, 3.92, and 4.93 MCR@5 points for action, noun, and verb anticipation. These results demonstrate transfer across diverse domains and suggest potential applications to robotic control and action planning.

Fourier-World also exhibits less degradation during rollout on VSPW: mIoU drops by 5.98 points from H_1 to H_3, compared with 7.87 for DeltaWorld and 6.60 for Copy Last. This supports its potential for longer-horizon prediction.

5.2 Architectural Effectiveness

To isolate predictor design from backbone differences, we compare `-style` reimplementations using the same frozen DINOv3 backbone, approximately 80M predictor parameters, and identical

Table 2: Main results across five benchmarks. H_1 and H_3 denote one- and three-step forecasting, respectively. Bold indicates the best result in each column, excluding Oracle. Since the EK100 requires classifying actual actions, "copy-last" is irrelevant and was not measured.

	Cityscapes mIoU ↑		VSPW mIoU ↑		KITTI RMSE ↓		TAP-Vid RGB-Stacking δ_{avg} ↑		EK100 MCR@5 ↑		
Method	H_1	H_3	H_1	H_3	H_1	H_3	H_1	H_3	Action	Noun	Verb
Copy Last	53.5	39.6	50.9	44.3	3.76	4.86	50.23	47.57	–	–	–
COSMOS 4B	54.4	48.5	49.2	44.7	3.78	4.15	48.97	42.59	22.23	35.36	41.14
COSMOS 12B	56.3	51.2	50.3	45.7	3.73	4.16	49.21	43.23	23.48	36.89	42.62
DeltaWorld	**64.12**	49.95	54.04	46.17	3.37	5.19	56.68	51	20.24	30.31	31.43
Fourier-World	63.19	**51.75**	**54.21**	**48.23**	**3.145**	**4.004**	**57**	**52.76**	**25.54**	**40.81**	**47.55**
Oracle	70.51	70.5	58.05	58.03	2.79	2.79	57.07	53.6	30.78	42.49	59.6

Table 3: Semantic segmentation results on VSPW and Cityscapes. H_1 and H_3 report mIoU; Drop denotes the decrease from H_1 to H_3. Bold indicates the best result in each column, excluding Copy Last and Oracle.

	VSPW			Cityscapes		
Method	H_1 ↑	H_3 ↑	Drop ↓	H_1 ↑	H_3 ↑	Drop ↓
Copy Last	50.944	44.298	6.646	53.557	39.6	13.957
DINO-Foresight-style	49.11	41.201	7.909	58.799	47.373	**11.426**
DINO-World-style	50.671	43.731	6.940	56.409	43.142	13.267
V-JEPA-style	47.73	37.6	10.130	56.78	43.5	13.280
DeltaWorld-style	50.563	42.91	7.653	58.919	46.526	12.393
Fourier-World-style	**52.6**	**45.99**	**6.610**	**62.93**	**50.224**	12.706
Oracle	58.05	58.03	–	70.51	70.51	–

training data from Cityscapes and VSPW. The parameter budget matches the original Fourier-World and DeltaWorld predictors.

Fourier-World-style achieves the highest mIoU at both horizons on both datasets (Table 3). Although it shares DINO-World's token-wise query-based attention formulation, its stronger performance under this common setup suggests the utility of explicitly organizing temporal context as Fourier memory.

With only four context frames, Fourier-World-style also achieves the smallest H_1-to-H_3 mIoU drop on VSPW, at 6.610 points, and a competitive relative decline of 20.19% on Cityscapes. These results suggest that Fourier decomposition supports longer-horizon prediction even with a short observation window. However, declines depend partly on starting scores, and Fourier-World does not achieve the smallest drop in every setting, motivating the following geometry analysis.

5.3 Autoregressive Rollout Stability

Autoregressive rollouts are susceptible to error accumulation. Cosine similarity alone cannot fully assess stability: predictions may remain directionally similar to oracle features while their magnitudes or broader embedding structure drift. We therefore evaluate cosine similarity, norm ratio, rank ratio, Spectrum JS, and Procrustes error at H_3 on VSPW and Cityscapes. These measurements complement downstream performance by assessing how well predicted features preserve the VFM embedding geometry after multi-step autoregressive rollout.

Across the two datasets, Fourier-World-style leads on nine of ten geometry measurements within the `-style` group, while pretrained Fourier-World outperforms DeltaWorld on eight of ten. Both versions achieve norm and rank ratios closer to one, lower Spectrum JS, and lower Procrustes error

Table 4: Feature-space geometry comparison at H_3 on VSPW and Cityscapes (City). Norm and rank ratios are ideally one. Bold indicates the best result within each comparison group, determined before rounding.

	Cosine ↑		Norm ratio (→ 1)		Rank ratio (→ 1)		Spectrum JS (×100) ↓		Procrustes ↓	
Method	VSPW	City	VSPW	City	VSPW	City	VSPW	City	VSPW	City
DINO-Foresight-style	0.86	0.91	0.97	0.96	0.65	0.69	1.66	1.43	0.32	0.29
DINO-World-style	0.90	0.91	0.97	0.96	0.66	0.64	1.53	1.70	0.31	0.30
V-JEPA-style	0.85	0.89	0.93	0.94	0.59	0.63	2.27	1.89	0.35	0.31
DeltaWorld-style	0.90	**0.92**	0.95	0.96	0.86	0.80	0.52	0.64	0.23	0.23
Fourier-World-style	**0.91**	0.92	**0.98**	**0.99**	**0.90**	**0.88**	**0.34**	**0.25**	**0.22**	**0.20**
DeltaWorld	**0.92**	**0.92**	0.90	0.89	0.70	0.68	1.61	1.42	0.29	0.28
Fourier-World	0.92	0.92	**0.97**	**0.96**	**0.85**	**0.81**	**0.50**	**0.54**	**0.22**	**0.22**

Table 5: Computational efficiency per H_1 prediction. GFLOPs count linear and attention matrix operations. Lower is better; bold indicates the minimum.

Method	Compute ↓ GFLOPs	Time ↓ ms	Memory ↓ GiB
DINO-Foresight-style	1015.205	27.486	0.465
DINO-World-style	416.746	**25.727**	**0.374**
V-JEPA-style	1836.287	35.326	0.465
DeltaWorld-style	6279.724	228.837	1.481
Fourier-World	**259.947**	33.258	0.421

Table 6: Fourier ablations at H_3. The full model reports absolute scores; other rows report changes (ablated minus full).

Variant	VSPW mIoU ↑	City mIoU ↑	KITTI RMSE ↓
Fourier-World	48.23	51.75	4.004
Imag. quadrature off	−3.134	−8.366	+0.612
AC off	−2.990	−7.778	+0.611
Nyquist off	−2.410	−7.565	+0.572
AC phase shuffle	−1.411	−4.216	+0.279
Complete update off	−3.930	−12.150	+0.856

than their DeltaWorld counterparts on both datasets. On Cityscapes, Fourier-World-style achieves a rank ratio of 0.8837 versus 0.7996 for DeltaWorld-style; the pretrained comparison similarly favors Fourier-World, at 0.8113 versus 0.6808.

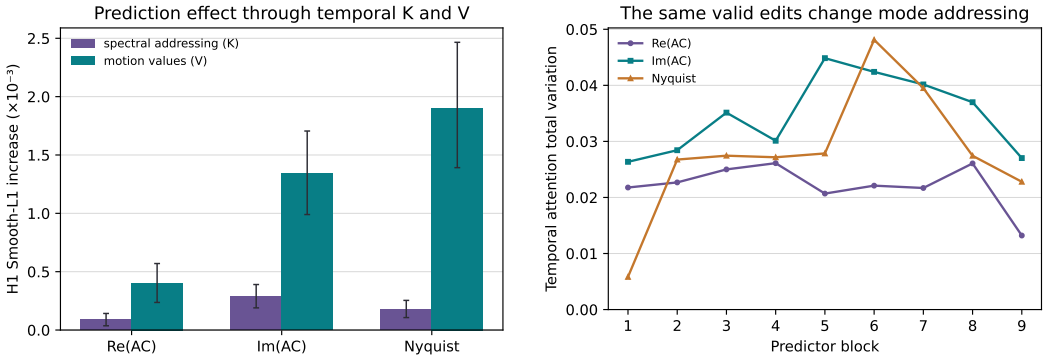
DeltaWorld-style retains higher cosine similarity on Cityscapes, while pretrained DeltaWorld has higher cosine similarity on both datasets. Fourier-World's advantages on the remaining measures highlight the importance of evaluating embedding structure alongside directional similarity. These results support its ability to preserve VFM feature geometry after autoregressive rollout across both visual domains.

5.4 Efficiency and Fourier-State Analysis

Efficiency analysis. We report per-sample matrix-operation counts, inference latency averaged over 64 samples, and peak inference memory. Latency for each sample is the median of 30 timed runs following five warmup calls. Predictor parameter counts are matched across methods; measurement details are provided in Appendix A.4.

Fourier-World requires the least computation, at 259.947 GFLOPs per H_1 prediction (Table 5). DeltaWorld-style and DINO-World-style require 24.16 and 1.60 times as much, respectively. Fourier-World takes 33.258 ms and uses 0.421 GiB of peak memory; DeltaWorld-style requires 6.88 times the inference time and 3.52 times the memory.

Despite its single-token predictor, DeltaWorld requires an encoder and decoder for feature compression and reconstruction. Fourier-World predicts directly in VFM feature space, with additional processing limited to CPU-side FFT and Fourier attention. This yields the lowest FLOPs in the comparison with competitive inference time and memory usage.



(a) Prediction sensitivity through K and V

(b) Changes in mode attention across blocks

Figure 3: **Fourier state utilization in temporal attention.** (a) Two-factor Shapley attribution of the H_1 Smooth L1 loss increase to the key (K) and value (V) pathways under Fourier-component interventions. Error bars indicate 95% confidence intervals from 2,000 bootstrap resamples clustered by source video. (b) Total variation distance between the original and edited temporal attention distributions, averaged across heads, spatial tokens, and clips for each predictor block.

Fourier-state analysis. Table 6 examines the contribution of Fourier state through inference-time ablations that disable or perturb memory components or prediction updates.

Removing AC or Nyquist substantially degrades prediction, reducing Cityscapes mIoU at H_3 by 7.778 and 7.565 points, respectively. Removing the imaginary quadrature reduces H_3 mIoU by 3.134 points on VSPW and 8.366 on Cityscapes, while increasing KITTI RMSE by 0.612. These results support the contribution of the spectral components and the complex information representing temporal amplitude and phase.

The full ablation results in Appendix B.3 show consistently larger downstream degradation at H_3 than at H_1. These larger effects at the later horizon support the role of Fourier state in multi-step prediction and our hypothesis that explicit temporal dynamics provide useful information for world-model forecasting.

To examine how the predictor uses Fourier state, we decompose intervention-induced prediction loss using two-factor Shapley attribution over the temporal key and value pathways (Figure 3). Across all three components, the value pathway receives a larger attribution of the H_1 loss increase than the key pathway, with the largest value-pathway contributions observed for Nyquist and the imaginary AC component.

The attention analysis further shows that these interventions change mode selection throughout the predictor. Imaginary AC edits produce larger attention changes than real AC edits across all blocks, while Nyquist edits have their strongest effect in the middle blocks. Together, these observations support the use of Fourier state both to determine attention weights and to supply information for prediction updates. Intervention and attribution details are provided in Appendix A.4.

6 Conclusion

We introduced Fourier-World to investigate whether established motion analysis techniques can benefit world models. Fourier-World uses FFT to construct temporal memory for attention-based prediction of future dense VFM features. Pretrained on large-scale video data, it achieves strong performance across five benchmarks spanning diverse domains, with substantially greater computational efficiency than DeltaWorld. Geometry analysis shows improved preservation of VFM embeddings during autoregressive rollouts, supporting its potential for stable long-term prediction. Component-wise ablations further demonstrate the contribution of temporal amplitude and phase to forecasting. Analysis of temporal attention supports the role of Fourier state in both mode selection and prediction updates. These findings suggest that preserving the current semantic state while explicitly modeling temporal variation in the frequency domain can support efficient and transferable world models.

AI USE STATEMENT

We used generative AI tools solely to assist with manuscript writing, including translation and language editing. All AI-assisted text was carefully reviewed by the authors. We take full responsibility for the final content of this paper, including its claims and conclusions.

ETHICS STATEMENT

The authors reviewed the ethical considerations associated with this work and identified no specific ethical concerns.

REPRODUCIBILITY STATEMENT

We will provide the source code as supplementary material to facilitate reproducibility. The experimental setup is described in the main text, with additional details on datasets, implementation, baselines, and evaluation protocols provided in Appendices A.1, A.2, A.3, and A.4, respectively.

REFERENCES

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.

Mahmoud Assran, Quentin Duval, Ishan Misra, Piotr Bojanowski, Pascal Vincent, Michael Rabbat, Yann LeCun, and Nicolas Ballas. Self-supervised learning from images with a joint-embedding predictive architecture. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 15619–15629, 2023.

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.

Federico Baldassarre, Marc Szafraniec, Basile Terver, Vasil Khalidov, Francisco Massa, Yann LeCun, Patrick Labatut, Maximilian Seitzer, and Piotr Bojanowski. Back to the features: Dino as a foundation for video world models. *arXiv preprint arXiv:2507.19468*, 2025.

Adrien Bardes, Quentin Garrido, Jean Ponce, Xinlei Chen, Michael Rabbat, Yann LeCun, Mahmoud Assran, and Nicolas Ballas. Revisiting feature prediction for learning visual representations from video. *arXiv preprint arXiv:2404.08471*, 2024.

Gabrijel Boduljak, Yushi Lan, Christian Rupprecht, and Andrea Vedaldi. Vfmf: World modeling by forecasting vision foundation model features. *arXiv preprint arXiv:2512.11225*, 2025.

Abdulkadir Canatar, Jenelle Feather, Albert Wakhloo, and SueYeon Chung. A spectral theory of neural prediction and alignment. *Advances in Neural Information Processing Systems*, 36:47052–47080, 2023.

Delong Chen, Mustafa Shukor, Theo Moutakanni, Willy Chung, Jade Yu, Tejaswi Kasarla, Yejin Bang, Allen Bolourchi, Yann LeCun, and Pascale Fung. Vl-jepa: Joint embedding predictive architecture for vision-language. *arXiv preprint arXiv:2512.10942*, 2025.

W.T. Cochran, J.W. Cooley, D.L. Favin, H.D. Helms, R.A. Kaenel, W.W. Lang, G.C. Maling, D.E. Nelson, C.M. Rader, and P.D. Welch. What is the fast fourier transform? *Proceedings of the IEEE*, 55(10):1664–1674, 1967. doi: 10.1109/PROC.1967.5957.

James W. Cooley. *Fast Fourier transform (FFT)*, pp. 695–698. John Wiley and Sons Ltd., GBR, 2003. ISBN 0470864125.

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 *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

Dima Damen, Hazel Doughty, Giovanni Maria Farinella, Antonino Furnari, Jian Ma, Evangelos Kazakos, Davide Moltisanti, Jonathan Munro, Toby Perrett, Will Price, and Michael Wray. Rescaling egocentric vision. *International Journal of Computer Vision*, 130(1):33–55, 2022.

Carl Doersch, Ankush Gupta, Larisa Markeeva, Adria Recasens, Lucas Smaira, Yusuf Aytar, Joao Carreira, Andrew Zisserman, and Yi Yang. Tap-vid: A benchmark for tracking any point in a video. *Advances in Neural Information Processing Systems*, 35:13610–13626, 2022.

Chelsea Finn and Sergey Levine. Deep visual foresight for planning robot motion. In *2017 IEEE international conference on robotics and automation (ICRA)*, pp. 2786–2793. IEEE, 2017.

Jean-Yves Franceschi, Edouard Delasalles, Mickaël Chen, Sylvain Lamprier, and Patrick Gallinari. Stochastic latent residual video prediction. In *International Conference on Machine Learning*, pp. 3233–3246. PMLR, 2020.

Chenxiao Gao, Haotian Sun, Na Li, Dale Schuurmans, and Bo Dai. Spectral representation-based reinforcement learning. *arXiv preprint arXiv:2512.15036*, 2025.

A Geiger, P Lenz, C Stiller, and R Urtasun. Vision meets robotics: The kitti dataset. *The International Journal of Robotics Research*, 32(11):1231–1237, 2013. doi: 10.1177/0278364913491297. URL https://doi.org/10.1177/0278364913491297.

John Gkountouras, Matthias Lindemann, Phillip Lippe, Efstratios Gavves, and Ivan Titov. Language agents meet causality–bridging llms and causal world models. *arXiv preprint arXiv:2410.19923*, 2024.

Amara Graps. An introduction to wavelets. *IEEE Computational Science and Engineering*, 2:50–61, 02 1995. doi: 10.1109/99.388960.

David Ha and Jürgen Schmidhuber. World models. *eprint arXiv: 1803.10122*, 2018.

Danijar Hafner, Timothy Lillicrap, Jimmy Ba, and Mohammad Norouzi. Dream to control: Learning behaviors by latent imagination. *arXiv preprint arXiv:1912.01603*, 2019a.

Danijar Hafner, Timothy Lillicrap, Ian Fischer, Ruben Villegas, David Ha, Honglak Lee, and James Davidson. Learning latent dynamics for planning from pixels. In *International conference on machine learning*, pp. 2555–2565. PMLR, 2019b.

Xun Huang, Zhengqi Li, Guande He, Mingyuan Zhou, and Eli Shechtman. Self forcing: Bridging the train-test gap in autoregressive video diffusion. In D. Belgrave, C. Zhang, H. Lin, R. Pascanu, P. Koniusz, M. Ghassemi, and N. Chen (eds.), *Advances in Neural Information Processing Systems*, volume 38, Main Conference, pp. 167283–167308. Curran Associates, Inc., 2025. doi: 10.52202/085713-5576. URL https://proceedings.neurips.cc/paper_files/paper/2025/file/f4823f831af67a3ef15e41a85434422a-Paper-Conference.pdf.

Dhruva Karkada, Daniel J Korchinski, Andres Nava, Matthieu Wyart, and Yasaman Bahri. Symmetry in language statistics shapes the geometry of model representations. *arXiv preprint arXiv:2602.15029*, 2026.

Efstathios Karypidis, Ioannis Kakogeorgiou, Spyros Gidaris, and Nikos Komodakis. Advancing semantic future prediction through multimodal visual sequence transformers. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 3793–3803, 2025a.

Efstathios Karypidis, Ioannis Kakogeorgiou, Spyros Gidaris, and Nikos Komodakis. DINO-foresight: Looking into the future with DINO. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025b.

Tommie Kerssies, Gabriele Berton, Ju He, Qihang Yu, Wufei Ma, Daan de Geus, Gijs Dubbelman, and Liang-Chieh Chen. A frame is worth one token: Efficient generative world modeling with delta tokens. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2026.

Fuchen Long, Zhaofan Qiu, Yingwei Pan, Ting Yao, Chong-Wah Ngo, and Tao Mei. Dynamic temporal filtering in video models. In *European Conference on Computer Vision*, pp. 475–492. Springer, 2022.

William Lotter, Gabriel Kreiman, and David Cox. Deep predictive coding networks for video prediction and unsupervised learning. *arXiv preprint arXiv:1605.08104*, 2016.

Pauline Luc, Natalia Neverova, Camille Couprie, Jakob Verbeek, and Yann LeCun. Predicting deeper into the future of semantic segmentation. In *Proceedings of the IEEE international conference on computer vision*, pp. 648–657, 2017.

Pauline Luc, Camille Couprie, Yann Lecun, and Jakob Verbeek. Predicting future instance segmentation by forecasting convolutional features. In *Proceedings of the european conference on computer vision (ECCV)*, pp. 584–599, 2018.

Lucas Maes, Quentin Le Lidec, Damien Scieur, Yann LeCun, and Randall Balestriero. Leworldmodel: Stable end-to-end joint-embedding predictive architecture from pixels. *arXiv preprint arXiv:2603.19312*, 2026.

Jiaxu Miao, Yunchao Wei, Yu Wu, Chen Liang, Guangrui Li, and Yi Yang. Vspw: A large-scale dataset for video scene parsing in the wild. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4133–4143, June 2021.

Malte Mosbach and Sven Behnke. Fourier-based video prediction through relational object motion. *arXiv preprint arXiv:2110.05881*, 2021.

Lorenzo Mur-Labadia, Matthew Muckley, Amir Bar, Mido Assran, Koustuv Sinha, Mike Rabbat, Yann LeCun, Nicolas Ballas, and Adrien Bardes. V-jepa 2.1: Unlocking dense features in video self-supervised learning. *arXiv preprint arXiv:2603.14482*, 2026.

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.

Enrico Pallotta, Sina Mokhtarzadeh Azar, Shuai Li, Olga Zatsarynna, and Juergen Gall. Syncvp: joint diffusion for synchronous multi-modal video prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13787–13797, 2025.

Jordan Peper, Zhenjiang Mao, Yuang Geng, Siyuan Pan, and Ivan Ruchkin. Four principles for physically interpretable world models. *arXiv preprint arXiv:2503.02143*, 2025.

Francois Porcher, Nicolas Carion, Karteek Alahari, and Shizhe Chen. Flow matching in feature space for stochastic world modeling. *arXiv preprint arXiv:2606.29059*, 2026.

Yongming Rao, Wenliang Zhao, Zheng Zhu, Jiwen Lu, and Jie Zhou. Global filter networks for image classification. *Advances in neural information processing systems*, 34:980–993, 2021.

Marc Rigter, Tarun Gupta, Agrin Hilmkil, and Chao Ma. Avid: Adapting video diffusion models to world models. *arXiv preprint arXiv:2410.12822*, 2024.

Daniel N. Rockmore. The fft: An algorithm the whole family can use. *Computing in Science and Engg.*, 2(1):60–64, January 2000. ISSN 1521-9615. doi: 10.1109/5992.814659. URL `https://doi.org/10.1109/5992.814659`.

Oriane Siméoni, Huy V Vo, Maximilian Seitzer, Federico Baldassarre, Maxime Oquab, Cijo Jose, Vasil Khalidov, Marc Szafraniec, Seungeun Yi, Michaël Ramamonjisoa, et al. Dinov3. *arXiv preprint arXiv:2508.10104*, 2025.

Lucas Smaira, João Carreira, Eric Noland, Ellen Clancy, Amy Wu, and Andrew Zisserman. A short note on the kinetics-700-2020 human action dataset. *arXiv preprint arXiv:2010.10864*, 2020.

Matthew Tancik, Pratul Srinivasan, Ben Mildenhall, Sara Fridovich-Keil, Nithin Raghavan, Utkarsh Singhal, Ravi Ramamoorthi, Jonathan Barron, and Ren Ng. Fourier features let networks learn high frequency functions in low dimensional domains. *Advances in neural information processing systems*, 33:7537–7547, 2020.

Robbyant Team, Zelin Gao, Qiuyu Wang, Yanhong Zeng, Jiapeng Zhu, Ka Leong Cheng, Yixuan Li, Hanlin Wang, Yinghao Xu, Shuailei Ma, et al. Advancing open-source world models. *arXiv preprint arXiv:2601.20540*, 2026.

Matthieu Tehenan, Christian Bolivar Moya, Tenghai Long, and Guang Lin. Linear spatial world models emerge in large language models. *arXiv preprint arXiv:2506.02996*, 2025.

Zhan Tong, Yibing Song, Jue Wang, and Limin Wang. Videomae: Masked autoencoders are data-efficient learners for self-supervised video pre-training. *Advances in neural information processing systems*, 35:10078–10093, 2022.

Carl Vondrick, Hamed Pirsiavash, and Antonio Torralba. Anticipating visual representations from unlabeled video. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 98–106, 2016.

Philipp Wu, Alejandro Escontrela, Danijar Hafner, Pieter Abbeel, and Ken Goldberg. Daydreamer: World models for physical robot learning. In *Conference on robot learning*, pp. 2226–2240. PMLR, 2023.

Haipeng Xiao, Yueyue Fang, Junzhe Fu, Lixin Guan, Mengshan Li, Xindi Yuan, Wenyi Zhou, and Jihong Zhu. Fc-font: Few-shot diffusion-based font generation via frequency-domain style compensation. *Computers & Graphics*, pp. 104570, 2026.

Shicheng Yin, Kaixuan Yin, Weixing Chen, Yang Liu, Guanbin Li, and Liang Lin. Ddp-wm: Disentangled dynamics prediction for efficient world models. *arXiv preprint arXiv:2602.01780*, 2026.

Afia Zafar, Muhammad Aamir, Nazri Mohd Nawi, Ali Arshad, Saman Riaz, Abdulrahman Alruban, Ashit Kumar Dutta, and Sultan Almotairi. A comparison of pooling methods for convolutional neural networks. *Applied Sciences*, 12(17):8643, 2022.

Peng-Fei Zhang, Ying Cheng, Xiaofan Sun, Shijie Wang, Fengling Li, Lei Zhu, and Heng Tao Shen. A step toward world models: A survey on robotic manipulation. *arXiv preprint arXiv:2511.02097*, 2025.

Anlin Zheng, Xin Wen, Xuanyang Zhang, Chuofan Ma, Tiancai Wang, Gang Yu, Xiangyu Zhang, and Xiaojuan Qi. Vision foundation models as effective visual tokenizers for autoregressive image generation. *arXiv preprint arXiv:2507.08441*, 2025.

Penghao Zhu, Salvatore Penachio, Kaustav Mukherjee, and Aneesh Jonelagadda. Spectral-target physical latent structuring for jepa-style world models. *arXiv preprint arXiv:2609.04264*, 2026.

APPENDIX

A EXPERIMENTAL SETUP

A.1 DATASETS

Data and temporal sampling. We pretrain on the training release of *Kinetics-700-2020* (Smaira et al., 2020). The first eight eligible entries in each of the 700 archives are excluded from training, providing 256 held-out clips for monitoring and 696 for Fourier component analysis. These subsets are distinct from the official validation split. For each training example, we sample eight timestamps with independent adjacent intervals uniformly drawn from $[1/25, 1/3]$ seconds and a uniformly sampled feasible starting time. Timestamps are mapped to the nearest available frames, retaining repeated selections. One of the four available four-context/one-target windows is then selected uniformly.

Downstream evaluation populations are summarized in Table 1. VSPW (Miao et al., 2021) uses 758 clips from 329 validation videos; KITTI (Geiger et al., 2013) uses 603 samples from 27 sequences in the filtered Eigen test split. TAP-Vid-RGB-Stacking (Doersch et al., 2022) comparisons use a common set of 1,210 queries from 50 videos, with 1,187 and 1,154 visible targets at H_1 and H_3, respectively.

Forecast alignment. Sampling intervals are three original frames for Cityscapes (Cordts et al., 2016) and VSPW, two for KITTI, and five for TAP-Vid-RGB-Stacking. These correspond to 0.1875 seconds for Cityscapes and 0.2 seconds for the other three datasets. EK100 (Damen et al., 2022) uses approximately one-third of a second, quantized to the video frame rate.

Cityscapes and KITTI evaluate both horizons against the same annotated endpoint. For sampled states $\mathbf{y}_0, \ldots, \mathbf{y}_6$, H_1 uses $\mathbf{y}_{2:5}$ as context, whereas H_3 rolls out from $\mathbf{y}_{0:3}$ to $\mathbf{y}_6$. VSPW and TAP-Vid-RGB-Stacking instead use the same initial context for both horizons. Intermediate ground-truth features are never supplied during autoregressive rollout.

Spatial preprocessing. Kinetics-700 training uses a temporally shared random resized crop with area scale $[0.6, 1.0]$, log-uniform aspect ratio $[3/4, 4/3]$, output resolution 512×512, and horizontal flipping with probability 0.5.

Cityscapes and VSPW images are resized with shorter side 512 and each dimension capped at 1,024, then cropped at opposite ends of the longer axis to obtain two 512×512 inputs. For KITTI, we first apply a bottom-center 352×1216 crop to RGB and depth, resize to 512×1024, and extract two 512×512 crops. TAP-Vid-RGB-Stacking uses direct resizing to 512×512, while EK100 uses two 512×512 crops per frame. Spatial preprocessing is consistent across time.

A.2 Implementation Details

Model configuration. We use the frozen DINOv3 (Siméoni et al., 2025) checkpoint `facebook/dinov3-vitb16-pretrain-lvd1689m`, extracting final-layer patch tokens without classification or register tokens. The predictor has nine blocks, hidden dimension 768, 12 attention heads, FFN intermediate dimension 3,376, and LayerScale initialized to 10^{-5}. Queries use a shared learnable mask initialized with standard deviation 0.02. The four-frame Fourier transform uses orthonormal normalization.

The pretrained model includes a $7 \to 16 \to 768$ timestamp-conditioning MLP for temporal keys. Its correction vanishes at uniformly spaced 0.2-second intervals; the FFT remains defined over sample indices without temporal resampling. The predictor contains approximately 85.7M parameters.

Pretraining. We use one-step, elementwise mean Smooth L1 loss with $\beta = 0.1$. AdamW uses learning rate 10^{-4}, betas $(0.9, 0.999)$, $\epsilon = 10^{-8}$, and weight decay 0.4 applied only to linear and convolutional weights. The schedule consists of 5,000 linear-warmup updates, a constant learning rate until update 55,000, and 10^{-5} for the final 5,000 updates.

Training uses four 24-GiB NVIDIA RTX A5000 GPUs, global batch size 1,024 accumulated over 13 microbatches per GPU, BF16 autocasting, FP32 loss, distributed data parallelism, and `torch.compile`. We use seed 0 and no gradient clipping.

Benchmark-specific training. Cityscapes and VSPW each provide 20,000 training examples and 1,000 internal validation examples, including crops. Segmentation annotations guide class-aware sampling, while predictors receive feature-space supervision. Scratch `-style` training uses 2,500 updates, batch size 32, learning rate 10^{-4}, and 625 warmup updates. Domain adaptation starts from pretrained predictors and uses 5,000 updates at 10^{-5} without warmup; the final checkpoint is evaluated.

Both protocols retain the pretraining AdamW settings and use seed 20260724. Only predictors are trained; the VFM, task heads, and DeltaWorld's pretrained DeltaTok codec remain frozen. Dense-feature predictors use Smooth L1 loss, whereas DeltaWorld-style retains its latent-space best-of-256 objective. Query-initialization variants are trained separately under the scratch protocol, with identical Fourier memory and latest-feature-anchored output formulations.

A.3 Baselines

Latent predictors. DeltaWorld uses `Amazon-FAR/deltaworld-kinetics` and the frozen `Amazon-FAR/deltatok-kinetics` codec. The latter contains approximately 170.4M parameters excluded from predictor-capacity matching. DeltaWorld-style initializes only the predictor

from scratch. For `BoM-mean 20`, Gaussian queries with standard deviation 0.02 generate 20 hypotheses, each maintaining an independent autoregressive history. Decoded feature maps are averaged at the evaluated horizon before applying the task head. Each forecasting context initializes its feature boundary using a black-image DINOv3 feature.

All `-style` predictors use hidden dimension 768 and 12 heads, with approximately 85.7M parameters. DINO-Foresight-style uses nine blocks and FFN width 3,028, with factorized temporal/spatial attention over context grids and a fully masked future grid. DINO-World-style uses 12 blocks and FFN width 3,072, with learned future queries, context cross-attention, and three-axis rotary positional encoding. V-JEPA-style uses 12 blocks and FFN width 3,041, joint context/mask processing, and fixed three-dimensional sinusoidal positions. These are architecture-style controls under frozen DINOv3 representations rather than reproductions of the original pretraining procedures.

RGB predictors. We directly evaluate the released `nvidia/Cosmos-Predict1-4B` and `nvidia/Cosmos-Predict1-12B` models without additional training. Generated frames are encoded with frozen DINOv3 and evaluated using the corresponding task heads or feature matcher.

A.4 Evaluation Details

Segmentation and depth. We use DeltaWorld's frozen Cityscapes, VSPW, and KITTI heads in FP32 evaluation mode. Segmentation logits are bilinearly resized to the annotation grid; overlapping crops are evaluated without double counting. mIoU is computed from the aggregated confusion matrix, excluding ignored pixels.

KITTI uses the bottom-center crop, Garg evaluation mask, and valid depths in $(10^{-3}, 80)$ meters. The released depth head is used without modification. We apply no median scaling and compute RMSE from the total squared error over all valid pixels.

EK100. Evaluation is aligned to action end time s: observed frames occur at $s-2, s-5/3, s-4/3, s-1$ seconds, and predicted frames at $s-2/3, s-1/3, s$. Unavailable history is padded with the first frame. This custom protocol can observe part of the action and differs from action-onset anticipation.

We average patch features within each crop and then across crops, concatenating seven 768-dimensional frame vectors into 5,376 dimensions. Features are standardized using training statistics, with standard deviation bounded below by 10^{-6}. Three linear classifiers predict 97 verbs, 289 nouns, and 3,568 actions using summed cross-entropy loss. Probe training uses AdamW, learning rate 10^{-3}, weight decay 10^{-4}, batch size 2,048, and 80 epochs.

Probes are trained on real observed and future features, then frozen and shared across models; predicted features replace future features during evaluation. We report MCR@5 across five probe seeds, 41–45; reported variability is their population standard deviation.

The 20,000 training trajectories are selected using validation action-class identities to prioritize coverage, initially selecting up to three training examples per class. Validation RGB/features are not used for fitting, but selection is label-informed. One of 8,294 evaluation clips has an action absent from probe training and is excluded, leaving 8,293 scored clips.

Future-point forecasting. TAP-Vid queries are visible, in-bounds annotated points in the last observed frame; selection does not use future visibility. Unit-normalized query descriptors are bilinearly sampled and matched by cosine similarity to normalized future feature tokens. The 32×32 similarity map is bilinearly upsampled to 256×256, and its argmax identifies the predicted pixel center. The matcher has no trainable parameters.

For each video, we compute localization accuracy on future-visible targets at pixel thresholds $\{1, 2, 4, 8, 16\}$ on the 256×256 grid. Averaging across thresholds and then videos yields δ_{avg}, reported as a percentage. Only the Oracle accesses actual future features; RGB baselines use features of generated frames.

Feature geometry. Targets are DINOv3 features of actual future frames. Norm ratio averages tokenwise predicted-to-target norm ratios. For spatial geometry, 32 evenly spaced tokens per crop are

concatenated into a 64×768 matrix and centered over tokens. Effective rank is $\exp(-\sum_i p_i \log p_i)$ for normalized covariance eigenvalues p_i. Norm and predicted-to-target rank ratios are ideally one.

Spectrum JS measures Jensen–Shannon divergence between normalized spatial-covariance eigenspectra using natural logarithms. Procrustes error is $\min_{\mathbf{R}^\top\mathbf{R}=\mathbf{I}} \|\mathbf{X}_c\mathbf{R} - \mathbf{Y}_c\|_F / \|\mathbf{Y}_c\|_F$, without scale fitting. Metrics are averaged equally across evaluation samples. Sampled-matrix cosine is distinguished from full-grid cosine; the covariance spectrum is distinct from temporal Fourier energy.

Efficiency. Measurements use one RTX A5000, PyTorch 2.6, CUDA 12.6, eager BF16 inference, batch size one, and pre-extracted features of shape $1 \times 4 \times 1024 \times 768$. After five warmup calls, we time 30 repetitions per sample with CUDA synchronization. Reported latency averages the per-sample medians over 64 samples.

Peak allocated CUDA memory includes model weights, inputs, intermediates, and outputs. VFM encoding and task heads are excluded. Latency and memory include FFT and other predictor operations, whereas GFLOPs count only linear and attention matrix operations, with multiply-add counted as two FLOPs.

DeltaWorld measurements include context encoding, 20-hypothesis prediction, and dense decoding, but exclude final feature averaging. Architecture-style controls are profiled with initialized weights; Fourier-World and DeltaWorld use saved weights.

Fourier interventions. Component ablations modify the frozen Kinetics-700 model without retraining. AC removal zeros both real and imaginary components; imaginary removal affects only the latter. Nyquist removal zeros its lifted token, and complete-update removal suppresses every block innovation.

Spatial phase shuffling retains local AC magnitudes and transfers unit phases by a one-token circular spatial shift, retaining the original phase for zero-magnitude donors. Mode slots and identities remain present; normalization scales and motion masks are computed before intervention. The intervention is repeated at every autoregressive step for H_3.

K/V pathway analysis. We use 96 VSPW clips from 82 videos, selected across quantiles of observed feature variation without using future targets. Each edit halves the real AC, imaginary AC, or Nyquist component. Before inverse FFT, DC is corrected to preserve the latest feature:

$$\mathbf{DC}' = \mathbf{DC} + (\mathbf{N}' - \mathbf{N}) - 2\big(\mathrm{Im}(\mathbf{AC}') - \mathrm{Im}(\mathbf{AC})\big), \tag{10}$$

where $\mathbf{N}$ denotes Nyquist. These are consistent feature-space edits, without assuming realizable RGB counterparts.

At all nine blocks, we substitute edited temporal keys, values, or both after projection, normalization, and positional encoding. For native, key-only, value-only, and joint H_1 Smooth L1 losses L_0, L_K, L_V, L_{KV}, pathway contributions are

$$\begin{aligned} \phi_K &= \tfrac{1}{2}[(L_K - L_0) + (L_{KV} - L_V)], \\ \phi_V &= \tfrac{1}{2}[(L_V - L_0) + (L_{KV} - L_K)]. \end{aligned} \tag{11}$$

Error bars are 95% intervals from 2,000 bootstrap resamples clustered by video. Attention changes are measured separately as $\frac{1}{2}\sum_k |\alpha_k^{\mathrm{edit}} - \alpha_k^{\mathrm{native}}|$, averaged over heads, tokens, and clips per block.

B Additional Analysis

B.1 Confidence Intervals

Table 7 reports asymmetric 95% confidence intervals across all eleven main evaluation settings. Using the full evaluation sets, we perform 10,000 cluster bootstrap resamples, sampling sequences, videos, or drives with replacement according to the dataset. The evaluation metric is recomputed for each resample, and the 2.5th and 97.5th percentiles of the resulting distribution define the interval bounds. The table reports signed lower and upper offsets from each model's point estimate, preserving interval asymmetry.

Table 7: Asymmetric 95% confidence intervals for the main results. Entries report signed offsets from each model's point estimate: for a score s, the interval is $[s + \Delta_{\text{lower}},\ s + \Delta_{\text{upper}}]$. Offsets are in metric points for mIoU, δ_{avg}, and MCR@5, and in meters for KITTI RMSE.

		Fourier-World		DeltaWorld	
Dataset	Horizon / target	Δ_{lower}	Δ_{upper}	Δ_{lower}	Δ_{upper}
Cityscapes	H_1	–1.854	+0.728	–1.457	+0.899
	H_3	–2.963	+0.733	–2.102	+0.778
VSPW	H_1	–4.429	+0.319	–4.360	+0.345
	H_3	–4.380	+0.421	–4.246	+0.481
KITTI	H_1	–0.186	+0.167	–0.180	+0.155
	H_3	–0.223	+0.213	–0.269	+0.250
TAP-Vid-RGB-Stacking	H_1	–1.830	+1.680	–1.611	+1.524
	H_3	–2.518	+2.294	–2.383	+2.191
EK100	Verb	–2.409	+4.800	–1.003	+3.547
	Noun	–1.467	+3.107	–1.096	+2.506
	Action	–0.637	+1.950	–0.399	+1.585

Table 8: Temporal representation analysis on four-frame DINO features from Cityscapes and VSPW. Profile cosine and TV compare normalized group-energy profiles before and after a context-window shift; correlation RMS measures correlation between non-DC coordinates. Ranges show the minimum and maximum set-level means across four evaluation sets, not confidence intervals. Bold indicates the best result for each metric.

Basis	Profile cosine ↑	Profile TV ↓	Correlation RMS ↓
Fourier	**0.9637–0.9690**	**0.1045–0.1143**	0.2224–0.2833
DCT	0.9530–0.9598	0.1193–0.1282	**0.0245–0.0327**
Haar	0.9289–0.9319	0.1627–0.1658	0.1593–0.1929

B.2 Temporal Stability of Fourier Representations

We examine the properties of temporal Fourier representations that motivate their use in FourierWorld. We compare Fourier, discrete cosine transform (DCT), and Haar decompositions of four-frame DINO feature sequences from Cityscapes and VSPW. For each dataset, we use two disjoint sets of 120 sources, yielding four evaluation sets in total. All decompositions are computed on CPU without neural network training. Reported ranges denote the minimum and maximum of the set-level means across these four sets, rather than confidence intervals.

Stability under window shifts. After excluding the DC component, we organize the remaining components into two groups. For Fourier, these are the AC quadrature pair and the Nyquist component; for Haar, the difference between the two temporal halves and the two within-half differences; for DCT, the first non-DC coefficient and the remaining two coefficients. We construct a normalized energy profile from the mean squared energy of each group and compare profiles before and after shifting the context window. Profile cosine therefore measures similarity between grouped energy distributions, rather than between feature vectors. We also report profile total variation (TV) distance and the root mean square (RMS) correlation between non-DC coordinates.

As shown in Table 8, Fourier achieves the highest profile cosine and lowest TV distance, indicating more stable group-level energy composition as the window shifts. DCT, however, exhibits the lowest correlation between non-DC coordinates, followed by Haar. Energy concentration in a single coordinate is also lower for Fourier (48.73–49.21%) than for Haar (52.45–54.16%) and DCT (55.03–57.26%). These results identify temporal stability of the grouped representation as a particular strength of Fourier, while DCT provides greater coordinate decorrelation and energy concentration.

Table 9: Representation stability with calibrated singleton-plus-pair groupings. Each evaluation set uses 60 calibration and 60 test sources. All bases decompose the same features predicted by a shared autoregressive rule. The H_3 cosine compares energy profiles of predicted and ground-truth contexts. Ranges summarize the four test-set means, not confidence intervals. Bold indicates the best result for each metric.

Basis	Circular-shift cross-group mixing ↓	H_3 energy-profile cosine ↑
Fourier	$\approx \mathbf{0}$	**0.9672–0.9757**
DCT	0.1357	0.8873–0.9154
Haar	0.3750	0.9058–0.9402

Table 10: Fourier-World ablations. The first row reports absolute scores; all other rows report changes relative to the full model (ablated minus full). Kinetics-700 reports Smooth L1 loss, with ablation entries denoting Δloss. Positive loss/RMSE changes and negative mIoU changes indicate degradation.

	Kinetics-700 Smooth L1 loss ↓		VSPW mIoU ↑		Cityscapes mIoU ↑		KITTI RMSE ↓	
Variant	H_1	H_3	H_1	H_3	H_1	H_3	H_1	H_3
Fourier-World	0.057947	0.095304	54.21	48.23	63.19	51.75	3.145	4.004
Imaginary quadrature off	+0.006014	+0.004213	−1.897	−3.134	−6.384	−8.366	+0.338	+0.612
AC off	+0.007759	+0.005807	−2.261	−2.990	−7.238	−7.778	+0.432	+0.611
Nyquist off	+0.004709	+0.005856	−1.653	−2.410	−5.480	−7.565	+0.329	+0.572
AC phase spatial shuffle	+0.002594	+0.002034	−0.679	−1.411	−2.994	−4.216	+0.121	+0.279
Complete update off	+0.013263	+0.01738	−2.766	−3.930	−8.128	−12.150	+0.514	+0.856

Stability under autoregressive prediction. To reduce dependence on manually specified groups, we conduct a second analysis in which a singleton-plus-pair grouping is selected separately for each basis using calibration data. Each 120-source set is divided into 60 calibration and 60 test sources. A shared lightweight autoregressive rule generates the predicted features, and every basis decomposes exactly the same predictions. This isolates differences between representations from differences in predictor quality.

Table 9 reports cross-group mixing under circular shifts and the energy-profile cosine between predicted and ground-truth contexts at H_3. Fourier exhibits approximately zero cross-group mixing, compared with 0.1357 for DCT and 0.3750 for Haar. Its H_3 profile cosine also remains higher (0.9672–0.9757) than that of DCT (0.8873–0.9154) and Haar (0.9058–0.9402).

Together, these analyses support using Fourier memory to provide temporal groups whose energy composition remains stable under window shifts and autoregressive perturbations. They motivate our representation choice, but do not establish that a trained Fourier-based predictor necessarily outperforms predictors using alternative bases.

B.3 Full Fourier Ablation Results

Table 10 reports the full inference-time ablation results at H_1 and H_3, including Kinetics-700 forecasting loss and downstream metrics.

B.4 Query Ablation

Fourier-World initializes its queries with a shared learnable mask embedding. Since the predicted feature update is computed from the difference between the final and initial queries, we examine how query initialization affects downstream performance. Table 11 compares four initializations: the latest observed feature, zero, the mean of the context features, and a learned mask.

Table 11: Query initialization ablation on Cityscapes and VSPW. We report mIoU at H_1 and H_3. Higher is better; bold indicates the best result in each column.

	Cityscapes (mIoU ↑)		VSPW (mIoU ↑)	
Query initialization	H_1	H_3	H_1	H_3
Latest feature	62.4702	49.7639	51.78	45.3005
Zero	62.842	48.9845	52.5706	45.7979
Context mean	62.3925	49.259	51.9206	45.2411
Learned mask	**62.9284**	**50.2242**	**52.6069**	**45.9865**

Table 12: Domain adaptation on VSPW and Cityscapes. We report mIoU before and after adaptation, with gains in mIoU points. Bold indicates the best post-adaptation performance within each dataset.

		Before adaptation		After adaptation		Gain	
Dataset	Method	H_1	H_3	H_1	H_3	H_1	H_3
VSPW	DeltaWorld (BoM-mean 20)	54.0509	46.1821	54.1226	47.1089	+0.0717	+0.9268
	Fourier-World	54.21	48.23	**54.5684**	**48.7483**	+0.3584	+0.5183
Cityscapes	DeltaWorld (BoM-mean 20)	64.1199	49.9497	65.2723	52.9327	+1.1524	+2.983
	Fourier-World	63.19	51.75	**66.2729**	**55.5796**	+3.0829	+3.8296

The learned mask achieves the highest mIoU at both horizons on Cityscapes and VSPW, although some margins are small. These results support our choice of a content-independent learnable initialization: directly incorporating the latest feature or the context mean into the initial query does not improve performance in this comparison.

B.5 Domain Adaptation

Beyond transfer without additional predictor training, practical applications may benefit from adapting a world model to a target domain. We therefore examine whether Fourier-World can further improve its predictive performance through domain adaptation. Table 12 compares Fourier-World and DeltaWorld before and after adaptation on VSPW and Cityscapes.

Both models improve at both forecasting horizons on both datasets. Fourier-World achieves the highest post-adaptation mIoU in all four settings. Its gains are particularly pronounced on Cityscapes, where adaptation improves mIoU by 3.0829 points at H_1 and 3.8296 points at H_3, compared with 1.1524 and 2.983 points for DeltaWorld, respectively. These improvements allow Fourier-World to surpass DeltaWorld at H_1 while extending its advantage at H_3.

On VSPW, Fourier-World improves by 0.3584 and 0.5183 points at H_1 and H_3, respectively. Although DeltaWorld obtains a larger gain at H_3, Fourier-World retains higher absolute performance at both horizons. These results show that Fourier-World benefits from target-domain adaptation in addition to its transfer capability, with the magnitude of improvement varying across datasets.

B.6 Prediction Consistency under Static Contexts

We examine whether world models preserve a static feature state when the observed context contains no temporal variation. For this controlled experiment, we construct each context by repeating the same VFM feature map across all four input positions. We measure the mean absolute latent drift between the predicted feature map and the repeated input feature at H_1 and H_3. This experiment tests whether a predictor introduces changes despite the absence of temporal variation in its input.

As shown in Figure 4, the compared baselines produce nonzero feature changes even under identical context inputs, with larger deviations at H_3 than at H_1. Fourier-World instead maintains zero drift at both horizons. The baseline results indicate that a static observed context does not necessarily lead to a static prediction in these learned predictors.

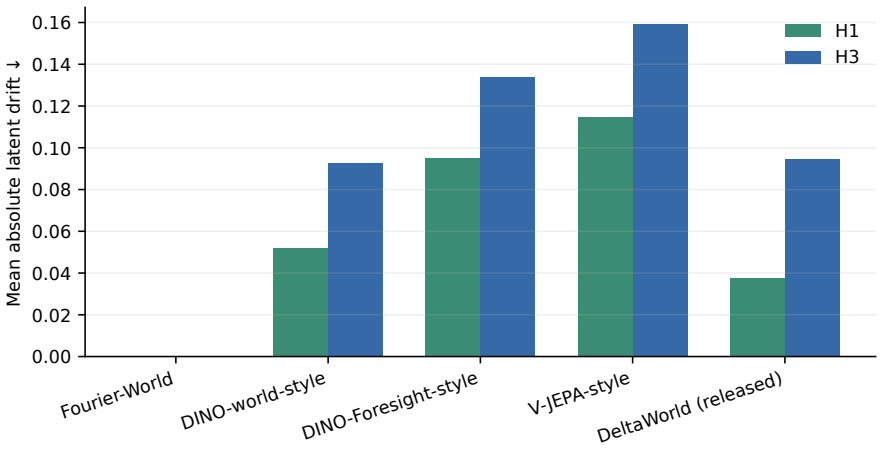


Figure 4: **Prediction drift under static contexts.** Each context consists of four identical VFM feature maps. We report mean absolute latent drift from the repeated input feature at H_1 and H_3; lower is better. Fourier-World preserves the input state at both horizons, whereas the compared baseline predictors exhibit nonzero drift that increases at H_3.

For Fourier-World, this behavior follows directly from its design. Identical context features yield zero non-DC Fourier coefficients. The spectral-activity gate therefore suppresses the block updates, leaving the final query equal to its initialization. Consequently, the predicted feature update is zero and the latest feature is preserved. Inserting this prediction back into the context maintains the same static sequence, so the property also holds under autoregressive rollout:

$$\mathbf{x}_{t-3} = \cdots = \mathbf{x}_t = \mathbf{x} \quad \Longrightarrow \quad \hat{\mathbf{x}}_{t+h} = \mathbf{x}, \qquad h \geq 1. \tag{12}$$

This static fixed-point property provides a structural consistency guarantee: in the absence of observed temporal variation, Fourier-World preserves the current state without introducing autoregressive drift.

C Qualitative Analysis

We present qualitative examples across five benchmarks, covering semantic segmentation, depth forecasting, stop-aligned action prediction, and future-point forecasting. These examples illustrate spatial prediction errors and changes in prediction quality over extended autoregressive rollouts. Reported ROI metrics describe the displayed regions rather than dataset-level performance.

C.1 Semantic Segmentation Forecasting

Figures 5 and 6 compare segmentation forecasts at H_3. In the Cityscapes example, Fourier-World more closely recovers the future bicycle region. In the VSPW example, it better matches the future person region around the hand and arm. Both examples show higher ROI class IoU than Copy-last and DeltaWorld.

C.2 Depth Forecasting

Figure 7 compares depth forecasts at H_3. Fourier-World achieves an ROI RMSE of 3.61 m, compared with 4.42 m for Copy-last and 10.37 m for DeltaWorld. The accompanying error maps show the spatial distribution of depth errors at valid LiDAR locations.

C.3 Extended Autoregressive Rollouts

Figure 8 extends the comparison to H_8, corresponding to 1.6 seconds after the last observation. Although Copy-last and DeltaWorld obtain lower ROI RMSE at H_1, Fourier-World maintains lower

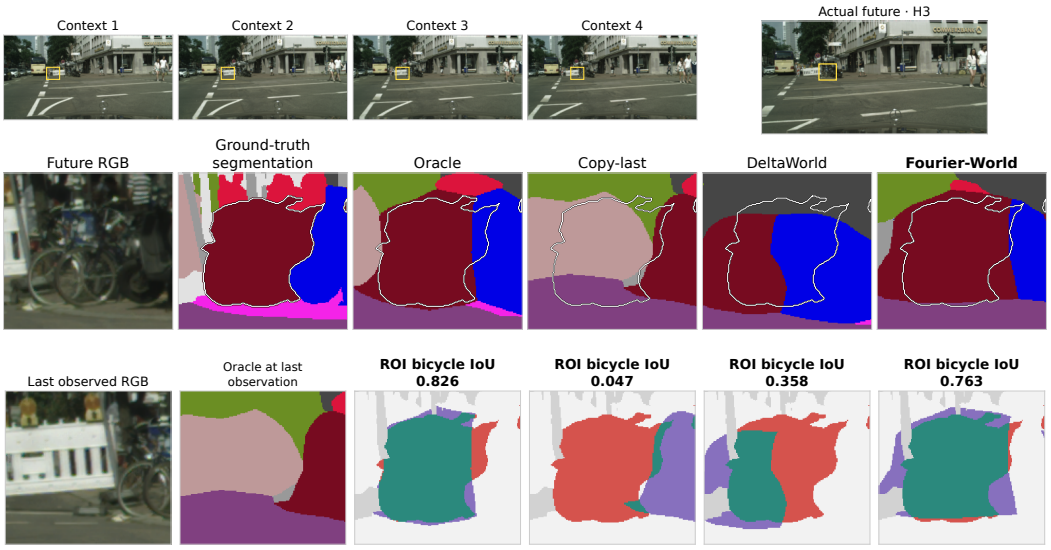


Figure 5: **Qualitative segmentation forecasting on Cityscapes at H_3.** Yellow boxes indicate the displayed region of interest (ROI), and white contours mark the ground-truth bicycle boundary at the future time step. The bottom-row error maps show true positives (teal), false negatives (red), and false positives (purple) for the bicycle class. Reported IoUs are computed within the displayed ROI. Oracle predictions are obtained by applying the segmentation head to actual VFM features at the indicated time step. Fourier-World more closely matches the future bicycle region than Copy-last and DeltaWorld in this example.

error at H_3, H_5, and H_8. At H_8, its ROI RMSE is 1.83 m, compared with 4.49 m for Copy-last and 6.40 m for DeltaWorld. This example illustrates the benefit of maintaining prediction quality over longer rollouts.

C.4 Stop-Aligned Action Prediction

Figure 9 illustrates stop-aligned action prediction on EK100. For the action *open oven door*, Fourier-World includes the ground-truth action among the top-five predictions for all five linear probes, as does the Oracle. DeltaWorld does not include it for any of the five probes in this example.

C.5 Future-Point Forecasting

Figure 10 compares future-point predictions on TAP-Vid-RGB-Stacking at H_3. Fourier-World obtains a lower reported endpoint error (EPE) than DeltaWorld in this example: 8.91 versus 13.04 pixels. The endpoint overlays show the spatial distribution of the remaining localization errors.

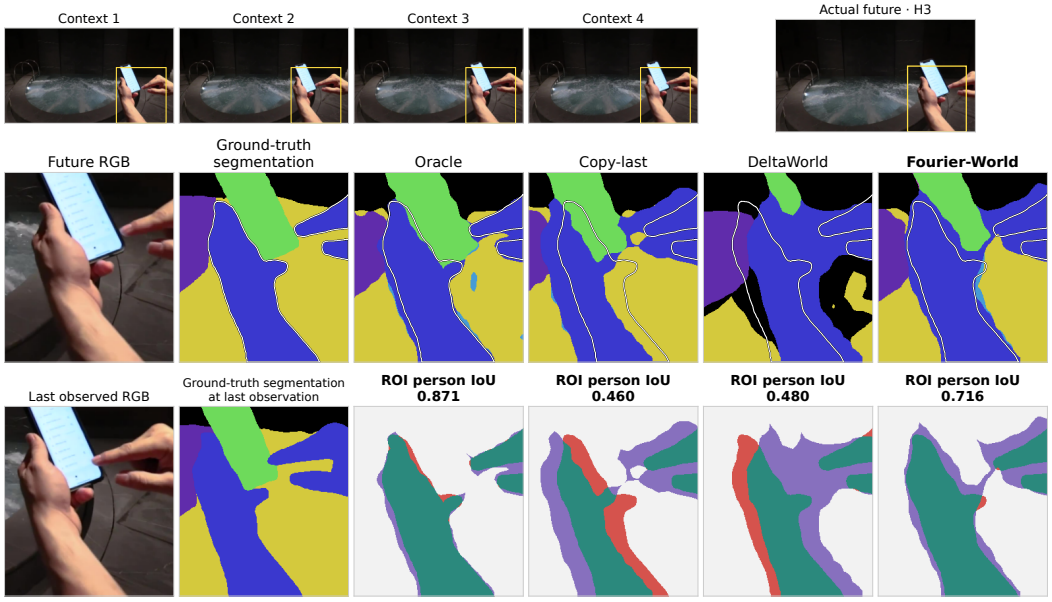


Figure 6: **Qualitative segmentation forecasting on VSPW at H_3.** Yellow boxes indicate the displayed region of interest (ROI), and white contours mark the ground-truth person boundary at the future time step. The bottom row includes the last observed RGB crop and its ground-truth segmentation, followed by person-class error maps. Teal, red, and purple indicate true positives, false negatives, and false positives, respectively. Reported IoUs are computed for the person class within the displayed ROI. The Oracle applies the segmentation head to actual future VFM features. Fourier-World more closely matches the future person region than Copy-last and DeltaWorld in this example.

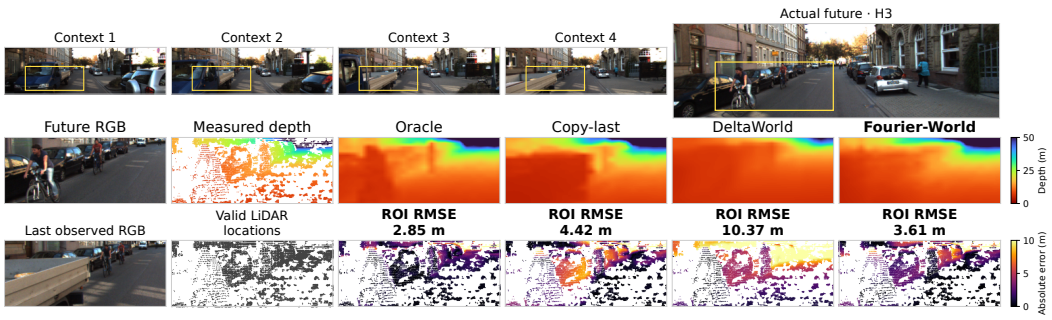


Figure 7: **Qualitative depth forecasting on KITTI at H_3.** The top row shows four observed frames and the actual future frame, with yellow boxes indicating the displayed region of interest (ROI). The middle row compares measured depth and depth estimates from Oracle, Copy-last, DeltaWorld, and Fourier-World features. The bottom row shows the last observed RGB crop, valid LiDAR locations, and absolute depth errors. Reported RMSE values are computed at valid depth locations within the displayed ROI. Depth and absolute-error color scales span 0–50 m and 0–10 m, respectively. The Oracle applies the depth head to actual future VFM features. Fourier-World achieves lower ROI RMSE than Copy-last and DeltaWorld in this example.

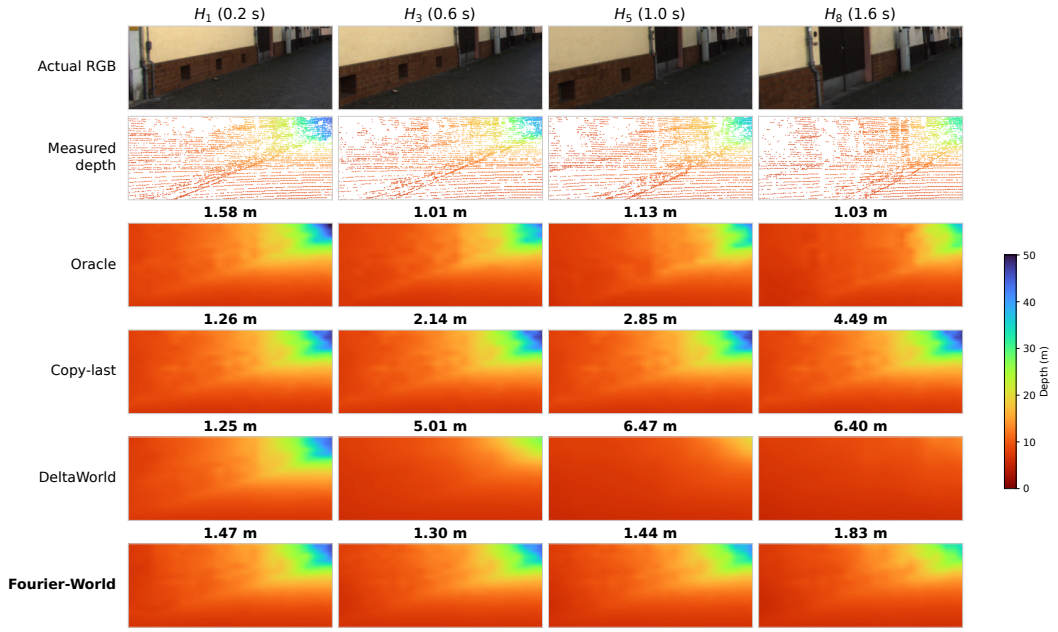


Figure 8: **Extended autoregressive depth forecasting on KITTI.** Columns show forecasts at H_1, H_3, H_5, and H_8, corresponding to 0.2, 0.6, 1.0, and 1.6 seconds after the last observation. The top two rows show actual future RGB frames and measured depth. Subsequent rows compare depth estimates from Oracle, Copy-last, DeltaWorld, and Fourier-World features using a shared depth color scale. Values above depth maps indicate ROI RMSE in meters, evaluated at valid depth locations. The Oracle uses actual future VFM features. Although Copy-last and DeltaWorld obtain lower error at H_1, Fourier-World maintains lower error at the subsequent displayed horizons in this example.

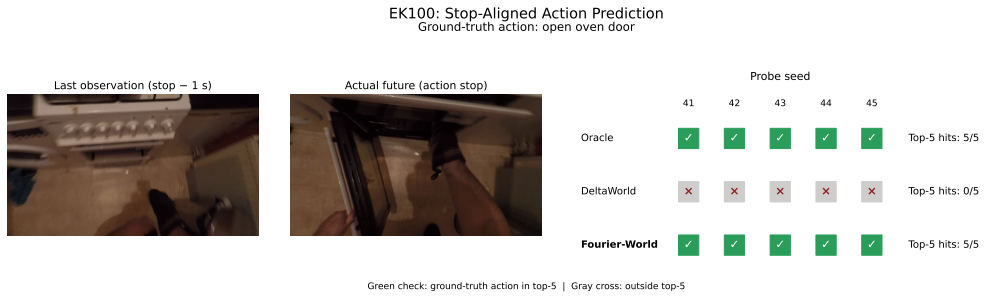


Figure 9: **Qualitative action prediction on EK100.** The left panels show the last observed frame, one second before the action ends, and the actual frame at action completion. Each column on the right corresponds to a separately trained linear probe with the indicated seed. A check indicates that the ground-truth action appears among the top-five predictions; a cross indicates that it does not. Classification uses four observed and three predicted frame representations, with the final prediction aligned to the action end. The Oracle instead uses actual future features. The displayed future RGB frame is provided for reference and is not an input to the forecasting models.

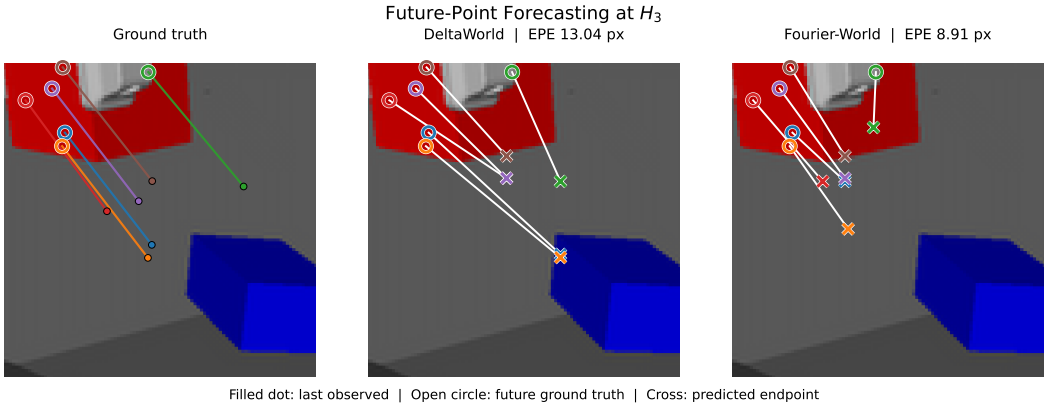


Figure 10: **Qualitative future-point forecasting on TAP-Vid-RGB-Stacking at H_3.** Filled dots indicate last observed point locations, open circles indicate future ground-truth locations, and crosses indicate predicted endpoints. Colors identify corresponding points across panels. In the ground-truth panel, colored segments connect observed and future locations. In the prediction panels, white segments connect predicted endpoints to their corresponding future ground-truth locations, showing localization errors. All panels use the same crop and spatial scale. Fourier-World achieves lower reported EPE than DeltaWorld in this example.