

Evaluating Early Multimodal Fusion in Discrete Latent World Models for Robot Navigation

Mohammad Safwan Mondal¹, Mohammad Nazeri¹, Ruben Hammele², Samira Huber², Alexandyr Card¹, Anuj Pokhrel¹, Daeun Song³, Sören Pirk², and Xuesu Xiao¹

Abstract—Multimodal navigation world models fuse heterogeneous modalities—such as visual observations, proprioceptive states, and action histories—to predict and plan future motion. This early fusion operator is not merely a feature aggregator; it fundamentally dictates the topological structure of the latent planning space, with cascading effects from perception input to continuous action execution. Yet, its impact on representational stability, particularly during sensor dropout, remains underexplored. In this study, we evaluate five multimodal fusion mechanisms—Shift, Scale, FiLM, Cross-Attention, and AdaLN-Zero—within Hydra, a discrete latent World Action Model. We assess these operators across offline predictive metrics utilizing guidance dropout, alongside 100 closed-loop physical robot trials evaluating input modality dropout. Our findings reveal an architectural trade-off: high-capacity operators like FiLM and Cross-Attention excel under offline guidance dropout but frequently encounter planning degradation during physical sensor dropout. We demonstrate that additive translation parameters induce a covariate shift within the pre-quantization latent space under input dropout, shifting representations across discrete boundaries and triggering the retrieval of physically incoherent dynamic intents. Conversely, restricting fusion to bounded element-wise magnitude scaling regularizes against this shift, preserving the topological integrity of the latent space and yielding smoother physical trajectories with up to an 80% success rate under sensor loss. Ultimately, these results offer an architectural insight for multimodal world models in navigation: robust closed-loop execution benefits from topology-preserving modulation, highlighting the representational vulnerabilities introduced by additive translation biases.

I. INTRODUCTION

Autonomous navigation requires predicting how robot actions affect both future motion and observations. World models learn these relationships from visual and kinodynamic histories by forecasting future states or observations [1]–[3]. Generative World Action Models further connect predictive modeling with action generation. Hydra [4], for example, performs discrete latent planning over candidate trajectories and decodes selected intents into continuous motion. The multimodal representation connecting perception and kinodynamics is therefore important for both predictive accuracy and closed-loop execution.

A central design choice is how heterogeneous modalities are fused. Visual observations provide spatially distributed environmental information, whereas actions and robot states encode compact kinodynamic signals. Early fusion combines these modalities before temporal prediction, providing a computationally compact way to form a shared representation without maintaining separate downstream processing

streams for each modality [4]. The fusion operator then determines how these signals interact within this shared representation. In discrete-latent world models, these interactions may also influence the geometry and stability of the representation presented to the planner.

To characterize this relationship, it is necessary to decouple a model’s conditional predictive capacity from its representational resilience. We approach this through a dual-evaluation paradigm. During offline dataset benchmarking, we assess conditional predictive capacity utilizing *guidance dropout* at the decoding stage, maintaining fully observable inputs during early fusion. Conversely, during physical robot deployment, we evaluate representational robustness by subjecting the network to *input modality dropout* prior to the fusion layer. This methodological distinction isolates a specific representational vulnerability: a fusion operator that effectively structures the latent space for downstream querying may simultaneously induce a covariate shift when sensor data is dropped from the input.

We investigate this dynamic within Hydra by comparing five fusion mechanisms—Shift, Scale, FiLM [5], Cross-Attention [6], and AdaLN-Zero [7]—while holding the modality encoders, discrete latent representations, and generative heads fixed. Evaluating these operators across offline predictive metrics and 100 physical robot trials reveals an architectural trade-off. High-capacity operators (e.g., FiLM, Cross-Attention) excel under offline guidance dropping but frequently encounter planning degradation during physical sensor dropout. We demonstrate that additive translation parameters induce a covariate shift within the pre-quantization latent space under input dropout, compromising the mapping integrity at the discrete Vector-Quantized (VQ) bottleneck. In contrast, restricting fusion to bounded element-wise magnitude scaling structurally regularizes against this shift, preserving topological integrity and yielding smoother physical trajectories under sensor loss. Ultimately, these findings indicate that robust closed-loop execution benefits from topology-preserving modulation, highlighting the representational vulnerabilities introduced by additive translation biases.

II. RELATED WORK

We review related work in robotics world models, fusion and conditioning, and multimodal inputs.

Robotic World Models. Robotic world models learn predictive representations for planning and control. DayDreamer learns latent dynamics directly on physical robots [1], Navi-

¹George Mason University ²Kiel University ³Ewha Womans University

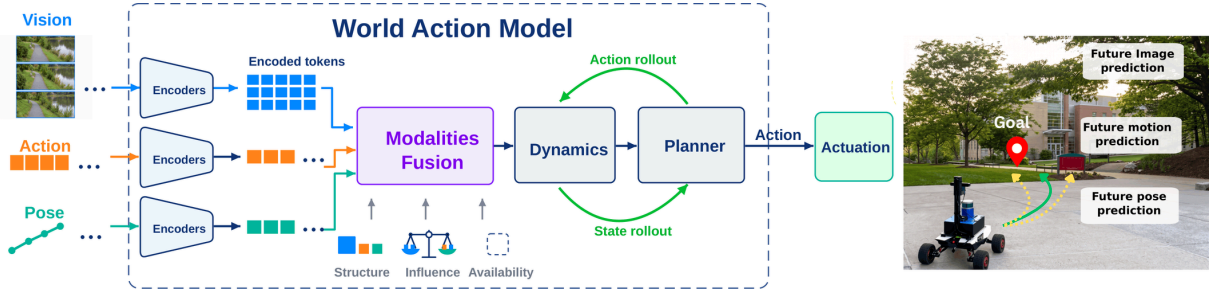


Fig. 1: **Multimodal prediction and planning in Hydra.** Visual observations, actions, and poses are encoded and fused before temporal prediction. The resulting representation supports latent prediction, planning, and continuous action generation.

gation World Models (NWM) predicts future egocentric observations conditioned on candidate actions [2], and DINO-WM predicts pretrained visual features using action and proprioceptive information [3]. Unified World Models (UWM) jointly models video and actions for forward dynamics, inverse dynamics, and policy inference [8].

Several models combine perception and motion before temporal prediction. VertFormer [9] fuses terrain, action, and pose information and supports multiple kinodynamic tasks through modality masking [9]. Hydra extends this paradigm with discrete intent codebooks and continuous Flow Matching heads for navigation planning and execution [4], [10]. We use Hydra as a fixed backbone to isolate the effect of early multimodal interaction.

Fusion and Conditioning Mechanisms. Multimodal information can be combined through feature interaction, conditional modulation, or attention. Conditional normalization modulates a representation using external information [11], while FiLM applies feature-wise affine conditioning without normalization [5]. Simpler variants include direct shifts and scale-only modulation; FiLM has also been used for language conditioning in robotic policies such as RT-1 [12].

Adaptive normalization provides a related conditioning mechanism. Diffusion Transformers compare Cross-Attention with adaptive LayerNorm and introduce adaLN-Zero, in which residual conditioning is initialized to preserve an unconditioned pathway [7]. Layer normalization follows the formulation of Ba et al. [13].

Attention [6] instead allows one representation to selectively retrieve information from another. Perceiver uses latent queries to aggregate high-dimensional inputs [14], Perceiver-Actor applies related ideas to robotic manipulation [15], and CReF combines proprioceptive conditioning, attention, and gating for locomotion [16].

These mechanisms also appear in world models: UWM compares AdaLN and Cross-Attention [8], NWM uses adaptive normalization for action conditioning [2], and Hydra uses attention and adaptive normalization in later temporal and Flow Matching components [4]. Our study instead isolates their use at the *early-fusion* interface before temporal processing.

Modality Dependence and Missing Inputs. Multimodal learning can be affected by unequal dependence on different modalities. Prior work has studied modality generalization [17], proprioception dependence in visuomotor

learning [18], [19], and spatial grounding of robot state in visual representations [20]. These results motivate examining how visual and kinodynamic information interact rather than assuming that all modalities contribute equally.

Related work also considers incomplete inputs. ModDrop randomly omits modalities during training [21], multimodal VAEs support different observed subsets [22], and MultiMAE learns from masked multimodal inputs [23]. Octo supports heterogeneous robot observations with absent channels [24], while MultiBench considers both predictive quality and robustness to missing or corrupted modalities [25]. Hydra and VertFormer similarly use modality masking [4], [9].

Our study adopts a common explicit missing-input representation across all fusion variants so that the comparison primarily reflects the fusion operator rather than differences in modality handling.

III. PRELIMINARIES AND PROBLEM FORMULATION

In this study, we use Hydra [4] as a common backbone and modify only the fusion operator that forms the multimodal history representation. We formulate the problem and introduce the fusion interface and available modalities.

A. Fusion Interface

At each timestep, visual observations are encoded into spatial tokens, while action and pose are mapped to compact kinodynamic embeddings. The fused history is then processed by the same temporal predictor, discrete representations, and continuous Flow Matching (FM) heads across all variants.

At timestep t , let $\mathbf{x}_t$ denote the visual observation, $\mathbf{p}_t$ the robot pose/state, and $\bar{\mathbf{a}}_t$ the action preceding that observation. Following Hydra’s history formulation, $\bar{\mathbf{a}}_t = \mathbf{a}_{t-1}$, the last excuted action. At the fusion boundary,

$$\begin{aligned}\mathbf{H}_{x,t} &= E_x(\mathbf{x}_t) \in \mathbb{R}^{N \times d}, \\ \mathbf{h}_{a,t} &= E_a(\bar{\mathbf{a}}_t) \in \mathbb{R}^d, \\ \mathbf{h}_{p,t} &= E_p(\mathbf{p}_t) \in \mathbb{R}^d,\end{aligned}$$

where N is the number of visual tokens and d is the shared embedding dimension. Thus, vision retains spatial structure through N tokens, while action and pose each contribute one embedding per timestep.

In the original Hydra formulation, broadcast action and pose embeddings are added to the visual tokens. We replace this interaction with each candidate fusion operator while

keeping the remaining architecture and training procedure unchanged.

B. Modality Availability

Vision is assumed available throughout, while action and pose may be missing. Their availability is represented by $r_{a,t}, r_{p,t} \in \{0, 1\}$, where one indicates that the corresponding measurement is present. For $j \in \{a, p\}$,

$$\tilde{\mathbf{h}}_{j,t} = \begin{cases} \mathbf{h}_{j,t}, & r_{j,t} = 1, \\ \mathbf{0}, & r_{j,t} = 0. \end{cases}$$

This separates modality availability from the physical value of a measurement. For example, a valid zero-valued action is encoded normally with $r_{a,t} = 1$, whereas an unavailable action is represented by $r_{a,t} = 0$ and then $\tilde{\mathbf{h}}_{a,t} = \mathbf{0}$.

We use this fixed-zero representation together with explicit availability indicators for every fusion mechanism. Note that these history availability indicators are distinct from the task-specific masking of future action or pose conditions used for inverse and forward kinodynamics [4], [9].

IV. MULTIMODAL FUSION FRAMEWORK

Figure 2 summarizes the five fusion mechanisms. At timestep t , the visual encoder produces $\mathbf{H}_{x,t} \in \mathbb{R}^{N \times d}$, while action and pose are represented by $\tilde{\mathbf{h}}_{a,t}, \tilde{\mathbf{h}}_{p,t} \in \mathbb{R}^d$. Each fusion operator produces N fused tokens, $\mathbf{Z}_t \in \mathbb{R}^{N \times d}$, compatible with the common temporal predictor. For elementwise fusion operators, kinodynamic embeddings are scaled as $\hat{\mathbf{h}}_{j,t} = \tilde{\mathbf{h}}_{j,t} / \sqrt{N}$ and broadcast over visual-token positions. Cross-Attention instead retains action and pose as separate tokens.

Shift directly offsets each visual token using the available kinodynamic embeddings, $\mathbf{u}_t^{(i)} = \mathbf{h}_{x,t}^{(i)} + \hat{\mathbf{h}}_{a,t} + \hat{\mathbf{h}}_{p,t}$.

Scale instead applies multiplicative conditioning. Learned projections $S_j : \mathbb{R}^d \rightarrow \mathbb{R}^d$ define $\gamma_t = \sum_{j \in \{a,p\}} r_{j,t} S_j(\tilde{\mathbf{h}}_{j,t})$, giving

$$\mathbf{u}_t^{(i)} = (\mathbf{1} + \gamma_t) \odot \mathbf{h}_{x,t}^{(i)}. \quad (1)$$

Thus, Scale modifies feature amplitudes without introducing an additive offset.

FiLM [5] combines multiplicative and additive conditioning. For each kinodynamic modality, $G_j(\tilde{\mathbf{h}}_{j,t}) = (\gamma_{j,t}, \beta_{j,t})$, and the available contributions are summed as $\gamma_t = \sum_j r_{j,t} \gamma_{j,t}$ and $\beta_t = \sum_j r_{j,t} \beta_{j,t}$. The fused token is

$$\mathbf{u}_t^{(i)} = (\mathbf{1} + \gamma_t) \odot \mathbf{h}_{x,t}^{(i)} + \beta_t. \quad (2)$$

FiLM therefore generalizes Scale by allowing kinodynamic context to both rescale and shift the visual representation.

Cross-Attention [6] enables token-dependent interaction between vision and kinodynamics. Visual tokens act as queries, while the action and pose tokens form keys and values. With $\mathbf{H}_{k,t} = [\tilde{\mathbf{h}}_{a,t}; \tilde{\mathbf{h}}_{p,t}]$, we define $\mathbf{Q}_t = W_Q(\mathbf{H}_{x,t})$, $\mathbf{K}_t = W_K \mathbf{H}_{k,t}$, and $\mathbf{V}_t = W_V \mathbf{H}_{k,t}$. For each attention head,

$$\mathbf{A}_t = \text{softmax} \left(\frac{\mathbf{Q}_t \mathbf{K}_t^\top}{\sqrt{d_h}} + \mathbf{M}_t \right) \mathbf{V}_t, \quad \mathbf{U}_t = \mathbf{H}_{x,t} + W_O \mathbf{A}_t, \quad (3)$$

where $\mathbf{M}_t$ masks unavailable kinodynamic tokens. If neither action nor pose is available, the attention contribution is set to zero. The implementation uses 12 attention heads, and W_O is zero-initialized so that training begins from the visual residual pathway.

AdaLN+Zero follows adaptive normalization and gated residual conditioning [7]. The conditioning vector $\mathbf{c}_t = [\tilde{\mathbf{h}}_{a,t}; \tilde{\mathbf{h}}_{p,t}; r_{a,t}; r_{p,t}]$ is processed by a two-layer MLP to produce $(\hat{\gamma}_t, \hat{\beta}_t, \hat{\mathbf{g}}_t)$. We suppress this branch when both kinodynamic modalities are absent using $m_t = 1 - (1 - r_{a,t})(1 - r_{p,t})$ and $(\gamma_t, \beta_t, \mathbf{g}_t) = m_t(\hat{\gamma}_t, \hat{\beta}_t, \hat{\mathbf{g}}_t)$. The resulting update is

$$\begin{aligned} \mathbf{v}_t^{(i)} &= (\mathbf{1} + \gamma_t) \odot \text{LN}(\mathbf{h}_{x,t}^{(i)}) + \beta_t, \\ \mathbf{u}_t^{(i)} &= \mathbf{h}_{x,t}^{(i)} + \mathbf{g}_t \odot \mathbf{v}_t^{(i)}. \end{aligned} \quad (4)$$

Here, LN is non-affine LayerNorm [13]. The final conditioning layer is zero-initialized and the gate is unconstrained, allowing signed residual corrections.

V. EXPERIMENTAL SETUP

A. Controlled Comparison

All experiments use the same Hydra backbone, modality encoders, temporal prediction architecture, discrete representations, training objectives, and prediction heads. Only the early-fusion operator is changed. This isolates the contribution of the multimodal interaction from changes elsewhere in the world model.

We evaluate the fusion mechanisms on held-out navigation data from SCAND [26], TartanDrive [27], and SAC-SoN/HuRoN [28]. The individual evaluations use the datasets for which the required modalities and complete temporal windows are available.

We also performed 100 robot trials and compared different sensor dropout combinations. For the experiment setup, the goal is set to 10 meters away, occluded by multiple obstacles. The robot needs to find a path to the goal without colliding with the obstacles (see Fig. 3).

B. Prediction Tasks

We consider three complementary capabilities. In inverse kinodynamics (IKD), future actions are masked while future robot motion is provided, requiring the model to infer the action sequence that corresponds to the specified motion. In forward kinodynamics (FKD), future poses are masked while future actions are provided, requiring prediction of the resulting robot motion. Visual evaluation measures both single-step image prediction and long-horizon autoregressive video generation.

These tasks probe different properties of the same fused representation: whether it preserves commanded motion, resulting robot dynamics, and visual information over both short and extended horizons.

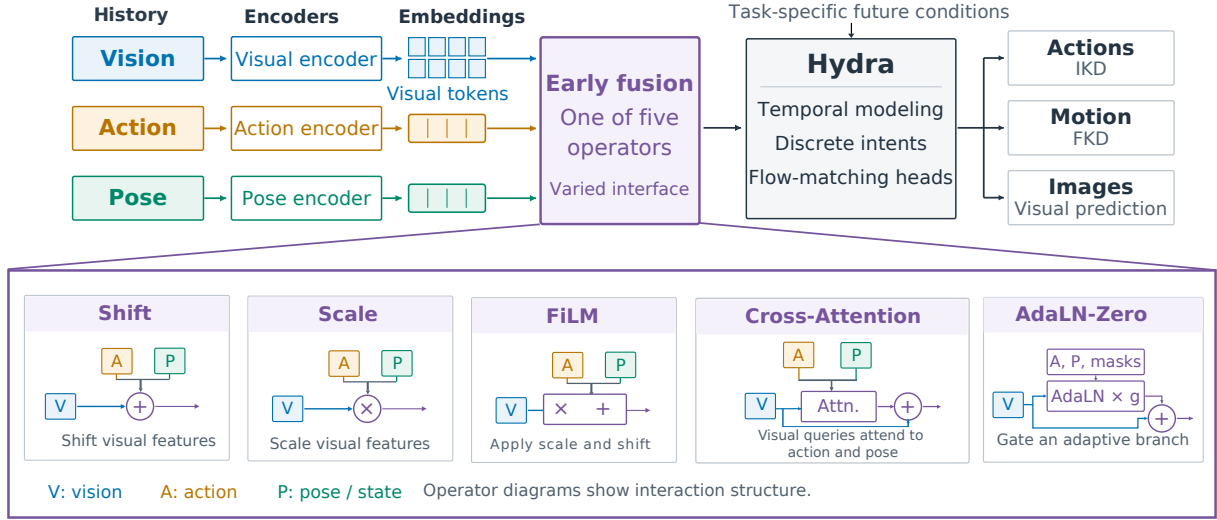


Fig. 2: **Controlled comparison of early-fusion mechanisms.** Shift, Scale, FiLM, Cross-Attention, and AdaLN+Zero operate at the same multimodal interface while the remaining World Action Model is unchanged.

C. Evaluation Protocol

Our evaluation protocol is divided into offline dataset benchmarking and physical robot trials, each utilizing a distinct masking strategy to isolate specific fusion capabilities.

During offline dataset evaluation, we assess task-specific predictive performance by masking guidance during the decoding phase. In this setting, the input modalities remain fully observable. Masking action guidance yields the FKD task, while masking pose guidance forms the IKD task. IKD is evaluated on held-out SCAND and TartanDrive windows. FKD is evaluated on SCAND, TartanDrive, and SACSoN/HuRoN (SACSoN dataset does not provide robot actions), with cross-dataset results weighted by the number of valid temporal windows.

Conversely, during closed-loop robot experiments, we evaluate architectural robustness to sensor loss by dropping modalities at the input level (e.g., masking the kinodynamic history before the early fusion operator). In these physical trials, task guidance is explicitly retained to ensure the policy maintains goal-directed navigation toward the destination.

Single-step visual prediction utilizes the same 19,143 held-out windows for all fusion mechanisms: 5,246 from SCAND, 3,053 from TartanDrive, and 10,844 from SACSoN/HuRoN. Long-horizon visual prediction employs a fixed set of 100 valid clips from SCAND and 100 from TartanDrive. Identical clips are reused across fusion mechanisms, and the reported FVD-R3D value represents the average of the two dataset-level scores.

For the computational comparison, trainable parameters are counted for the complete model. Encoder and fusion latency is measured on TartanDrive with a batch size of 1 on an NVIDIA A100-40GB GPU. We report the mean and standard deviation over 100 timed runs following 20 warm-up iterations. This timing excludes subsequent temporal prediction and Flow Matching rollout, thereby measuring the encoder/fusion path rather than end-to-end inference.

VI. RESULTS

We evaluate the five fusion mechanisms through complementary prediction tasks spanning inverse kinodynamics, forward kinodynamics, and visual forecasting. These tasks probe different properties of the same fused representation: recovering actions from future motion, predicting motion from supplied actions, and preserving visual information over short and long horizons. Rather than seeking a single universally superior fusion operator, we examine how different fusion mechanisms align with different prediction requirements.

A. Action Prediction through Inverse Kinodynamics

We first evaluate whether the fused representation retains sufficient kinodynamic information to recover the actions associated with a specified future robot motion. Future action tokens are masked while the corresponding future poses remain available.

We evaluate both the deterministic auxiliary head and the final Flow Matching prediction. Let $\mathbf{a}_{n,t} = [v_{n,t}, \omega_{n,t}]$ denote the ground-truth linear and angular velocity. For the deterministic prediction $\hat{\mathbf{a}}_{n,t}^{\text{det}}$,

$$\mathcal{E}_{\text{pre}} = \frac{1}{MTd_a} \sum_{n=1}^M \sum_{t=1}^T \left\| \hat{\mathbf{a}}_{n,t}^{\text{det}} - \mathbf{a}_{n,t} \right\|_2^2, \quad (5)$$

where M is the number of evaluation windows, T is the prediction horizon, and $d_a = 2$. The corresponding Flow Matching error is

$$\mathcal{E}_{\text{post}} = \frac{1}{MTd_a} \sum_{n=1}^M \sum_{t=1}^T \left\| \hat{\mathbf{a}}_{n,t}^{\text{FM}} - \mathbf{a}_{n,t} \right\|_2^2. \quad (6)$$

Both metrics use the recorded action representation without additional normalization or channel rescaling, and only complete temporal windows are included.

FiLM produces the lowest error in both stages (Table I), with MSE values of 0.1951 before and 0.1701 after Flow

TABLE I: Inverse-kinodynamic action prediction.

Fusion Method	Pre-FM MSE ↓	Post-FM MSE ↓
Shift	0.3063	1.2487
Scale	0.4489	1.4820
FiLM	0.1951	0.1701
Cross-Attention	0.6078	1.2611
AdaLN+Zero	0.4810	1.3532

Matching. Relative to Shift, these correspond to reductions of 36.3% and 86.4%, respectively.

The result suggests that feature-wise affine conditioning is particularly effective when multimodal information must support the recovery of a compact control signal. FiLM applies action- and pose-dependent scale and shift parameters consistently across the visual tokens, providing a direct global conditioning pathway between motion context and action prediction.

The pre- and post-FM results also show that the downstream generative head does not remove differences introduced by the fusion operator. For Shift, Scale, Cross-Attention, and AdaLN+Zero, error increases after Flow Matching, whereas FiLM improves slightly. This suggests that the quality of the fused representation influences how effectively the shared continuous decoder can recover actions. Whether this reflects differences in latent smoothness or conditioning geometry requires additional analysis.

B. Future Motion Prediction

We next evaluate the complementary FKD problem: given the observed history and a future action sequence, can the world model predict the resulting robot motion? Future pose tokens are masked while future actions remain available. Prediction is performed over $T = 12$ steps, corresponding to a 3 s horizon at 4 Hz.

Translational motion is represented as step-to-step increments in a common ego-aligned frame. Predicted and ground-truth trajectories are reconstructed by cumulative integration,

$$\hat{\mathbf{p}}_{n,t} = \sum_{\tau=1}^t \Delta \hat{\mathbf{p}}_{n,\tau}, \quad \mathbf{p}_{n,t} = \sum_{\tau=1}^t \Delta \mathbf{p}_{n,\tau}, \quad (7)$$

where $\mathbf{p}_{n,t} \in \mathbb{R}^2$ denotes planar displacement.

We report Average Displacement Error (ADE), Final Displacement Error (FDE), and translational Relative Pose Error (RPE). Orientation is evaluated separately by integrating predicted yaw increments,

$$\hat{\psi}_{n,t} = \text{wrap} \left(\sum_{\tau=1}^t \Delta \hat{\psi}_{n,\tau} \right), \quad \psi_{n,t} = \text{wrap} \left(\sum_{\tau=1}^t \Delta \psi_{n,\tau} \right). \quad (8)$$

where $\text{wrap}(\cdot)$ maps angles to $[-\pi, \pi)$. The mean accumulated yaw error is

$$E_{\text{yaw}} = \frac{1}{MT} \sum_{n=1}^M \sum_{t=1}^T |\text{wrap}(\hat{\psi}_{n,t} - \psi_{n,t})|. \quad (9)$$

The ordering changes for forward kinodynamics. Cross-Attention achieves the lowest ADE (0.6224 m), FDE

TABLE II: Forward-kinodynamic motion prediction.

Fusion Method	ADE (m) ↓	FDE (m) ↓	RPE (m) ↓	Yaw Error (rad) ↓
Shift	1.2073	2.3563	0.2143	0.2640
Scale	0.9317	1.7881	0.1758	0.2611
FiLM	0.8268	1.6172	0.1632	0.2714
Cross-Attention	0.6224	1.2710	0.1373	0.2845
AdaLN+Zero	0.9573	1.8465	0.1801	0.2662

TABLE III: Single-step visual prediction.

Fusion Method	DreamSim ↓	LPIPS ↓	PSNR (dB) ↑
Shift	0.2528	0.3950	14.751
Scale	0.2088	0.3420	15.450
FiLM	0.2089	0.3424	16.122
Cross-Attention	0.2330	0.3747	15.077
AdaLN+Zero	0.2109	0.3441	15.317

(1.2710 m), and RPE (0.1373 m), corresponding to improvements of 48.4%, 46.1%, and 35.9% over Shift. FiLM ranks second on all three translational metrics.

This contrast with IKD highlights a difference between token-shared modulation and token-dependent interaction. FiLM applies the same kinodynamic modulation to every visual token, whereas Cross-Attention allows individual visual tokens to query the action and pose context. The FKD results are consistent with such selective interaction being particularly useful when predicting how commanded motion evolves through spatial scene structure.

The same preference does not hold for orientation. Scale achieves the lowest accumulated yaw error (0.2611 rad), whereas Cross-Attention has the highest (0.2845 rad). Thus, even within FKD, translational and rotational prediction can favor different fusion structures.

C. Visual Prediction Quality

Kinodynamic accuracy alone does not establish whether a fusion mechanism preserves the visual information required by a world model. We therefore evaluate both immediate image prediction and long-horizon autoregressive visual prediction.

1) Single-Step Image Prediction

For single-step evaluation, the model receives visual, action, and pose histories together with the future action sequence, while future pose tokens are masked. Only the first predicted future observation, $\hat{I}_{t+1}$, is decoded for evaluation.

We report DreamSim [29], LPIPS [30], and PSNR. DreamSim and LPIPS are perceptual distances for which lower values are better, whereas higher PSNR indicates better pixel-level fidelity.

Scale and FiLM provide the strongest immediate visual predictions (Table III). Their perceptual scores are nearly identical: Scale obtains DreamSim of 0.2088 and LPIPS of 0.3420, compared with 0.2089 and 0.3424 for FiLM. FiLM achieves the highest PSNR at 16.122 dB, while AdaLN+Zero remains close under the perceptual metrics.

These results suggest that simple feature-wise modulation is sufficient for accurate one-step visual reconstruction. Cross-Attention, despite its advantage in translational FKD, is weaker under the immediate perceptual metrics, further

TABLE IV: **Long-horizon autoregressive visual prediction.**

Fusion Method	FVD-R3D ↓
Shift	50.3751
Scale	66.6345
FiLM	38.5068
Cross-Attention	39.0519
AdaLN+Zero	67.2436

indicating that motion prediction and visual reconstruction favor different properties of the fused representation.

2) Video Prediction

We next evaluate visual stability during autoregressive prediction. Each model generates a 16s rollout at 4Hz, corresponding to 64 future frames. Predicted visual latents are fed back into the observation history, while ground-truth future actions and poses are retained to isolate visual drift from kinodynamic prediction error.

Video quality is measured using a Fréchet distance in a pretrained R3D-18 [31] feature space. Inputs are normalized using $\mu = (0.43216, 0.394666, 0.37645)$, and $\sigma = (0.22803, 0.22145, 0.216989)$. Let (μ_r, Σ_r) and (μ_g, Σ_g) denote the empirical feature statistics of reference and generated videos. We compute

$$D_{FV} = \|\mu_r - \mu_g\|_2^2 + \text{Tr}(\Sigma_r + \Sigma_g - 2(\Sigma_r \Sigma_g)^{1/2}).$$

We refer to this metric as FVD-R3D to distinguish the R3D-18 feature representation from the conventional I3D-based FVD formulation [32]. Table IV reports the macro-average over fixed SCAND and TartanDrive evaluation subsets.

The long-horizon ordering differs markedly from the single-step results. FiLM obtains the lowest FVD-R3D (38.5068), closely followed by Cross-Attention (39.0519). Both substantially outperform Shift, Scale, and AdaLN+Zero.

Scale provides the clearest contrast: it performs strongly under single-step perceptual metrics but degrades considerably during autoregressive rollout. Cross-Attention exhibits the opposite tendency, with weaker immediate visual fidelity but substantially better long-horizon consistency. Single-step reconstruction quality is therefore not sufficient to characterize the suitability of a fused representation for long-horizon world modeling.

D. Robot Evaluation under Missing Modalities

While our offline dataset evaluations measure task-specific predictive capacity by masking task guidance during decoding, our closed-loop robot trials assess structural resilience by dropping modalities directly at the model input. To investigate how these distinct occlusion stages affect physical deployment, each of the five fusion variants—Shift, Scale, FiLM, Cross-Attention, and AdaLN-Zero—is deployed using an identical navigation pipeline. We evaluate closed-loop performance under four input-availability conditions: (i) all modalities available, (ii) action unavailable, (iii) pose unavailable, and (iv) both action and pose unavailable. Vision remains available across all conditions.

For a missing modality, its history input is removed according to the same availability convention used during training, while all other model and navigation components are kept unchanged. This produces a controlled 5×4 evaluation in which only the fusion mechanism and the available kinodynamic information vary.

Navigation performance is measured using success rate and traversal time. Success rate measures whether the robot completes the navigation task under the specified modality condition, while traversal time measures the time required to reach the goal for successful trials. Each configuration is repeated over multiple physical trials, and traversal time is reported as mean $\pm$ standard deviation.

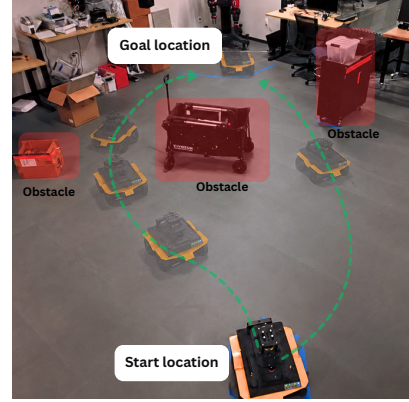


Fig. 3: **Physical robot evaluation under missing modalities.** The five fusion mechanisms are evaluated in closed-loop navigation under four modality-availability conditions: all modalities available, action missing, pose missing, and both action and pose missing.

VII. DISCUSSION

A. Guidance Dropout with Offline Evaluations

The offline dataset evaluations isolate the model’s conditional predictive capacity by masking task guidance signals (actions or poses) during the decoding phase. Because the multimodal inputs remain fully observable during early fusion in these offline tasks, the performance of the fusion operator reflects how effectively it structures the continuous latent space for downstream conditional querying by the guidance signal.

The combined results indicate that early fusion operators condition the latent geometry in fundamentally different ways, exposing distinct relationships between the fused representations and the late guidance signals.

For action prediction (IKD), where future pose guidance is utilized to decode actions, FiLM yields the strongest results. Actions in navigation, such as linear and angular velocities, represent global, low-dimensional control states. FiLM’s token-shared affine modulation (Eq. 2) applies a spatially uniform, continuous transformation across the visual manifold. This global affine shift appears highly suitable for aligning the latent representation with low-dimensional action distributions, making it readily accessible when queried by the downstream pose guidance. Furthermore, the substantial

TABLE V: **Robot navigation performance under different modality-availability conditions.** Success rate (SR) and traversal time are reported for each fusion mechanism with all modalities available, action history missing, pose history missing, and both kinodynamic modalities missing. Traversal time is computed over successful trials only.

Fusion Method	All Modalities		Action Missing		Pose Missing		Action + Pose Missing	
	SR (%) $\uparrow$	Time (s) $\downarrow$	SR (%) $\uparrow$	Time (s) $\downarrow$	SR (%) $\uparrow$	Time (s) $\downarrow$	SR (%) $\uparrow$	Time (s) $\downarrow$
Shift	20	33.03	20	37.33	40	33.25 ± 1.95	20	34.30
Scale	60	32.61 ± 1.81	80	33.84 ± 2.28	60	34.76 ± 0.28	60	34.88 ± 1.22
FiLM	40	34.33 ± 2.87	20	37.01	0	–	0	–
Cross-Attention	20	33.67	20	36.50	0	–	0	–
AdaLN-Zero	60	33.67 ± 4.45	40	32.73 ± 3.37	60	35.29 ± 1.69	20	37.70

differences between pre- and post-FM errors demonstrate that the generative head relies heavily on this globally conditioned continuous representation.

Conversely, FKD favors a spatially selective interaction. In this task, action guidance is used to predict future physical trajectories (poses). Because future trajectories are physically grounded in specific regions of the visual observation (e.g., navigating a right-hand turn requires querying the right-side environmental geometry), Cross-Attention excels. By computing spatial routing based on query-key similarities (Eq. 3), Cross-Attention preserves spatially localized feature segregation. This allows the downstream action guidance to selectively attend to the specific regions of the latent space necessary for accurate trajectory forecasting. This structure proves particularly effective for ADE, FDE, and RPE metrics.

A similar task-dependent distinction emerges in visual prediction. Feature-wise magnitude modulation (Scale) and affine translations (FiLM) are highly effective for immediate single-step reconstruction, likely because they preserve global image fidelity without aggressively redistributing spatial features. However, for long-horizon visual rollouts conditioned on action guidance, the spatial routing of Cross-Attention and the expressive capacity of FiLM provide superior temporal stability.

Taken together, these offline evaluations suggest that early fusion mechanisms do not strictly form a hierarchy of performance; rather, they condition the latent space to favor specific types of downstream guidance. FiLM facilitates global state mapping (ideal for IKD and short-term fidelity), whereas Cross-Attention preserves the spatial heterogeneity required for targeted trajectory querying (ideal for FKD). This indicates that the choice of multimodal fusion operator is best guided by the specific conditional querying requirements of the downstream world-model application.

B. Latent Covariate Shift in Physical Robot Deployment

Our physical robot evaluations reveal an architectural trade-off in multimodal world models: high-capacity fusion operators that excel under offline guidance masking (e.g., FiLM, Cross-Attention) frequently encounter planning degradation when subjected to physical sensor dropout. Conversely, structurally constrained mechanisms, such as Scale, demonstrate substantial real-world resilience. We attribute this performance divergence to an induced covariate shift within the pre-quantization continuous latent space, which

compromises the mapping integrity at the discrete VQ bottleneck.

Discrete Latent Imposed Limitation. Because Hydra compresses the fused representations into a shared kinodynamic codebook learned across heterogeneous datasets (e.g., low-speed SCAND and high-velocity TartanDrive), the flow-matching execution policy relies entirely on the topological stability of the pre-quantization latent z_e .

When kinodynamic modalities drop out, high-capacity operators distort this topology. For instance, FiLM (Eq. 2) relies on unregularized affine translations β_i . When the modulating modality is missing, the collapse of these affine parameters shifts z_e across the discrete latent boundaries separating distinct physical domains. Similarly, Cross-Attention (Eq. 3) dynamically re-routes the latent manifold, causing a representation shift. On a physical robot, this geometric distortion forces the VQ bottleneck to sample out-of-distribution discrete intents, minimizing the distance $\|z_e - e_k\|_2^2$ to incorrect codebook vectors e_k . This manifests as the observed “jerky” control behavior during physical experiments.

The Topological Invariance of Scale. In contrast, we observe that the Scale mechanism (Eq. 1) regularizes against this domain shift. By restricting multimodal interaction to element-wise magnitude modulation γ_i without the additive translation β_i seen in FiLM or AdaLN-Zero (Eq. 4), Scale preserves the boundaries of the latent embeddings.

Consequently, z_e remains securely within the correct codebook vector even under total kinodynamic dropout. Our robot trials validate this structural advantage: Scale consistently produces the smoothest navigation trajectories during sensor dropout, maintaining physical execution characteristics nearly equivalent to the full-modality baseline. This suggests that real-world robot planning benefits from topology-preserving modulation, and that unconstrained additive biases degrade planning reliability under sensor loss.

Limitations. Although the robot evaluation includes 100 physical trials, the results should be interpreted as empirical observations rather than statistically conclusive evidence across deployment conditions. The present study also focuses on navigation-oriented tasks within a single discrete-latent world action model architecture, Hydra. Future work will examine whether the observed fusion trends extend to other task domains, world-model formulations, and broader hardware settings to further establish the generality of these findings.

VIII. CONCLUSION

We presented a controlled study of early multimodal fusion in a navigation World Action Model, comparing Shift, Scale, FiLM, Cross-Attention, and AdaLN+Zero while holding the surrounding Hydra architecture fixed. The results show that different fusion operators favor different aspects of prediction: Cross-Attention is strongest for translational forward dynamics, whereas FiLM provides the strongest inverse-kinodynamic and long-horizon visual prediction and remains competitive across the remaining metrics. Taken together, these results identify FiLM as the most promising overall strategy in the evaluated setting and demonstrate that multimodal fusion should be assessed jointly through perceptual and kinodynamic objectives. Future work can extend this analysis to explicit missing-modality robustness, additional navigation domains, and the effect of fusion choice on closed-loop planning performance.

REFERENCES

- [1] P. Wu, A. Escontrela, D. Hafner, P. Abbeel, and K. Goldberg, “Day-Dreamer: World models for physical robot learning,” in *Proceedings of the 6th Conference on Robot Learning*, ser. Proceedings of Machine Learning Research, vol. 205. PMLR, 2023, pp. 2226–2240.
- [2] A. Bar, G. Zhou, D. Tran, T. Darrell, and Y. LeCun, “Navigation world models,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2025, pp. 15 791–15 801.
- [3] G. Zhou, H. Pan, Y. LeCun, and L. Pinto, “DINO-WM: World models on pre-trained visual features enable zero-shot planning,” in *Proceedings of the 42nd International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, vol. 267. PMLR, 2025, pp. 79 115–79 135.
- [4] M. Nazeri, A. Card, S. Huber, A. Pokhrel, Y. Wang, R. Hammele, D. Song, S. Pirk, and X. Xiao, “Hydra: A navigation world action model with discrete latent planning and continuous flow-matching execution,” 2026.
- [5] E. Perez, F. Strub, H. de Vries, V. Dumoulin, and A. C. Courville, “FiLM: Visual reasoning with a general conditioning layer,” in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 32, no. 1, 2018, pp. 3942–3951.
- [6] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, “Attention is all you need,” in *Advances in Neural Information Processing Systems*, vol. 30, 2017.
- [7] W. Peebles and S. Xie, “Scalable diffusion models with transformers,” in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 4195–4205.
- [8] C. Zhu, R. Yu, S. Feng, B. Burchfiel, P. Shah, and A. Gupta, “Unified world models: Coupling video and action diffusion for pretraining on large robotic datasets,” in *Proceedings of Robotics: Science and Systems*, Los Angeles, CA, USA, June 2025.
- [9] M. Nazeri, A. Pokhrel, A. Card, A. Datar, G. Warnell, and X. Xiao, “VertiFormer: A data-efficient multi-task transformer for off-road robot mobility,” 2025.
- [10] Y. Lipman, R. T. Q. Chen, H. Ben-Hamu, M. Nickel, and M. Le, “Flow matching for generative modeling,” in *International Conference on Learning Representations*, 2023.
- [11] H. de Vries, F. Strub, J. Mary, H. Larochelle, O. Pietquin, and A. C. Courville, “Modulating early visual processing by language,” in *Advances in Neural Information Processing Systems*, vol. 30, 2017, pp. 6594–6604.
- [12] A. Brohan, N. Brown, J. Carbajal, Y. Chebotar, J. Dabis, C. Finn, K. Gopalakrishnan, K. Hausman, A. Herzog, J. Hsu, J. Ibarz, B. Ichter, A. Irpan, T. Jackson, S. Jesmonth, N. J. Joshi, R. Julian, D. Kalashnikov, Y. Kuang, I. Leal, K.-H. Lee, S. Levine, Y. Lu, U. Malla, D. Manjunath, I. Mordatch, O. Nachum, C. Parada, J. Peralta, E. Perez, K. Pertsch, J. Quiambao, K. Rao, M. S. Ryoo, G. Salazar, P. R. Sanketi, K. Sayed, J. Singh, S. Sontakke, A. Stone, C. Tan, H. Tran, V. Vanhoucke, S. Vega, Q. H. Vuong, F. Xia, T. Xiao, P. Xu, S. Xu, T. Yu, and B. Zitkovich, “RT-1: Robotics transformer for real-world control at scale,” in *Proceedings of Robotics: Science and Systems*, Daegu, Republic of Korea, July 2023.
- [13] J. L. Ba, J. R. Kiros, and G. E. Hinton, “Layer normalization,” *arXiv preprint arXiv:1607.06450*, 2016.
- [14] A. Jaegle, F. Gimeno, A. Brock, O. Vinyals, A. Zisserman, and J. Carreira, “Perceiver: General perception with iterative attention,” in *Proceedings of the 38th International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, vol. 139. PMLR, 2021, pp. 4651–4664.
- [15] M. Shridhar, L. Manuelli, and D. Fox, “Perceiver-actor: A multi-task transformer for robotic manipulation,” in *Proceedings of the 6th Conference on Robot Learning*, ser. Proceedings of Machine Learning Research, vol. 205. PMLR, 2023, pp. 785–799.
- [16] Y. Hao, R. Yu, S. Luo, G. Zhang, J. Wu, and Q. Zhu, “CReF: Cross-modal and recurrent fusion for depth-conditioned humanoid locomotion,” 2026.
- [17] W. Wang, D. Tran, and M. Feiszli, “What makes training multimodal classification networks hard?” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 12 695–12 705.
- [18] J. Zhao, W. Lu, D. Zhang, Y. Liu, Y. Liang, T. Zhang, Y. Cao, J. Xie, Y. Hu, S. Wang, J. Guo, D. Wang, and Y. Gao, “Do you need proprioceptive states in visuomotor policies?” 2025.
- [19] J. Lu, W. Xia, Y. Wu, Z. Lu, and D. Hu, “When would vision-proprioception policies fail in robotic manipulation?” in *International Conference on Learning Representations*, 2026.
- [20] G. Zhao, Q. Qian, G. Zhang, W. Li, J. Wang, X. Lu, D. Zhao, and R. Xu, “GeoProp: Grounding robot state in vision for generalist manipulation,” 2026.
- [21] N. Neverova, C. Wolf, G. W. Taylor, and F. Nebout, “ModDrop: Adaptive multi-modal gesture recognition,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 38, no. 8, pp. 1692–1706, 2016.
- [22] M. Wu and N. Goodman, “Multimodal generative models for scalable weakly-supervised learning,” in *Advances in Neural Information Processing Systems*, vol. 31, 2018.
- [23] R. Bachmann, D. Mizrahi, A. Atanov, and A. Zamir, “MultiMAE: Multi-modal multi-task masked autoencoders,” in *European Conference on Computer Vision*, 2022, pp. 348–367.
- [24] Octo Model Team, D. Ghosh, H. Walke, K. Pertsch, K. Black, O. Mees, S. Dasari, J. Hejna, T. Kreiman, C. Xu, J. Luo, Y. L. Tan, L. Y. Chen, P. Sanketi, Q. Vuong, T. Xiao, D. Sadigh, C. Finn, and S. Levine, “Octo: An open-source generalist robot policy,” in *Proceedings of Robotics: Science and Systems*, Delft, Netherlands, 2024.
- [25] P. P. Liang, Y. Lyu, X. Fan, Z. Wu, Y. Cheng, J. Wu, L. Y. Chen, P. Wu, M. A. Lee, Y. Zhu, R. Salakhutdinov, and L.-P. Morency, “MultiBench: Multiscale benchmarks for multimodal representation learning,” in *Advances in Neural Information Processing Systems Datasets and Benchmarks Track*, vol. 34, 2021.
- [26] H. Karnan, A. Nair, X. Xiao, G. Warnell, S. Pirk, A. Toshev, J. W. Hart, J. Biswas, and P. Stone, “Socially compliant navigation dataset (SCAND): A large-scale dataset of demonstrations for social navigation,” *IEEE Robotics and Automation Letters*, vol. 7, no. 4, pp. 11 807–11 814, 2022.
- [27] S. Triest, M. Sivaprakasam, S. J. Wang, W. Wang, A. M. Johnson, and S. Scherer, “TartanDrive: A large-scale dataset for learning off-road dynamics models,” in *2022 International Conference on Robotics and Automation*, 2022, pp. 2546–2552.
- [28] N. Hirose, D. Shah, A. Sridhar, and S. Levine, “SACSoN: Scalable autonomous control for social navigation,” *IEEE Robotics and Automation Letters*, vol. 9, no. 1, pp. 49–56, 2024.
- [29] S. Fu, N. Tamir, S. Sundaram, L. Chai, R. Zhang, T. Dekel, and P. Isola, “DreamSim: Learning new dimensions of human visual similarity using synthetic data,” in *Advances in Neural Information Processing Systems*, vol. 36, 2023, pp. 50 742–50 768.
- [30] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang, “The unreasonable effectiveness of deep features as a perceptual metric,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 586–595.
- [31] D. Tran, H. Wang, L. Torresani, J. Ray, Y. LeCun, and M. Paluri, “A closer look at spatiotemporal convolutions for action recognition,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018.

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