DEPHY: DECOUPLED PHYSICS-AWARE PREDICTIVE MODELS FOR STATE ESTIMATION, FORECASTING AND DATA ASSIMILATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Physics-aware state estimation from noisy observations requires both robust measurement representations and dynamics that support physically plausible reconstruction. Existing physics-aware methods typically rely on reconstruction-driven representation learning, with physical consistency introduced through additional regularization or post-hoc correction. As a result, measurement-related nuisance variation can remain entangled with the states used for physical propagation. We introduce **DePhy**, a decoupled latent state-space framework that learns physical state propagation from measurement latent without reconstruction supervision. DePhy uses mask-based self-supervised learning to obtain representations that are robust to corrupted or incomplete measurements, while physics-mediated future latent prediction preserves information needed for physical state evolution. The resulting representations are mapped to a physical state space governed by an augmented physics-based model that combines analytical mechanics with bounded learned residuals. Amortized inference of trajectory-specific physical factors further enables instance-specific adaptation without per-instance optimization. To reconcile physics-based predictions with new observations, a Kalman-inspired neural assimilation module fuses the propagated state prior with current measurement evidence, yielding measurement-consistent states for reconstruction. By decoupling predictive representation and dynamics learning from reconstruction, DePhy learns states that are both robust to noisy measurements and suitable for stable physical propagation. Experiments on monocular human motion benchmarks demonstrate improved state estimation and long-horizon forecasting, with DePhy reducing Human3.6M MPJPE from 54.8 to 47.3 mm and acceleration error from 8.4 to 3.9 mm/s^2 relative to OSDCap, while improving 1-s forecasting MPJPE from 103.9 to 90.8 mm relative to PhysMoP.

1 INTRODUCTION

Estimating the physical state of a dynamical system from monocular observations is a fundamental problem in computer vision and robotics (Agarwal et al., 2013; Talbot et al., 2025; Hayakawa & Dariush, 2019; Šajina & Ivašić-Kos, 2022; Qin et al., 2018). The standard methodology used for the above heavily relies on first inferring 3D pose and motion using modern monocular pose and mesh estimators (Sun et al., 2023), which are then used as measurements for estimating state factors like contact states, joint torques, and external forces. However, because of environmental factors like depth ambiguity, self-occlusion, and limited observability, pose and motion predictions are often noisy and temporally inconsistent, leading to incorrect state estimation. Physics-aware methods address this issue through losses, simulation, or sequential correction (Shimada et al., 2020; Yuan et al., 2021; Zhang et al., 2024b; Le et al., 2024). However, learning a state representation that remains reliable under severe measurement corruption and useful for physical prediction remains challenging. Specifically, these methods rely on raw measurement reconstruction objectives to enforce physics plausibility. While this approach works for simple scenarios, in real-world settings, these methods often fail due to extreme measurement corruptions.

We introduce DePhy, a decoupled state-space framework that separates the learning of a predictive physical prior from the incorporation of measurement evidence. Noisy observations are first mapped

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.

to a predictive representation and then to an explicit physical state, which is propagated through an instance-adaptive hybrid dynamics model. The resulting prior is subsequently combined with incoming measurements through latent data assimilation. The predictive representation and physical transition are learned before reconstruction and remain fixed while the assimilation mechanism is trained, preventing measurement reconstruction from reshaping the learned physical prior.

Learning such a prior requires a representation that remains stable under measurement corruption while preserving information useful for predicting the system's future evolution. Joint-Embedding Predictive Architectures (JEPAs) provide a natural mechanism for learning these properties through prediction in latent space (LeCun, 2022; Assran et al., 2023; Bardes et al., 2024). We combine masked latent prediction, which promotes robustness to measurement corruption, with future latent prediction, which encourages the representation to retain information relevant to physical propagation. The resulting representation is mapped to an explicit physical state where known mechanical structure can guide its evolution.

Analytical mechanics provides valuable structure for physical propagation, but relevant quantities may be unobserved and some effects omitted or only approximately modeled. DePhy therefore augments the analytical dynamics rather than replacing them. Amortized inference estimates trajectory-dependent physical quantities with known roles in the equations, such as damping, restitution, body properties, contact forces, and joint torques (Auzina et al., 2023), while a bounded and regularized residual captures remaining model discrepancy (Imbiriba et al., 2022; 2024; Yin et al., 2021). This yields an instance-adaptive transition that preserves mechanistic structure while accounting for trajectory-specific variability and unmodeled effects.

Model mismatch and accumulated prediction error can still make the propagated state inconsistent with new observations. Data assimilation addresses this by updating a model-based prior with incoming measurement evidence (Särkkä, 2013; Revach et al., 2022; Buchnik et al., 2024). In DePhy, a learned latent gain combines the dynamics-informed propagated prior with the current measurement representation, allowing measurement-based correction while leaving the predictive representation and physical transition unchanged.

Our main contributions are:

- A decoupled formulation for physical state estimation in which a predictive physical prior is learned from noisy observations without reconstruction supervision and later combined with measurement evidence through data assimilation.
- A predictive representation learned through complementary masked and future latent-prediction objectives, promoting robustness to measurement corruption while retaining information needed for physical propagation.
- An instance-adaptive augmented physics transition that combines analytical dynamics, amortized inference of unobserved physical quantities, and a bounded residual for remaining model discrepancy, together with a learned latent assimilation mechanism for incorporating incoming measurement evidence.
- Experiments on bouncing-ball and damped-pendulum dynamics and monocular human motion demonstrate improved state estimation, physical factor recovery, and medium- to long-horizon forecasting. On Human3.6M, DePhy reduces MPJPE from 54.8 to 47.3 mm and acceleration error from 8.4 to 3.9 mm/s^2 relative to OSDCap (Le et al., 2024).

2 Related Work

Physics-aware human state estimation from video. Physics-aware human-motion methods use mechanical structure to improve the temporal consistency and physical plausibility of monocular kinematic estimates. Simulation-based approaches enforce physical constraints through control or simulation (Shimada et al., 2020; Yuan et al., 2021), while differentiable-physics and temporal methods incorporate dynamics, physical supervision, or structured motion priors (Koleini et al., 2025; Gärtner et al., 2022a; Shin et al., 2024; Zhang et al., 2024a). PhysMoP combines physics-based and data-driven predictors (Zhang et al., 2024b), and OSDCap sequentially refines monocular estimates using simulated dynamics and Kalman-style state estimation (Le et al., 2024). QuaMo extends this online setting with quaternion joint states integrated on the unit sphere and a meta-PD controller that

tracks upstream pose estimates, yielding a kinematic transition without explicit forces or contacts (Le et al., 2026). DePhy instead learns the representation and instance-adaptive physical transition as a predictive prior without reconstruction supervision, and only then incorporates measurement evidence through latent data assimilation.

JEPA and predictive latent representations. Joint-Embedding Predictive Architectures (JEPAs) learn representations by predicting target embeddings from incomplete or transformed context (LeCun, 2022; Assran et al., 2023; Bardes et al., 2024; Assran et al., 2025) and have been extended to temporal and structured data (Abdelfattah & Alahi, 2024; Yee & Koh, 2026; Nam et al., 2026; Lanji et al., 2026). For noisy physical state estimation, the representation should remain robust to measurement corruption while retaining information useful for dynamical prediction. DePhy addresses these two requirements through masked latent prediction and physics-mediated future latent prediction, respectively.

Hybrid physics and instance-adaptive dynamics. Hybrid dynamical models retain known analytical structure while learning missing or uncertain components (Imbiriba et al., 2022; 2024; Yin et al., 2021), while MoNODE uses amortized inference of latent factors to capture trajectory-level variation (Auzina et al., 2023). In our setting, some unobserved quantities have a known role in the governing equations, whereas other effects are not represented by the analytical model. DePhy therefore infers trajectory-specific physical quantities to condition the analytical transition and uses a bounded residual for the remaining model discrepancy.

Neural filtering and latent data assimilation. Learned filtering methods combine model-based prediction with measurement-dependent correction, with KalmanNet-style approaches learning components of the Kalman update (Revach et al., 2022) and latent filtering methods operating directly in representation space (Guen & Thome, 2020; Buchnik et al., 2024). DePhy learns the physics-structured predictive prior first and freezes it during assimilation training, allowing a learned latent gain to incorporate new measurement evidence without reshaping the representation or physical transition.

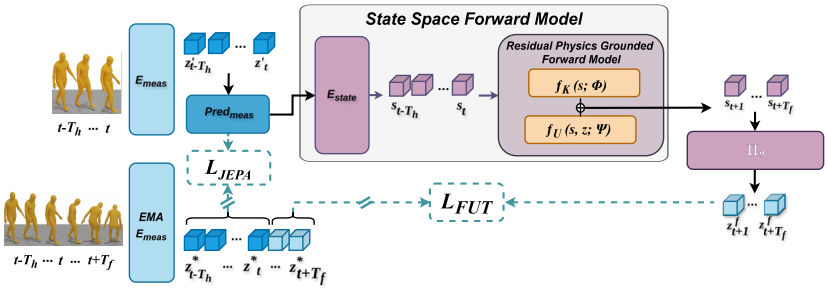


Figure 1: **State-space forward dynamics in DePhy.** The measurement encoder E_{meas} and measurement predictor $\text{Pred}_{\text{meas}}$ produce predictive latents z^m from noisy measurements. The state encoder E_{state} maps z^m to a physical state, which is propagated by known dynamics f_K plus a gated residual f_U. The latent projection head Π_z maps the propagated state back to latent space for future prediction.

3 Method

3.1 Problem Setup and Overview

We observe structured, noisy measurements $\tilde{y}_{1:T}$, $\tilde{y}_t \in \mathcal{Y} \subseteq \mathbb{R}^{d_y}$, and infer a physical state trajectory $s_{1:T}$, $s_t \in \mathcal{S} \subseteq \mathbb{R}^{d_s}$. For monocular human motion, $\tilde{y}_t$ contains 3D joints, pose parameters, and generalized kinematics from a fixed upstream estimator (Kocabas et al., 2020; Li et al., 2021; 2022b; Shin et al., 2024), with errors due to jitter, occlusion, global-position uncertainty, or bias. For synthetic systems, $\tilde{y}_t$ is the simulated observation after measurement noise.

The framework separates three roles (Figs. 1–2). The measurement encoder E_{meas} and predictor $\text{Pred}_{\text{meas}}$ produce predictive measurement latents $z_t^m \in \mathcal{Z}$, which E_{state} maps to the physical state s_t. An augmented physics model propagates s_t using analytical dynamics f_K and a bounded

residual f_U, while Π_z projects the propagated state to the latent prior z_t^p. Finally, data assimilation, Figure 2, combines z_t^p and z_t^m, and D_{meas} reconstructs the measurement from the assimilated latent.

Training mirrors this separation. Stage 1 learns the measurement representation and transition prior using the complementary latent objectives in Secs. 3.2 and 3.4. Stage 2 freezes this prior and trains only assimilation and reconstruction, so reconstruction gradients never update the learned representation or dynamics.

3.2 Predictive Measurement Latents

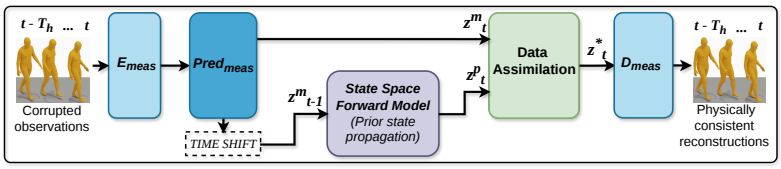


Figure 2: **Latent data assimilation.** The assimilation network A_η fuses the propagated prior z_t^p with measurement evidence z_t^m using gain α_t; D_{meas} reconstructs the measurement from the assimilated latent z_t^*.

Given a history of T_h measurements $\tilde{y}_{t-T_h+1:t}$, we form an additionally corrupted student view $\bar{y} = \mathcal{C}(\tilde{y}; m, \epsilon)$ using masking m and noise ϵ. The student predicts measurement latents from the corrupted history, while the EMA encoder provides targets from the unaugmented history:

$$z^m_{t-T_h+1:t} = \text{Pred}_{\text{meas}}(E_{\text{meas}}(\bar{y}_{t-T_h+1:t}, m_{t-T_h+1:t})), \quad z^\star_{t-T_h+1:t} = \text{sg}\big[E^{\text{EMA}}_{\text{meas}}(\tilde{y}_{t-T_h+1:t}, \mathbf{1})\big]. \tag{1}$$

Here $\text{sg}[\cdot]$ denotes stop-gradient. Both pathways originate from the same upstream measurements, and only the student receives additional corruption, so $z^\star$ does not provide clean-state supervision. We optimize

$$\mathcal{L}_{\text{JEPA}} = \frac{1}{T_h} \sum_{i=t-T_h+1}^{t} \|z_i^m - z_i^\star\|_2^2. \tag{2}$$

This objective promotes robustness to measurement nuisance without requiring clean supervision.

3.3 Physical State and Amortized Physical Inference

The predictive measurement latent is mapped to an explicit physical state, $s_t = E_{\text{state}}(z_t^m)$. For human motion, $s_t = [q_t, \dot{q}_t]$; for the bouncing ball it contains position, translational velocity, and spin; and for the pendulum, angle and angular velocity.

The analytical dynamics may depend on physical quantities whose role is known but whose values are unobserved and can vary across trajectories. Rather than using fixed global values or performing per-trajectory optimization at test time, we infer these quantities amortized from the observed history, following the trajectory-level modulation principle of MoNODE (Auzina et al., 2023). Specifically, M_{traj} predicts trajectory-specific factors $\hat{\phi}$ that remain fixed over the rollout, while M_{dyn} predicts time-varying latent inputs:

$$\hat{\phi} = M_{\text{traj}}(z^m_{t-T_h+1:t}, s_{t-T_h+1:t}), \qquad \hat{u}_\tau = M_{\text{dyn}}(s_\tau, z_t^m, \hat{\phi}), \quad \tau \geq t. \tag{3}$$

This provides instance-specific adaptation in a single forward pass while enforcing trajectory-level constancy by construction. Each system uses only the inference heads required by its analytical model. For human motion, we can follow PhysPT (Zhang et al., 2024a), for providing the unobserved generalized actuation and contact forces supervision to M_{dyn}; see supplement.

3.4 Latent Augmented-Physics Propagation

We propagate the physical state with an augmented physics-based model (APBM) (Imbiriba et al., 2022; 2024; Yin et al., 2021):

$$\dot{s}_\tau = f_K(s_\tau; \phi_{\text{fix}}, \hat{\phi}, \hat{u}_\tau) + \gamma_\tau \odot f_U(s_\tau, z_t^m; \Psi), \qquad \gamma_\tau = \gamma_{\max}\sigma(g_\gamma(s_\tau, z_t^m)). \tag{4}$$

The analytical branch f_K contains the specified mechanics conditioned on the inferred physical quantities, while f_U captures remaining model discrepancy. The gate γ_τ bounds the residual scaling. Let $r_\tau = \gamma_\tau \odot f_U(s_\tau, z_t^m; \Psi)$. We regularize its magnitude relative to the analytical branch using

$$\mathcal{R}_{\text{res}} = \frac{1}{T_f}\sum_{k=0}^{T_f-1} \frac{\|r_{t+k}\|_2^2}{\|f_K(s_{t+k}^p; \phi_{\text{fix}}, \hat{\phi}, \hat{u}_{t+k})\|_2^2 + \varepsilon} + \beta_\gamma \frac{1}{T_f}\sum_{k=0}^{T_f-1} \|\gamma_{t+k}\|_1, \qquad \beta_\gamma = 0.2. \tag{5}$$

This keeps the residual corrective rather than dominant. We do not assume identifiability between the analytical and learned components; the prescribed mechanics and bounded residual instead provide an interpretable decomposition and a bias toward stable extrapolation. Sensitivity to $\gamma_{\max}$ and together with empirical distributions of γ_t and the relative residual magnitude $\rho_t = \|\gamma_t \odot f_U\|_2/(\|f_K\|_2 + \varepsilon)$, are reported in the supplement.

Starting from s_t, the APBM produces the propagated physical state and latent prior:

$$s_{t+k}^p = \text{ODESolve}(f_{\text{APBM}}, s_t, [t, t+k\Delta t]), \qquad z_{t+k}^p = \Pi_z(s_{t+k}^p). \tag{6}$$

Future EMA targets supervise this propagation:

$$z_{t+1:t+T_f}^\star = \text{sg}\left[E_{\text{meas}}^{\text{EMA}}(\tilde{y}_{t+1:t+T_f}, \mathbf{1})\right], \qquad \mathcal{L}_{\text{FUT}} = \frac{1}{T_f}\sum_{k=1}^{T_f} \|z_{t+k}^p - z_{t+k}^\star\|_2^2. \tag{7}$$

The future targets are used only for supervision and never enter the vector field. Hence $\mathcal{L}_{\text{JEPA}}$ promotes robustness to measurement corruption, while $\mathcal{L}_{\text{FUT}}$ promotes physics-mediated temporal predictability.

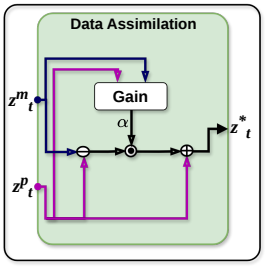


Figure 3: **Assimilation module.**

3.5 Latent Data Assimilation for Reconstruction

For sequential state estimation, z_t^m is computed from the unaugmented measurement history using the measurement pathway with $m = \mathbf{1}$. The propagated prior and predictive measurement latent are combined as

$$\alpha_t = \text{sigmoid}(A_\eta(z_t^p, z_t^m, z_t^m - z_t^p)), \qquad \alpha_t \in [0,1]^{d_z}, \tag{8}$$

$$z_t^* = z_t^p + \alpha_t \odot (z_t^m - z_t^p), \qquad \hat{y}_t = D_{\text{meas}}(z_t^*). \tag{9}$$

This follows the prediction–correction structure of learned data assimilation (Särkkä, 2013; Revach et al., 2022; Buchnik et al., 2024), with α_t learned directly rather than computed from state and measurement covariances. Measurement-specific correction is therefore confined to the assimilation pathway while the physical transition remains unchanged. For human-motion forecasting, rollouts are autoregressive and each forecast is fed back through the measurement pathway to construct the next predictive measurement latent.

3.6 Decoupled Training

The formulation uses three task losses, $\mathcal{L}_{\text{JEPA}}$, $\mathcal{L}_{\text{FUT}}$, and $\mathcal{L}_{\text{rec}}$, optimized in two stages.

Stage 1: predictive-prior learning. We optimize

$$\mathcal{J}_{\text{PhyCore}} = \mathcal{L}_{\text{JEPA}} + \mathcal{L}_{\text{FUT}} + \mathcal{R}_{\text{res}}. \tag{10}$$

Here $\mathcal{R}_{\text{res}}$ is a structural regularizer rather than an additional prediction task. Stage 1 trains the measurement and state encoders, measurement predictor, active physical-inference heads, residual dynamics and gate, and latent projection head; the target encoder $E_{\text{meas}}^{\text{EMA}}$ is updated by momentum. Neither A_η nor D_{meas} is trained, and no reconstruction loss is applied. For human motion, M_{dyn} is warm-initialized with PhysPT-style physical pseudo-labels before Stage 1; this initialization is detailed in the supplement and is not part of $\mathcal{J}_{\text{PhyCore}}$.

Table 1: **Human motion-estimation** results on Human3.6M under the OSDCap protocol (Le et al., 2024). Related methods are separated into kinematics-based (top) and physics-/dynamics-based (bottom) approaches. QuaMo$_{\mathrm{TRACE}}$ uses TRACE as its upstream kinematic estimator for direct comparison. FS denotes foot-skating rate (%); lower is better.

Method	Phys.	Adapt.	MPJPE↓ [mm]	MPJPE-G↓ [mm]	MPJPE-PA↓ [mm]	GRP↓ [mm]	Accel↓ [mm/s^2]	FS↓ [%]
VNECT (Mehta et al., 2017)	×	✓	89.6	–	62.7	185.1	–	–
HMMR (Kanazawa et al., 2019)	×	✓	79.4	–	55.0	231.1	–	–
HMR (Kanazawa et al., 2018)	×	✓	78.9	–	54.3	204.2	–	–
TRACE (Sun et al., 2023)	×	✓	78.1	152.7	62.5	125.9	19.2	–
VIBE (Kocabas et al., 2020)	×	✓	68.6	207.7	43.6	–	23.4	27.4
PhysCap (Shimada et al., 2020)	✓	✓	97.4	–	65.1	182.6	–	–
TrajOpt (Gärtner et al., 2022b)	✓	×	84.0	143.0	56.0	–	–	**4.0**
DiffPhy (Gärtner et al., 2022a)	✓	×	81.7	139.1	55.6	–	–	7.4
NeurPhys (Shimada et al., 2021)	✓	✓	76.5	–	58.2	–	–	–
Xie et al. (Xie et al., 2021)	✓	✓	68.1	–	–	85.1	–	–
IPMAN-R (Tripathi et al., 2023)	✓	✓	60.7	–	41.1	–	–	–
SimPoE (Yuan et al., 2021)	✓	✓	56.7	–	41.6	–	6.7	–
PhysPT (Zhang et al., 2024a)	✓	×	52.7	–	36.7	–	–	–
DnD (Li et al., 2022a)	✓	×	52.5	–	35.5	–	–	–
OSDCap (Le et al., 2024)	✓	✓	54.8	132.8	39.8	119.1	8.4	15.2
QuaMo$_{\mathrm{TRACE}}$ (Le et al., 2026)	✓	✓	51.4	116.9	37.5	101.8	5.7	6.6
DePhy (Ours)	✓	✓	**47.3**$_{\pm 0.14}$	**116.4**$_{\pm 0.18}$	**32.7**$_{\pm 0.11}$	103.5$_{\pm 0.21}$	**3.9**$_{\pm 0.02}$	6.2

Table 2: **Cross-dataset human motion-estimation** results on Fit3D and SportsPose under the OSDCap protocol (Le et al., 2024). QuaMo$_{\mathrm{TRACE}}$ uses TRACE as its upstream estimator. FS denotes foot-skating rate (%); lower is better.

Dataset	Method	MPJPE↓ [mm]	MPJPE-G↓ [mm]	MPJPE-PA↓ [mm]	GRP↓ [mm]	Accel↓ [mm/s^2]	FS↓ [%]
Fit3D (Fieraru et al., 2021)	TRACE (Sun et al., 2023)	85.4	131.2	65.2	178.1	20.2	–
	OSDCap (Le et al., 2024)	58.7	73.8	42.6	47.2	8.2	25.9
	QuaMo (Le et al., 2026)	50.3	68.8	**35.6**	45.2	**3.8**	16.3
	DePhy (Ours)	**50.1**$_{\pm 0.60}$	**67.5**$_{\pm 0.43}$	40.8$_{\pm 0.29}$	**41.2**$_{\pm 0.27}$	4.4$_{\pm 0.06}$	**15.9**$_{\pm 0.24}$
SportsPose (Ingwersen et al., 2023)	TRACE (Sun et al., 2023)	97.3	361.9	71.1	333.0	15.8	–
	OSDCap (Le et al., 2024)	71.7	113.6	52.4	90.2	10.9	38.0
	QuaMo (Le et al., 2026)	71.4	112.2	48.7	**82.3**	5.3	24.1
	DePhy (Ours)	**67.5**$_{\pm 0.62}$	**105.6**$_{\pm 0.87}$	**42.8**$_{\pm 0.25}$	86.5$_{\pm 0.37}$	**4.8**$_{\pm 0.09}$	**22.6**$_{\pm 0.47}$

Stage 2: assimilation and reconstruction. The PhyCore is frozen and only A_η and D_{meas} are trained:

$$\mathcal{J}_{\mathrm{DA}} = \mathcal{L}_{\mathrm{rec}} + \mathcal{R}_{\mathrm{stay}}, \quad \mathcal{L}_{\mathrm{rec}} = \frac{1}{T}\sum_{t=1}^{T} \|\hat{y}_t - y_t^{\mathrm{ref}}\|_2^2, \quad \mathcal{R}_{\mathrm{stay}} = \frac{1}{T}\sum_{t=1}^{T} \|z_t^* - z_t^p\|_2^2. \tag{11}$$

When no separate reference is available, $\tilde{y}_t$ is used as the reconstruction target. The stay term has unit weight in all experiments. Because the PhyCore is frozen, reconstruction gradients cannot reshape the predictive representation or physical transition.

4 Experiments

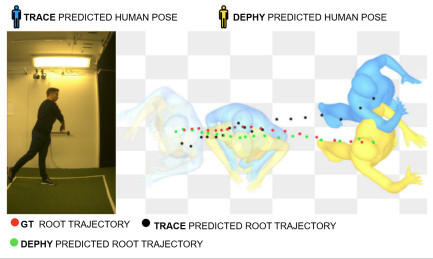
We evaluate DePhy on four tasks spanning real-world and synthetic dynamical systems: 3D human motion estimation and forecasting from monocular video, and 3D bouncing-ball and planar damped-pendulum dynamics. The human-motion tasks test reconstruction from noisy kinematic measurements and long-horizon rollout stability, while the synthetic benchmarks test recovery of trajectory-specific physical factors and extrapolation beyond the training horizon.

4.1 Human Motion Estimation from Monocular Videos

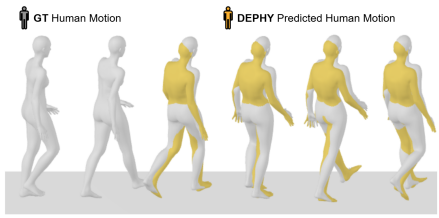
This task evaluates whether DePhy's assimilation mechanism can reconcile a physics-propagated prior with noisy kinematic measurements. We use TRACE (Sun et al., 2023) as the fixed upstream estimator and refine its outputs through causal measurement updates. The assimilation module is trained offline and updates the propagated state using the current measurement without parameter updates at test time. Thus, "Adapt." in Tab. 1 indicates adaptation to the current measurement. Fig. 4a illustrates the resulting correction of upstream kinematic errors.

Table 3: **Human3.6M long-horizon forecasting under the PhysMoP protocol** (Zhang et al., 2024b). MPJPE is reported at standard future horizons used in prior work for both Human3.6M-P1 and Human3.6M-P2.

Method	80	160	320	400	560	720	880	1000
Human3.6M-P1 MPJPE (mm)								
HisRep (Mao et al., 2020)	10.4	22.6	47.1	58.3	77.3	91.8	104.1	112.1
ST-DGCN (Ma et al., 2022)	10.6	23.1	47.1	57.9	76.3	90.7	102.4	109.7
siMLPe (Guo et al., 2023)	9.6	21.7	46.3	57.3	75.7	90.1	101.8	109.4
EqMotion (Xu et al., 2023)	9.1	20.1	43.7	55.0	73.4	–	–	106.9
PhysMoP (Zhang et al., 2024b)	2.1	7.6	28.4	43.8	72.9	86.2	96.1	103.9
DePhy (w/o DA)	2.7	7.2	23.1	34.4	64.8	80.5	89.3	97.7
DePhy (w DA)	2.5	5.3	21.2	30.9	59.7	73.4	83.9	90.8
Human3.6M-P2 MPJPE (mm)								
STG-GCN (Zhong et al., 2022)	10.1	16.9	32.5	38.5	50.0	–	–	72.9
siMLPe (Guo et al., 2023)	4.5	9.8	22.0	28.1	39.3	49.2	57.8	63.7
PhysMoP (Zhang et al., 2024b)	1.4	3.7	11.6	17.2	30.5	41.9	51.1	57.1
DePhy (w/o DA)	2.4	5.1	10.5	14.8	26.6	34.8	46.2	52.4
DePhy (w DA)	2.3	4.2	8.3	12.7	24.1	30.3	41.6	48.9



(a) Reconstruction result on SportsPose (Ingwersen et al., 2023). Shows the error between the inaccurate kinematic estimator and DePhy. DePhy tracks follow much more closely to the ground-truth tracks.



(b) Forecasting on Human3.6M. The first two frames denote history, and the subsequent frames show forecast and ground-truth overlap over a {300, 600, 900 and 1200} ms forecast horizon.

Figure 4: **Qualitative results** for human motion reconstruction and motion forecasting.

Following OSDCap (Le et al., 2024), we evaluate on Human3.6M (Ionescu et al., 2014), Fit3D (Fieraru et al., 2021), and SportsPose (Ingwersen et al., 2023) using their corresponding protocols; implementation details are in Supp. C. We report MPJPE, MPJPE-G, MPJPE-PA, global-root error, acceleration error, and foot-skating rate. On Human3.6M, DePhy improves over OSDCap across all reported metrics, reducing MPJPE from 54.8 to 47.3 mm and acceleration error from 8.4 to 3.9 mm/s^2 (Tab. 1). These gains persist on Fit3D (58.7 $\rightarrow$ 50.1 mm MPJPE; 8.2 $\rightarrow$ 4.4 mm/s^2 Accel) and SportsPose (71.7 $\rightarrow$ 67.5 mm; 10.9 $\rightarrow$ 4.8 mm/s^2), demonstrating robust state estimation under dataset and measurement-domain shifts (Tab. 2).

4.2 Human Motion Forecasting

We next evaluate the learned transition prior under autoregressive forecasting without future ground-truth measurements. DePhy w/o DA performs a pure transition-prior rollout, whereas DePhy w DA recursively assimilates only model-generated measurement-side evidence. We also show data assimilation results under different observation noise in supplementary. Their comparison isolates whether prediction–correction improves long-horizon stability without future observations. Fig. 4b shows forecast (gold) and ground-truth (grey) overlap.

Following PhysMoP (Zhang et al., 2024b), we evaluate on Human3.6M up to 1 s (Guo et al., 2023; Mao et al., 2020; 2019; Martinez et al., 2017; Tanke et al., 2021) under P1 (MPJPE at the target horizon) and P2 (MPJPE averaged up to the target) (Zhang et al., 2024b). Tab. 3 shows that DePhy with assimilation improves over PhysMoP from 160 ms onward on P1 and 320 ms onward on P2, reaching 90.8 versus 103.9 mm at 1 s on P1 and 48.9 versus 57.1 mm on P2. PhysMoP performs better at the shortest horizons, but the trend reverses as the forecast horizon increases. Assimilation

Table 4: **Bouncing ball results.** Left: factor identification. Right: within-horizon and extrapolation forecasting.

Method	g	C_D	C_L	c_ω	e	μ	k_ω	**Mean**
NODE (Chen et al., 2018)	-0.31	-0.44	-0.29	-0.18	-0.22	-0.35	-0.27	-0.29
MoNODE (Auzina et al., 2023)	0.61	0.42	0.38	0.29	0.52	0.44	0.47	0.45
DePhy-NoRes	0.81	0.73	0.76	0.69	0.67	0.64	0.62	0.70
DePhy (Ours)	**0.89**	**0.84**	**0.87**	**0.81**	**0.79**	**0.77**	**0.80**	**0.82**

(a) **FoV identification (R^2) on the bouncing ball dataset.** Higher is better.

Method	**MSE** (T=20)↓	**MSE** (T=40)↓
NODE (Chen et al., 2018)	$0.0199^{(0.001)}$	$0.0631^{(0.003)}$
MoNODE (Auzina et al., 2023)	$0.0164^{(0.001)}$	$0.0512^{(0.002)}$
DePhy-NoRes	$0.0138^{(0.001)}$	$0.0421^{(0.002)}$
DePhy (Ours)	$\mathbf{0.0112}^{(0.001)}$	$\mathbf{0.0317}^{(0.002)}$

(b) **Trajectory forecasting MSE on the bouncing ball dataset.** Lower is better.

Table 5: **Pendulum results.** Left: factor identification. Right: forecasting and parameter error.

Method	$R^2(\ell)$	$R^2(g)$	$R^2(c)$	Mean
GRU	-0.19	-0.24	-0.33	-0.25
NODE	-0.14	-0.17	-0.28	-0.20
LNN	0.31	0.27	0.19	0.26
APHYNITY	0.71	0.68	0.59	0.66
MoNODE	0.63	0.58	0.51	0.57
DePhy-NoRes	0.81	0.76	0.70	0.76
DePhy	**0.93**	**0.89**	**0.84**	**0.89**

(a) **FoV identification (R^2) on the pendulum dataset.** Higher is better.

Method	T_f=20	T_f=40	ϵ_ϕ(%)
Known ODE (fitted)	0.0014	0.0038	1.9
GRU	0.0091	0.0487	–
NODE	0.0074	0.0391	–
LNN	0.0061	0.0312	–
APHYNITY	0.0023	0.0097	2.3
Phys-VAE	0.0027	0.0118	2.8
MoNODE	0.0049	0.0219	–
DePhy-NoRes	0.0019	0.0081	2.1
DePhy	**0.0015**	**0.0059**	**1.7**

(b) **Trajectory forecasting on the pendulum dataset.** Lower is better.

also improves the same DePhy prior at 1 s from 97.7 to 90.8 mm on P1 and 52.4 to 48.9 mm on P2, supporting the separation between the predictive prior and learned correction mechanism.

4.3 Synthetic Physics Benchmarks

The synthetic benchmarks test recovery of trajectory-specific physical factors and extrapolation beyond the training horizon. We use a 3D bouncing ball and planar damped pendulum, both simulated at 10 Hz. We compare against NODE (Chen et al., 2018) and MoNODE (Auzina et al., 2023) on both benchmarks; pendulum additionally includes GRU (Chung et al., 2014), LNN (Cranmer et al., 2020), APHYNITY (Yin et al., 2021), Phys-VAE (Takeishi & Kalousis, 2021), and a fitted Known-ODE reference. We use official implementations for NODE, LNN, and APHYNITY and re-implement MoNODE in PyTorch. APHYNITY uses the authors' defaults, while other trainable baselines are tuned by grid search. We also evaluate DePhy-NoRes, which removes the learned residual.

Following MoNODE (Auzina et al., 2023), FoV identification fits a linear regressor from learned trajectory-level representations to ground-truth physical parameters and reports R^2 (Auzina et al., 2023; Eastwood & Williams, 2018). Forecasting uses trajectory MSE; for pendulum methods with explicit parameter estimates, we additionally report normalized parameter error ϵ_ϕ following APHYNITY (Yin et al., 2021).

Bouncing Ball. Following (Auzina et al., 2023), we vary $\phi = [g, C_D, C_L, c_\omega, e, \mu, k_\omega]$ and generate $1{,}000/100/100$ train/validation/test trajectories. Models condition on the first 10 observed steps and are evaluated at T=20 and the extrapolation horizon T=40. Observations are noisy 3D positions (σ=0.02), with results averaged over three seeds; further details are in Supp. D. Tab. 4 shows that DePhy achieves mean FoV R^2=0.82, compared with 0.45 for MoNODE and 0.70 for DePhy-NoRes. It also achieves MSEs of 0.0112 at T=20 and 0.0317 at T=40, compared with $0.0138/0.0421$ for DePhy-NoRes and $0.0164/0.0512$ for MoNODE. The improvement over DePhy-NoRes shows that the bounded residual benefits both factor recovery and long-horizon prediction.

Planar Damped Pendulum. We use state $s_t = [\theta_t, \dot{\theta}_t]$ with factors $\phi = [\ell, g, c]$, sampling $\ell \in [0.5, 2.0]$, $g \in [7.0, 12.0]$, and $c \in [0.0, 0.5]$. We generate $800/100/100$ train/validation/test trajectories at Δt=0.1 s with Gaussian observation noise (σ=0.01). Training uses T_h=20 context and T_f=20 prediction steps; at test time, we additionally roll out for T_f=40 to evaluate extrapolation. Results are averaged over three seeds. Tab. 5 shows that DePhy achieves mean FoV R^2=0.89, com-

Table 6: **Training-objective ablations for sequential human state estimation.** All variants use the same fixed TRACE measurements as input. The "Rec. updates PhyCore?" column indicates whether $\mathcal{L}_{\mathrm{rec}}$ is allowed to backpropagate through the measurement representation and physics-structured transition during prior learning. In DePhy, PhyCore is first trained with $\mathcal{L}_{\mathrm{JEPA}}$ and $\mathcal{L}_{\mathrm{FUT}}$ and then frozen before assimilation and reconstruction training. $\mathcal{R}_{\mathrm{stay}}$ is the fixed-weight prior-anchoring regularizer used only during Stage 2. Lower is better.

Dataset	Variant	$\mathcal{L}_{\mathrm{JEPA}}$	$\mathcal{L}_{\mathrm{FUT}}$	Rec. updates PhyCore?	$\mathcal{R}_{\mathrm{stay}}$	MPJPE↓ [mm]	MPJPE-G↓ [mm]	Accel↓ [mm/s^2]
Human3.6M	TRACE input	–	–	–	–	78.1	152.7	19.2
	w/o $\mathcal{L}_{\mathrm{JEPA}}$	✗	✓	✗	✓	50.6	123.8	4.7
	with only $\mathcal{L}_{\mathrm{JEPA}}$	✓	✗	✗	✗	76.2	138.7	16.8
	Rec.-supervised future prior	✓	✗	✓	✓	52.6	128.2	9.6
	with only $\mathcal{L}_{\mathrm{FUT}}$	✗	✓	✗	✗	50.9	124.3	5.1
	Coupled w/ $\mathcal{L}_{\mathrm{rec}}$	✓	✓	✓	✓	49.8	125.5	8.8
	w/o $\mathcal{R}_{\mathrm{stay}}$	✓	✓	✗	✗	50.2	122.9	4.9
	DePhy	✓	✓	✗	✓	**47.3**	**116.4**	**3.9**
SportsPose	TRACE input	–	–	–	–	97.3	361.9	15.8
	w/o $\mathcal{L}_{\mathrm{JEPA}}$	✗	✓	✗	✓	75.6	122.9	6.1
	with only $\mathcal{L}_{\mathrm{JEPA}}$	✓	✗	✗	✗	95.2	315.2	14.8
	Rec.-supervised future prior	✓	✗	✓	✓	76.3	135.7	12.5
	with only $\mathcal{L}_{\mathrm{FUT}}$	✗	✓	✗	✗	73.5	117.9	6.5
	Coupled w/ $\mathcal{L}_{\mathrm{rec}}$	✓	✓	✓	✓	77.2	142.6	12.8
	w/o $\mathcal{R}_{\mathrm{stay}}$	✓	✓	✗	✗	71.5	113.3	6.3
	DePhy	✓	✓	✗	✓	**67.5**	**105.6**	**4.8**

pared with 0.66 for APHYNITY and 0.57 for MoNODE. At $T_f{=}40$, its MSE is 0.0059, compared with 0.0097 for APHYNITY, 0.0118 for Phys-VAE, and 0.0081 for DePhy-NoRes. Its parameter error is 1.7%, versus 1.9% for the fitted Known-ODE reference, although Known-ODE retains the lowest extrapolation MSE (0.0038). These results indicate accurate recovery of trajectory-specific physical parameters while retaining flexibility beyond the nominal dynamics.

Ablation. Tab. 6 evaluates the objectives for learning the measurement representation, physics-structured transition prior, and assimilation stage using fixed TRACE measurements. Removing $\mathcal{L}_{\mathrm{JEPA}}$ increases MPJPE from 47.3 to 50.6 mm on Human3.6M and from 67.5 to 75.6 mm on SportsPose, supporting masked latent prediction for corruption-robust representations. Using only $\mathcal{L}_{\mathrm{JEPA}}$ or $\mathcal{L}_{\mathrm{FUT}}$ also underperforms the full objective, indicating complementary supervision. A reconstruction-supervised future prior yields substantially larger acceleration error (9.6 versus 3.9 mm/s^2 on Human3.6M; 12.5 versus 4.8 on SportsPose), indicating less stable dynamics.

The ablation also supports decoupled training. Allowing reconstruction gradients to update the PhyCore gives 49.8 mm MPJPE and 8.8 mm/s^2 acceleration error on Human3.6M, versus 47.3 mm and 3.9 mm/s^2 with two-stage training; on SportsPose, the corresponding values are 77.2 versus 67.5 mm and 12.8 versus 4.8 mm/s^2. Removing R_{stay} also degrades performance. Finally, with a fixed transition prior, assimilation reduces 1-s P1 MPJPE from 97.7 to 90.8 mm and P2 from 52.4 to 48.9 mm. Together, these results support complementary latent objectives, decoupled predictive-prior learning, and a separate assimilation mechanism. Further details are in Supp. E.5.

5 Conclusion

We presented DePhy, a decoupled state-space framework for physics-aware state estimation from noisy or partially observed measurements. DePhy learns a predictive measurement representation, maps it to an explicit physical state, and propagates that state with an augmented physics model that combines analytical dynamics with learned corrections. Measurement evidence is incorporated separately through latent data assimilation, allowing the predictive representation and physical transition to remain independent of reconstruction supervision. Across synthetic and real-world benchmarks, DePhy improves state estimation, long-horizon forecasting, and recovery of trajectory-specific physical factors. Future work will extend the framework to less-specified dynamical systems through latent system identification, combine the physics-structured prior with generative decoders for motion synthesis, and develop learned observation models for severe occlusion and sensor degradation. Additionally extend the current framework to learn across dynamical systems.

AI USE STATEMENT

Generative AI tools were used only for writing assistance, including improving the clarity, grammar, and presentation of parts of the manuscript, and for limited code generation during implementation. Generative AI was not used to conceive the research idea, formulate the scientific contributions, or conduct the literature review. All AI-assisted text and code were reviewed by the authors, and generated code was checked and tested before being used in the experiments. The authors take full responsibility for the final content of the paper, including all text, claims, experimental results, and code produced with the assistance of generative AI.

ETHICS STATEMENT

We do not identify any specific ethical concerns arising from this work. The study uses existing datasets and synthetic dynamical systems and does not involve the collection of new human-subject data. The proposed methods are intended for research on physics-aware state estimation, reconstruction, and forecasting, and we are not aware of additional concerns related to privacy, fairness, security, or harmful applications that require specific discussion.

REPRODUCIBILITY STATEMENT

We provide the complete model formulation, training objectives, and evaluation protocols in the main paper, with additional architectural, optimization, preprocessing, dataset, and system-specific implementation details in the appendix, see C.8 C.7. The supplementary material also documents the physical dynamics, warm initialization procedure, ablations, robustness analyses, and compute requirements used in our experiments see E.7.

REFERENCES

Mohamed Abdelfattah and Alexandre Alahi. S-JEPA: A joint embedding predictive architecture for skeletal action recognition. In *European Conference on Computer Vision*, pp. 367–384. Springer, 2024.

Priyanshu Agarwal, Suren Kumar, Julian Ryde, Jason J Corso, and Venkat N Krovi. Estimating dynamics on-the-fly using monocular video for vision-based robotics. *IEEE/ASME Transactions on Mechatronics*, 19(4):1412–1423, 2013.

Emre Aksan, Manuel Kaufmann, Peng Cao, and Otmar Hilliges. A spatio-temporal transformer for 3D human motion prediction. In *Proceedings of the International Conference on 3D Vision (3DV)*, 2021.

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.

Ilze Amanda Auzina, Çağatay Yıldız, Sara Magliacane, Matthias Bethge, and Efstratios Gavves. Modulated neural ODEs. *Advances in Neural Information Processing Systems*, 36:44572–44594, 2023.

Adrien Bardes, Jean Ponce, and Yann LeCun. MC-JEPA: A joint-embedding predictive architecture for self-supervised learning of motion and content features. *arXiv preprint arXiv:2307.12698*, 2023.

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.

Itay Buchnik, Dan Steger, Guy Revach, Ruud J.G. van Sloun, Tirza Routtenberg, and Nir Shlezinger. Latent-KalmanNet: Learned Kalman filtering for tracking from high-dimensional signals. *IEEE Transactions on Signal Processing*, 2024.

Ricky TQ Chen, Yulia Rubanova, Jesse Bettencourt, and David K Duvenaud. Neural ordinary differential equations. *Advances in Neural Information Processing Systems*, 31, 2018.

Sheng Cheng, Deqian Kong, Jianwen Xie, Kookjin Lee, Ying Nian Wu, and Yezhou Yang. Latent space energy-based neural ODEs. *arXiv preprint arXiv:2409.03845*, 2024.

Junyoung Chung, Caglar Gulcehre, KyungHyun Cho, and Yoshua Bengio. Empirical evaluation of gated recurrent neural networks on sequence modeling. *arXiv preprint arXiv:1412.3555*, 2014.

Miles Cranmer, Sam Greydanus, Stephan Hoyer, Peter Battaglia, David Spergel, and Shirley Ho. Lagrangian neural networks. *arXiv preprint arXiv:2003.04630*, 2020.

Edward De Brouwer, Jaak Simm, Adam Arany, and Yves Moreau. GRU-ODE-Bayes: Continuous modeling of sporadically-observed time series. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

Cian Eastwood and Christopher K.I. Williams. A framework for the quantitative evaluation of disentangled representations. In *International Conference on Learning Representations*, 2018.

Mihai Fieraru, Mihai Zanfir, Silviu-Cristian Pirlea, Vlad Olaru, and Cristian Sminchisescu. AiFiT: Automatic 3d human-interpretable feedback models for fitness training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021.

Katerina Fragkiadaki, Sergey Levine, Panna Felsen, and Jitendra Malik. Recurrent network models for human dynamics. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 4346–4354, 2015.

Erik Gärtner, Mykhaylo Andriluka, Erwin Coumans, and Cristian Sminchisescu. Differentiable dynamics for articulated 3D human motion reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 13190–13200, 2022a.

Erik Gärtner, Mykhaylo Andriluka, Hongyi Xu, and Cristian Sminchisescu. Trajectory optimization for physics-based reconstruction of 3D human pose from monocular video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 13106–13115, 2022b.

Vincent Le Guen and Nicolas Thome. Disentangling physical dynamics from unknown factors for unsupervised video prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 11474–11484, 2020.

Wen Guo, Yuming Du, Xi Shen, Vincent Lepetit, Xavier Alameda-Pineda, and Francesc Moreno-Noguer. Back to MLP: A simple baseline for human motion prediction. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 4809–4819, 2023.

Jun Hayakawa and Behzad Dariush. Ego-motion and surrounding vehicle state estimation using a monocular camera. In *2019 IEEE Intelligent Vehicles Symposium (IV)*, pp. 2550–2556. IEEE, 2019.

Tales Imbiriba, Ahmet Demirkaya, Jindřich Duník, Ondřej Straka, Deniz Erdoğmuş, and Pau Closas. Hybrid neural network augmented physics-based models for nonlinear filtering. In *Proceedings of the 25th International Conference on Information Fusion*, pp. 1–6, 2022.

Tales Imbiriba, Ondřej Straka, Jindřich Duník, and Pau Closas. Augmented physics-based machine learning for navigation and tracking. *IEEE Transactions on Aerospace and Electronic Systems*, 60(3):2692–2704, 2024.

Christian Keilstrup Ingwersen, Christian Møller Mikkelstrup, Janus Nørtoft Jensen, Morten Rieger Hannemose, and Anders Bjorholm Dahl. SportsPose: A dynamic 3d sports pose dataset. In *IEEE/CVF International Workshop on Computer Vision in Sports*, 2023.

Catalin Ionescu, Dragos Papava, Vlad Olaru, and Cristian Sminchisescu. Human3.6M: Large scale datasets and predictive methods for 3D human sensing in natural environments. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 36(7):1325–1339, 2014.

Angjoo Kanazawa, Michael J. Black, David W. Jacobs, and Jitendra Malik. End-to-end recovery of human shape and pose. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

Angjoo Kanazawa, Jason Y. Zhang, Panna Felsen, and Jitendra Malik. Learning 3d human dynamics from video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

Patrick Kidger, James Morrill, James Foster, and Terry Lyons. Neural controlled differential equations for irregular time series. In *Advances in Neural Information Processing Systems*, volume 33, pp. 6696–6707, 2020.

Muhammed Kocabas, Nikos Athanasiou, and Michael J. Black. VIBE: Video inference for human body pose and shape estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5253–5263, 2020.

Farnoosh Koleini, Hongfei Xue, Ahmed Helmy, and Pu Wang. Monomsk: Monocular 3d musculoskeletal dynamics estimation. *arXiv preprint arXiv:2511.19326*, 2025.

An Lanji, Dawei Liu, Jin Li, Haoran Xu, Mei Chen, and Yu Tian. Unijepa: A unified joint-embedding predictive architecture for task-agnostic visual world modeling. *arXiv preprint arXiv:2608.07409*, 2026.

Cuong Le, Viktor Johansson, Manon Kok, and Bastian Wandt. Optimal-state dynamics estimation for physics-based human motion capture from videos. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024.

Cuong Le, Pavlo Melnyk, Urs Waldmann, Mårten Wadenbäck, and Bastian Wandt. Quamo: Quaternion motions for vision-based 3d human kinematics capture. In *The Fourteenth International Conference on Learning Representations*, 2026.

Yann LeCun. A path towards autonomous machine intelligence. *OpenReview preprint*, 2022.

Jiefeng Li, Chao Xu, Zhicun Chen, Siyuan Bian, Lixin Yang, and Cewu Lu. HybriK: A hybrid analytical-neural inverse kinematics solution for 3D human pose and shape estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3383–3393, 2021.

Jiefeng Li, Siyuan Bian, Chao Xu, Gang Liu, Gang Yu, and Cewu Lu. D&D: Learning human dynamics from dynamic camera. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 479–496, 2022a.

Lihuan Li, Hao Xue, Shuang Ao, Yang Song, and Flora Salim. HiT-JEPA: A hierarchical self-supervised trajectory embedding framework for similarity computation. *arXiv preprint arXiv:2507.00028*, 2025.

Zhihao Li, Jianzhuang Liu, Zhensong Zhang, Songcen Xu, and Youliang Yan. CLIFF: Carrying location information in full frames into human pose and shape estimation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 590–606, 2022b.

Michael Lutter, Christian Ritter, and Jan Peters. Deep Lagrangian networks: Using physics as model prior for deep learning. *arXiv preprint arXiv:1907.04490*, 2019.

Tiezheng Ma, Yongwei Nie, Chengjiang Long, Qing Zhang, and Guiqing Li. Progressively generating better initial guesses towards next stages for high-quality human motion prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

Wei Mao, Miaomiao Liu, Mathieu Salzmann, and Hongdong Li. Learning trajectory dependencies for human motion prediction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 9489–9497, 2019.

Wei Mao, Miaomiao Liu, and Mathieu Salzmann. History repeats itself: Human motion prediction via motion attention. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 474–489, 2020.

Julieta Martinez, Michael J. Black, and Javier Romero. On human motion prediction using recurrent neural networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2891–2900, 2017.

Dushyant Mehta, Srinath Sridhar, Oleksandr Sotnychenko, Helge Rhodin, Mohammad Shafiei, Hans-Peter Seidel, Weipeng Xu, Dan Casas, and Christian Theobalt. VNect: Real-time 3d human pose estimation with a single rgb camera. *ACM Transactions on Graphics*, 36(4), 2017.

Heejeong Nam, Quentin Le Lidec, Lucas Maes, Yann LeCun, and Randall Balestriero. Causal-jepa: Learning world models through object-level latent masking. *arXiv preprint arXiv:2602.11389*, 2026.

Alexander Norcliffe, Cristian Bodnar, Ben Day, Nikola Simidjievski, and Pietro Liò. On second order behaviour in augmented neural ODEs. *Advances in Neural Information Processing Systems*, 33:5911–5921, 2020.

Tong Qin, Peiliang Li, and Shaojie Shen. Vins-mono: A robust and versatile monocular visual-inertial state estimator. *IEEE transactions on robotics*, 34(4):1004–1020, 2018.

Guy Revach, Nir Shlezinger, Xiaoyong Ni, Adria Lopez Escoriza, Ruud J.G. van Sloun, and Yonina C. Eldar. KalmanNet: Neural network aided Kalman filtering for partially known dynamics. *IEEE Transactions on Signal Processing*, 70:1532–1547, 2022.

Yulia Rubanova, Ricky T. Q. Chen, and David K. Duvenaud. Latent ordinary differential equations for irregularly-sampled time series. *Advances in Neural Information Processing Systems*, 32, 2019.

Romeo Šajina and Marina Ivašić-Kos. 3d pose estimation and tracking in handball actions using a monocular camera. *Journal of imaging*, 8(11):308, 2022.

Simo Särkkä. *Bayesian Filtering and Smoothing*. Cambridge University Press, 2013.

Soshi Shimada, Vladislav Golyanik, Weipeng Xu, and Christian Theobalt. PhysCap: Physically plausible monocular 3D motion capture in real time. *ACM Transactions on Graphics*, 39(6), 2020.

Soshi Shimada, Vladislav Golyanik, Weipeng Xu, Patrick Pérez, and Christian Theobalt. Neural monocular 3D human motion capture with physical awareness. *Transactions on Graphics*, 2021.

Soyong Shin, Juyong Kim, Eni Halilaj, and Michael J. Black. WHAM: Reconstructing world-grounded humans with accurate 3D motion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024.

Shengkai Sun, Zefan Zhang, Jianfeng Dong, Zhiyong Cheng, Xiaojun Chang, and Meng Wang. Towards efficient general feature prediction in masked skeleton modeling. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 12212–12221, October 2025.

Yu Sun, Qian Bao, Wu Liu, Tao Mei, and Michael J Black. Trace: 5d temporal regression of avatars with dynamic cameras in 3d environments. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8856–8866, 2023.

Naoya Takeishi and Alexandros Kalousis. Physics-integrated variational autoencoders for robust and interpretable generative modeling. *Advances in Neural Information Processing Systems*, 34: 14809–14821, 2021.

Hiromu Taketsugu, Takeru Oba, Takahiro Maeda, Shohei Nobuhara, and Norimichi Ukita. Physical plausibility-aware trajectory prediction via locomotion embodiment. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025.

William Talbot, Julian Nubert, Turcan Tuna, Cesar Cadena, Frederike Dümbgen, Jesus Tordesillas, Timothy D Barfoot, and Marco Hutter. Continuous-time state estimation methods in robotics: A survey. *IEEE Transactions on Robotics*, 2025.

Julian Tanke, Chintan Zaveri, and Juergen Gall. Intention-based long-term human motion anticipation. In *International Conference on 3D Vision*, 2021.

Shashank Tripathi, Lea M"uller, Chun-Hao P. Huang, Omid Taheri, Michael J. Black, and Dimitrios Tzionas. 3d human pose estimation via intuitive physics. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023.

Yibing Wei, Abhinav Gupta, and Pedro Morgado. Towards latent masked image modeling for self-supervised visual representation learning. In *European Conference on Computer Vision*, pp. 1–17. Springer, 2024.

Kevin Xie, Tingwu Wang, Umar Iqbal, Yunrong Guo, Sanja Fidler, and Florian Shkurti. Physics-based human motion estimation and synthesis from videos. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021.

Chenxin Xu, Robby T Tan, Yuhong Tan, Siheng Chen, Yu Guang Wang, Xinchao Wang, and Yanfeng Wang. Eqmotion: Equivariant multi-agent motion prediction with invariant interaction reasoning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 1410–1420, 2023.

Brandon Yee and Pairie Koh. Pi-jepa: Label-free surrogate pretraining for coupled multiphysics simulation via operator-split latent prediction. *arXiv preprint arXiv:2604.01349*, 2026.

Çağatay Yıldız, Markus Heinonen, and Harri Lähdesmäki. ODE2VAE: Deep generative second order ODEs with Bayesian neural networks. *Advances in Neural Information Processing Systems*, 32, 2019.

Yuan Yin, Vincent Le Guen, Jérémie Dona, Emmanuel de Bézenac, Ibrahim Ayed, Nicolas Thome, and Patrick Gallinari. Augmenting physical models with deep networks for complex dynamics forecasting. In *International Conference on Learning Representations (ICLR)*, 2021.

Ye Yuan, Shih-En Wei, Tomas Simon, Kris Kitani, and Jason Saragih. SimPoE: Simulated character control for 3D human pose estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021.

Yufei Zhang, Jeffrey O. Kephart, Zijun Cui, and Qiang Ji. PhysPT: Physics-aware pretrained transformer for estimating human dynamics from monocular videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024a.

Yufei Zhang, Jeffrey O. Kephart, and Qiang Ji. Incorporating physics principles for precise human motion prediction. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 6164–6174, 2024b.

Chongyang Zhong, Lei Hu, Zihao Zhang, Yongjing Ye, and Shihong Xia. Spatio-temporal gating-adjacency GCN for human motion prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6447–6456, 2022.

Yaofeng Desmond Zhong, Biswadip Dey, and Amit Chakraborty. Symplectic ODE-Net: Learning Hamiltonian dynamics with control. *arXiv preprint arXiv:1909.12077*, 2019.

A Extended Motivation

The main paper motivates DePhy around a separation between three operations that are often optimized jointly in physics-aware state estimation: learning a representation of imperfect measurements, learning how a physical state evolves, and reconciling a propagated state with new observations. This section expands the motivation for that separation and clarifies the role of each component.

A.1 Why Decouple Predictive Dynamics from Reconstruction?

Monocular motion estimates are not direct observations of physical state. They inherit depth ambiguity, self-occlusion, global-position drift, temporal jitter, and estimator-specific bias from the upstream vision model. A reconstruction loss alone therefore creates an ambiguous training signal: it rewards agreement with the available measurement stream even when that stream contains nuisance variation that should not be propagated by the physical model. This is particularly problematic for long-horizon prediction, where small frame-level errors can be repeatedly integrated into the transition.

DePhy addresses this by learning the predictive prior before reconstruction. The measurement representation is trained to be stable to additional corruption through $\mathcal{L}_{\mathrm{JEPA}}$, while the physical transition is trained through future latent prediction $\mathcal{L}_{\mathrm{FUT}}$. Reconstruction is introduced only after the representation and transition have been learned and frozen. The resulting design does not assume that the upstream measurements are clean; instead, it asks the representation to preserve the components of the measurement history that are useful for predicting future evolution while suppressing nuisance variation that is unstable under masking and corruption.

This distinction is important for interpreting the ablations in the main paper. The reconstruction-supervised future-prior variant and the fully coupled reconstruction variant both achieve competitive local pose error, but their acceleration errors increase sharply. The failure is therefore not simply underfitting. Rather, allowing decoded measurement error to shape the transition encourages a prior that can reproduce local measurements while becoming less stable as a dynamical model. The full objective avoids this coupling by making latent predictability, rather than measurement reconstruction, the criterion used to shape the physical prior.

A.2 Why Combine Same-Time and Future Latent Prediction?

The two latent objectives address different failure modes. Same-time masked prediction, $\mathcal{L}_{\mathrm{JEPA}}$, aligns the student representation of an additionally corrupted history with an EMA-teacher representation of the same upstream measurement stream. It therefore promotes invariance to masking and injected nuisance without requiring clean-state labels. Future prediction, $\mathcal{L}_{\mathrm{FUT}}$, instead asks a state propagated through the physical transition to agree with future EMA targets. It forces the learned representation to retain information that is dynamically predictive rather than merely sufficient for reconstructing the current frame.

Neither objective alone is sufficient in the human-motion experiments. A model trained only with $\mathcal{L}_{\mathrm{JEPA}}$ lacks an explicit temporal-prediction signal, while a model trained only with $\mathcal{L}_{\mathrm{FUT}}$ loses the explicit denoising constraint on the measurement encoder. Their combination therefore implements the two requirements stated in the main paper: robustness to measurement nuisance and preservation of information required for physical propagation.

Although JEPA-style training is often discussed in the context of large-scale foundation models (LeCun, 2022; Assran et al., 2023; Bardes et al., 2024; Assran et al., 2025), that is not the objective here. DePhy is trained for a specified dynamical family with an explicit state definition and analytical transition. The role of joint-embedding prediction is to provide a reconstruction-free learning signal for that system-specific predictive state, not to learn a universal dynamics model across unrelated physical systems.

A.3 Why Hybrid Analytical–Learned Dynamics?

Known mechanics provide useful inductive structure but are rarely complete. In human motion, generalized equations specify how inertia, gravity, actuation, and contact forces determine acceleration, yet the required actuation and contact quantities are not directly observed from monocular kinematics. In the synthetic systems, the form of the dynamics is specified but physical coefficients vary from trajectory to trajectory. At the same time, a purely analytical model can be misspecified because of simplified contact, omitted drag or damping terms, inaccurate body parameters, or other modeling approximations.

We therefore separate uncertainty into two forms. Quantities whose *role* in the governing equations is known are inferred amortizedly and inserted into the analytical branch. Remaining discrepancy is represented by a learned residual. The residual is gated and regularized so that it acts as a correction to the known vector field rather than an unconstrained replacement. This yields a model that can exploit specified mechanics while retaining enough capacity to account for effects that the nominal equations do not capture.

The synthetic benchmarks isolate this design. The factor-identification experiments test whether the trajectory-level representation retains the varying physical factors, whereas the extrapolation experiments test whether the resulting transition remains stable beyond its training horizon. The human-motion residual-capacity ablation provides the complementary result: physics-only propagation is too restrictive, but a high-capacity residual or a neural-only transition also degrades long-horizon performance. The best result occurs at an intermediate bounded residual capacity.

A.4 Why Separate Data Assimilation from the Prior?

Even a well-trained transition accumulates error, and incoming measurements can contain information that is unavailable to the model during propagation. Classical filtering addresses this mismatch through prediction–correction: first propagate a prior, then update it using the observation. DePhy follows this structure in latent space (Revach et al., 2022; Buchnik et al., 2024). The learned gain α_t controls how strongly measurement-side evidence moves the propagated latent prior toward the current predictive measurement latent.

Crucially, assimilation is trained after the predictive prior is frozen. Measurement-specific correction is therefore confined to the assimilation path and cannot reshape the transition itself. The additional robustness experiments support this separation: the learned gain changes with measurement quality, and explicit assimilation remains substantially more robust than directly teacher-forcing the transition when forecast-time measurements are sparse or noisy.

B Extended Related Work

B.1 Physics-Based Human Motion Modeling

Physics-aware human-motion methods differ in where physical structure enters the pipeline, ranging from simulation and sequential physical correction to learned models regularized by mechanics (Zhang et al., 2024a;b; Le et al., 2024). PhysPT (Zhang et al., 2024a) derives force, contact, and Euler–Lagrange losses for pretraining a Transformer that refines kinematic motion and predicts motion forces. In contrast, the analytical equations in DePhy are part of the forward transition itself; PhysPT-style force/contact information is used only to warm-initialize the human dynamical modulator, while the core transition is subsequently learned through latent future prediction.

PhysMoP (Zhang et al., 2024b) combines physics-based and data-driven predictors for human-motion forecasting. Its objective is directly aligned with long-horizon motion prediction, but the data-driven component is not explicitly bounded relative to the physics branch. DePhy instead uses a gated residual whose capacity is controlled by $\gamma_{\max}$ and regularized relative to the analytical vector field. OSDCap (Le et al., 2024) combines a differentiable rigid-body simulator with sequential neural state correction and is the closest state-estimation comparison in the main experiments. DePhy differs in learning the predictive representation and physical prior directly from the measurement stream and in freezing that prior before learning measurement assimilation.

Other related work uses physical plausibility as an optimization or evaluation signal. EmLoco (Taketsugu et al., 2025) employs a differentiable surrogate for physical evaluation, while related physics-aware reconstruction methods use contact or interaction constraints to improve pose plausibility. These methods motivate the value of mechanical structure but do not address the specific combination studied here: corruption-robust predictive representation learning, instance-adaptive physical propagation, bounded model discrepancy, and latent data assimilation.

B.2 Kinematic Motion Prediction

Human-motion forecasting has been studied with recurrent models (Martinez et al., 2017; Fragkiadaki et al., 2015; Tanke et al., 2021), graph convolutional networks (Mao et al., 2019; Zhong et al., 2022), Transformers (Aksan et al., 2021; Mao et al., 2020), and MLP-based architectures (Guo et al., 2023). These methods learn strong statistical priors from pose sequences and remain important baselines for short- and medium-horizon prediction. Their transition, however, is defined primarily in observation or learned feature space, so mechanical structure must be recovered implicitly from the training distribution. The forecasting experiments in the main paper are therefore intended to test a complementary question: whether explicitly structured propagation can retain accuracy as the rollout horizon grows.

B.3 Latent Dynamics, Hybrid Models, and Modulation

Neural ODEs (Chen et al., 2018) and variants based on second-order, Hamiltonian, Lagrangian, controlled, recurrent, or energy-based dynamics (Yıldız et al., 2019; Norcliffe et al., 2020; Zhong et al., 2019; Lutter et al., 2019; Kidger et al., 2020; Rubanova et al., 2019; De Brouwer et al., 2019; Cheng et al., 2024) provide continuous-time parameterizations for learned state evolution. Hybrid models instead retain a mechanistic component and learn only missing or uncertain dynamics. APHYNITY (Yin et al., 2021) and augmented-physics-based models (Imbiriba et al., 2022; 2024) are representative examples of this grey-box formulation.

MoNODE (Auzina et al., 2023) is particularly relevant to the trajectory-specific inference component of DePhy. MoNODE amortizes latent modulators that explain between-trajectory variation and evaluates them using factor-of-variation probes (Eastwood & Williams, 2018). DePhy adopts the same broad principle of amortized instance-specific inference but places the inferred factors inside a known-physics transition and augments that transition with a bounded residual. This separates parameters that have a specified mechanistic role from discrepancy that is not represented by the nominal equations.

B.4 Joint-Embedding Predictive Representation Learning

Joint-Embedding Predictive Architectures learn representations by predicting target embeddings from incomplete, transformed, or temporally displaced context (LeCun, 2022; Assran et al., 2023; Bardes et al., 2024; Assran et al., 2025). Related work extends this principle to motion/content decomposition, skeleton representation learning, latent masked modeling, and structured temporal data (Bardes et al., 2023; Abdelfattah & Alahi, 2024; Wei et al., 2024; Sun et al., 2025; Li et al., 2025). DePhy uses joint-embedding prediction for a different purpose: the prediction target supervises a system-specific physical transition. Same-time prediction regularizes the measurement representation, while future latent prediction is reached only through the state encoder and augmented-physics rollout. Thus, the representation objective and the physical transition are coupled through prediction but remain decoupled from reconstruction.

B.5 Neural Filtering and Data Assimilation

Classical state estimation alternates model-based prediction and measurement update. Learned filtering methods retain this structure while replacing selected components with neural networks. KalmanNet (Revach et al., 2022) learns components of the Kalman update, and latent filtering methods perform prediction–correction in learned representation spaces (Buchnik et al., 2024; Guen & Thome, 2020). DePhy follows this family but differs in how the prior is obtained: the propagated latent is produced by a JEPA-trained, physics-structured transition that is frozen before the correc-

tion module is optimized. This makes the assimilation stage explicitly corrective rather than another route by which reconstruction gradients can reshape the prior.

C HUMAN MOTION AND DEPHY

This section gives the system-specific human-motion formulation, PhysPT-style warm initialization, training protocol, evaluation setup, and an expanded discussion of the human-motion results in the main paper.

C.1 NOTATION AND HUMAN STATE

Throughout the paper, M_{dyn} denotes the learned time-varying dynamical modulation network. In the multibody equations below, $\mathbf{M}(q;\hat{\psi})$ denotes the generalized inertia matrix. These symbols are unrelated. We use boldface $\mathbf{M}$ for the inertia matrix in the supplement to make the distinction explicit; the main-paper notation is unchanged.

For human motion, the predictive measurement latent is mapped to the generalized kinematic state

$$s_t^{\mathrm{H}} = \begin{bmatrix} q_t \\ \dot{q}_t \end{bmatrix} = E_{\mathrm{state}}(z_t^m), \tag{12}$$

where

$$q_t = \{r_t, R_t, q_t^r\}. \tag{13}$$

Here $r_t \in \mathbb{R}^3$ is global root translation, $R_t \in \mathbb{R}^3$ is the minimal root-rotation coordinate used by the kinematic representation, q_t^r contains articulated joint coordinates, and $\dot{q}_t$ is the corresponding generalized velocity. The state encoder therefore estimates kinematics; the dynamical modulator described next estimates quantities that determine how those kinematics evolve.

C.2 AMORTIZED PHYSICAL INFERENCE

Human dynamics contain unobserved quantities whose roles in the mechanical equations are known. We separate them into trajectory-level and time-varying variables.

Trajectory-level factors. The trajectory modulator predicts approximately sequence-stationary body and dynamics factors,

$$\hat{\psi} = M_{\mathrm{traj}}\left(z^m_{t-T_h+1:t}, s^{\mathrm{H}}_{t-T_h+1:t}\right). \tag{14}$$

The exact active entries depend on the analytical model and can include quantities controlling body scale, inertia, stiffness, damping, or contact response. The estimate is fixed for the rollout initialized at time t.

Time-varying dynamical quantities. The dynamical modulator predicts inputs whose values change over time but whose location in the governing equations is specified:

$$\hat{u}_\tau^{\mathrm{H}} = M_{\mathrm{dyn}}\left(s_\tau^{\mathrm{H}}, z_t^m, \hat{\psi}\right) = \{\hat{\tau}_\tau, \hat{\lambda}_\tau, \hat{c}_\tau\}, \qquad \tau \geq t. \tag{15}$$

Here $\hat{\tau}_\tau$ is generalized internal actuation, $\hat{\lambda}_\tau$ contains external/contact forces, and $\hat{c}_\tau$ parameterizes the active contact configuration. During open-loop propagation, z_t^m and $\hat{\psi}$ remain fixed context while M_{dyn} is reevaluated from the evolving state. No future teacher or ground-truth latent is supplied to the vector field.

C.3 GENERALIZED-COORDINATE FORWARD DYNAMICS

Conditioned on the inferred physical quantities, the known human branch follows generalized Newton–Euler dynamics,

$$\mathbf{M}(q_\tau;\beta,\hat{\psi})\ddot{q}_\tau + C(q_\tau,\dot{q}_\tau;\beta,\hat{\psi}) + g(q_\tau;\beta,\hat{\psi}) = \hat{\tau}_\tau + J_C(q_\tau,\hat{c}_\tau)^\top \hat{\lambda}_\tau, \tag{16}$$

where $\mathbf{M}$ is the generalized inertia matrix, C contains Coriolis and centrifugal terms, g is the generalized gravity term, and J_C is the contact Jacobian. The generalized-actuation vector $\hat{\tau}_\tau$ is zero on the unactuated floating-base coordinates.

Analytical body-dynamics terms. We instantiate the rigid-body terms following the Phys-SMPL formulation of PhysPT (Zhang et al., 2024a). The SMPL shape coefficients β determine the subject-specific geometry of the articulated body. PhysPT partitions the SMPL body into 24 rigid segments and derives their physical properties from the corresponding body geometry. In particular, the segment masses $m_n(\beta)$ and inertia tensors $I_n(\beta)$ are obtained from the shape-dependent segment volumes and mass distribution. Together with the current generalized configuration, these quantities analytically determine the rigid-body terms in Eq. equation 16. The generalized inertia matrix is assembled from the translational and rotational contributions of all body segments,

$$\mathbf{M}(q_\tau;\beta) = \sum_{n=1}^{24} \left[J_{S,n}(q_\tau)^\top m_n(\beta) J_{S,n}(q_\tau) + J_{R,n}(q_\tau)^\top R_n(q_\tau) I_n(\beta) R_n(q_\tau)^\top J_{R,n}(q_\tau) \right], \tag{17}$$

where $J_{S,n}$ and $J_{R,n}$ denote the translational and rotational Jacobians of body segment n, respectively, and R_n maps its local inertia tensor to the world frame. The generalized gravity term is similarly computed from the segment masses and translational Jacobians,

$$g(q_\tau;\beta) = -\sum_{n=1}^{24} J_{S,n}(q_\tau)^\top m_n(\beta)\, \mathbf{g}, \tag{18}$$

with $\mathbf{g}$ the gravity vector. The Coriolis and centrifugal term $C(q_\tau, \dot{q}_\tau; \beta)$ follows analytically from the same shape-dependent masses and inertias, body Jacobians, and their time derivatives under the Euler–Lagrange formulation. The contact Jacobian $J_C(q_\tau, \hat{c}_\tau)$ is obtained from the articulated SMPL kinematic chain at the active candidate contact locations.

Thus, the quantities $\mathbf{M}$, C, g, and J_C are not unconstrained network predictions: they are differentiable analytical quantities constructed from the current articulated state and the subject-specific SMPL body model. The learned dynamical modulator instead estimates the quantities not directly available from monocular kinematics,

$$\hat{u}_\tau^{\mathrm{H}} = M_{\mathrm{dyn}}\left(s_\tau^{\mathrm{H}}, z_t^m, \hat{\psi}\right) = \{\hat{\tau}_\tau, \hat{\lambda}_\tau, \hat{c}_\tau\}, \tag{19}$$

namely generalized actuation, external contact forces, and contact state. Additional trajectory-level quantities represented by $\hat{\psi}$ may condition the analytical model when they are not fixed by the SMPL body description. We refer the reader to PhysPT (Zhang et al., 2024a), in particular its Phys-SMPL construction and derivation of the Euler–Lagrange dynamics, for the full construction of the shape-dependent rigid-body quantities.

Notation. We use M_{dyn} for the learned time-varying dynamical modulation network, whereas $\mathbf{M}(q;\beta,\hat{\psi})$ denotes the generalized inertia (or mass) matrix in the equations of motion. These are unrelated quantities despite sharing the symbol M.

Solving Eq. equation 16 for generalized acceleration gives

$$\ddot{q}_\tau = \mathbf{M}(q_\tau;\beta,\hat{\psi})^{-1}\left[\hat{\tau}_\tau + J_C(q_\tau,\hat{c}_\tau)^\top \hat{\lambda}_\tau - C(q_\tau,\dot{q}_\tau;\beta,\hat{\psi}) - g(q_\tau;\beta,\hat{\psi})\right]. \tag{20}$$

The resulting known vector field is

$$f_K^{\mathrm{H}}\left(s_\tau^{\mathrm{H}};\beta,\hat{\psi},\hat{u}_\tau^{\mathrm{H}}\right) = \begin{bmatrix} \dot{q}_\tau \\ \mathbf{M}(q_\tau;\beta,\hat{\psi})^{-1}\left[\hat{\tau}_\tau + J_C(q_\tau,\hat{c}_\tau)^\top \hat{\lambda}_\tau - C(q_\tau,\dot{q}_\tau;\beta,\hat{\psi}) - g(q_\tau;\beta,\hat{\psi})\right] \end{bmatrix}. \tag{21}$$

For K candidate contact points, we stack the predicted per-contact forces as

$$\hat{\lambda}_\tau = \begin{bmatrix} \hat{f}_{1,\tau}^\top & \cdots & \hat{f}_{K,\tau}^\top \end{bmatrix}^\top, \tag{22}$$

with block contact Jacobian

$$J_C = \begin{bmatrix} J_1^\top & \cdots & J_K^\top \end{bmatrix}^\top. \tag{23}$$

The corresponding generalized contact force is therefore

$$J_C^\top \hat{\lambda}_\tau = \sum_{k=1}^{K} J_k^\top \hat{f}_{k,\tau}. \tag{24}$$

A total ground-reaction force may be reported as $\sum_k \hat{f}_{k,\tau}$, but the forward dynamics use the contact-wise generalized-force mapping above.

C.4 Augmented-Physics Transition and Learning without Direct Force Labels

The complete human transition is

$$\dot{s}_\tau^{\mathrm{H}} = f_K^{\mathrm{H}}\left(s_\tau^{\mathrm{H}}; \hat{\psi}, \hat{u}_\tau^{\mathrm{H}}\right) + \gamma_\tau \odot f_U\left(s_\tau^{\mathrm{H}}, z_t^m; \Psi\right), \tag{25}$$

where

$$\gamma_\tau = \gamma_{\max}\sigma\big(g_\gamma(s_\tau^{\mathrm{H}}, z_t^m)\big). \tag{26}$$

The learned branch captures discrepancy not represented by the nominal multibody model, including simplified contact and omitted musculoskeletal effects. Measurement noise is handled by the predictive measurement representation and assimilation pathway rather than being assigned to the residual.

A central point is that M_{dyn} does not require direct force labels in the core training objective. Its outputs enter the differentiable analytical dynamics in Eq. equation 16; these dynamics determine the propagated state, which is projected back to latent space and compared with future EMA targets. Consequently, $\mathcal{L}_{\mathrm{FUT}}$ supplies an indirect learning signal to M_{dyn}: incorrect actuation or contact estimates induce incorrect accelerations, which induce an incorrect future state and therefore a larger future-latent prediction loss. In shorthand,

$$M_{\mathrm{dyn}} \to (\hat{\tau}, \hat{\lambda}, \hat{c}) \to f_K \to s_{t+k}^p \to z_{t+k}^p \to \mathcal{L}_{\mathrm{FUT}}. \tag{27}$$

This is why the physical transition remains trainable even when no direct supervision is applied to the latent physical inputs. Direct physical supervision is used only as an optional warm initialization, described next.

C.5 PhysPT-Style Warm Initialization of the Dynamical Modulator

For the human-motion experiments, we warm-initialize M_{dyn} before Stage 1 using PhysPT-style physical pseudo-labels (Zhang et al., 2024a). This warm start is not part of $\mathcal{J}_{\mathrm{PhyCore}}$ in the main paper. After initialization, the transition is trained by the reconstruction-free Stage 1 objective.

Offline force pseudo-labels. Following PhysPT Zhang et al. (2024a), reference kinematics $\{q_t\}_{t=1}^T$ are differentiated to obtain $\dot{q}_t$ and $\ddot{q}_t$. PhysPT parameterizes contact forces using a continuous spring–damper model and solves a constrained inverse-dynamics problem of the form

$$\begin{aligned}(\bar{x}_t, \bar{\tau}_t) = \arg\min_{x_t, \tau_t} \ & \left\|\mathbf{M}_t\ddot{q}_t + C_t + g_t - J_{C,t}^\top A_t x_t - \tau_t\right\|_2^2, \\ \text{s.t. } & 0 < x_t < x_{\max},\end{aligned} \tag{28}$$

with recovered contact-force pseudo-label $\bar{\lambda}_t = A_t\bar{x}_t$. We use these quantities as weak physical targets; they are inferred from reference motion and the specified body/contact model rather than measured force-plate ground truth.

Force supervision. The direct supervision of the time-varying physical inputs is

$$\mathcal{L}_{\mathrm{force}} = \frac{1}{T}\sum_{t=1}^{T}\left[\gamma_\tau\|\hat{\tau}_t - \bar{\tau}_t\|_1 + \gamma_\lambda\|\hat{\lambda}_t - \bar{\lambda}_t\|_1\right]. \tag{29}$$

This term directly biases M_{dyn} toward the PhysPT-derived actuation and contact-force solution before the predictive-prior training stage.

Contact supervision. The recovered force solution also determines the set $\mathcal{C}_t$ of contact vertices. Following the PhysPT contact objective, we penalize motion of predicted contact vertices and their distance from the ground,

$$\mathcal{L}_{\mathrm{contact}} = \frac{1}{T}\sum_{t=1}^{T}\frac{1}{|\mathcal{C}_t|}\sum_{i\in\mathcal{C}_t}\left[\gamma_v\|v_{i,t}\|_1 + \gamma_z|h_{i,t}|\right], \tag{30}$$

where $v_{i,t}$ is the predicted contact-vertex velocity and $h_{i,t}$ is its vertical distance to the ground. During this warm-start phase the loss is used to shape the dynamical modulator through the differentiable human transition.

We define

$$\mathcal{L}_{\text{physup}} = \lambda_f \mathcal{L}_{\text{force}} + \lambda_c \mathcal{L}_{\text{contact}}. \tag{31}$$

PhysPT additionally uses an Euler–Lagrange consistency loss. We do not add it as a separate term in $\mathcal{L}_{\text{physup}}$ because the generalized equation of motion is already the forward operator of DePhy rather than a soft reconstruction regularizer. The ablation in Sec. E.1 evaluates the two warm-start terms used in our implementation.

C.6 Residual Regularization

Let

$$r_\tau = \gamma_\tau \odot f_U(s_\tau, z_t^m; \Psi) \tag{32}$$

denote the effective residual vector field. We regularize it as

$$\mathcal{R}_{\text{res}} = \frac{1}{T_f} \sum_{k=0}^{T_f-1} \frac{\|r_{t+k}\|_2^2}{\|f_K(s_{t+k}^p; \phi_{\text{fix}}, \hat{\phi}, \hat{u}_{t+k})\|_2^2 + \epsilon} + \beta_\gamma \frac{1}{T_f} \sum_{k=0}^{T_f-1} \|\gamma_{t+k}\|_1, \qquad \beta_\gamma = 0.2. \tag{33}$$

The first term penalizes a large learned correction relative to the analytical vector field, while the second discourages persistently open residual gates. This is a soft inductive bias, not a hard guarantee that $\|r_\tau\| < \|f_K\|$ at every frame.

C.7 Training Procedure

Warm initialization. For human motion only, M_{dyn} is initialized with the PhysPT-style objective in Eq. equation 31. This phase supplies a physically meaningful starting point for generalized actuation and contact inference, but it is not required by the core formulation and is ablated separately.

Stage 1: predictive-prior learning. We then optimize the core objective from the main paper,

$$\mathcal{J}_{\text{PhyCore}} = \mathcal{L}_{\text{JEPA}} + \mathcal{L}_{\text{FUT}} + \mathcal{R}_{\text{res}}. \tag{34}$$

This stage updates the measurement and state encoders, measurement predictor, active physical-inference modules, residual dynamics and gate, and latent projection head. The EMA measurement encoder is updated by momentum. Neither the assimilation network A_η nor decoder D_{meas} is trained, and no reconstruction loss is applied.

Stage 2: assimilation and reconstruction. The complete Stage 1 prior is frozen. We optimize only A_η and D_{meas} with

$$\mathcal{J}_{\text{DA}} = \mathcal{L}_{\text{rec}} + \mathcal{R}_{\text{stay}}. \tag{35}$$

Because the prior is frozen, reconstruction gradients cannot reshape the predictive representation, state encoder, physical-inference modules, or transition.

C.8 Architecture and Implementation Details

Input measurements and preprocessing. Following OSDCap (Le et al., 2024), we use TRACE (Sun et al., 2023) as the fixed upstream monocular motion estimator. The visual front-end is not updated during DePhy training, so improvements are attributable to the state-estimation model rather than changes in the image-to-motion estimator. The resulting structured kinematic sequence is treated as the measurement stream $\tilde{y}_{1:T}$. Estimated motions are downsampled from 100 Hz to 50 Hz, aligned to the world origin at the first frame, and divided into 100-frame subsequences for training and evaluation. The contextual measurement model operates on windows of length $T_h = 16$.

Predictive measurement representation. For human motion, E_{meas} follows the PhysPT-style temporal Transformer architecture (Zhang et al., 2024a). The input kinematic vector at each time step is linearly projected to a 1024-dimensional embedding and augmented with sinusoidal temporal positional encodings. The contextual backbone consists of 6 Transformer layers with 8 attention heads, embedding dimension 1024, and dropout 0.1.

We train the measurement representation using a standard JEPA student–teacher construction (LeCun, 2022; Assran et al., 2023; Bardes et al., 2024). Given an upstream measurement history $\tilde{y}_{t-T_h+1:t}$, the student receives an additionally corrupted view

$$\bar{y} = \mathcal{C}(\tilde{y}; m, \epsilon), \tag{36}$$

where m denotes structured masking and ϵ denotes additional measurement noise. The masking and noising severity are increased progressively during training using a curriculum. The target encoder $E^{\text{EMA}}_{\text{meas}}$ has the same architecture as the student but receives the corresponding upstream measurement sequence without the additional student-side corruption. Hence, the teacher is not given ground-truth motion: both pathways originate from the same imperfect upstream measurements. The target encoder is not optimized by gradient descent and is updated as an exponential moving average of the student parameters,

$$\theta_{\text{EMA}} \leftarrow \mu\theta_{\text{EMA}} + (1-\mu)\theta_{\text{meas}}, \tag{37}$$

with stop-gradient applied to its output.

The student-side measurement predictor is a lightweight token-wise MLP,

$$\text{Pred}_{\text{meas}} : 1024 \to 512 \to 1024, \tag{38}$$

with LayerNorm at the input and GELU activation in the hidden layer. The predictive measurement latent is therefore

$$z^m_t = \text{Pred}_{\text{meas}}\left(E_{\text{meas}}(\bar{y}_{t-T_h+1:t})\right), \tag{39}$$

whereas the target representation is taken directly from the EMA encoder. No additional temporal model is used in $\text{Pred}_{\text{meas}}$, since temporal contextualization is already performed by E_{meas}.

Physical-state encoder. The predictive measurement representation is lifted into the explicit generalized physical state

$$s^{\text{H}}_t = E_{\text{state}}(z^m_t) = [q_t, \dot{q}_t]. \tag{40}$$

E_{state} consists of a shared $1024 \to 512 \to 512$ MLP trunk with LayerNorm at the input and GELU nonlinearities, followed by separate linear heads for q_t and $\dot{q}_t$:

$$q_t = W_q h^s_t + b_q, \qquad \dot{q}_t = W_{\dot{q}} h^s_t + b_{\dot{q}}. \tag{41}$$

The generalized configuration contains the floating-base motion and articulated joint coordinates used by the human dynamics model. No temporal layers are introduced inside E_{state}; temporal information is supplied by the contextual measurement representation, while subsequent temporal evolution is handled by the physical transition.

Trajectory-level physical modulation. M_{traj} estimates physical quantities that are approximately constant over a rollout. For each context frame, the predictive measurement latent and physical state are independently projected to 256 dimensions,

$$h^z_i = \text{GELU}\left(W_z\,\text{LN}(z^m_i)\right), \qquad h^s_i = \text{GELU}\left(W_s s^{\text{H}}_i\right), \tag{42}$$

and concatenated as $h_i = [h^z_i, h^s_i]$. We summarize the context using element-wise temporal mean and standard deviation,

$$h_{\text{traj}} = \left[\text{Mean}_i(h_i), \text{Std}_i(h_i)\right]. \tag{43}$$

The resulting 1024-dimensional summary is processed by

$$M_{\text{traj}} : 1024 \to 512 \to 256 \to d_\psi, \tag{44}$$

with GELU activations, producing

$$\hat{\psi} = M_{\text{traj}}\left(z^m_{t-T_h+1:t}, s^{\text{H}}_{t-T_h+1:t}\right). \tag{45}$$

$\hat{\psi}$ is held fixed during the rollout initialized from the corresponding context.

Time-varying dynamical modulation. The dynamical modulator estimates quantities whose mechanical role is known but whose instantaneous values are not directly observed:

$$\hat{u}_\tau^{\mathrm{H}} = M_{\mathrm{dyn}}\left(s_\tau^{\mathrm{H}}, z_t^m, \hat{\psi}\right) = \{\hat{\tau}_\tau, \hat{\lambda}_\tau, \hat{c}_\tau\}. \tag{46}$$

The current state s_τ^{H}, fixed measurement context z_t^m, and trajectory factor $\hat{\psi}$ are projected to 256, 256, and 128 dimensions, respectively. Their concatenation is processed by a shared $640 \to 512 \to 512$ MLP with GELU activations. Separate output heads predict generalized actuation, contact forces, and contact probability:

$$\hat{\tau}_\tau = W_\tau h_\tau^{\mathrm{dyn}} + b_\tau, \tag{47}$$

$$\tilde{\lambda}_\tau = W_\lambda h_\tau^{\mathrm{dyn}} + b_\lambda, \tag{48}$$

$$\hat{c}_\tau = \sigma(W_c h_\tau^{\mathrm{dyn}} + b_c). \tag{49}$$

For K candidate contact locations, the predicted contact force is gated by the corresponding contact probability,

$$\hat{f}_{k,\tau} = \hat{c}_{k,\tau}\tilde{f}_{k,\tau}, \qquad \hat{\lambda}_\tau = [\hat{f}_{1,\tau}^\top, \ldots, \hat{f}_{K,\tau}^\top]^\top. \tag{50}$$

The generalized-actuation entries corresponding to the unactuated floating base are fixed to zero. Unlike M_{traj}, M_{dyn} is reevaluated from the evolving state at each ODE evaluation point; z_t^m and $\hat{\psi}$ remain fixed over the rollout.

Analytical human dynamics. The known branch is instantiated using the Phys-SMPL rigid-body construction of PhysPT (Zhang et al., 2024a). SMPL shape parameters determine the subject-specific body geometry from which segment masses and inertia tensors are obtained. Together with the current generalized configuration, these quantities determine the generalized inertia matrix $\mathbf{M}(q)$, Coriolis/centrifugal term $C(q, \dot{q})$, generalized gravity $g(q)$, and contact Jacobians $J_C(q)$ analytically. Thus, $\mathbf{M}$, C, g, and J_C are not predicted by neural networks; M_{dyn} instead estimates the unobserved actuation and contact quantities entering these equations. We refer to PhysPT for the complete Phys-SMPL and Euler–Lagrange derivation.

We use M_{dyn} to denote the learned dynamical modulation network and $\mathbf{M}(q)$ to denote the generalized inertia matrix; these are unrelated quantities despite the shared letter M.

Augmented-physics transition. The analytical transition is augmented with a learned residual vector field,

$$\dot{s}_\tau = f_K\left(s_\tau; \hat{\psi}, \hat{u}_\tau\right) + \gamma_\tau \odot f_U(s_\tau, z_t^m; \Psi), \tag{51}$$

where f_U is implemented as a lightweight MLP with two 512-dimensional hidden layers and GELU activations. The residual gate is predicted by a separate MLP g_γ from the current state and measurement latent,

$$\gamma_\tau = \gamma_{\max}\sigma\left(g_\gamma(s_\tau, z_t^m)\right), \tag{52}$$

with $\gamma_{\max} = 0.30$ for the human-motion experiments. We additionally regularize the effective residual relative to the analytical vector field and penalize persistently open gates, using $\beta_\gamma = 0.2$. The resulting state derivative is propagated with a differentiable ODE solver.

Latent projection. After physical propagation, Π_z maps the propagated state back into the same latent space used by the predictive measurement representation:

$$\Pi_z : d_s \to 512 \to 1024, \qquad z_{t+k}^p = \Pi_z(s_{t+k}^p). \tag{53}$$

Π_z uses a GELU activation in its hidden layer. Future EMA representations supervise z_{t+k}^p through $\mathcal{L}_{\mathrm{FUT}}$, but these future targets never enter the physical vector field or rollout state.

Physical supervision of the dynamical modulator. For human motion, M_{dyn} is warm-initialized for 20 epochs using PhysPT-style physical pseudo-supervision. Reference kinematics are processed with the Phys-SMPL inverse-dynamics formulation to obtain supervisory actuation/contact quantities. We write

$$\mathcal{L}_{\mathrm{physup}} = \lambda_{\mathrm{force}}\mathcal{L}_{\mathrm{force}} + \lambda_{\mathrm{contact}}\mathcal{L}_{\mathrm{contact}}, \tag{54}$$

where $\mathcal{L}_{\mathrm{force}}$ supervises the inferred generalized actuation and contact forces and $\mathcal{L}_{\mathrm{contact}}$ encourages physically consistent contact predictions. This supervision initializes the physical-inference pathway but is separate from the final DePhy predictive objective and is not required at inference.

Stage 1: predictive-prior learning. Stage 1 is trained for 150 epochs using curriculum optimization. Training initially emphasizes the JEPA measurement representation under progressively stronger masking and noising, after which the physical-inference modules are introduced and the analytical transition is trained through future latent prediction. Gradients from $\mathcal{L}_{\mathrm{FUT}}$ pass through the differentiable known-physics transition into E_{state}, M_{traj}, M_{dyn}, and Π_z; the analytical form of f_K itself remains fixed. The residual branch is introduced only after the known-physics pathway has been established, after which both analytical and residual pathways participate jointly in predictive-prior learning. We use curriculum loss weights for the future prediction and residual terms rather than applying their terminal weights from the beginning of optimization. The resulting Stage 1 objective is

$$\mathcal{J}_{\mathrm{PhyCore}} = \mathcal{L}_{\mathrm{JEPA}} + \lambda_{\mathrm{FUT}}\mathcal{L}_{\mathrm{FUT}} + \lambda_{\mathrm{res}}\mathcal{R}_{\mathrm{res}}. \tag{55}$$

The EMA teacher is updated throughout Stage 1 by momentum. No reconstruction loss is allowed to update the predictive representation or transition prior. At convergence, the complete PhyCore—E_{meas}, $\mathrm{Pred}_{\mathrm{meas}}$, E_{state}, M_{traj}, M_{dyn}, f_U, g_γ, and Π_z—is frozen.

Assimilation and reconstruction. Stage 2 trains only the latent assimilation module A_η and measurement decoder D_{meas}. Given the propagated prior and current measurement representation, the assimilation input is

$$h_t^A = [z_t^p,\, z_t^m,\, z_t^m - z_t^p]. \tag{56}$$

A_η is a lightweight MLP operating on this concatenated representation and predicts an element-wise gain

$$\alpha_t = \sigma\left(A_\eta(h_t^A)\right), \qquad \alpha_t \in [0,1]^{1024}. \tag{57}$$

The assimilated latent is

$$z_t^* = z_t^p + \alpha_t \odot (z_t^m - z_t^p). \tag{58}$$

D_{meas} is a $1024 \to 512 \to d_y$ MLP with GELU activation and maps z_t^* back to the structured kinematic measurement space. Stage 2 optimizes

$$\mathcal{J}_{\mathrm{DA}} = \mathcal{L}_{\mathrm{rec}} + \mathcal{R}_{\mathrm{stay}}, \tag{59}$$

with unit terminal weight on the prior-anchoring term. Since the complete PhyCore is frozen, reconstruction gradients cannot reshape either the learned measurement representation or the physical transition.

Inference. No parameters are updated at test time. For sequential motion estimation, the current unaugmented TRACE measurements produce z_t^m, while the frozen augmented-physics transition produces z_t^p; the two are combined by A_η before decoding. For long-horizon forecasting, the transition is rolled out autoregressively. The w/o-DA variant performs a pure transition-prior rollout, while the DA variant recursively feeds model-generated measurement-side evidence through the frozen measurement pathway before applying the same prediction–correction update. No future ground-truth measurements are used.

Dataset-specific implementation and compute. For SportsPose, the Human3.6M-trained model is used for initialization and fine-tuned for 15 epochs following the OSDCap protocol. The upstream TRACE model remains fixed across all datasets. The complete Human3.6M model contains approximately 178M parameters and trains in approximately 14 hours on a single consumer GPU; differentiable ODE integration is the primary computational bottleneck.

C.9 Human-Motion Datasets and Protocols

Human3.6M. Human3.6M (Ionescu et al., 2014) contains seven actors performing 15 everyday actions. Following OSDCap and established human-motion protocols (Martinez et al., 2017; Tanke et al., 2021; Shimada et al., 2021; Li et al., 2022a; Le et al., 2024), we retain action sequences with valid foot–ground contact; S1, S5, S6, S7, and S8 are used for training and S9 and S11 for evaluation.

Fit3D. For Fit3D (Fieraru et al., 2021), we follow the OSDCap dataset protocol: sequences with foot–ground contact from s03, s04, s05, s07, s08, and s10 are used for training, while s09 and s11 are used for evaluation. We therefore use Fit3D as a domain-shift benchmark under its corresponding training/evaluation split rather than describing it as a zero-shot transfer experiment.

SportsPose. For SportsPose (Ingwersen et al., 2023), S02, S03, S05, S06, S07, S08, and S09 are used for fine-tuning, and S12, S13, and S14 for evaluation, following OSDCap. The model is initialized from Human3.6M and fine-tuned for 15 epochs. SportsPose produces substantially noisier upstream monocular measurements than Human3.6M or Fit3D and therefore provides the strongest measurement-domain shift among the human benchmarks.

C.10 METRICS

For state estimation we report MPJPE, global MPJPE (MPJPE-G), Procrustes-aligned MPJPE (MPJPE-PA), global-root error, acceleration error (Accel), and foot-skating rate (FS), following OSDCap (Le et al., 2024). MPJPE evaluates joint accuracy after root alignment, MPJPE-G evaluates global joint accuracy, MPJPE-PA evaluates pose after rigid alignment, the global-root metric evaluates root-translation accuracy, Accel measures temporal jitter through second-order differences, and FS measures implausible foot motion during contact. The main Human3.6M table denotes the root metric as GRP, whereas the cross-dataset table uses GRE; we refer to both generically as global-root error in the supplementary discussion.

For forecasting, we follow PhysMoP (Zhang et al., 2024b) and report MPJPE at 80, 160, 320, 400, 560, 720, 880, and 1000 ms. Human3.6M-P1 reports error at the target horizon, while P2 averages error over the trajectory up to the target horizon.

C.11 FORECASTING PROTOCOL AND THE ROLE OF ASSIMILATION

The standard forecasting results in Table 3 use no future ground-truth measurements. DePhy w/o DA performs an autoregressive rollout of the frozen transition prior. For DePhy w DA, the current model-generated forecast is fed back through the measurement pathway to construct the next predictive measurement latent, and the learned assimilation update is applied recursively. Thus the w-DA result does not receive external future observations; it tests whether the learned prediction–correction loop can stabilize a rollout using only model-generated measurement-side evidence.

This protocol is distinct from the additional noisy/sparse-measurement experiment in Sec. E.4, where forecast-time measurements are deliberately supplied at selected time steps to study robustness of data assimilation under partial observation.

C.12 EXPANDED DISCUSSION OF MAIN HUMAN-MOTION RESULTS

Sequential state estimation. Table 1 shows that DePhy improves over OSDCap on every reported Human3.6M metric. MPJPE decreases from 54.8 to 47.3 mm, MPJPE-G from 132.8 to 116.4 mm, MPJPE-PA from 39.8 to 32.7 mm, and acceleration error from 8.4 to 3.9 mm/s^2. The simultaneous reduction in pose and acceleration error is important: the improvement is not obtained by smoothing at the expense of positional accuracy. The foot-skating rate also decreases from 15.2% to 6.2%, consistent with the contact-aware physical transition and PhysPT-style warm initialization.

Relative to QuaMoTRACE, DePhy improves MPJPE, MPJPE-G, MPJPE-PA, Accel, and FS, while QuaMoTRACE has a slightly lower Human3.6M root-position error. This suggests that the largest gains of the proposed model occur in articulated pose, temporal smoothness, and contact plausibility rather than from uniformly improving every global-trajectory quantity.

Fit3D and SportsPose. Table 2 shows that the same behavior persists under dataset and measurement shifts. On Fit3D, DePhy reduces OSDCap MPJPE from 58.7 to 50.1 mm and Accel from 8.2 to 4.4 mm/s^2; on SportsPose the corresponding values decrease from 71.7 to 67.5 mm and from 10.9 to 4.8 mm/s^2. The improvement in acceleration is especially large on SportsPose, where the upstream measurements are the noisiest. Against QuaMo, DePhy gives better MPJPE, MPJPE-G, MPJPE-PA, and FS on SportsPose, while QuaMo has lower GRE. On Fit3D, QuaMo has lower MPJPE-PA and Accel, whereas DePhy improves MPJPE, MPJPE-G, GRE, and FS. These comparisons indicate that the advantage is not a uniform dominance of every metric, but a consistent improvement in the overall state-estimation tradeoff under corrupted kinematics.

Long-horizon forecasting. Table 3 separates short-horizon fit from rollout stability. PhysMoP is better at the earliest horizons, but DePhy w DA becomes better from 160 ms onward on P1 and from

320 ms onward on P2. At 1000 ms, P1 MPJPE is 90.8 mm versus 103.9 mm for PhysMoP, and P2 is 48.9 mm versus 57.1 mm. The pure DePhy prior already reaches 97.7 mm on P1 and 52.4 mm on P2, showing that the transition itself accounts for most of the long-horizon improvement. Recursive assimilation further reduces the 1-s error to 90.8 and 48.9 mm, respectively.

D Synthetic Setups

The synthetic benchmarks isolate two properties that are difficult to measure directly in monocular human motion: whether the trajectory representation contains the physical factors that vary between sequences, and whether the learned transition extrapolates beyond the horizon used during training. Both systems are simulated at 10 Hz.

D.1 Evaluation Protocol and Metrics

We compare with NODE (Chen et al., 2018) and MoNODE (Auzina et al., 2023) on both benchmarks. The pendulum experiment additionally includes GRU (Chung et al., 2014), LNN (Cranmer et al., 2020), APHYNITY (Yin et al., 2021), Phys-VAE (Takeishi & Kalousis, 2021), and a fitted Known-ODE reference. We use official implementations for NODE, LNN, and APHYNITY where available and reimplement MoNODE in PyTorch for the present setting. APHYNITY uses the authors' default configuration; remaining trainable baselines are selected by validation-set grid search. DePhy-NoRes removes the learned residual and propagates with the analytical branch alone.

Forecasting is evaluated against the clean simulated trajectory using

$$\mathrm{MSE}(H) = \frac{1}{N|\mathcal{T}_H|d_y} \sum_{n=1}^{N} \sum_{t\in\mathcal{T}_H} \|\hat{y}_{n,t} - y_{n,t}\|_2^2, \tag{60}$$

where $\mathcal{T}_H$ denotes predicted indices up to horizon H and d_y is observation dimension.

For factor-of-variation (FoV) evaluation, the model is frozen and a linear probe is fit from the learned trajectory-level representation to the known simulator factors using the training/validation data; the untouched test set is used only for evaluation. We report factor-wise

$$R_j^2 = 1 - \frac{\sum_n (\phi_{n,j} - \hat{\phi}_{n,j})^2}{\sum_n (\phi_{n,j} - \bar{\phi}_j)^2}. \tag{61}$$

This follows the representation-evaluation protocol of MoNODE and disentanglement-style probing (Auzina et al., 2023; Eastwood & Williams, 2018).

For methods that explicitly estimate pendulum parameters, we additionally report

$$\epsilon_\phi = \frac{100}{|\mathcal{P}|} \sum_{j\in\mathcal{P}} \frac{|\hat{\phi}_j - \phi_j|}{|\phi_j|}, \tag{62}$$

with the same definition used in the main-paper table. Lower values indicate more accurate parameter estimates.

D.2 Three-Dimensional Bouncing Ball

State and observations. The simulator state is

$$s_t^{\mathrm{B}} = [p_t, v_t, \omega_t] \in \mathbb{R}^9, \qquad p_t, v_t, \omega_t \in \mathbb{R}^3, \tag{63}$$

where p_t is position, v_t translational velocity, and ω_t angular velocity. Only position is observed:

$$y_t = p_t, \qquad \tilde{y}_t = y_t + \epsilon_t, \qquad \epsilon_t \sim \mathcal{N}(0, 0.02^2 I_3). \tag{64}$$

Velocity and spin therefore influence the trajectory without being directly observed.

Trajectory-specific factors and dynamics. Each trajectory uses

$$\phi^{\mathrm{B}} = [g, C_D, C_L, c_\omega, e, \mu, k_\omega], \tag{65}$$

where g is gravity, C_D quadratic drag, C_L lift/Magnus coefficient, c_ω spin decay, e restitution, μ tangential collision friction, and k_ω collision-induced spin coupling. Mass and radius are fixed.

Between contacts,

$$\dot{p}_t = v_t, \tag{66}$$
$$m\dot{v}_t = F_g + F_D + F_M, \tag{67}$$
$$\dot{\omega}_t = -c_\omega \omega_t, \tag{68}$$

with

$$F_g = [0, 0, -mg]^\top, \tag{69}$$
$$F_D = -\frac{1}{2}\rho C_D A \|v_t\|_2 v_t, \tag{70}$$
$$F_M = \frac{1}{2}\rho C_L A r(\omega_t \times v_t), \tag{71}$$

where $A = \pi r^2$. Trajectories are integrated with a Runge–Kutta solver and recorded at 10 Hz.

Collision model. The table is the plane $p_z = 0$. At downward contact, define $n = [0, 0, 1]^\top$, $v_n = (v \cdot n)n$, and $v_\tau = v - v_n$. We apply

$$v_n^+ = -e v_n, \tag{72}$$

$$a_{\mathrm{fric}} = \min\left(\frac{\mu(1+e)|v \cdot n|}{\|v_\tau\|_2 + \epsilon}, 1\right), \qquad v_\tau^+ = (1 - a_{\mathrm{fric}})v_\tau, \tag{73}$$

and

$$\omega^+ = \omega + k_\omega n \times (v_\tau - v_\tau^+). \tag{74}$$

The post-contact velocity is $v^+ = v_n^+ + v_\tau^+$. No time-varying M_{dyn} head is required for this system; the amortized inference problem is trajectory-level factor estimation.

Data and forecasting. We generate $1{,}000/100/100$ train/validation/test trajectories. Training and validation trajectories have length $T = 20$. Test trajectories are evaluated at $T = 20$ and extended to $T = 40$ for extrapolation. Each model receives the first 10 observed steps as context. Results are averaged over three random seeds.

Expanded result discussion. Table 4 shows that DePhy recovers all seven varying factors more strongly than NODE and MoNODE, with mean FoV $R^2 = 0.82$ compared with 0.45 for MoNODE. The factor-wise gains are not concentrated in a single easy parameter: DePhy obtains R^2 values between 0.77 and 0.89 across gravity, aerodynamic, spin, restitution, and friction factors. Removing the learned residual reduces the mean to 0.70, indicating that modest model-discrepancy capacity improves the trajectory representation even when the nominal equations already contain the dominant physical structure.

The forecasting result shows the same ordering. At the training horizon $T = 20$, DePhy reaches MSE 0.0112, compared with 0.0138 for DePhy-NoRes and 0.0164 for MoNODE. At the extrapolation horizon $T = 40$, the gap grows to 0.0317 versus 0.0421 and 0.0512, respectively. The larger separation outside the training horizon supports the intended role of the analytical branch and bounded residual as a structured extrapolative prior rather than only an in-distribution sequence predictor.

D.3 Planar Damped Pendulum

The state is

$$s_t^{\mathrm{P}} = [\theta_t, \omega_t], \qquad \omega_t = \dot{\theta}_t, \tag{75}$$

Table 7: **Ablation of PhysPT-style physical supervision on Human3.6M.** The core DePhy objective is fixed across all variants; only the components of the warm-start objective $\mathcal{L}_{\text{physup}}$ are varied. Lower is better.

$\mathcal{L}_{\text{physup}}$		Motion Estimation			
$\mathcal{L}_{\text{force}}$	$\mathcal{L}_{\text{contact}}$	MPJPE↓ [mm]	MPJPE-G↓ [mm]	Accel↓ [mm/s^2]	FS↓ [%]
×	×	49.8	121.5	6.4	8.8
✓	×	48.5	118.1	5.2	7.1
✓	✓	47.3	116.4	3.9	6.2

and the nominal dynamics are

$$\dot{\theta}_t = \omega_t, \qquad \dot{\omega}_t = -\frac{g}{\ell}\sin\theta_t - c\omega_t. \tag{76}$$

For each trajectory we sample

$$\ell \sim \mathcal{U}(0.5, 2.0), \qquad g \sim \mathcal{U}(7.0, 12.0), \qquad c \sim \mathcal{U}(0.0, 0.5). \tag{77}$$

Observations are

$$y_t = [\theta_t, \omega_t], \qquad \tilde{y}_t = y_t + \epsilon_t, \qquad \epsilon_t \sim \mathcal{N}(0, 0.01^2 I_2). \tag{78}$$

Trajectories are integrated at $\Delta t = 0.1$ s. We generate $800/100/100$ train/validation/test trajectories. Training uses $T_h = 20$ context steps and $T_f = 20$ prediction steps; at test time we additionally roll out for $T_f = 40$ future steps. Results are averaged over three seeds.

Identifiability note. Equation equation 76 depends on g and ℓ through the ratio g/ℓ. Accordingly, the factor-wise probe scores in Table 5 should be interpreted as empirical recoverability of the simulator generator variables $[\ell, g, c]$ under the finite benchmark distribution, not as a claim of structural identifiability of g and ℓ from the angular ODE alone. The physical propagation uses the corresponding dynamical combination, while the FoV probe evaluates what information about the sampled generator factors is retained in the learned trajectory representation.

Expanded result discussion. Table 5 shows a mean FoV $R^2 = 0.89$ for DePhy, compared with 0.66 for APHYNITY, 0.57 for MoNODE, and 0.76 for DePhy-NoRes. The improvement over the no-residual variant mirrors the bouncing-ball result, suggesting that a bounded discrepancy model improves factor-aware representation learning even when the analytical dynamics are strong.

For forecasting, DePhy reaches MSE 0.0015 at $T_f = 20$, nearly matching the fitted Known-ODE reference at 0.0014, and obtains 0.0059 at the $T_f = 40$ extrapolation horizon. The fitted Known-ODE remains best at long horizon with 0.0038, as expected when the model class matches the nominal simulator closely. Among learned baselines, however, DePhy has the lowest extrapolation error and the lowest reported parameter error, 1.7%, compared with 1.9% for the fitted Known-ODE, 2.3% for APHYNITY, and 2.8% for Phys-VAE. Together with the no-residual comparison, this indicates that the learned residual provides useful flexibility without erasing the trajectory-level physical information captured by the analytical branch.

E Additional Experiments

This section expands the ablations summarized in the main paper and analyzes the behavior of the physical-supervision warm start, residual branch, and assimilation module.

E.1 PhysPT-Style Physical-Supervision Ablation

Table 7 clarifies the role of the PhysPT-style warm start. Even without $\mathcal{L}_{\text{physup}}$, the model reaches 49.8 mm MPJPE because M_{dyn} continues to receive a learning signal through the differentiable transition and $\mathcal{L}_{\text{FUT}}$. Adding force supervision improves both positional and temporal metrics, reducing MPJPE to 48.5 mm and Accel to 5.2 mm/s^2. Adding the contact term gives the full 47.3 mm

Table 8: **Ablation of transition structure and residual capacity on Human3.6M forecasting.** The neural-only variant removes the known-physics branch. For physics-based variants, $\dot{s}_t = f_K(s_t;\Phi) + \gamma_t \odot f_U(s_t, z_t^m;\Psi)$ with $\gamma_t = \gamma_{\max}\sigma(g_\gamma(s_t, z_t^m))$. $\gamma_{\max} = 0$ is physics-only propagation; increasing $\gamma_{\max}$ increases residual capacity. MPJPE is reported in mm. Lower is better.

Variant	**80**	**160**	**320**	**400**	**560**	**720**	**880**	**1000**
Neural-only (MLP forward)	3.5	8.6	31.4	45.7	80.6	96.8	112.7	125.4
$\gamma_{\max} = 0.00$ (physics only)	4.5	12.4	27.8	40.9	76.6	91.7	105.5	119.2
$\gamma_{\max} = 0.15$	3.2	6.5	23.9	34.7	64.8	76.5	90.6	96.8
$\gamma_{\max} = 0.30$ (DePhy)	**2.5**	**5.3**	**21.2**	**30.9**	**59.7**	**73.4**	**83.9**	**90.8**
$\gamma_{\max} = 0.50$	3.0	6.1	22.4	34.9	64.1	78.9	93.5	101.7
High residual capacity, $f_K + f_U$	3.9	7.6	25.8	38.9	70.1	84.5	109.4	123.1

Table 9: **Residual and assimilation diagnostics on Human3.6M.** Distributional statistics are computed over the evaluation set. ρ_t is the magnitude of the effective learned residual relative to the analytical vector field.

Quantity	**Mean**	P_{25}	**Median**	P_{75}	P_{95}	$P(>1)$
Residual gate γ_t	0.13	0.06	0.11	0.15	0.24	–
Assimilation gain α_t	0.38	0.25	0.34	0.51	0.78	–
Relative residual ρ_t	0.234	0.05	0.19	0.48	1.10	0.08

MPJPE, 3.9 mm/s^2 acceleration error, and 6.2% foot-skating rate. The larger relative improvement in Accel and FS is consistent with the purpose of the auxiliary supervision: it provides a better initialization for the unobserved actuation/contact variables, while the subsequent Stage 1 objective remains reconstruction-free.

E.2 Transition Structure and Residual Capacity

Table 8 shows a non-monotonic relationship between residual capacity and forecasting accuracy. Physics-only propagation is more stable than the neural-only transition at 1 s (119.2 versus 125.4 mm) but is too restrictive at short and medium horizons. Introducing a bounded residual substantially improves performance, with $\gamma_{\max} = 0.30$ giving the best result across all reported horizons. Increasing the bound to 0.50 degrades the 1-s error to 101.7 mm, while removing the effective bound through a high-capacity $f_K + f_U$ transition degrades it further to 123.1 mm. The result supports the intended hybrid interpretation: the learned branch is useful when it corrects model discrepancy, but excessive residual capacity weakens the stabilizing bias of the analytical dynamics.

E.3 Residual and Assimilation Diagnostics

We define

$$\rho_t = \frac{\|\gamma_t \odot f_U(s_t, z_t^m;\Psi)\|_2}{\|f_K(s_t;\Phi)\|_2 + \varepsilon}. \tag{79}$$

Table 9 shows that the residual gate is usually small (mean 0.13, median 0.11, $P_{95} = 0.24$). The relative residual has mean 0.234 and median 0.19; 75% of frames satisfy $\rho_t \le 0.48$, and only 8% exceed 1. Thus, although the residual can become locally strong when the analytical model is insufficient, it is typically a correction rather than the dominant vector field.

The assimilation gain has a broader distribution (mean 0.38, median 0.34, $P_{25} = 0.25$, $P_{75} = 0.51$, $P_{95} = 0.78$). This spread is expected for a learned prediction–correction rule: the model does not use a fixed interpolation coefficient between the propagated prior and measurement evidence, but changes the correction strength across frames.

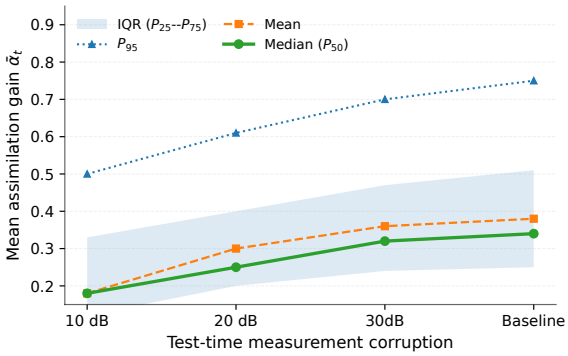


Figure 5: **Assimilation gain under increasing test-time measurement corruption.** Distribution of the learned assimilation gain α_t for the baseline measurements and additive noise levels of 30, 20, and 10 dB. The shaded region denotes the interquartile range (P_{25}–P_{75}); curves show the mean, median (P_{50}), and P_{95}. As corruption increases, the gain distribution shifts downward, showing that the assimilation network suppresses measurement-side correction when the supplied evidence becomes less reliable.

Table 10: **Robustness of data assimilation to noisy and sparse forecast-time measurements.** MPJPE at the 1000 ms horizon while varying measurement quality and forecast-time availability. Both variants receive identical measurements at identical time steps. *Prior only (TF)* directly teacher-forces the transition with available measurements, whereas *w/ DA* fuses them with the propagated prior. Lower is better.

SNR	Method	Forecast-time measurement availability			
		10%	**20%**	**40%**	**60%**
Clean	Prior only (TF)	98.4	90.2	78.5	70.9
	w/ DA	**78.5**	**69.2**	**60.8**	**52.6**
30 dB	Prior only (TF)	104.9	100.3	90.7	83.6
	w/ DA	**79.6**	**71.1**	**62.8**	**53.7**
20 dB	Prior only (TF)	114.2	111.7	106.3	97.5
	w/ DA	**80.7**	**73.4**	**67.4**	**58.2**
10 dB	Prior only (TF)	131.5	125.2	117.6	108.1
	w/ DA	**83.1**	**79.6**	**73.2**	**67.5**

E.4 Assimilation under Measurement Corruption

Figure 5 provides a direct behavioral test of the learned gain. Because $z_t^* = z_t^p + \alpha_t \odot (z_t^m - z_t^p)$, smaller α_t places less weight on the measurement-side correction. The observed downward shift under lower SNR is therefore consistent with uncertainty-aware behavior even though the model is not supplied an explicit measurement covariance.

Table 10 shows that direct teacher forcing becomes increasingly brittle as measurement quality decreases. At 10% availability and 10 dB SNR, Prior only (TF) reaches 131.5 mm MPJPE, whereas data assimilation gives 83.1 mm. The advantage persists at every SNR and availability level. The gap is particularly informative because the two variants receive the same measurements at the same time steps; the improvement therefore comes from how the evidence is incorporated, not from access to additional observations. Together with Fig. 5, the result shows that the learned correction can attenuate unreliable measurements rather than forcing the dynamical state to follow them directly.

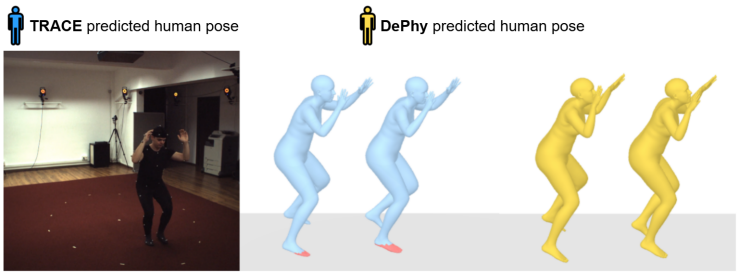


Figure 6: **Qualitative result** on the SportsPose dataset. TRACE Sun et al. (2023) produces floor penetration (red mesh), which DePhy corrects to a physically plausible pose.

E.5 Training-Objective Ablations

Table 6 ablates the objectives used to learn the measurement representation, transition prior, and assimilation stage while holding the TRACE measurement stream fixed. The full model uses $\mathcal{L}_{\text{JEPA}}$ and $\mathcal{L}_{\text{FUT}}$ to train the prior, freezes that prior, and then trains assimilation with $\mathcal{L}_{\text{rec}} + \mathcal{R}_{\text{stay}}$.

Complementary latent objectives. Removing $\mathcal{L}_{\text{JEPA}}$ increases Human3.6M MPJPE from 47.3 to 50.6 mm and SportsPose MPJPE from 67.5 to 75.6 mm. The SportsPose drop is larger, consistent with the role of same-time masked prediction in making the measurement representation robust to corrupted upstream kinematics. Training with only $\mathcal{L}_{\text{JEPA}}$ performs poorly (76.2 mm on Human3.6M and 95.2 mm on SportsPose), because the state representation is not explicitly required to support future physical propagation. Conversely, training with only $\mathcal{L}_{\text{FUT}}$ reaches 50.9 and 73.5 mm, respectively, but remains worse than the full model. The two latent objectives therefore provide distinct and complementary constraints.

Why reconstruction should not train the prior. The reconstruction-supervised future-prior variant replaces EMA-latent future prediction with a decoded reconstruction signal while retaining JEPA supervision. It reaches 52.6 mm MPJPE on Human3.6M, but acceleration error rises to 9.6 mm/s^2; on SportsPose the corresponding values are 76.3 mm and 12.5 mm/s^2. The deterioration is stronger in temporal smoothness than local pose error, indicating that reconstruction provides a poor objective for shaping a rollout prior even when individual decoded frames remain plausible.

The fully coupled variant allows $\mathcal{L}_{\text{rec}}$ to update the measurement encoder, state encoder, physical transition, assimilation module, and decoder jointly. It obtains 49.8 mm MPJPE on Human3.6M but 8.8 mm/s^2 Accel; on SportsPose it degrades to 77.2 mm MPJPE, 142.6 mm MPJPE-G, and 12.8 mm/s^2 Accel. These results provide the clearest empirical support for the decoupled design: allowing reconstruction pressure into the prior can preserve local fit while substantially weakening global and temporal behavior.

Prior anchoring during assimilation. Removing $\mathcal{R}_{\text{stay}}$ yields 50.2 mm MPJPE and 4.9 mm/s^2 Accel on Human3.6M, and 71.5 mm and 6.3 mm/s^2 on SportsPose. Thus the assimilation network benefits from an explicit bias toward the propagated prior rather than being free to follow the current measurement latent arbitrarily. The full objective provides the best overall combination of position accuracy, global consistency, and temporal smoothness.

E.6 Qualitative Results

Figure 6 illustrates a characteristic state-estimation failure of the fixed upstream kinematic estimator: floor penetration. DePhy corrects the physically implausible configuration while preserving the observed motion, consistent with the quantitative reductions in acceleration error and foot-skating rate. Figure 7 shows long-horizon Human3.6M forecasts, with ground truth in gray and DePhy in gold. The qualitative rollouts complement Table 3 by showing that the medium- and long-horizon numerical gains correspond to coherent whole-body motion rather than only lower average joint error.

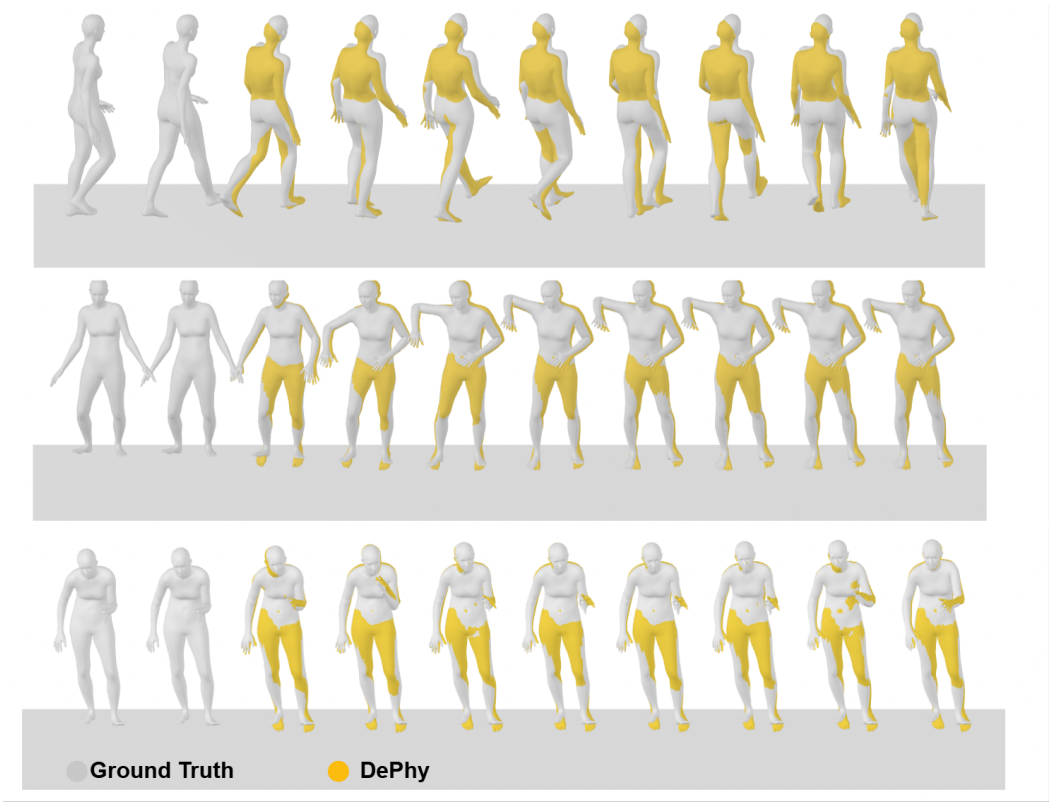


Figure 7: **Qualitative result** on human motion forecasting on Human3.6M. Gray represents the ground-truth motion, and gold represents the DePhy forecast prediction of the human motion. The first two frames represent the input motion.

E.7 Compute Resources

The Human3.6M model contains approximately 178M parameters and trains in approximately 14 hours on a single consumer GPU, with ODE integration constituting the dominant computational cost. The bouncing-ball and damped-pendulum experiments use the same compute environment. Training the proposed models and accompanying synthetic baselines required roughly 10 hours in total for the two synthetic benchmarks. Dataset and derived-result storage was approximately 80 GB.

E.8 Limitations and Future Directions

The framework still assumes that a useful analytical model can be specified. The bounded residual reduces sensitivity to moderate model mismatch, but it does not remove the need to define state variables and a mechanistic backbone. An important extension is to combine the current framework with system identification so that missing governing terms can be discovered rather than manually specified. A second limitation is that the current reconstruction decoder is deterministic; coupling the physics-structured prior with a generative decoder could extend the method from state estimation and forecasting to multimodal motion synthesis. Finally, the measurement encoder is learned without an explicit sensor likelihood. Learned observation models or uncertainty-calibrated measurement representations could provide a more principled interface between severe sensing corruption and the latent assimilation gain.