SYMPWORLD-4D: UNIFYING DYNAMICS AND SPATIOTEMPORAL REPRESENTATION

Anonymous authors
Paper under double-blind review

ABSTRACT

A 4D world model should describe a scene's spatial structure and the dynamics of its interactions. Repeated interactions reveal persistent world properties that can inform a new trajectory. 4D World Identification (4DWI) evaluates this reuse by predicting future geometry and motion from same-world support interactions and an observed query prefix. Branch4D organizes interactions by physical world, providing controlled support comparisons and aligned tasks for reconstruction, scene flow, future prediction (Future4D), and interaction understanding. SympWorld-4D unifies dynamics and continuous spatiotemporal representation through a phase trajectory of configuration and momentum. Inferred world conditions govern its constrained port-Hamiltonian evolution, while the query prefix determines the initial state and latent interaction drive. The evolving configuration and its rate generate geometry and motion. Rendering and temporal entity relations provide RGB-D predictions and interaction understanding from the same representation. Correct-world support improves validation F-score by 8.751 percentage points over wrong-world support. Equal-budget validation comparisons show a Future4D F-score gain of 12.828 percentage points over Neural ODE. One set of weights serves all five tasks, leading all four 4DWI test metrics and the evaluated reconstruction and Future4D F-scores. These results support world reuse and structured evolution as a common basis for 4D generation and understanding.

1 INTRODUCTION

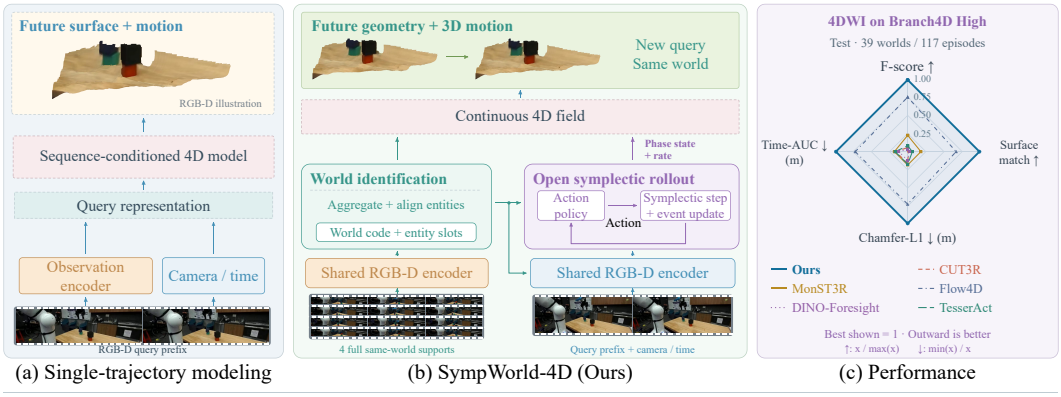
A 4D world model should recover scene structure, anticipate change, and support understanding of the interactions that cause it. When a hand opens a drawer, the drawer's shape, displacement, and changing relation to the hand describe the same physical process. Predicting the drawer's geometry and understanding the interaction both require following its configuration through time. A common representation can express these changes through geometry, motion, and entity relations.

Continuous fields and 4D models recover time-varying geometry and motion (Du et al., 2021; Karhade et al., 2026; Liu et al., 2026), while spatiotemporal visual representations support dynamic scene understanding (Zhou & Lee, 2025). Structured neural dynamics describe how states evolve under energy, constraints, and input (Desai et al., 2021; Schulze et al., 2025; Luan & Shi, 2026). Predicting a new interaction in a familiar world requires a geometric representation whose evolution reflects that world's dynamics. Across repeated interactions, the drawer retains its articulation and surrounding geometry, while its trajectory varies with its state and the applied interaction input, or drive. Persistent world structure can therefore be shared across interactions, while each interaction's state and drive select a phase trajectory of configuration and momentum. Configuration describes the arrangement of entities, while momentum and drive govern its change. That trajectory can therefore supply both the scene's geometry and its motion.

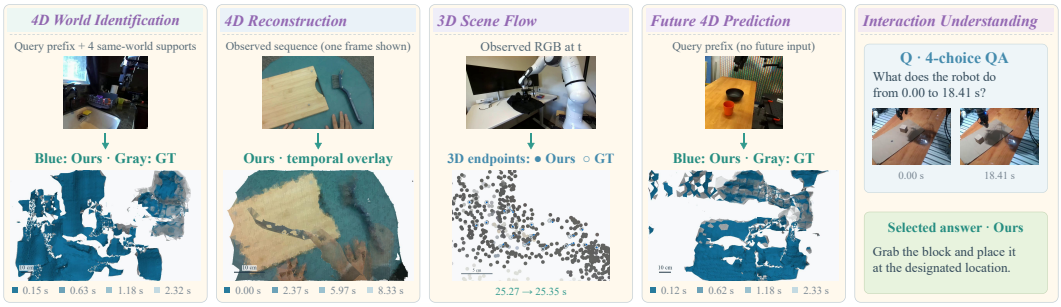
4D World Identification (4DWI) evaluates how information from prior interactions supports prediction in the same world. Given complete RGB-D support interactions and an observed query prefix, the task predicts the query's future metric surfaces and three-dimensional motion (Figure 1). Removing support or changing its world while fixing the query and targets measures the benefit of prior experience and its dependence on world identity.

Branch4D makes this comparison possible across real interactions by grouping RH20T, DROID, HOI4D, and TACO according to persistent entities and scene structure (Fang et al., 2024; Khaz-

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.



(a) Single-trajectory modeling (b) SympWorld-4D (Ours) (c) Performance



(d) Five 4D tasks

Figure 1. SympWorld-4D models a world across interactions. (a,b) Shared world structure and query-specific phase evolution determine a continuous 4D representation; Section 4.2 specifies the constrained discrete-gradient rollout. (c) Four 4DWI test metrics normalized to the best value among the displayed methods. Larger radii indicate better performance; Table 1 reports absolute values. (d) Geometric generation and interaction understanding across five tasks.

atsky et al., 2024; Liu et al., 2022; 2024). Each group belongs to one training, validation, or test split. Its sequences provide calibrated geometry, motion correspondences, and event annotations for 4DWI, reconstruction, scene flow, future prediction (Future4D), and interaction understanding. These aligned tasks evaluate geometric generation and interaction understanding within the same worlds.

SympWorld-4D unifies dynamics and continuous spatiotemporal representation through a shared phase trajectory. Repeated interactions supply shared world conditions. Under these conditions, the query prefix determines the query's initial state and latent drive. Learned symplectic geometry and a Hamiltonian describe the conservative response, while constraints, dissipation, drive, and contact events determine the full evolution. Configuration generates a continuous geometric field, and its rate induces motion. Rendering and temporal entity relations express the same evolution as RGB-D predictions and interaction outputs. Supervising the generated geometry consequently trains both the geometric field and its underlying dynamics.

Branch4D experiments show gains across all four 4DWI test metrics. In equal-budget validation comparisons, structured evolution improves F-score over Neural ODE by 19.112 percentage points on 4DWI, with gains also observed on Future4D. One set of weights supports all five test tasks, achieving higher reconstruction and future-prediction F-scores than the evaluated baselines while supporting motion estimation and interaction understanding.

Our contributions are:

- **We formulate 4DWI**, a task that makes the predictive value of prior interactions measurable through controlled comparisons of support from the same and different physical worlds.
- **We construct Branch4D**, a benchmark whose organization by physical world and aligned geometric and semantic supervision support the study of reusable world information across five tasks.

- **We propose SympWorld-4D**, a method that unifies dynamics and spatiotemporal representation through constrained Hamiltonian phase evolution, from which continuous geometry, motion, and temporal entity relations are generated.
- **We demonstrate experimentally that shared world information and structured evolution improve 4D prediction.** One model supports generation and understanding across five tasks, leading all four 4DWI metrics and the reported reconstruction and future-prediction F-scores.

2 Related Work

Dynamic geometry and prediction. Occupancy Flow transports implicit geometry with a learned velocity field (Niemeyer et al., 2019); continuous fields and dynamic Gaussians recover geometry, appearance, and motion (Du et al., 2021; Luiten et al., 2023; Wu et al., 2024; Cong et al., 2024). Any4D jointly predicts metric geometry and dense motion, and Point4Cast extends reconstruction to forecasting (Karhade et al., 2026; Liu et al., 2026). VGGT estimates cameras, depth, point maps, and tracks from multiple views (Wang et al., 2025a). Sharing information across repeated interactions further requires separating persistent world properties from the current trajectory.

Learning worlds through interaction. DSR-Net builds scene representations through interaction, while Meta-SysId learns identification and prediction from support trajectories (Xu et al., 2021; Park et al., 2022). ACID and PointWorld model dynamics through geometric representations (Shen et al., 2022; Huang et al., 2026). DIO predicts occupancy and flow; TesserAct and RynnWorld-4D generate multimodal futures (Diehl et al., 2025; Zhen et al., 2025; Zhao et al., 2026). LLaVA-4D embeds spatiotemporal coordinates into visual features for scene understanding (Zhou & Lee, 2025). 4DWI evaluates world reuse through support-identity comparisons with a fixed query.

Structured dynamics. Hamiltonian neural networks learn energy-based evolution, and port-Hamiltonian networks incorporate forcing and dissipation (Greydanus et al., 2019; Desai et al., 2021). Context-Aware DeLaN infers environment embeddings from state and torque histories for control (Schulze et al., 2025). NeHaD uses Hamiltonian Gaussian deformation, PH-Dreamer learns port-Hamiltonian visual dynamics, and NGFF evolves Gaussian scenes through learned force fields (Qin et al., 2025; Luan & Shi, 2026; Li et al., 2026). Physically Native World Models discusses Hamiltonian world modeling (Cui & Ma, 2026). SympWorld-4D infers shared world conditions from repeated RGB-D interactions to couple phase evolution with continuous geometry and motion. Alongside 4DWorldBench and RoboWM-Bench (Lu et al., 2026; Jiang et al., 2026), Branch4D evaluates the predictive value of support identity.

3 Task and Benchmark

3.1 4D World Identification

A physical world g contains persistent entities, geometric structure, and shared interaction constraints. Its interactions vary in initial state and drive. Each observation $O_t = (I_t, D_t, K_t, T_t, t)$ records single-view RGB-D, camera intrinsics, the camera-to-world transform, and time. Calibration places surfaces and motion in a common metric frame despite camera motion. Given K complete support interactions $\mathcal{S}_g$ and a distinct query's observed prefix $O_q^{0:k}$, prediction is

$$(\hat{\mathcal{G}}_q^{k+1:T}, \hat{\mathcal{V}}_q^{k:T}) = f_\theta(\mathcal{S}_g, O_q^{0:k}), \tag{1}$$

where $\mathcal{G}$ denotes metric surfaces and $\mathcal{V}$ their three-dimensional motion between evaluation times. The main protocol uses four supporting sequences and one query. Query observation ends after the ongoing drive has produced visible geometric motion. Native action and event annotations determine supervision and evaluation; the predictive input consists of the permitted RGB-D observations, calibration, and timestamps.

3.2 Building worlds from repeated interactions

World grouping. Branch4D groups RH20T, DROID, HOI4D, and TACO interactions by source, scene, persistent entities, and stable geometry (Fang et al., 2024; Khazatsky et al., 2024; Liu et al., 2022; 2024). Entire worlds enter training, validation, or test before task construction, so test queries belong to unseen worlds.

Geometric and interaction targets. Depth and calibration place surfaces in a common world frame and lift image correspondences into three-dimensional motion. Each sequence's action, entity-relation, and event annotations ground interaction targets and four-choice questions, aligning geometry, motion, and understanding within the same sequence. Appendix A.1 details validity and question construction.

Support-query construction. Each interaction has a drive signature defined by its relation type and interacting entity categories. Supports are selected to cover these signatures. A distinct interaction supplies the query, whose prefix extends through observable motion onset. Correct, absent, and matched wrong-world support are evaluated against the same query and future target. Figure 2 summarizes this comparison and the four downstream tasks; Table A1 specifies their observation access.

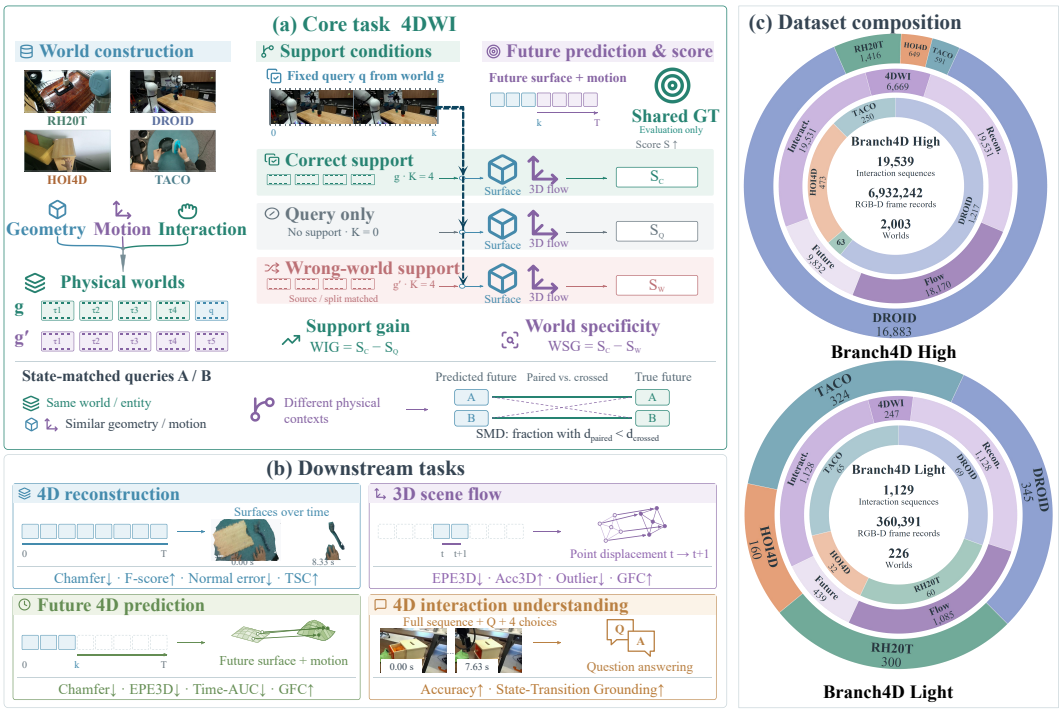


Figure 2. Branch4D. (a) Three support conditions share the same query and future target; WIG and WSG measure support benefit and world specificity. The state-matched protocol compares paired and crossed futures. (b) Four downstream tasks evaluate geometry, motion, prediction, and understanding. (c) High corpus and compact Light subset; rings distinguish source sequences, task instances, and physical worlds.

3.3 Measuring reusable world information

All conditions share the query prefix. Correct support adds same-world interactions; Query Only adds none. Wrong-world support matches source, split, geometry, and sequence statistics. For a higher-is-better score S, World Identification Gain (WIG) and World Specificity Gap (WSG) measure the gains over these controls:

$$\mathrm{WIG} = S_{\mathrm{correct}} - S_{\mathrm{query}}, \qquad \mathrm{WSG} = S_{\mathrm{correct}} - S_{\mathrm{wrong}}. \tag{2}$$

Chamfer-L1 measures surface distance, F-score precision and recall within 1 cm, EPE3D displacement error, and Time-AUC geometric error over physical time. Geometry-flow consistency (GFC) compares a surface advanced by predicted motion with the next predicted surface. Scores aggregate within worlds, then equally across sources. Appendix A specifies metrics, confidence intervals, and state matching.

4 SympWorld-4D

SympWorld-4D couples a world-conditioned phase trajectory to continuous geometry and motion (Figure 3). *World and State Inference* estimates shared world structure and the query state and history. *Constrained Phase Evolution* evolves this state to generate configuration q_t and its rate $\dot{q}_t$, which the *Unified 4D Representation* maps to geometry and motion. Temporal entity relations describe the interaction along this trajectory.

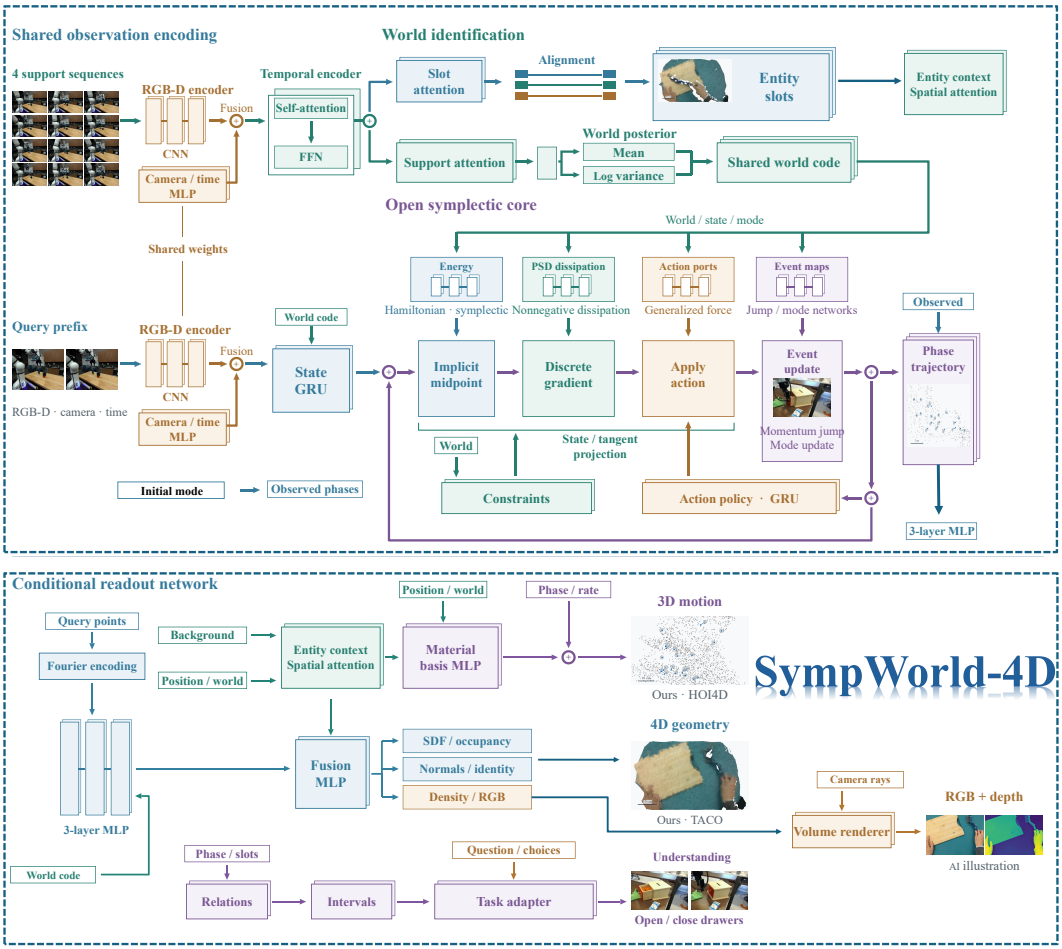


Figure 3. Shared RGB-D features infer world, entity, and query phase representations. Hamiltonian evolution, constraints, dissipation, drive, and events generate the trajectory; configuration and its rate determine geometry and motion. Rendering and temporal relations provide RGB-D and interaction outputs. "Action policy" predicts latent drive, applied through "action ports." Appendix B.3 specifies discrete-gradient rollout.

4.1 World and State Inference

Persistent entities provide a common reference across interactions. Depth and calibration locate each pixel feature at $x_w = T_t D_t(u) K_t^{-1} \tilde{u}$, where $\tilde{u}$ is homogeneous and T_t acts on homogeneous

coordinates. Entity slots summarize these features and observed motion; Sinkhorn assignment (Cuturi, 2013) aligns them to canonical entities across interactions. Aggregation infers the world code ξ_g through posterior $q_\psi(\xi_g \mid \mathcal{S}_g)$ and shared entity features E_g, which condition geometry and dynamics. Query Only and Future4D instead infer ξ_g, E_g from the observed query prefix.

The world-conditioned query prefix determines phase state $z_k = (q_k, p_k)$, constraint mode m_k, and interaction history h_k. Configuration q_k determines geometry; momentum p_k and latent drive govern evolution. History h_k informs drive prediction, while m_k selects the active constraints.

4.2 Constrained phase evolution

The shared world specifies how a state responds to an interaction. Its conservative response follows a world- and mode-conditioned Hamiltonian,

$$H_g(q, p, m) = \tfrac{1}{2} p^\top M_g^{-1}(q, m) p + V_g(q, m), \qquad M_g^{-1} \succ 0. \tag{3}$$

The inverse mass matrix M_g^{-1} and potential V_g define kinetic and potential energy. Energy gradients generate conservative motion through an antisymmetric, nondegenerate matrix $J_g(m)$, constant in phase state for each world and mode, with symplectic form $\Omega_g = J_g^{-1}$. Contact and articulation impose configuration constraints $\phi_m(q) = 0$ and tangency conditions, collected in $\psi_m(z)$ with Jacobian $C_m = D_z \psi_m$. On a regular active chart, the Dirac construction (van der Schaft & Jeltsema, 2014) gives

$$J_g^D = J_g - J_g C_m^\top (C_m J_g C_m^\top)^{-1} C_m J_g. \tag{4}$$

The identity $C_m J_g^D = 0$ makes the conservative velocity tangent to the active constraint. The tangent projector P_C restricts the dissipation factor L_R and input port B_g to admissible directions, giving $R_g^D = P_C L_R L_R^\top P_C^\top$ and $B_g^D = P_C B_g$. This construction makes dissipation positive semidefinite and preserves tangency under the drive. Combining these terms gives

$$\dot{z}_t = (J_g^D - R_g^D) \nabla H_g(z_t, m_t) + B_g^D a_q(z_t, h_k; \xi_g). \tag{5}$$

The bounded latent drive a_q depends on the visible query history and evolving state. Between events, with world and mode fixed, the energy rate is $\dot{H}_g = -\nabla H_g^\top R_g^D \nabla H_g + (B_g^{D\top} \nabla H_g)^\top a_q$: dissipation removes energy, and the port supplies or extracts energy through the drive. Appendix B.2 specifies the structures and derives this relation.

Contact onset and release also change the active mode. An event updates momentum through an impulse and projects it for the new mode, while configuration remains continuous. Between events, constrained discrete-gradient steps (Gonzalez, 1996) compose conservative evolution, dissipation, and input. The rollout supplies configuration q_t and its rate $\dot{q}_t$ at the requested physical timestamps for the geometric outputs below.

4.3 Unified 4D representation

Geometry. The evolving configuration places the world's entities in their current arrangement. For spatial query x, the implicit field evaluates

$$\mathcal{F}_g(x, t) := F_\theta(x, q_t; \xi_g, E_g) = (s_t, \rho_t, c_t, n_t, e_t)(x), \tag{6}$$

giving signed distance, density, color, normal, and entity assignment. The field's temporal dependence comes from q_t, while attention over entity features supplies spatial context. The zero set of s_t defines the surface at time t.

Configuration-induced motion. A spatially varying basis $\mathcal{B}_\theta$ defines local displacement and velocity at spatial query x,

$$\Pi_\theta(x, q_t) = x + \mathcal{B}_\theta(x; \xi_g, E_g) q_t, \qquad v_t(x) = \mathcal{B}_\theta(x; \xi_g, E_g) \dot{q}_t. \tag{7}$$

The Jacobian $D_q \Pi_\theta$ thus maps configuration rate into three-dimensional motion. At a fixed spatial query, mapping differences between adjacent timestamps give scene-flow displacements. Future field evaluations use a smooth local configuration chart around the terminal observed configuration; velocity uses its corresponding derivative (Appendix B.4).

Rendering and interaction. Integrating field density and color along calibrated camera rays produces RGB-D predictions. Along the same trajectory, temporal entity and mode representations supply relations and event intervals for comparing the four candidate answers in interaction understanding.

4.4 Learning the dynamical representation

Geometric supervision. Geometry, motion, RGB-D, and event targets jointly supervise the inferred world representation, phase evolution, and geometric field. Surface, signed distance field (SDF), occupancy, color, and normal objectives supervise the field. Displacement targets supervise configuration-induced motion, while ray rendering relates the field to RGB-D observations. A surface-transport loss penalizes disagreement between the predicted velocity and the change of the distance field,

$$\mathcal{L}_{\text{transport}} = \mathbb{E}_{x,t} \left| \nabla_x s_t(x)^\top v_t(x) + \nabla_q s_t(x)^\top \dot{q}_t \right|. \tag{8}$$

World consistency. Symmetric Kullback-Leibler (KL) divergence aligns world posteriors from disjoint support subsets. An identity objective aligns the query posterior with correct-world support and separates it from wrong-world support, using the same encoder. Future geometry and motion supervise the correct-world trajectory (Appendix B.1).

Drive and events. Projecting the residual between observed training transitions and the integrator's zero-input predictions onto the input port yields a regularized, bounded drive target. This detached target supervises drive prediction from query history. Native event labels anchor its semantics; norm, temporal-smoothness, and port-normalization penalties regularize its effect. Event losses supervise mode changes, constraint compatibility, and tangency. At inference, the observed prefix initializes state and history; predicted drive and events then govern evolution. Appendix B.5 specifies target computation and Appendix C the objectives and optimization.

5 Experiments

The experiments evaluate 4D world identification, examine support identity and dynamical structure, and assess the shared representation across four downstream tasks.

Setup. Branch4D High uses frozen world-level splits. The 4.35M-parameter model observes up to four supports with eight frames each and eight query-prefix frames for 4DWI; Table A1 defines other tasks' inputs. Prediction spans 32 native frames. All five tasks share weights trained for 360 AdamW updates with effective batch size 64 and peak learning rate 2×10^{-5}. Support and structural analyses use validation worlds; task comparisons use test panels. Appendices C and D provide training details and complete results, including horizon and support-quantity analyses.

4D world identification. SympWorld-4D leads all four metrics on the 39 test worlds in Table 1. Relative to Copy-Last, it reduces Chamfer-L1 by 3.31% and Time-AUC by 8.44%, while increasing F-score by 1.32 percentage points and surface match by 2.263 points. These gains jointly establish improved surface proximity and recovery over static persistence.

5.1 Reusing information from the same world

Support identity. On identical queries, correct-world support improves F-score by 9.138 percentage points over Query Only and 8.751 points over matched wrong-world support (Table 2). It also lowers Chamfer-L1 and EPE3D, tying improved surface and motion prediction to the supporting world's identity.

Table 1. 4DWI on 39 test worlds and 117 episodes. CD and Time-AUC are in meters; F-score and Surface match use 1 cm and 5 mm. Bold marks the best score.

Method	CD ↓	F-score ↑	AUC ↓	Surface match ↑
Copy-Last	0.0353	0.3907	0.0372	0.5019
Constant-Velocity	0.1055	0.2129	0.1054	0.2897
CUT3R + static extrapolation	0.2057	0.0231	0.2058	0.0056
MonST3R + static extrapolation	0.1836	0.0917	0.1836	0.0967
Flow4D + fixed rollout	0.0466	0.3072	0.0466	0.3541
DINO-Foresight + depth adapter	0.2814	0.0132	0.2813	0.0153
TesserAct + 4D adapter	0.1948	0.0328	0.1947	0.0345
RynnWorld-4D + 4D adapter	0.2848	0.0388	0.2869	0.0888
SympWorld-4D	**0.03413**	**0.40390**	**0.03406**	**0.52453**

Table 2. Support conditions on the final 10% validation subset. All conditions share queries and targets. CD and EPE3D are in meters; F-score and GFC are fractions.

Support condition	CD ↓	F-score ↑	EPE3D ↓	Time-AUC ↓	GFC ↑
Query only	0.04141	0.15043	0.01923	0.04232	0.95542
Matched wrong world	0.04249	0.15430	0.01865	0.04259	0.95280
Correct world	**0.04088**	**0.24181**	**0.01784**	**0.04097**	**0.95717**

5.2 Turning world information into geometric evolution

Representation and dynamics. Table 3 compares five independently trained models using identical data, 360 updates, and effective batch size 64. Free-time field replaces configuration with time encoding; Context field fixes query-terminal configuration, zeros its velocity, and disables SDF position correction. These variants test geometry's dependence on configuration. Neural ODE and Residual temporal instead test configuration evolution, replacing the structured generator with a generic continuous vector field and a learned discrete update, respectively.

Table 3. Independently trained structures with identical data and budget, evaluated on validation worlds. Context uses fixed query-terminal configuration.

Model	4DWI		Future4D	
	CD ↓	F-score ↑	CD ↓	F-score ↑
Context field	0.02990	0.53574	0.03640	0.66463
Free-time field	0.06156	0.09695	0.08686	0.10254
Neural ODE	0.03316	0.40314	0.03873	0.54660
Residual temporal	0.03315	0.40331	0.03873	0.54661
Full model	**0.02420**	**0.59426**	**0.03146**	**0.67488**

The full model outperforms time-conditioned geometry and generic continuous or residual dynamics, supporting configuration-conditioned geometry with structured state evolution. Compared with the Context field, it improves F-score by 5.852 percentage points on 4DWI and 1.025 points on Future4D.

Constraints, drive, and events. Fixed-weight interventions examine the trained generator's dependence on its dynamics. Removing constraints has the largest Future4D F-score effect, a decrease of 2.975 percentage points with a paired 95% interval of [2.245, 3.682]. Removing drive or event updates lowers F-score by 1.729 or 1.064 points. These decreases associate future surface recovery with constraint enforcement, ongoing input, and contact-mode changes. Appendix D.5 reports all mechanisms and intervals; D.6 and D.8 give coordinate and 4DWI interventions.

5.3 4D GENERATION AND UNDERSTANDING

SympWorld-4D leads reconstruction and Future4D in Chamfer-L1 and F-score and achieves the highest strict scene-flow accuracy (Table 4). Its understanding accuracy and type-macro accuracy exceed V-JEPA2 and VideoMamba, extending the same world representation from geometric generation to interaction semantics. All four tasks use the same weights as 4DWI; Appendix D reports full metrics and nonvisual controls.

Table 4. Four downstream tasks with shared SympWorld-4D weights. CD and EPE3D are in meters; scores are fractions. Strict: strict scene-flow accuracy; Macro: type-macro QA accuracy. Visual baselines use fixed QA predictors. Bold marks the best score per panel; full results appear in Appendix D.

(a) Reconstruction (198 worlds)

Method	CD ↓	F-score ↑
CUT3R	0.2465	0.0614
MonST3R	0.2218	0.0881
TesserAct	0.2503	0.0592
SympWorld-4D	**0.02154**	**0.78939**

(b) Scene flow (186 worlds)

Method	EPE3D ↓	Strict ↑
SeFlow	0.0303	0.8419
Flow4D	**0.0196**	0.9007
DeltaFlow	0.0264	0.8980
SympWorld-4D	0.02317	**0.90688**

(c) Future4D (95 worlds)

Method	CD ↓	F-score ↑
DINO-Foresight	0.2755	0.0234
TesserAct	0.2353	0.0900
RynnWorld-4D	0.3143	0.0693
SympWorld-4D	**0.04151**	**0.44887**

(d) Understanding (198 worlds)

Method	Acc. ↑	Macro ↑
V-JEPA2	0.6086	0.6013
VideoMamba	0.5631	0.5569
InternVideo2-B/14	**0.6843**	**0.6760**
SympWorld-4D	0.62401	0.61660

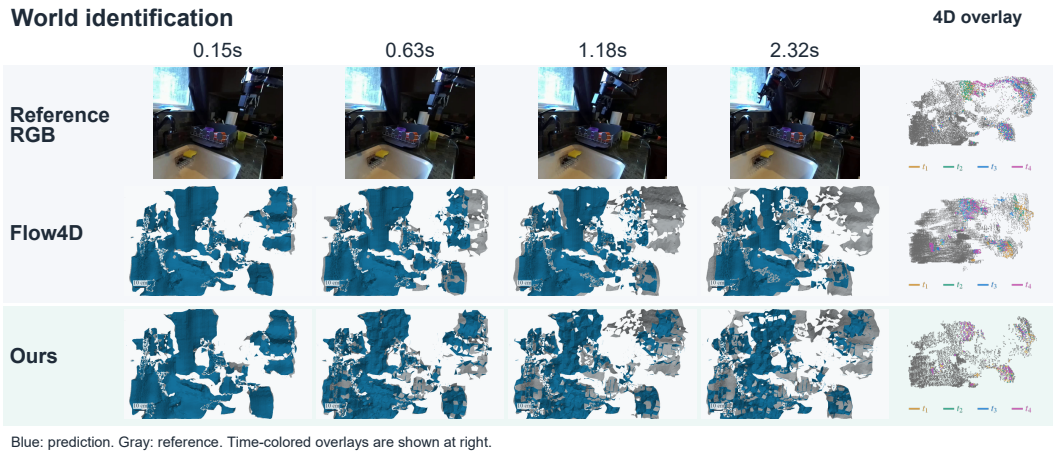


Figure 4. World identification at matched future times. Blue denotes prediction and gray the reference; the right column overlays four times. SympWorld-4D retains greater overlap around the manipulated region at later times than Flow4D. Appendix E presents nine further cases.

6 CONCLUSION

SympWorld-4D unifies dynamics and spatiotemporal representation through world-conditioned phase evolution. Branch4D shows gains over static persistence on all four 4DWI metrics, supported by correct-world experience and structured evolution. The same representation generates geometry, motion, and temporal entity relations for five 4D generation and understanding tasks.

AI USE STATEMENT

Generative AI tools assisted with synthetic data generation and language polishing. The authors take responsibility for the final data and manuscript.

REPRODUCIBILITY STATEMENT

Appendix A specifies task construction, permitted observations, metrics, and world-level aggregation. Appendix B details the dynamics, numerical integration, geometric mapping, and training targets. Appendix C provides model dimensions, optimization settings, and loss weights. Appendix D gives the complete quantitative comparisons and interventions.

REFERENCES

Mahmoud Assran, Adrien Bardes, David Fan, Quentin Garrido, Russell Howes, Mojtaba Komeili, Matthew Muckley, Ammar Rizvi, Claire Roberts, Koustuv Sinha, Artem Zholus, Sergio Arnaud, Abha Gejji, Ada Martin, Francois Robert Hogan, Daniel Dugas, Piotr Bojanowski, Vasil Khalidov, Patrick Labatut, Francisco Massa, Marc Szafraniec, Kapil Krishnakumar, Yong Li, Xiaodong Ma, Sarath Chandar, Franziska Meier, Yann LeCun, Michael Rabbat, and Nicolas Ballas. V-JEPA 2: Self-Supervised Video Models Enable Understanding, Prediction and Planning. *arXiv preprint arXiv:2506.09985*, 2025. URL https://arxiv.org/abs/2506.09985.

Xiaoyan Cong, Haitao Yang, Liyan Chen, Kaifeng Zhang, Li Yi, Chandrajit Bajaj, and Qixing Huang. 4DRecons: 4D Neural Implicit Deformable Objects Reconstruction from a single RGB-D Camera with Geometrical and Topological Regularizations. *arXiv preprint arXiv:2406.10167*, 2024. URL https://arxiv.org/abs/2406.10167.

Sen Cui and Jingheng Ma. Physically Native World Models: A Hamiltonian Perspective on Generative World Modeling. *arXiv preprint arXiv:2605.00412*, 2026. URL https://arxiv.org/abs/2605.00412.

Marco Cuturi. Sinkhorn Distances: Lightspeed Computation of Optimal Transportation Distances. In *Advances in Neural Information Processing Systems*, volume 26, pp. 2292–2300, 2013. URL https://arxiv.org/abs/1306.0895.

Shaan A. Desai, Marios Mattheakis, David Sondak, Pavlos Protopapas, and Stephen J. Roberts. Port-Hamiltonian Neural Networks for Learning Explicit Time-Dependent Dynamical Systems. *Physical Review E*, 104(3):034312, 2021. doi: 10.1103/PhysRevE.104.034312. URL https://journals.aps.org/pre/abstract/10.1103/PhysRevE.104.034312.

Christopher Diehl, Quinlan Sykora, Ben Agro, Thomas Gilles, Sergio Casas, and Raquel Urtasun. DIO: Decomposable Implicit 4D Occupancy-Flow World Model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 27456–27466, 2025. URL https://openaccess.thecvf.com/content/CVPR2025/html/Diehl_DIO_Decomposable_Implicit_4D_Occupancy-Flow_World_Model_CVPR_2025_paper.html.

Yilun Du, Yinan Zhang, Hong-Xing Yu, Joshua B. Tenenbaum, and Jiajun Wu. Neural Radiance Flow for 4D View Synthesis and Video Processing. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 14324–14334, 2021. URL https://openaccess.thecvf.com/content/ICCV2021/html/Du_Neural_Radiance_Flow_for_4D_View_Synthesis_and_Video_Processing_ICCV_2021_paper.html.

Hao-Shu Fang, Hongjie Fang, Zhenyu Tang, Jirong Liu, Chenxi Wang, Junbo Wang, Haoyi Zhu, and Cewu Lu. RH20T: A Comprehensive Robotic Dataset for Learning Diverse Skills in One-Shot. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 653–660, 2024. URL https://rh20t.github.io/.

Oscar Gonzalez. Time Integration and Discrete Hamiltonian Systems. *Journal of Nonlinear Science*, 6:449–467, 1996. URL https://web.ma.utexas.edu/users/og/PUBLICATIONS/paper_DisHamSys.pdf.

Samuel Greydanus, Misko Dzamba, and Jason Yosinski. Hamiltonian Neural Networks. In *Advances in Neural Information Processing Systems*, volume 32, 2019. URL https://papers.nips.cc/paper_files/paper/2019/hash/26cd8ecadce0d4efd6cc8a8725cbd1f8-Abstract.html.

Wenlong Huang, Yu-Wei Chao, Arsalan Mousavian, Ming-Yu Liu, Dieter Fox, Kaichun Mo, and Li Fei-Fei. PointWorld: Scaling 3D World Models for In-The-Wild Robotic Manipulation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 20765–20779, 2026. URL https://openaccess.thecvf.com/content/CVPR2026/html/Huang_PointWorld_Scaling_3D_World_Models_for_In-The-Wild_Robotic_Manipulation_CVPR_2026_paper.html.

Feng Jiang, Yang Chen, Kyle Xu, Yuchen Liu, Haifeng Wang, Zhenhao Shen, Jasper Lu, Shengze Huang, Yuanfei Wang, Chen Xie, and Ruihai Wu. RoboWM-Bench: A Benchmark for Evaluating World Models in Robotic Manipulation. *arXiv preprint arXiv:2604.19092*, 2026. URL https://arxiv.org/abs/2604.19092.

Jay Karhade, Nikhil Keetha, Yuchen Zhang, Tanisha Gupta, Akash Sharma, Sebastian Scherer, and Deva Ramanan. Any4D: Unified Feed-Forward Metric 4D Reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14578–14589, 2026. URL https://openaccess.thecvf.com/content/CVPR2026/html/Karhade_Any4D_Unified_Feed-Forward_Metric_4D_Reconstruction_CVPR_2026_paper.html.

Efstathios Karypidis, Ioannis Kakogeorgiou, Spyros Gidaris, and Nikos Komodakis. DINO-Foresight: Looking into the Future with DINO. *arXiv preprint arXiv:2412.11673*, 2024. URL https://arxiv.org/abs/2412.11673.

Alexander Khazatsky, Karl Pertsch, Suraj Nair, Ashwin Balakrishna, Sudeep Dasari, Siddharth Karamcheti, Soroush Nasiriany, Mohan Kumar Srirama, Lawrence Yunliang Chen, Kirsty Ellis, Peter David Fagan, Joey Hejna, Masha Itkina, Marion Lepert, Yecheng Jason Ma, Patrick Tree Miller, Jimmy Wu, Suneel Belkhale, Shivin Dass, Huy Ha, Arhan Jain, Abraham Lee, Youngwoon Lee, Marius Memmel, Sungjae Park, Ilija Radosavovic, Kaiyuan Wang, Albert Zhan, Kevin Black, Cheng Chi, Kyle Beltran Hatch, Shan Lin, Jingpei Lu, Jean Mercat, Abdul Rehman, Pannag R Sanketi, Archit Sharma, Cody Simpson, Quan Vuong, Homer Rich Walke, Blake Wulfe, Ted Xiao, Jonathan Heewon Yang, Arefeh Yavary, Tony Z. Zhao, Christopher Agia, Rohan Baijal, Mateo Guaman Castro, Daphne Chen, Qiuyu Chen, Trinity Chung, Jaimyn Drake, Ethan Paul Foster, Jensen Gao, Vitor Guizilini, David Antonio Herrera, Minho Heo, Kyle Hsu, Jiaheng Hu, Muhammad Zubair Irshad, Donovon Jackson, Charlotte Le, Yunshuang Li, Kevin Lin, Roy Lin, Zehan Ma, Abhiram Maddukuri, Suvir Mirchandani, Daniel Morton, Tony Nguyen, Abigail O'Neill, Rosario Scalise, Derick Seale, Victor Son, Stephen Tian, Emi Tran, Andrew E. Wang, Yilin Wu, Annie Xie, Jingyun Yang, Patrick Yin, Yunchu Zhang, Osbert Bastani, Glen Berseth, Jeannette Bohg, Ken Goldberg, Abhinav Gupta, Abhishek Gupta, Dinesh Jayaraman, Joseph J Lim, Jitendra Malik, Roberto Martín-Martín, Subramanian Ramamoorthy, Dorsa Sadigh, Shuran Song, Jiajun Wu, Michael C. Yip, Yuke Zhu, Thomas Kollar, Sergey Levine, and Chelsea Finn. DROID: A Large-Scale In-The-Wild Robot Manipulation Dataset. *arXiv preprint arXiv:2403.12945*, 2024. URL https://arxiv.org/abs/2403.12945.

Jaeyeul Kim, Jungwan Woo, Ukcheol Shin, Jean Oh, and Sunghoon Im. Flow4D: Leveraging 4D Voxel Network for LiDAR Scene Flow Estimation. *arXiv preprint arXiv:2407.07995*, 2024. URL https://arxiv.org/abs/2407.07995.

Kunchang Li, Xinhao Li, Yi Wang, Yinan He, Yali Wang, Limin Wang, and Yu Qiao. VideoMamba: State Space Model for Efficient Video Understanding. *arXiv preprint arXiv:2403.06977*, 2024. URL https://arxiv.org/abs/2403.06977.

Shiqian Li, Ruihong Shen, Junfeng Ni, Chang Pan, Chi Zhang, and Yixin Zhu. Learning Physics-Grounded 4D Dynamics with Neural Gaussian Force Fields. *arXiv preprint arXiv:2602.00148*, 2026. URL https://arxiv.org/abs/2602.00148.

Xinhang Liu, Pedro Miraldo, Suhas Lohit, Huaizu Jiang, Naoko Sawada, Yu-Wing Tai, Chi-Keung Tang, and Moitreya Chatterjee. Point4Cast: Streaming Dynamic Scene Reconstruction

and Forecasting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14600–14611, 2026. URL https://openaccess.thecvf.com/content/CVPR2026/html/Liu_Point4Cast_Streaming_Dynamic_Scene_Reconstruction_and_Forecasting_CVPR_2026_paper.html.

Yun Liu, Haolin Yang, Xu Si, Ling Liu, Zipeng Li, Yuxiang Zhang, Yebin Liu, and Li Yi. TACO: Benchmarking Generalizable Bimanual Tool-ACtion-Object Understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 21740–21751, 2024. URL https://openaccess.thecvf.com/content/CVPR2024/html/Liu_TACO_Benchmarking_Generalizable_Bimanual_Tool-ACtion-Object_Understanding_CVPR_2024_paper.html.

Yunze Liu, Yun Liu, Che Jiang, Kangbo Lyu, Weikang Wan, Hao Shen, Boqiang Liang, Zhoujie Fu, He Wang, and Li Yi. HOI4D: A 4D Egocentric Dataset for Category-Level Human-Object Interaction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 21013–21022, 2022. URL https://openaccess.thecvf.com/content/CVPR2022/html/Liu_HOI4D_A_4D_Egocentric_Dataset_for_Category-Level_Human-Object_Interaction_CVPR_2022_paper.html.

Yiting Lu, Wei Luo, Peiyan Tu, Haoran Li, Hanxin Zhu, Zihao Yu, Xingrui Wang, Xinyi Chen, Xinge Peng, Xin Li, and Zhibo Chen. 4DWorldBench: A Comprehensive Evaluation Framework for 3D/4D World Generation Models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 34322–34332, 2026. URL https://openaccess.thecvf.com/content/CVPR2026/html/Lu_4DWorldBench_A_Comprehensive_Evaluation_Framework_for_3D4D_World_Generation_Models_CVPR_2026_paper.html.

Xueyu Luan and Chenwei Shi. PH-Dreamer: A Physics-Driven World Model via Port-Hamiltonian Generative Dynamics. *arXiv preprint arXiv:2605.18303*, 2026. URL https://arxiv.org/abs/2605.18303.

Jonathon Luiten, Georgios Kopanas, Bastian Leibe, and Deva Ramanan. Dynamic 3D Gaussians: Tracking by Persistent Dynamic View Synthesis. *arXiv preprint arXiv:2308.09713*, 2023. URL https://arxiv.org/abs/2308.09713.

Michael Niemeyer, Lars Mescheder, Michael Oechsle, and Andreas Geiger. Occupancy Flow: 4D Reconstruction by Learning Particle Dynamics. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 5379–5389, 2019. URL https://openaccess.thecvf.com/content_ICCV_2019/html/Niemeyer_Occupancy_Flow_4D_Reconstruction_by_Learning_Particle_Dynamics_ICCV_2019_paper.html.

Junyoung Park, Federico Berto, Arec Jamgochian, Mykel J. Kochenderfer, and Jinkyoo Park. Meta-SysId: A Meta-Learning Approach for Simultaneous Identification and Prediction. *arXiv preprint arXiv:2206.00694*, 2022. URL https://arxiv.org/abs/2206.00694.

Hai-Long Qin, Sixian Wang, Guo Lu, and Jincheng Dai. Neural Hamiltonian Deformation Fields for Dynamic Scene Rendering. *arXiv preprint arXiv:2512.10424*, 2025. URL https://arxiv.org/abs/2512.10424.

Lucas Schulze, Jan Peters, and Oleg Arenz. Context-Aware Deep Lagrangian Networks for Model Predictive Control. In *2025 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 6939–6946, 2025. doi: 10.1109/IROS60139.2025.11246292. URL https://arxiv.org/abs/2506.15249.

Bokui Shen, Zhenyu Jiang, Christopher Choy, Leonidas J. Guibas, Silvio Savarese, Anima Anandkumar, and Yuke Zhu. ACID: Action-Conditional Implicit Visual Dynamics for Deformable Object Manipulation. *arXiv preprint arXiv:2203.06856*, 2022. URL https://arxiv.org/abs/2203.06856.

Zachary Teed and Jia Deng. RAFT: Recurrent All-Pairs Field Transforms for Optical Flow. *arXiv preprint arXiv:2003.12039*, 2020. URL https://arxiv.org/abs/2003.12039.

Arjan van der Schaft and Dimitri Jeltsema. Port-Hamiltonian Systems Theory: An Introductory Overview. *Foundations and Trends in Systems and Control*, 1(2–3):173–378, 2014. doi: 10.1561/2600000002. URL https://doi.org/10.1561/2600000002.

Jianyuan Wang, Minghao Chen, Nikita Karaev, Andrea Vedaldi, Christian Rupprecht, and David Novotny. VGGT: Visual Geometry Grounded Transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2025a. URL https://arxiv.org/abs/2503.11651.

Qianqian Wang, Yifei Zhang, Aleksander Holynski, Alexei A. Efros, and Angjoo Kanazawa. Continuous 3D Perception Model with Persistent State. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10510–10522, 2025b. URL https://openaccess.thecvf.com/content/CVPR2025/html/Wang_Continuous_3D_Perception_Model_with_Persistent_State_CVPR_2025_paper.html.

Yi Wang, Kunchang Li, Xinhao Li, Jiashuo Yu, Yinan He, Chenting Wang, Guo Chen, Baoqi Pei, Ziang Yan, Rongkun Zheng, Jilan Xu, Zun Wang, Yansong Shi, Tianxiang Jiang, Songze Li, Hongjie Zhang, Yifei Huang, Yu Qiao, Yali Wang, and Limin Wang. InternVideo2: Scaling Foundation Models for Multimodal Video Understanding. *arXiv preprint arXiv:2403.15377*, 2024. URL https://arxiv.org/abs/2403.15377.

Guanjun Wu, Taoran Yi, Jiemin Fang, Lingxi Xie, Xiaopeng Zhang, Wei Wei, Wenyu Liu, Qi Tian, and Xinggang Wang. 4D Gaussian Splatting for Real-Time Dynamic Scene Rendering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 20310–20320, 2024. URL https://openaccess.thecvf.com/content/CVPR2024/html/Wu_4D_Gaussian_Splatting_for_Real-Time_Dynamic_Scene_Rendering_CVPR_2024_paper.html.

Zhenjia Xu, Zhanpeng He, Jiajun Wu, and Shuran Song. Learning 3D Dynamic Scene Representations for Robot Manipulation. In *Proceedings of the 2020 Conference on Robot Learning*, volume 155 of *Proceedings of Machine Learning Research*, pp. 126–142. PMLR, 2021. URL https://proceedings.mlr.press/v155/xu21b.html.

Junyi Zhang, Charles Herrmann, Junhwa Hur, Varun Jampani, Trevor Darrell, Forrester Cole, Deqing Sun, and Ming-Hsuan Yang. MonST3R: A Simple Approach for Estimating Geometry in the Presence of Motion. *arXiv preprint arXiv:2410.03825*, 2024a. URL https://arxiv.org/abs/2410.03825.

Qingwen Zhang, Yi Yang, Peizheng Li, Olov Andersson, and Patric Jensfelt. SeFlow: A Self-Supervised Scene Flow Method in Autonomous Driving. *arXiv preprint arXiv:2407.01702*, 2024b. URL https://arxiv.org/abs/2407.01702.

Qingwen Zhang, Xiaomeng Zhu, Yushan Zhang, Yixi Cai, Olov Andersson, and Patric Jensfelt. DeltaFlow: An Efficient Multi-frame Scene Flow Estimation Method. *arXiv preprint arXiv:2508.17054*, 2025. URL https://arxiv.org/abs/2508.17054.

Haoyu Zhao, Xingyue Zhao, Siteng Huang, Xin Li, Deli Zhao, and Zhongyu Li. RynnWorld-4D: 4D Embodied World Models for Robotic Manipulation. *arXiv preprint arXiv:2607.06559*, 2026. URL https://arxiv.org/abs/2607.06559.

Haoyu Zhen, Qiao Sun, Hongxin Zhang, Junyan Li, Siyuan Zhou, Yilun Du, and Chuang Gan. TesserAct: Learning 4D Embodied World Models. *arXiv preprint arXiv:2504.20995*, 2025. URL https://arxiv.org/abs/2504.20995.

Hanyu Zhou and Gim Hee Lee. LLaVA-4D: Embedding SpatioTemporal Prompt into LMMs for 4D Scene Understanding. *arXiv preprint arXiv:2505.12253*, 2025. URL https://arxiv.org/abs/2505.12253.

APPENDIX A. TASK AND EVALUATION DETAILS

A.1 OBSERVATION AND TARGET ORGANIZATION

The five tasks use calibrated RGB-D observations in a shared world frame, with observation access determined by the prediction target (Table A1). Reconstruction uses designated observed frames, and scene flow uses an observed pair or short window. Future4D uses the query prefix; 4DWI augments that prefix with complete support interactions. Interaction understanding uses the observed sequence and a question with four candidate answers.

Table A1. Task-specific observations, outputs, and test-panel sizes. Worlds, sequences, episodes, and task records are separate counting units. The 4DWI comparison contains 117 episodes. Corpus composition in Figure 2 counts sequences and worlds; the High 4DWI ring contains 6,669 correct-support comparisons, each paired with query-only and wrong-world conditions.

Task	Available observations	Evaluation output	Test worlds
4DWI	Same-world support and a query prefix	Future metric surface and 3D motion	39
Reconstruction	Designated observed RGB-D frames	Surface, color, depth, and normal	198
Scene flow	Designated observed frame pairs	3D point displacement	186
Future prediction	Query prefix	Geometry and motion over 32 future frames	95
Interaction understanding	Observed sequence and a four-choice question	Answer index	198

A world group contains persistent entities and stable scene structure across complete sequences. TACO groups tool and target instances; HOI4D groups object category, instance, and room; RH20T groups scene and configuration; DROID groups laboratory, building, scene, and persistent entities. Actions and trajectories vary within these groups. World assignment precedes task construction, so every sequence from a world belongs to the same split. Core 4DWI worlds contain at least five eligible interactions for four supports and a distinct query.

Branch4D includes a compact Light subset organized by world and a larger High corpus containing every Light sequence. High selects complete sequences from these world groups, starting from a 20% sampling budget per source and expanding it to cover task, action, object, tool, layout, configuration, and scene categories. Additional sequences bring each source's category distributions within Jensen–Shannon divergence 0.10 of the full available collection. All selected sequences retain their native duration and annotations.

For each selected sequence, depth and camera calibration determine metric surface samples in the shared world frame. DROID depth is recovered from its original ZED stereo recordings using the official ZED SDK. Each surface sample includes color, normal, entity identity, and validity information. Motion targets are derived from bidirectional RAFT-Large correspondences (Teed & Deng, 2020) and measured depth, with native object poses and mesh annotations providing additional constraints where available. Visibility and bidirectional correspondence consistency determine point validity. A maximum of 8,192 sampled surface points represents each evaluated surface under the specified output setting.

The geometric targets are paired with event annotations that associate active and affected entities with semantic descriptions and time intervals. Frame-level motion locates each interaction's observable onset, while its relation type and entity categories define the drive signature. Support selection covers these signatures, and the query prefix extends through motion onset to reveal the current drive. These annotations determine sample boundaries and supervision; the observation encoder receives RGB-D, cameras, and timestamps.

The same event annotations ground interaction-understanding questions. A program determines the answer and distractors, and Qwen3.5-27B selects the question wording from verified equivalent formulations while the options and answer remain fixed.

Comparisons use CUT3R (Wang et al., 2025b) and MonST3R (Zhang et al., 2024a) for geometry; SeFlow (Zhang et al., 2024b), Flow4D (Kim et al., 2024), and DeltaFlow (Zhang et al., 2025) for motion; and DINO-Foresight (Karypidis et al., 2024), TesserAct (Zhen et al., 2025), and RynnWorld-4D (Zhao et al., 2026) for future prediction. These methods are fine-tuned on Branch4D with the same task-specific observation access and geometric preprocessing and postprocessing as SympWorld-4D. The interaction panel includes V-JEPA 2 (Assran et al., 2025), VideoMamba (Li et al., 2024), and InternVideo2 (Wang et al., 2024) with fixed question-answering predictors trained and selected on Branch4D training worlds using the same data records as the main experiment. Table 1 names the adapters and extrapolation used for 4DWI.

TesserAct and RynnWorld-4D receive an empty text condition, so their prediction is conditioned on the permitted visual observations. The RynnWorld-4D adapter decodes generated flow colors into pixel displacements using a fixed, known magnitude scale.

A.2 Geometry and motion metrics

For predicted and reference point sets P and G, Chamfer-L1 is

$$\mathrm{CD}_1(P,G) = \frac{1}{2}\left[\frac{1}{|P|}\sum_{p\in P}\min_{g\in G}\|p-g\|_2 + \frac{1}{|G|}\sum_{g\in G}\min_{p\in P}\|g-p\|_2\right]. \tag{A1}$$

Chamfer-L2 replaces Euclidean distances with squared distances. Thresholded surface agreement is measured by F-score and Surface match. F-score is the harmonic mean of geometric precision and recall at 0.01 m; Surface match is the fraction of valid reference points whose nearest predicted point lies within 0.005 m. Time-AUC integrates geometric error over the actual timestamps and divides by the evaluated duration.

Scene-flow evaluation matches each valid reference source point to the nearest predicted source-surface point and compares their attached displacements. EPE3D averages the resulting Euclidean displacement errors. Strict accuracy accepts an absolute error below 0.05 m or a relative error below 5%; relaxed accuracy uses 0.10 m or 10%. An outlier has absolute error above 0.30 m and relative error above 10%.

Geometry-flow consistency (GFC) measures agreement after surface points are advanced by predicted motion. For tasks predicting both geometry and motion, GFC compares the advanced predicted surface with the next predicted surface. For scene flow, GFC compares predicted and reference endpoints of the same valid source points. Temporal Surface Consistency (TSC) advances predicted geometry through valid reference motion correspondences. TSC is reported with verified-motion coverage, the proportion of points with accepted correspondences. Both consistency scores apply $\exp(-\bar{d}/\sigma)$ to the mean valid propagation error $\bar{d}$ in meters, with $\sigma = 0.05\,\mathrm{m}$. Distances remain in metric world coordinates without entity-scale normalization.

Depth RMSE is computed over valid matched pixels.

A.3 Interaction metrics and state-matched evaluation

Question-answering accuracy measures the fraction of correct options. Type-macro accuracy gives equal weight to interaction-class, temporal-localization, event-order, and spatial-motion questions. State-Transition Grounding evaluates questions requiring temporal, ordering, or spatial transition reasoning.

As an additional benchmark protocol, state-matched evaluation pairs queries from the same world and entity instance whose terminal geometry and motion are close, while their interaction context and subsequent trajectory differ. Matching uses geometry, velocity, joint state, and visibility, with thresholds determined on validation worlds. For a set of pairs $\mathcal{P}$, State-Matched Disambiguation is

$$\begin{aligned}\mathrm{SMD} &= \frac{1}{|\mathcal{P}|}\sum_{(a,b)\in\mathcal{P}}\mathbf{1}[\delta_{\mathrm{self}} < \delta_{\mathrm{swap}}],\\ \delta_{\mathrm{self}} &= d(\hat{\mathcal{G}}_a,\mathcal{G}_a) + d(\hat{\mathcal{G}}_b,\mathcal{G}_b),\\ \delta_{\mathrm{swap}} &= d(\hat{\mathcal{G}}_a,\mathcal{G}_b) + d(\hat{\mathcal{G}}_b,\mathcal{G}_a).\end{aligned} \tag{A2}$$

Here d compares valid surface regions in a shared metric world frame. The pair construction also accounts for entity extent and visibility and applies to 4DWI and Future4D.

A.4 Aggregation and paired intervals

Scores are aggregated within physical worlds and averaged equally over sources with the corresponding metric. This source-macro calculation preserves the contribution of each data source. A paired comparison uses the same worlds and direction-normalized score differences. Confidence intervals resample physical worlds within each source 10,000 times and recompute the source-macro difference. They describe variation across evaluated worlds for a fixed model. The intervals in Table D6 are pointwise intervals for the listed metrics.

The support and horizon panels evaluate the final model on the final ten percent validation subset. Coordinate and mechanism interventions retain the trained weights and change the specified inference operation. The structural panel evaluates five models independently trained for 360 updates under matched configurations. All panels retain the frozen world-level train/validation/test assignments.

Appendix B. Representation, Dynamics, and Learning

B.1 World and query representations

Each support sequence produces local entity slots from calibrated RGB-D features. Sinkhorn assignment aligns these slots to shared entities before aggregation into the world posterior and entity features. The model uses a 128-dimensional world code, eight 96-dimensional entity slots, and a 64-dimensional phase state comprising 32-dimensional configuration and momentum. The query encoder conditions its state, mode, and history estimates on the world representation. Query Only and Future4D obtain the world code and entity slots from their observed query prefix.

Posteriors inferred from disjoint subsets of the same world's supports are matched with symmetric KL divergence. The identity objective passes the query prefix, correct support, and wrong support through the same world encoder. For their normalized posterior means μ_q, μ_+, μ_- and temperature $\tau = 0.1$,

$$\mathcal{L}_{\mathrm{id}} = -\log \frac{\exp(\mu_q^\top \mu_+/\tau)}{\exp(\mu_q^\top \mu_+/\tau) + \exp(\mu_q^\top \mu_-/\tau)}. \tag{B1}$$

The symmetric KL objective encourages a stable world posterior across supporting trajectories, while the identity objective associates the query with its world. Wrong-world support supplies the negative identity comparison. Future geometry and motion supervision are attached to the correct-world trajectory.

B.2 Phase-space and constraint structure

The Hamiltonian in Equation (3) combines positive-definite kinetic energy with configuration potential. Its gradients generate conservative motion through the matrix

$$J_g(m) = \begin{bmatrix} 0 & A_g(m) \\ -A_g(m)^\top & \Gamma_g(m) \end{bmatrix}, \qquad A_g = \exp L_g, \qquad \Gamma_g^\top = -\Gamma_g. \tag{B2}$$

Here L_g is a learned square matrix and Γ_g is antisymmetric. For a fixed world and mode, J_g is constant in the phase state. If $J_g[u; v] = 0$, its upper block gives $A_g v = 0$, hence $v = 0$; the lower block then gives $A_g^\top u = 0$, hence $u = 0$. The matrix exponential therefore makes J_g nondegenerate. Its inverse $\Omega_g = J_g^{-1}$ is a constant, closed, nondegenerate two-form, defining the symplectic geometry of the latent phase space.

Let $\phi_m(q) = 0$ describe the active configuration constraint and $\dot{q}_H = [J_g \nabla H_g]_q$ its conservative configuration velocity. The position and tangency conditions are

$$\psi_m(z) = \begin{bmatrix} \phi_m(q) \\ D\phi_m(q)\dot{q}_H \end{bmatrix}, \qquad C_m = D_z \psi_m. \tag{B3}$$

On a regular active constraint chart, the Dirac structure is

$$J_g^D = J_g - J_g C_m^\top (C_m J_g C_m^\top)^{-1} C_m J_g. \tag{B4}$$

Multiplication by C_m gives $C_m J_g^D = 0$, so conservative motion is tangent to the active constraint. With tangent projector P_C, the dissipative matrix is $R_g^D = P_C L_R L_R^\top P_C^\top$, and the input port is projected into the same admissible directions. For any w,

$$w^\top R_g^D w = \|L_R^\top P_C^\top w\|_2^2 \geq 0. \tag{B5}$$

Between events, with the world and mode fixed, differentiating H_g along Equation (5) eliminates the quadratic form involving antisymmetric J_g^D. With $y_g = (B_g^D)^\top \nabla H_g$, the remaining energy rate is

$$\dot{H}_g = -\nabla H_g^\top R_g^D \nabla H_g + y_g^\top a_q. \tag{B6}$$

The conservative contribution evolves state along the active constraint, dissipation removes energy, and the input port exchanges energy with the ongoing interaction.

B.3 Discrete evolution and events

For step duration h, $\Phi_{H,C}^h$, $\Phi_{R,C}^h$, and $\Phi_{a,C}^h$ denote constrained conservative, dissipative, and input updates. Their composition is followed by the event transition Δ_{e_n}:

$$z_{n+1} = \Delta_{e_n} \circ \Phi_{a,C}^h \circ \Phi_{R,C}^h \circ \Phi_{H,C}^h(z_n). \tag{B7}$$

The conservative and dissipative updates use the Gonzalez discrete gradient (Gonzalez, 1996). For $d = z' - z$ and midpoint $\bar{z} = (z + z')/2$, its nonzero-displacement form is

$$\bar{\nabla} H(z, z') = \nabla H(\bar{z}) + \frac{H(z') - H(z) - \nabla H(\bar{z})^\top d}{\|d\|_2^2} d. \tag{B8}$$

The derivative limit defines zero displacement. The secant identity $\bar{\nabla} H^\top d = H(z') - H(z)$ gives the energy relation for an exactly solved substep before chart projection. Structure coefficients are evaluated at each substep's start and reused through its fixed-point iterations. Each continuous substep ends with projection to the active constraint chart. The evaluated configuration uses four solver iterations, maximum substep length 0.125 s, at most 128 substeps, and four constraint-projection iterations with tolerance 10^{-5}.

An event changes the mode and updates momentum through a learned impulse, followed by momentum projection for the new active mode. Configuration remains continuous across the jump. Event-constraint supervision penalizes position compatibility and tangency residuals. The six event states are free, onset, sustained, release, compound, and transition.

B.4 Configuration chart, motion, and rendering

The field in Equation (6) predicts signed distance, density derived from distance, color, normal, and entity assignment at spatial queries. Attention over background and entity slots supplies spatially relevant features. Surface extraction projects spatial samples toward the zero level set and returns normal, color, and entity information. Density and color integration along calibrated rays gives RGB-D predictions.

Future queries use a smooth chart around the terminal observed configuration. With $d_t = q_t - q_k$, configuration dimension d_q, and radius r,

$$\tilde{q}_t = q_k + \frac{d_t}{\sqrt{1 + \|d_t / r\|_2^2 / d_q}}. \tag{B9}$$

With radius $r = 0.1$, the field receives the chart coordinate $\tilde{q}_t$, and configuration-induced velocity uses $\dot{\tilde{q}}_t = D_q \tilde{q}_t \, \dot{q}_t$. The phase state continues to evolve in (q_t, p_t); the chart determines the configuration coordinate and rate used for geometry and motion.

At a fixed spatial query x, the local configuration-conditioned mapping gives displacement $\Pi_\theta(x, \tilde{q}_{t+1}) - \Pi_\theta(x, \tilde{q}_t)$. Its configuration derivative gives instantaneous velocity at the spatial query. The flow loss supervises these local motion predictions, and the transport penalty in Equation (8) encourages agreement between predicted velocity and the field's temporal change. Thus geometry–motion consistency is learned through supervision.

Alongside geometric prediction, temporal entity and mode representations provide relations and event intervals for comparing the four candidate answers in interaction understanding.

B.5 Drive supervision and joint objectives

Metric surface, displacement, RGB-D, and interaction targets jointly supervise the generated trajectory. Geometry objectives supervise distance, occupancy, surface position, normal, and color. Motion objectives supervise flow and state-to-geometry transitions. Constraint, power, transport, and phase-state objectives shape the dynamics, while the world objectives associate trajectories with a reusable world representation.

The drive target is computed from an observed training transition. Starting at z_t, the same integrator predicts zero-input next state $\bar{z}_{t+1}$. For input port B, timestep Δt, regularizer ϵ, and magnitude limit $a_{\max}$,

$$a_t^{\text{post}} = a_{\max} \tanh\left[\frac{(B^\top B + \epsilon I)^{-1} B^\top (z_{t+1} - \bar{z}_{t+1})}{a_{\max} \Delta t}\right]. \tag{B10}$$

The detached target supervises the drive inferred from visible query history through a KL objective. Native event labels additionally anchor the drive's semantics. Norm, temporal-smoothness, and port-normalization penalties regularize the drive and its effect on state. The evaluated magnitude limit is 2.0. At inference, the observed prefix initializes state and history, and the learned drive and event predictions evolve with the future state. Table C2 lists the joint objective weights.

Appendix C. Model and Optimization Details

C.1 Evaluated configuration

Table C1. Architecture and geometric evaluation settings. Short prefixes can repeat sampled frames to form the input tensor. Future timestamps retain the source sequence's physical timing.

Quantity	Value
Trainable parameters	4,346,358
World code / phase state / hidden width	128 / 64 / 192
Configuration / momentum dimensions	32 / 32
Entity slots	8 slots, 96 dimensions each
Drive dimension / magnitude limit	32 / 2.0
Modes / configuration constraint rank	6 / 8
Temporal Transformer	3 layers, 8 attention heads
Relation vocabulary	1,325 classes
Field frequencies / chart radius	6 / 0.1
Input image resolution	128×128
Support	Up to 4 sequences, 8 sampled frames each
Query encoding	8 frames sampled from its observed prefix
Future output	32 frames
Training field samples	4,096 per frame
Surface band	0.02 m
Surface extraction	Up to 8,192 points; grid resolution 48
Ray rendering	256 rays, 64 samples per ray
Ray near / far distances	0.05 m / 10 m

C.2 Optimization

SympWorld-4D is trained for 360 optimizer updates, and the five test tasks share the resulting weights. Training uses the High correct-support records and the paired world-identity objective. Coverage-first sampling traverses the 5,413 records before subsequent sampling; at effective batch size 64, one unrepeated pass requires 85 updates. Training uses AdamW with weight decay 10^{-4}, gradient clipping at 1.0, and peak learning rate 2×10^{-5}, with random seed 2026 and bfloat16 computation. A 20-step warmup precedes cosine decay to a minimum learning-rate ratio of 0.1. Four observed auxiliary frames receive a supervision weight of 0.5.

Each structural variant is independently trained for 360 updates. The full model uses eight GPUs and each alternative uses two, with effective batch size 64. Unchanged encoders, field, renderer, and relation modules share their initialization across runs; newly introduced layers use Xavier initialization. Data, shared losses, and their weights remain fixed.

The Free-time field substitutes a time encoding of the same dimension for the geometric field's configuration input. Neural ODE uses a state vector field with two hidden layers of width 192 and SiLU activations. Residual temporal uses a residual state update of the same capacity. Drive inputs, constraint projection, and event updates remain active. Module-specific objectives follow the corresponding replacement, while all shared objectives retain the weights in Table C2.

The fixed-weight intervention panels retain the final model parameters and change only the specified inference operation. The surface-coordinate intervention preserves the initial observation and uses SDF projection gain 0.25.

Public comparison methods are fine-tuned on Branch4D for 360 updates under the common task protocol.

C.3 Loss weights

Table C2. Weights used by the evaluated optimization run.

Loss term	Weight	Loss term	Weight
SDF	1.0	Surface	0.5
Occupancy	0.5	Eikonal	0.1
Color	0.2	Normal	0.2
Flow	1.0	Relation	0.5
Transport	0.2	Power	0.1
Constraint	0.1	Poisson structure	0.01
Dirac tangency	0.1	Phase geometry	0.2
Phase transition	4.0	Rendered RGB	0.2
Rendered depth	0.5	World KL	0.001
World reuse	0.1	World identification	0.1
Drive KL	0.5	Drive anchor	0.2
Drive norm	0.01	Drive smoothness	0.01
Port normalization	0.05	Event	0.1
Event constraint	0.1	Observed supervision	0.5

APPENDIX D. COMPLETE QUANTITATIVE RESULTS

The following results expand the main comparisons from geometric accuracy to motion, interaction understanding, and the role of the learned dynamics. Tables D1-D4 compare task performance; Tables D5-D9 examine the dynamical and representational choices; Tables D10-D12 examine world reuse and prediction horizon.

D.1 RECONSTRUCTION

SympWorld-4D's reconstruction advantage extends across surface distance, F-score, and depth error. Its higher temporal consistency score accompanies greater verified-motion coverage, measuring agreement over a larger proportion of the evaluated geometry.

Table D1. Reconstruction on 198 test worlds. Verified motion reports accepted correspondence coverage and accompanies the temporal consistency score. The listed reconstruction comparisons have approximately 0.4–0.6% verified motion coverage; SympWorld-4D has 22.333%.

Method	Chamfer-L1 ↓	F-score ↑	Depth RMSE ↓	TSC ↑	Verified motion ↑
CUT3R	0.2465	0.0614	0.3907	0.2914	0.0041
MonST3R	0.2218	0.0881	0.3876	0.3831	0.0061
TesserAct	0.2503	0.0592	0.5457	0.3532	0.0040
SympWorld-4D	**0.02154**	**0.78939**	**0.05061**	**0.53404**	**0.22333**

D.2 SCENE FLOW

SympWorld-4D achieves the highest strict and relaxed accuracy and the lowest outlier rate in this panel, whereas Flow4D has lower average displacement error and higher endpoint consistency. These metrics distinguish accuracy within specified tolerances from average displacement error over all valid points.

Table D2. Scene flow on 186 test worlds. EPE3D measures average displacement error, while strict and relaxed accuracy report the fraction within their respective tolerances.

Method	EPE3D ↓	Strict accuracy ↑	Relaxed accuracy ↑	Outlier ↓	GFC ↑
SeFlow	0.0303	0.8419	0.9250	0.0103	0.7470
Flow4D	**0.0196**	0.9007	0.9688	0.0063	**0.7970**
DeltaFlow	0.0264	0.8980	0.9454	0.0200	0.7783
SympWorld-4D	0.02317	**0.90688**	**0.97073**	**0.00623**	0.74677

D.3 FUTURE PREDICTION

The geometric advantage persists across all three prediction horizons. SympWorld-4D's Chamfer-L1 increases with the prediction horizon but remains below that of every compared method at H8, H16, and H32.

Table D3. Future prediction on 95 test worlds. Horizon columns report Chamfer-L1 over the corresponding nested prediction prefixes.

Method	Chamfer-L1 ↓	F-score ↑	H8 Chamfer ↓	H16 Chamfer ↓	H32 Chamfer ↓	Time-AUC ↓
DINO-Foresight	0.2755	0.0234	0.2821	0.2811	0.2755	0.2756
TesserAct	0.2353	0.0900	0.2232	0.2277	0.2353	0.2356
RynnWorld-4D	0.3143	0.0693	0.2771	0.2964	0.3143	0.3165
SympWorld-4D	**0.04151**	**0.44887**	**0.03288**	**0.03521**	**0.04151**	**0.04149**

D.4 Interaction understanding

The question-only control establishes how much of the task is predictable from language alone. SympWorld-4D improves overall and type-macro accuracy over the question-only control but scores lower on the state-transition subset. InternVideo2 leads all three understanding metrics.

Table D4. Interaction understanding on 198 test worlds. Type macro averages the four question categories. STG evaluates the subset requiring state-transition reasoning.

Method	Accuracy ↑	Type macro ↑	STG ↑
Majority control	0.2421	0.2406	0.1877
Question-only control	0.6077	0.6010	0.6258
Video-only control	0.1915	0.1930	0.2105
V-JEPA2 + fixed QA predictor	0.6086	0.6013	0.5675
VideoMamba + fixed QA predictor	0.5631	0.5569	0.5986
InternVideo2-B/14 + fixed QA predictor	**0.6843**	**0.6760**	**0.7308**
SympWorld-4D	0.62401	0.61660	0.58115

D.5 Future4D fixed-weight mechanism interventions

Removing each mechanism lowers future-surface F-score, with the largest decrease from the constraint term. The paired intervals quantify these changes across worlds. With weights fixed, the decreases show how the learned generator depends on constraint enforcement, interaction drive, and event updates.

Table D5. The same trained parameters are used in every row. The paired F-score differences and confidence intervals are reported in Table D6.

Active configuration	Chamfer-L1 ↓	F-score ↑	EPE3D ↓	Time-AUC ↓	GFC ↑
Full model	0.045348	0.331798	0.014600	0.045437	0.975897
Without conservative term	0.045476	0.323731	0.014622	0.045566	0.975453
Without constraint term	0.045926	0.302051	0.014788	0.046022	0.975072
Without dissipation	0.045925	0.330921	0.014693	0.045512	0.973418
Without drive	0.045900	0.314508	0.014958	0.046982	0.975132
Without event update	0.045490	0.321162	0.014609	0.045579	0.975438

Table D6. Paired Future4D F-score advantages of the full model, in percentage points. The 95% intervals use 10,000 source-stratified paired-world bootstrap samples.

Disabled mechanism	F-score decrease (pp)	95% interval (pp)
Conservative evolution	0.807	[0.437, 1.231]
Constraint term	2.975	[2.245, 3.682]
Dissipation	0.088	[0.001, 0.173]
Interaction drive	1.729	[1.087, 2.472]
Event update	1.064	[0.559, 1.668]

D.6 Fixed-weight coordinate interventions

Replacing the field's configuration coordinates with time degrades both geometry and motion. Static context retains more predictive information than time, while evolving configuration gives the strongest results for both tasks in this panel.

Table D7. Field coordinates are changed at inference with the model weights fixed. Phase denotes the evolving configuration supplied by the phase trajectory. Phase-to-time F-score differences are 18.871 points for 4DWI and 27.083 for Future4D. Phase-to-context differences are 0.235 and 0.541 points. These comparisons measure the response of the trained field to the supplied coordinates.

Task	Field coordinate	Chamfer-L1 ↓	F-score ↑	EPE3D ↓	Time-AUC ↓	GFC ↑
4DWI	Phase	0.040875	0.241812	0.014424	0.040945	0.971978
4DWI	Time	0.088601	0.053106	0.079454	0.089344	0.882390
4DWI	Context	0.047019	0.239464	0.014679	0.041788	0.970348
Future4D	Phase	0.045348	0.331798	0.014600	0.045437	0.975897
Future4D	Time	0.106470	0.060969	0.068467	0.107278	0.867922
Future4D	Context	0.045801	0.326393	0.014832	0.046884	0.973247

D.7 Full structural comparison

Independent training tests the field and evolution choices under the same optimization budget. The full model improves geometry and motion over each replacement, while the Context field remains the closest alternative. These comparisons support the joint use of configuration-conditioned geometry and structured phase evolution.

Table D8. Complete metrics for the independently trained variants in Table 3.

Task	Structure	Chamfer-L1 ↓	F-score ↑	EPE3D ↓	Time-AUC ↓	GFC ↑
4DWI	**Full model**	0.02420	0.59426	0.01350	0.02412	0.996691
4DWI	Context field	0.02990	0.53574	0.01438	0.02983	0.996660
4DWI	Free-time field	0.06156	0.09695	0.03607	0.06191	0.995650
4DWI	Neural ODE	0.03316	0.40314	0.01446	0.03309	0.996607
4DWI	Residual temporal	0.03315	0.40331	0.01446	0.03307	0.996607
Future4D	**Full model**	0.03146	0.67488	0.01381	0.03143	0.996599
Future4D	Context field	0.03640	0.66463	0.01453	0.03636	0.996561
Future4D	Free-time field	0.08686	0.10254	0.04972	0.08743	0.971089
Future4D	Neural ODE	0.03873	0.54660	0.01470	0.03869	0.996509
Future4D	Residual temporal	0.03873	0.54661	0.01470	0.03869	0.996509

D.8 4DWI MECHANISM DIAGNOSTICS

Table D9. Fixed-weight interventions on the 4DWI full-model setting. Constraint projection has the largest F-score effect in this panel, while disabling the drive has a smaller F-score effect and increases geometric and motion error.

Active configuration	Chamfer-L1 ↓	F-score ↑	EPE3D ↓	Time-AUC ↓	GFC ↑	Depth RMSE ↓
Full model	0.02420	0.59426	0.01350	0.02412	0.996691	0.15591
Without conservative flow	0.04425	0.34156	0.01504	0.04425	0.996474	0.22000
Without constraint projection	0.05281	0.24758	0.01604	0.05277	0.996313	0.26433
Without dissipation	0.04399	0.37102	0.01510	0.04396	0.996461	0.22133
Without drive	0.03769	0.58781	0.01460	0.03765	0.996530	0.18762
Without event update	0.04306	0.41042	0.01493	0.04304	0.996487	0.21249

D.9 SUPPORT GAINS AND PREDICTION HORIZON

Correct support improves all direction-normalized scores over Query Only and matched wrong-world support (Table D10). For the resulting predictions, a longer horizon lowers surface F-score while increasing GFC (Table D11). These opposing trends distinguish agreement within the generated trajectory from accuracy against the observed future.

Table D10. Direction-normalized support gains. Positive Chamfer and EPE3D gains denote reductions in error.

Gain	Chamfer-L1	F-score	EPE3D	GFC
WIG	+0.00053	+0.09138	+0.00139	+0.00175
WSG	+0.00161	+0.08751	+0.00081	+0.00437

Table D11. Complete 4DWI horizon metrics.

Future frames	Mean duration (s)	Chamfer-L1 ↓	F-score ↑	EPE3D ↓	Time-AUC ↓	GFC ↑
8	0.503	0.03173	0.383506	0.01746	0.03194	0.92457
16	1.005	0.03528	0.302395	0.01777	0.03544	0.94723
32	2.049	0.04088	0.24181	0.01784	0.04097	0.95717

Future4D shows the same decline in surface F-score: 0.498717, 0.403518, and 0.331798 at H8, H16, and H32, respectively. Its mean H32 duration is 2.109 s.

D.10 SUPPORT QUANTITY

Additional support improves surface prediction for a fixed set of validation queries. Table D12 gives the full support-count panel.

Table D12. Support quantity on the final 10% validation subset. The same tasks are used for every support count.

Support K	Chamfer-L1 ↓	F-score ↑	EPE3D ↓	GFC ↑
1	0.06762	0.17816	0.01824	0.92568
2	0.05986	0.18039	0.01786	0.94232
4	**0.04088**	**0.24181**	**0.01784**	**0.95717**

APPENDIX E. ADDITIONAL QUALITATIVE RESULTS

These nine cases complement Figure 4 with world identification, reconstruction, scene flow, future prediction, and interaction understanding. Reference observations appear above the compared outputs. Each caption specifies the visual convention and the displayed interaction.

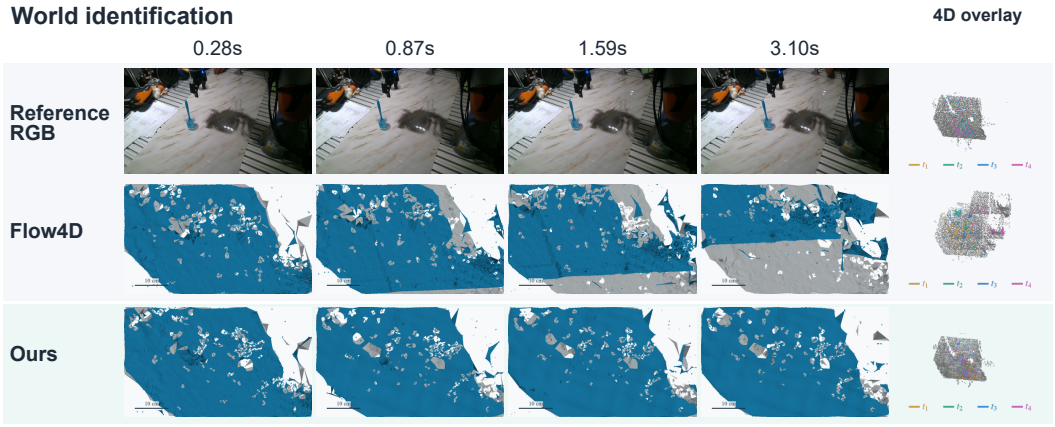


Figure E1. World identification at 0.28, 0.87, 1.59, and 3.10 s. Blue denotes predicted surfaces and gray denotes reference surfaces. At the later times, Flow4D separates from a broad reference region where SympWorld-4D retains more overlap. The right column overlays all four times.

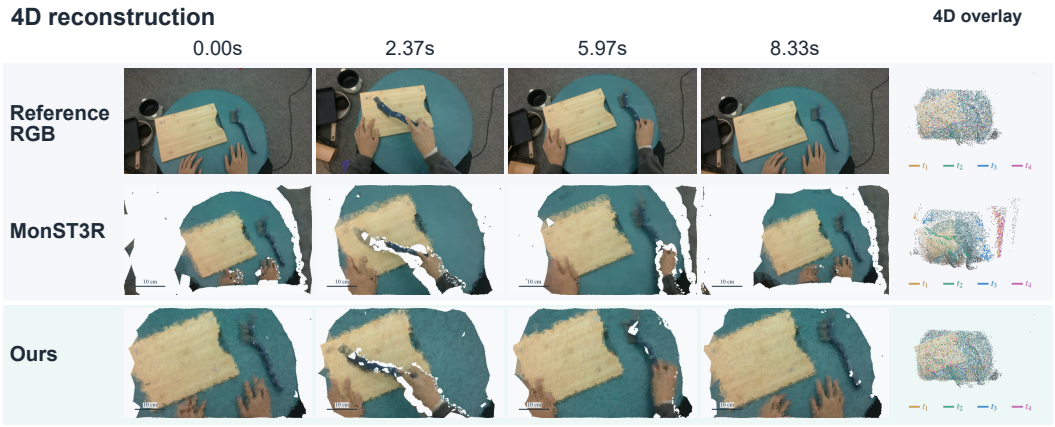


Figure E2. Reconstruction at 0.00, 2.37, 5.97, and 8.33 s. MonST3R and SympWorld-4D show the table, tool, and changing hand configurations as RGB-colored surfaces. Overlays colored by time locate changes across the four observations.

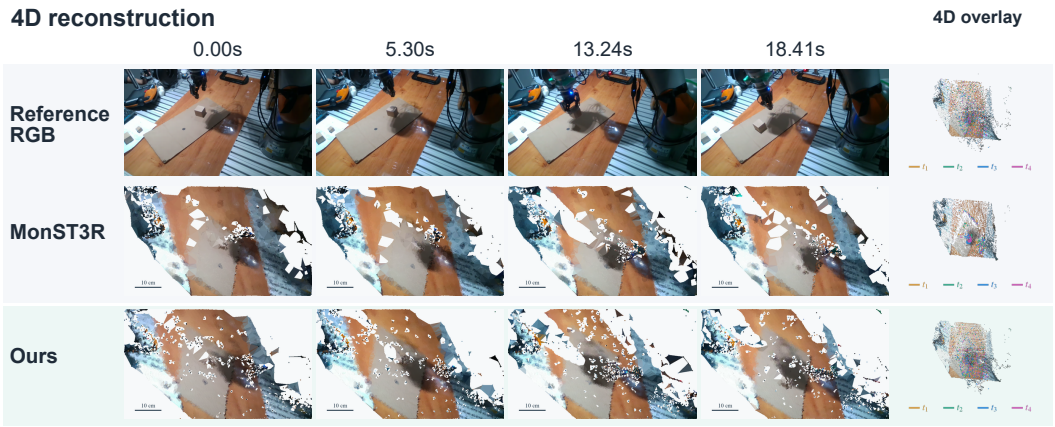


Figure E3. Reconstruction during block manipulation at 0.00, 5.30, 13.24, and 18.41 s. RGB-colored surfaces show the table, work mat, object, and robot. Overlays colored by time locate changes in the reconstructed scene.

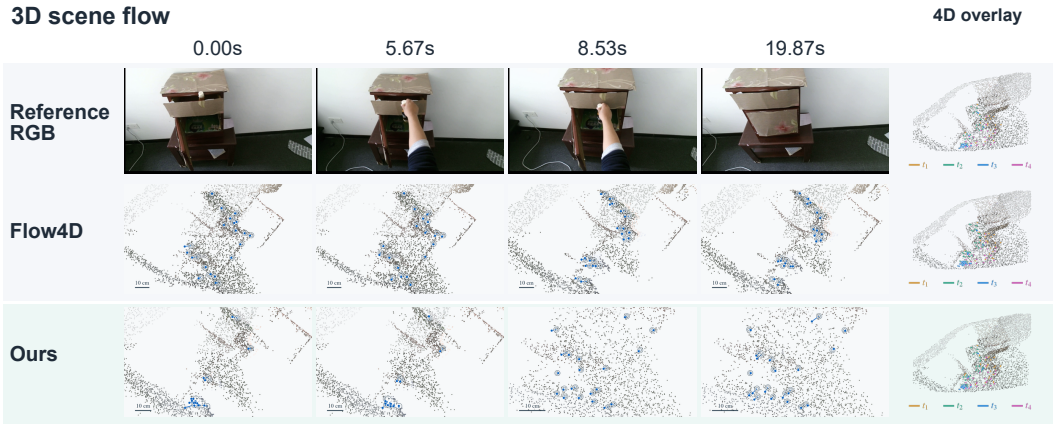


Figure E4. Scene flow during an articulated-object interaction. White rings mark reference endpoints and blue dots mark predicted endpoints at the original motion scale. Four observations and their overlay relate local motion to the moving hand and furniture.

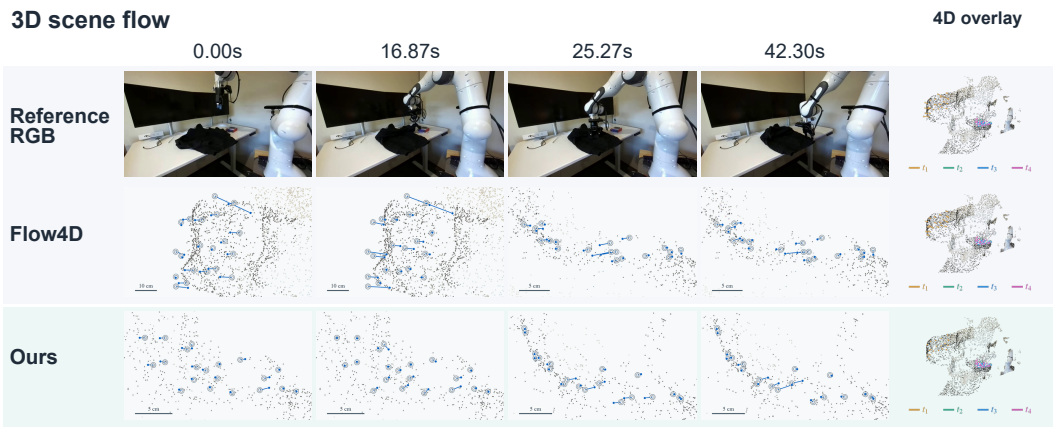


Figure E5. Scene flow at 0.00, 16.87, 25.27, and 42.30 s. White rings denote reference endpoints and blue dots denote predicted endpoints for both Flow4D and SympWorld-4D. Their separation shows the displacement error.

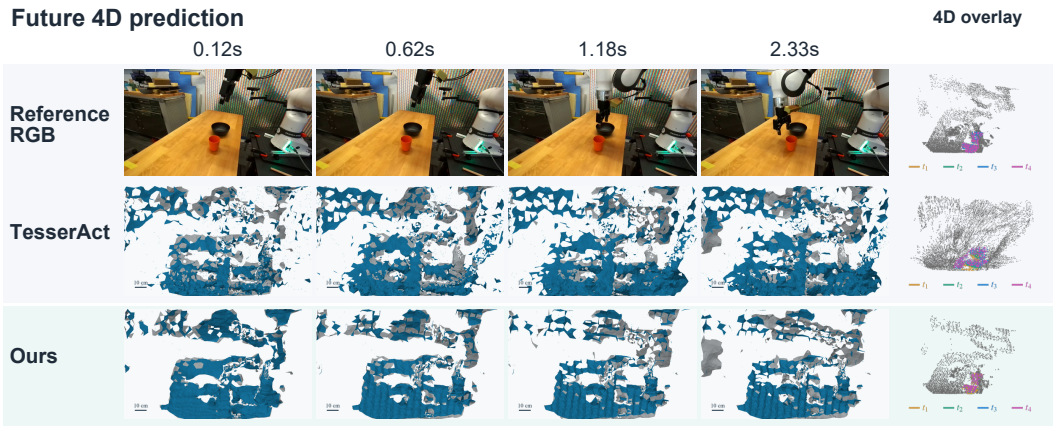


Figure E6. Future geometry at 0.12, 0.62, 1.18, and 2.33 s. Blue predictions are overlaid on gray reference geometry. The SympWorld-4D surfaces preserve a more coherent layout across the displayed times; TesserAct contains more fragmented geometric structure.

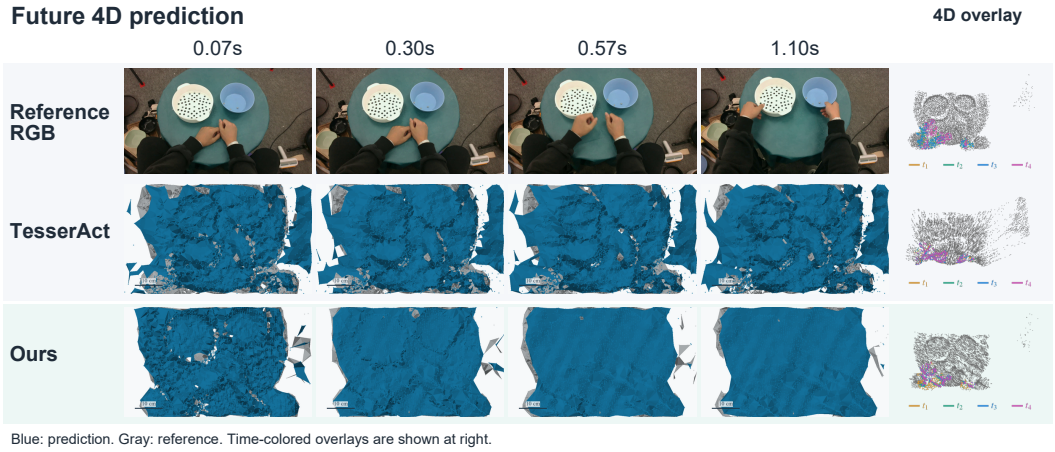
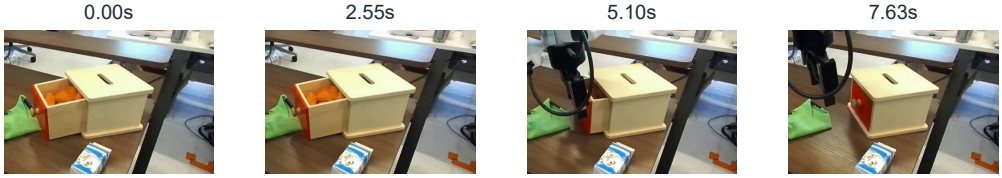


Figure E7. Future geometry at 0.07, 0.30, 0.57, and 1.10 s. Blue predictions are overlaid on gray reference geometry. At later times, SympWorld-4D retains a more continuous layout, while TesserAct contains more fragmented structure.

4D interaction understanding



Q: What interaction occurs between the robot and the scene during 0.00-7.63 s?

A. Two tasks: clean up the table twice.

B. Open or close slidable objects (e.g., drawers or sliding doors).

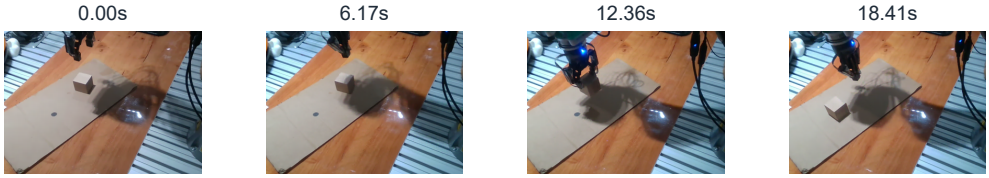
C. Three tasks: pick and place granola from the table to a cabinet.

D. Three tasks: move an object into or out of a container.

InternVideo2-B/14: Choice D (incorrect)	SympWorld-4D: Choice B (correct)

Figure E8. Interaction understanding over 0.00–7.63 s. SympWorld-4D selects opening or closing a slidable object, matching the reference answer. InternVideo2 selects a container-transfer description.

4D interaction understanding



Q: What interaction occurs between the robot and the scene during 0.00-18.41 s?

A. Take the cup off the cup rack.

B. Take everything out of the gift box.

C. Grab the block and place it at the designated location.

D. Stir the beaker with a glass rod.

InternVideo2-B/14: Choice D (incorrect)	SympWorld-4D: Choice C (correct)

Figure E9. Interaction understanding over 0.00–18.41 s. SympWorld-4D selects "Grab the block and place it at the designated location," matching the reference answer. InternVideo2 selects stirring a beaker.