ABDUCTIVE WORLD MODELING VIA CAUSAL REPRESENTATION LEARNING

Anonymous authors
Paper under double-blind review

ABSTRACT

The central challenge of world modeling is to learn representations that capture how the world evolves. However, existing world models predominantly represent future states without explicitly capturing the latent causes underlying their evolution, limiting their ability to reason about why and how the world changes. To address this limitation, we propose *Abductive World Modeling (AWM)*, a framework that learns structured causal representations by abductively inferring latent causes from predicted futures. Specifically, we realize AWM through the Hierarchical Abductive State Pyramid (HASP), which organizes the inferred world state into three complementary components—*Entity*, *Dynamic*, and *Relation*—capturing what exists, how it changes, and how entities interact, respectively. By jointly reasoning over the current observation and its predicted future, HASP abductively infers these latent factors and integrates them into a structured state representation for downstream reasoning. To the best of our knowledge, AWM is the first framework to introduce abductive state inference into latent-space world modeling for learning structured representations of world dynamics. Experiments across physical prediction, causal reasoning, and action understanding demonstrate the effectiveness of our approach. Compared with V-JEPA 2, a state-of-the-art latent-space world model, AWM improves physical prediction AUROC by 10.7%, causal reasoning accuracy by 16.8%, and action Top-1 accuracy by 68.0%.

1 INTRODUCTION

World models aim to compress observations into internal representations that enable agents to anticipate future outcomes, evaluate the consequences of different actions, and generalize across tasks(Ha & Schmidhuber, 2018; Hafner et al., 2019; 2020; 2025; Assran et al., 2025). A central challenge in world modeling is therefore to learn an effective representation of the underlying world state, since the structure and quality of this representation fundamentally determine how well a model can characterize and reason about an evolving environment.

Existing world models learn such representations through several complementary strategies. Generative and reconstruction-based approaches learn latent states by modeling future experience (Ha & Schmidhuber, 2018; Hafner et al., 2019; 2020), with later extensions scaling this paradigm to broader control and interactive environments (Hafner et al., 2025; Bruce et al., 2024). Latent predictive methods instead learn by predicting future states directly in representation space (Assran et al., 2023; Bardes et al., 2024; Assran et al., 2025), including recent visual world models operating over learned features (Zhou et al., 2025; Wang et al., 2026a). Object-centric methods organize representations around individual entities (Locatello et al., 2020; Kipf et al., 2022; Seitzer et al., 2023), while structured dynamics models explicitly capture object interactions (Wu et al., 2023; Battaglia et al., 2016; Kipf et al., 2018; 2020).

However, these representations are primarily optimized to predict what will happen, rather than to explicitly infer the latent structure that explains why and how the world evolves. A representation can therefore accurately predict a future outcome while remaining an entangled encoding of predictive cues, without explicitly exposing the entities, dynamics, and interactions. This limits its usefulness as a structured world state for reasoning beyond prediction itself.

To address this limitation, we propose **Abductive World Modeling (AWM)**, a world modeling framework that constructs structured representations by abductively inferring the latent factors un-

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.

derlying predicted world evolution. The central principle of AWM is *predict forward, then abduce backward*. Given a current observation, a predictive video backbone first produces a latent prediction of the future. Rather than treating this predicted future solely as a target to be matched, AWM treats it as evidence and jointly reasons over it with the current observation to infer a structured latent state. In this way, future prediction provides evidence about what entities are present, how they are changing, and which interactions may account for the predicted evolution. AWM therefore shifts the focus of world modeling from representing only what is likely to happen toward also recovering a structured account of the factors that explain how the world evolves.

We realize AWM through the **Hierarchical Abductive State Pyramid (HASP)**, which progressively constructs the abductive state through three specialized Attributors. The Entity Attributor first organizes visual evidence into object-centered *Entity* states, representing what exists in the scene. Conditioned on these Entity states, the Dynamic Attributor incorporates temporal evidence to construct *Dynamic* states, representing how entities change over time. The Relation Attributor subsequently reasons over entity pairs across time to construct *Relation* states, representing how entities interact. These levels form a hierarchical abductive state that organizes predictive evidence at the entity, temporal, and interaction granularities. By preserving lower-level information while progressively introducing higher-order structure, HASP transforms an entangled predictive representation into a structured state that can be directly examined and exploited for downstream reasoning.

Experiments across physical prediction (Physion++), event reasoning (CLEVRER), and action understanding (EK100) consistently outperform the V-JEPA 2 backbone-only baseline, improving Physion++ AUROC by 10.7%, CLEVRER question accuracy by 16.8%, and EK100 action Top-1 accuracy by 68.0%. Further analyses show that Entity, Dynamic, and Relation states capture complementary object, motion, and interaction information, while targeted interventions confirm selective dependence on task-relevant Entity, temporal, and Relation evidence.

Our contributions are as follows:

- We introduce Abductive World Modeling (AWM), a new perspective on world-state representation that uses predicted futures as evidence to infer the latent structure underlying world evolution through a predict-forward, abduce-backward process.
- We realize AWM through the Hierarchical Abductive State Pyramid (HASP), which hierarchically organizes predictive evidence into Entity, Dynamic, and Relation states at the object, temporal, and interaction levels. We further show that these states expose their intended factors and exhibit selective responses under targeted interventions.
- We extensively evaluate AWM across physical prediction, event reasoning, and action understanding, demonstrating that its structured abductive states provide consistent benefits across diverse video reasoning tasks.

2 Related Work

Latent predictive world models. Latent predictive world models learn representations by predicting future states rather than reconstructing future pixels. JEPA-style methods show that prediction in representation space can produce effective visual representations (Assran et al., 2023; Bardes et al., 2024; Assran et al., 2025). Latent-dynamics models such as PlaNet and Dreamer support prediction, imagination, and control (Hafner et al., 2019; 2020; 2025), while recent methods model dynamics directly in visual feature spaces (Zhou et al., 2025; Wang et al., 2026a). These approaches establish future prediction as an effective signal for world modeling, but their latent states are typically unified representations without explicit entity, dynamic, and relational structure.

Object-centric and relational representations. Object-centric learning organizes visual observations around individual entities. Methods such as IODINE, Slot Attention, and DINOSAUR learn object-centered representations (Greff et al., 2019; Locatello et al., 2020; Seitzer et al., 2023), while video extensions capture such representations over time (Kipf et al., 2022; Zadaianchuk et al., 2023). Other approaches model object-centric dynamics (Jiang et al., 2020; Wu et al., 2023; Song et al., 2025) or explicit interactions between entities (Battaglia et al., 2016; Kipf et al., 2018). Structured dynamics models further use object-level relations to predict physical evolution (Watters et al., 2017; Kipf et al., 2020; Sanchez-Gonzalez et al., 2020). These methods introduce useful structure,

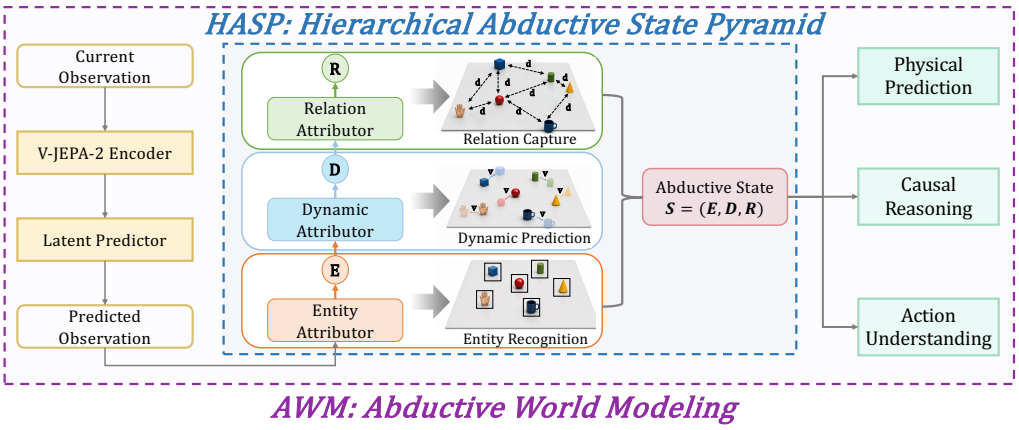


Figure 1: Overview of the AWM. Given an observed context, a frozen predictive visual backbone first predicts its future latent state. The HASP reasons over the current and predicted future representations to infer Entity, Dynamic, and Relation states for downstream prediction and reasoning.

but typically focus on individual aspects such as objects, dynamics, or interactions rather than jointly organizing all three.

Causal representation learning and abduction. Causal representation learning seeks latent variables that reflect underlying causal factors and mechanisms (Schölkopf et al., 2021; Khemakhem et al., 2020), while invariant and independent-mechanism approaches study factors that remain stable across environments (Parascandolo et al., 2018; Arjovsky et al., 2019). Temporal causal representation methods further recover causal factors from sequential observations and interventions (Lippe et al., 2022; 2023). Abductive reasoning instead focuses on inferring latent explanations from observations and their consequences (Peirce, 1997; Zhou, 2019). These directions provide important foundations, but generally do not infer structured causes of world evolution from predicted states.

3 Method

We propose **Abductive World Modeling (AWM)**, which transforms predictive video representations into structured world states through a *predict forward, then abduce backward* process. Given an observed context, a pretrained predictive backbone first estimates its future latent evolution, which AWM then treats as evidence for structured inference. Specifically, the **Hierarchical Abductive State Pyramid (HASP)** organizes visual evidence into *Entity*, *Dynamic*, and *Relation* states: it extracts entities from the observed scene, attributes predicted changes to individual entities, and derives pairwise relations from their temporal evolution.

3.1 From Future Prediction to Abductive State Inference

Let $X_{1:t}$ denote the observed video context. A frozen predictive visual backbone consists of a visual encoder E and latent predictor P:

$$Z_c = E(X_{1:t}), \qquad \widehat{Z}_f = P(Z_c), \tag{1}$$

where Z_c represents the observed world state and $\widehat{Z}_f$ represents its predicted future evolution in latent space. Importantly, $\widehat{Z}_f$ is generated solely from the observed context and never accesses ground-truth future frames.

Although Z_c and $\widehat{Z}_f$ contain rich predictive information, they remain distributed over spatial and temporal visual tokens. Such representations indicate *what future is likely*, but do not explicitly specify which entity, temporal change, or interaction accounts for that prediction. AWM therefore introduces an abductive mapping

$$A : (Z_c, \widehat{Z}_f) \to S = \left(S^E, S^D, S^R\right), \tag{2}$$

where S^E, S^D, and S^R denote Entity, Dynamic, and Relation states.

Rather than estimating these factors independently from the same backbone representation, HASP constructs them hierarchically, progressively organizing distributed patch evidence into entities, entity-specific temporal states, and finally entity-pair temporal states. This ordering reflects the dependency structure of the reasoning problem: temporal change can only be assigned after the corresponding entity is identified, while an interaction can only be described after both entities and their temporal evolution are represented. HASP therefore performs *progressive attribution*, with each level inheriting the structural units from the previous level and attributing additional predictive evidence to them.

3.2 Hierarchical Abductive State Pyramid

At a high level, HASP constructs the three states as

$$S^E = F_E(Z_c), \tag{3}$$

$$S^D = F_D(S^E, Z_c, \widehat{Z}_f) + G_D(S^E, Z_c), \tag{4}$$

$$S^R = F_R(S^E, S^D) + G_R(S^D, Z_c), \tag{5}$$

where F_E, F_D, and F_R are the three Attributors. G_D and G_R are residual information paths that preserve lower-level evidence while higher-order structure is introduced.

A critical property of this hierarchy is that the predicted future $\widehat{Z}_f$ does not directly define all three states. Entity Attribution is grounded in the observed scene, while predicted future evidence is introduced when reasoning about change in the Dynamic Attributor. Future-conditioned information is then propagated from Dynamic to Relation reasoning. Consequently, the hierarchy separates three questions: *what is currently present?*, *what change is implied by the predicted future?*, and *which pairwise interactions are consistent with these changes?*

3.2.1 Entity Attributor: Attributing Patch Evidence to Entities

The first stage converts visual evidence into explicit entity-level units. Direct temporal or relational reasoning on patch tokens is undesirable because the same physical entity may occupy many spatial tokens and move across locations over time. The Entity Attributor therefore establishes a persistent object-centered coordinate system on which subsequent reasoning can operate.

We initialize K learnable entity queries,

$$Q_E^{(0)} = \mathrm{LearnedQueries}, \tag{6}$$

which attend to Z_c and iteratively gather entity-specific visual evidence. At attribution layer l, each query updates its representation through cross-attention:

$$A_E^{(l)} = \mathrm{CrossAttn}\left(\mathrm{LN}(Q_E^{(l)}), \mathrm{LN}(Z_c)\right), \tag{7}$$

followed by iterative query refinement. After the final attribution layer,

$$S^E = Q_E^{(L_E)}. \tag{8}$$

The resulting $S^E \in \mathbb{R}^{K\times d}$ replaces the original patch organization with a fixed set of object-centered states. Conceptually, this stage answers the first abductive question:

Which parts of the distributed visual evidence can be attributed to the same entity?

Only the current representation Z_c is required at this stage because identifying what exists should not depend on hallucinating an entity from a predicted future. This also establishes a stable entity basis before future-dependent reasoning is introduced.

3.2.2 Dynamic Attributor: Attributing Predicted Change to Entities

Entity states alone characterize *what exists*, but do not explain how individual entities account for the predicted transition. The Dynamic Attributor therefore lifts every entity into an entity-specific temporal representation and attributes current–future differences to that entity.

For each Entity state, we construct temporal queries

$$Q_D^{(0)} = \phi_E(S^E) + E_{\text{time}} + E_{\text{future}}, \tag{9}$$

where $\phi_E(S^E)$ preserves entity identity, E_{time} specifies temporal position, and E_{future} explicitly distinguishes predicted-future positions from observed ones.

These entity-conditioned queries jointly attend to the current and predicted future evidence:

$$H_D = \text{CrossAttn}\left(Q_D^{(0)}, [Z_c; \widehat{Z}_f]\right). \tag{10}$$

Temporal dependencies are subsequently integrated along each entity trajectory,

$$S^D = \text{TemporalTransformer}(H_D) + G_D(S^E, Z_c). \tag{11}$$

This design is central to AWM's abductive interpretation. The predicted future provides evidence of what must change, while temporal queries initialized from S^E preserve entity-specific attribution.

The main attribution path therefore captures future-conditioned temporal change, whereas G_D retains entity and current-state information that should not be discarded when modeling motion. The resulting S^D is organized at the entity–time level and answers the second abductive question:

Given an entity, what temporal change is supported by the observed and predicted future states?

This is also the point at which predicted-future evidence first enters the hierarchy. As a result, HASP uses the future specifically to explain dynamics rather than to redefine entity identity.

3.2.3 Relation Attributor: Attributing Dynamics to Interactions

Changes of individual entities are still insufficient to represent many world dynamics. Collision, contact, manipulation, pursuit, and other interactions are inherently defined between entities. The Relation Attributor therefore converts entity-specific dynamics into explicit pairwise temporal states.

For each candidate entity pair (i, j) and temporal position t, we construct

$$r_{ij,t}^{(0)} = \left[S_i^E + S_j^E,\ |S_i^E - S_j^E|,\ S_{i,t}^D + S_{j,t}^D,\ |S_{i,t}^D - S_{j,t}^D|\right]. \tag{12}$$

The shared terms describe properties jointly expressed by the pair, whereas the difference terms explicitly capture their relative entity and dynamic states. This representation is subsequently projected and temporally aggregated:

$$S_{ij,1:T}^R = \text{TemporalAttention}\left(\text{MLP}_R(r_{ij,1:T}^{(0)})\right) + G_R(S^D, Z_c). \tag{13}$$

Relation reasoning is deliberately performed after Dynamic attribution. Instead of attempting to infer interactions directly from raw visual tokens, each relation is constructed from two already-grounded entities and their entity-specific temporal evolution. Thus, interaction evidence is represented in a pair-specific coordinate system.

Because S^D already incorporates evidence from $\widehat{Z}_f$, future information is propagated naturally into Relation inference without requiring the Relation Attributor to independently decode the predicted future representation. The resulting S^R is organized at the entity-pair–time level and answers the third abductive question:

Which interaction between two entities can account for their attributed temporal evolution?

The explicit pair organization additionally makes individual relations addressable. A particular (i, j) state can therefore be independently probed, masked, or intervened upon, which is important for the factor-specific analyses introduced in Sec. 4.2.

3.3 Training Objective

HASP is trained with supervision aligned to the native granularity of each state:

$$\mathcal{L}_{\text{HASP}} = \mathcal{L}_E + \mathcal{L}_D + \mathcal{L}_R + \lambda_{\text{task}}\mathcal{L}_{\text{task}}, \tag{14}$$

Table 1: Main comparison across datasets. All entries are percentage values reported without the percent sign, and the best value in each metric column is boldfaced. "Bal. Acc." and "Acc." denote balanced accuracy and accuracy, respectively, while EK100 results are reported as Top-1/Top-5. Orca and VideoMAE v2 are marked as "—" on CLEVRER because their released models do not provide the predictive module required by the CLEVRER evaluation pipeline.

	Physion++			CLEVRER		EK100 (T1/T5)		
Method	**AUROC**	**Bal. Acc.**	**Acc.**	**Option Acc.**	**Question Acc.**	**Verb**	**Noun**	**Action**
V-JEPA 2	65.20	61.57	61.60	65.46	42.20	55.74/86.81	34.65/62.67	25.67/46.94
Orca	59.15	55.94	56.64	—	—	33.75/74.39	21.54/45.36	13.08/30.46
VideoMAE v2	67.61	62.60	62.86	—	—	59.54/82.57	46.37/68.79	35.96/54.57
AWM (Ours)	**72.19**	**66.17**	**65.63**	**70.89**	**49.31**	**69.05/91.27**	**47.59/75.05**	**43.12/67.34**

where $\mathcal{L}_E$, $\mathcal{L}_D$, and $\mathcal{L}_R$ supervise Entity, Dynamic, and Relation states at the object, entity-time, and pairwise interaction levels, respectively, while $\mathcal{L}_{\text{task}}$ provides optional downstream task supervision.

This factor-aligned supervision is important because HASP is not intended merely to increase representation capacity. Instead, each level is encouraged to expose information at the structural granularity it is designed to represent. Entity supervision encourages object-centered attribution, Dynamic supervision encourages entity-specific temporal attribution, and Relation supervision encourages pair-specific interaction attribution.

Throughout training, the pretrained encoder E and latent predictor P remain frozen, while only HASP and the task-specific readouts are optimized. This setup isolates the contribution of abductive state construction: any performance gains come from reorganizing the predictive evidence already provided by the backbone into structured Entity, Dynamic, and Relation states, rather than from further adapting or improving the underlying future predictor. Additional tensor dimensions and HASP implementation details are provided in Appendix A.1 and Appendix A.2, while factor-specific training objectives are described in Appendix A.5.

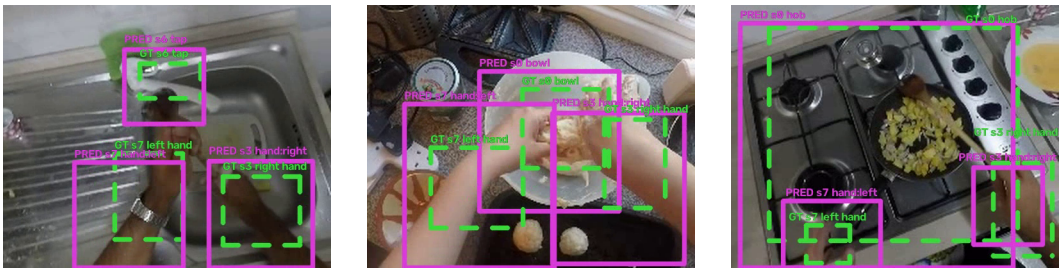
4 RESULTS

We evaluate AWM from three complementary aspects. **Experiment 1** compares its downstream performance with baselines across physical prediction, event reasoning, and action recognition. **Experiment 2** analyzes whether the Entity, Dynamic, and Relation states capture the object, motion, and interaction information associated with their respective Attributors. **Experiment 3** tests whether predictions depend selectively on the corresponding object, temporal, and relational evidence.

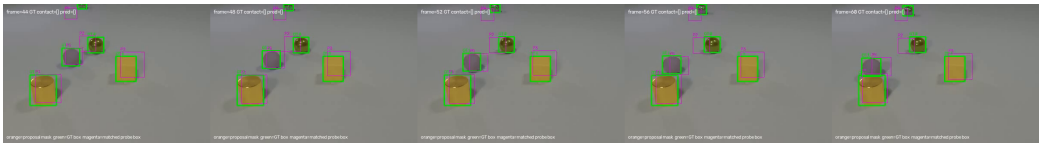
Tasks and Datasets. We evaluate AWM on three complementary video benchmarks: Physion++ (Tung et al., 2023) for physical prediction, CLEVRER (Yi et al., 2020) for future-event and causal reasoning, and EPIC-KITCHENS-100 (EK100) (Damen et al., 2022) for fine-grained action recognition. Dataset details are provided in Appendix C.

Experimental Setup. AWM is built on the V-JEPA 2 predictive video backbone, with HASP producing Entity, Dynamic, and Relation states. The backbone and HASP are frozen during evaluation, and only lightweight probes or task-specific readouts are trained. Experiments are conducted on 32 NVIDIA A800 80GB GPUs. Additional implementation and baseline evaluation details are provided in Appendix D.1.

Baselines. We compare AWM with V-JEPA 2 (Assran et al., 2025), Orca (Wang et al., 2026b), and VideoMAE v2 (Wang et al., 2023). V-JEPA 2 serves as the matched predictive baseline, sharing the same backbone and input protocol with AWM, while Orca and VideoMAE v2 provide external video representation baselines. All methods are evaluated under the same task-specific protocol whenever applicable.More details are provided in Appendix D.1.



(a) Egocentric Entity grounding. (b) Egocentric Entity grounding. (c) Egocentric Entity grounding.



(d) Temporal grounding across consecutive frames.

Figure 2: Qualitative visualization of Entity grounding on EK100 and CLEVRER. (a)–(c) show representative EK100 scenes, where predicted Entity regions align with ground-truth hands and manipulated objects. (d) shows consecutive CLEVRER frames, where matched Entity predictions remain aligned with the corresponding ground-truth objects over time. Green boxes denote ground-truth regions and magenta boxes denote matched predicted regions.

4.1 Experiment 1. Main Results

In this subsection, we compare AWM with V-JEPA 2, Orca, and VideoMAE v2 across physical prediction, event reasoning, and egocentric action recognition. As shown in Table 1, AWM achieves the best overall performance across all evaluated task spaces. Relative to the matched V-JEPA 2 baseline, AWM improves Physion++ AUROC by 10.7%, balanced accuracy by 7.5%, and accuracy by 6.5%, outperforming all three baselines. On CLEVRER, AWM improves option accuracy by 8.3% and question accuracy by 16.8% over V-JEPA 2. Orca and VideoMAE v2 are not evaluated on CLEVRER because their released models do not provide a predictive module required by the CLEVRER evaluation pipeline. On EK100, AWM improves verb Top-1/Top-5 accuracy by 23.9%/5.1%, noun Top-1/Top-5 accuracy by 37.3%/19.8%, and action Top-1/Top-5 accuracy by 68.0%/43.5% over V-JEPA 2, consistently outperforming all baselines. These results demonstrate that the proposed state interface provides strong and consistent gains across heterogeneous video understanding tasks. The following experiments further analyze whether these improvements can be attributed to the intended Entity, Dynamic, and Relation structure.

4.2 Experiment 2. Analysis of Attributor Interpretability

We analyze whether the Entity, Dynamic, and Relation states expose the object, temporal, and interaction information associated with their respective Attributors.

Native Factor Readability. We evaluate the three state levels on Physion++ at their native granularities: Entity for object-level factors, Dynamic for motion, and Relation for pairwise interactions. As shown in Table 2, Entity strongly captures object presence and spatial extent, Dynamic captures motion factors, and Relation achieves high contact readability while retaining distance and time-to-contact information. Together, these results show that the three state levels expose complementary information at their intended entity, temporal, and pairwise granularities. Additional analyses are provided in Appendix B.1.

Table 2: Native factor readability on Physion++.

State	Factor	Metric	Score
Entity	Presence	AUROC	0.9634
	Spatial extent	R^2	0.5316
Dynamic	Speed	R^2	0.7198
	Signed velocity	R^2	0.3805
Relation	Contact	AUROC	0.9789
	Pairwise distance	R^2	0.5815
	Time-to-contact	R^2	0.5542

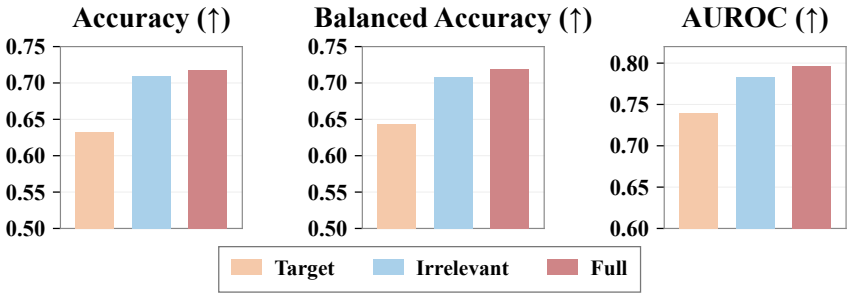


Figure 4: Entity intervention on Physion++ OCP prediction. Target masks the target Entity slot, Irrelevant masks a non-target slot, and Full retains all Entity slots. Performance is evaluated by accuracy, balanced accuracy, and AUROC (↑), showing how prediction quality changes when target-relevant or irrelevant entity information is removed.

Qualitative Entity Grounding. Beyond factor-level probes, we qualitatively examine whether Entity states capture localized and temporally coherent object information. As shown in Figure 2, predicted Entity regions align with ground-truth objects in EK100, including hands and manipulated objects, while remaining consistently aligned with the same objects across consecutive CLEVRER frames. These results provide qualitative evidence that Entity states preserve object-level grounding across both real and synthetic scenes.

Relation Interaction. We examine whether interaction information is specifically concentrated in the Relation state. On CLEVRER, we compare task readouts based on Entity, Dynamic, Relation, and the full state representation for contact prediction and TTC estimation. As shown in Figure 3, Relation performs close to the full representation on both interaction-related tasks, while Entity and Dynamic are less effective when used alone. This suggests that most pairwise interaction information is already captured at the Relation level, whereas the lower-level states primarily encode complementary object and motion information. The result is consistent with the hierarchical design of HASP, where Entity establishes object-level structure, Dynamic introduces entity-specific temporal evolution, and Relation integrates these cues into explicit pairwise interaction states.

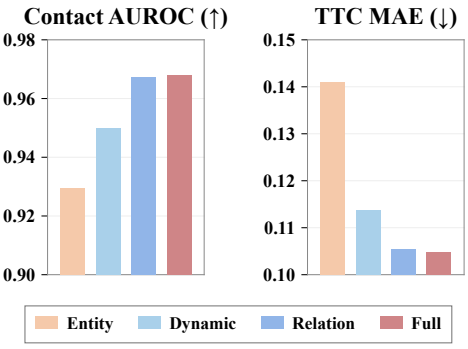


Figure 3: Interaction prediction with individual HASP states on CLEVRER. *Full* uses all three states.

4.3 Experiment 3. Ablation Study on Intervention Consistency

The previous analysis shows that Entity, Dynamic, and Relation states expose readable object, motion, and interaction information. We further test whether downstream predictions depend selectively on the corresponding evidence.

Entity Intervention. We test whether physical outcome prediction on Physion++ depends selectively on object-level evidence represented by the Entity state. As shown in Figure 4, masking the target-object slot reduces OCP accuracy from 0.7175 to 0.6325 and AUROC from 0.7962 to 0.7391. In contrast, masking an irrelevant-object slot retains substantially higher performance, with 0.7088 accuracy and 0.7834 AUROC. The larger degradation under target-object masking indicates that downstream prediction depends more strongly on task-relevant Entity evidence than on arbitrary object information. Additional confusion-matrix and paired-logit analyses of this intervention are reported in Appendix B.2.

Dynamic Intervention. We examine whether temporal variation in the predictive representation is functionally important for Dynamic reasoning. We construct a temporal-static input by averaging the

Table 3: Temporal-static latent intervention on Physion++. *AWM* uses the original context and future features. *AWM w/o Temporal* temporally averages and repeats these features at each time step, removing temporal variation while keeping the model and readouts frozen.

Condition	Speed MAE (↓)	Contact AUROC (↑)	OCP AUROC (↑)
AWM	**0.0670**	**0.9837**	**0.7907**
w/o Temporal	0.1120	0.9485	0.5646

Table 4: Relation-pair intervention on CLEVRER. *AWM* uses all Relation pairs; *AWM w/o Contact Pair* removes the target contact pair; and *AWM w/o Non-contact Pair* removes an unrelated pair. We report geometric and interaction prediction performance.

Condition	Pair-dist. MAE (↓)	TTC MAE (↓)	Contact AUROC (↑)
AWM	**0.0196**	**0.1306**	**0.9633**
w/o Contact Pair	0.0225	0.1376	0.9488
w/o Non-contact Pair	0.0208	0.1349	0.9518

context and predicted future features over time and repeating the resulting features at every temporal position, while keeping all model components and readouts fixed. As shown in Table 3, removing temporal variation increases Speed MAE from 0.0670 to 0.1120, decreases Contact AUROC from 0.9837 to 0.9485, and reduces OCP AUROC from 0.7907 to 0.5646. The consistent degradation across motion, interaction, and physical outcome prediction shows that temporal variation provides important evidence for downstream reasoning.

Relation Intervention. We test whether interaction prediction depends selectively on the Relation state associated with the relevant entity pair. On CLEVRER, we compare masking the interacting pair with masking a non-contact pair. As shown in Table 4, masking the relevant contact pair reduces Contact AUROC from 0.9633 to 0.9488 and increases TTC MAE from 0.1306 to 0.1376. Masking a non-contact pair causes a smaller change, yielding 0.9518 Contact AUROC and 0.1349 TTC MAE. This stronger sensitivity to the interacting pair indicates that Relation states capture pair-specific information that is directly used for contact and TTC prediction.

Together, these interventions provide functional evidence for the three levels of HASP: downstream predictions selectively depend on task-relevant Entity evidence, temporal variation associated with Dynamic reasoning, and interaction-specific Relation states. Additional query-conditioned object intervention results on CLEVRER are provided in Appendix B.3.

5 Conclusion

Existing world models are effective at predicting future states, but their representations often lack explicit causal structure for explaining how the world evolves. In this paper, we proposed **Abductive World Modeling (AWM)**, which follows the principle of *predict forward, then abduce backward* to infer latent causes from predicted future states. AWM instantiates this process with the Hierarchical Abductive State Pyramid (HASP), which organizes world dynamics into complementary Entity, Dynamic, and Relation factors that capture what exists, how it changes, and how entities interact. Experiments across physical prediction, causal reasoning, and action understanding demonstrate the effectiveness of the resulting structured causal representations, improving over the V-JEPA 2 backbone by 10.7% in Physion++ AUROC, 16.8% in CLEVRER question accuracy, and 68.0% in EK100 action Top-1 accuracy.

AI Use Statement

Generative AI tools were used to assist with literature organization and editorial drafting. All technical claims, equations, experimental numbers, and code references were checked against the project files by the authors, who take responsibility for the final manuscript.

Reproducibility Statement

The source code, dataset-isolated training protocols, analysis scripts, and experiment notes are maintained in the accompanying `recipe_formal` directory. The analysis scripts export reusable state features and reproduce the factor probes and intervention tables without modifying training checkpoints.

References

Martin Arjovsky et al. Invariant risk minimization. *arXiv preprint arXiv:1907.02893*, 2019.

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 *CVPR*, 2023.

Mido Assran et al. V-jepa 2: Self-supervised video models enable understanding, prediction and planning. *arXiv preprint arXiv:2506.09985*, 2025.

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

Peter Battaglia et al. Interaction networks for learning about objects, relations and physics. In *NeurIPS*, 2016.

Jake Bruce, Michael D. Dennis, Ashley Edwards, Jack Parker-Holder, Yuge Shi, Edward Hughes, Matthew Lai, Aditi Mavalankar, Richie Steigerwald, Chris Apps, Yusuf Aytar, Sarah Maria Elisabeth Bechtle, Feryal Behbahani, Stephanie C. Y. Chan, Nicolas Heess, Lucy Gonzalez, Simon Osindero, Sherjil Ozair, Scott Reed, Jingwei Zhang, Konrad Zolna, Jeff Clune, Nando de Freitas, Satinder Singh, and Tim Rocktäschel. Genie: Generative interactive environments. In *ICML*, 2024.

Dima Damen, Hazel Doughty, Giovanni Maria Farinella, et al. Rescaling egocentric vision: Collection, pipeline and challenges for epic-kitchens-100. *International Journal of Computer Vision*, 130:33–55, 2022.

Klaus Greff, Raphaël Lopez Kaufman, Rishabh Kabra, Nick Watters, Christopher Burgess, Daniel Zoran, Loic Matthey, Matthew Botvinick, and Alexander Lerchner. Multi-object representation learning with iterative variational inference. In *ICML*, 2019.

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

Danijar Hafner, Timothy Lillicrap, Ian Fischer, Ruben Villegas, David Ha, Honglak Lee, and James Davidson. Learning latent dynamics for planning from pixels. In *ICML*, 2019.

Danijar Hafner, Jurgis Pasukonis, Jimmy Ba, and Timothy Lillicrap. Mastering diverse control tasks through world models. *Nature*, 640(8059):647–653, 2025.

Danijar Hafner et al. Dream to control: Learning behaviors by latent imagination. In *ICLR*, 2020.

Jindong Jiang, Sepehr Janghorbani, Gerard de Melo, and Sungjin Ahn. Scalor: Generative world models with scalable object representations. In *ICLR*, 2020.

Ilyes Khemakhem, Diederik P. Kingma, Ricardo Monti, and Aapo Hyvärinen. Variational autoencoders and nonlinear ica: A unifying framework. In *AISTATS*, 2020.

Thomas Kipf, Ethan Fetaya, Kuan-Chieh Wang, Max Welling, and Richard Zemel. Neural relational inference for interacting systems. In *ICML*, 2018.

Thomas Kipf, Elise van der Pol, and Max Welling. Contrastive learning of structured world models. In *ICLR*, 2020.

Thomas Kipf et al. Conditional object-centric learning from video. In *ICLR*, 2022.

Phillip Lippe, Sara Magliacane, Sindy Löwe, Yuki M. Asano, Taco Cohen, and Stratis Gavves. Citris: Causal identifiability from temporal intervened sequences. In *ICML*, 2022.

Phillip Lippe, Sara Magliacane, Sindy Löwe, Yuki M. Asano, Taco Cohen, and Efstratios Gavves. Causal representation learning for instantaneous and temporal effects in interactive systems. In *ICLR*, 2023.

Francesco Locatello et al. Object-centric learning with slot attention. In *NeurIPS*, 2020.

Giambattista Parascandolo, Niki Kilbertus, Mateo Rojas-Carulla, and Bernhard Schölkopf. Learning independent causal mechanisms. In *ICML*, 2018.

Charles Sanders Peirce. *Pragmatism as a Principle and Method of Right Thinking: The 1903 Harvard Lectures on Pragmatism*. State University of New York Press, 1997.

Alvaro Sanchez-Gonzalez, Jonathan Godwin, Tobias Pfaff, Rex Ying, Jure Leskovec, and Peter Battaglia. Learning to simulate complex physics with graph networks. In *ICML*, 2020.

Bernhard Schölkopf et al. Toward causal representation learning. *Proceedings of the IEEE*, 2021.

Maximilian Seitzer et al. Bridging the gap to real-world object-centric learning. In *ICLR*, 2023.

Yeon-Ji Song, Jaein Kim, Suhyung Choi, Jin-Hwa Kim, and Byoung-Tak Zhang. Ock: Unsupervised dynamic video prediction with object-centric kinematics. In *ICCV*, 2025.

Hsiao-Yu Tung, Mingyu Ding, Zhenfang Chen, Daniel Bear, Chuang Gan, Joshua B. Tenenbaum, Daniel L. K. Yamins, Judith E. Fan, and Kevin A. Smith. Physion++: Evaluating physical scene understanding that requires online inference of different physical properties. In *NeurIPS*, 2023.

Chenting Wang, Yuhan Zhu, Yicheng Xu, Jiange Yang, Ziang Yan, Yali Wang, Yi Wang, and Limin Wang. Internvideo-next: Towards world-understanding video models. In *CVPR*, 2026a.

Limin Wang, Bingkun Huang, Zhiyu Zhao, Zhan Tong, Yinan He, Yi Wang, Yali Wang, and Yu Qiao. Videomae v2: Scaling video masked autoencoders with dual masking. In *CVPR*, 2023.

Yihao Wang et al. Orca: The world is in your mind. *arXiv preprint arXiv:2606.30534*, 2026b.

Nicholas Watters, Andrea Tacchetti, Théophane Weber, Razvan Pascanu, Peter Battaglia, and Daniel Zoran. Visual interaction networks: Learning a physics simulator from video. In *NeurIPS*, 2017.

Ziyi Wu et al. Slotformer: Unsupervised visual dynamics simulation with object-centric models. In *ICLR*, 2023.

Kexin Yi et al. Clevrer: Collision events for video representation and reasoning. In *ICLR*, 2020.

Andrii Zadaianchuk, Maximilian Seitzer, and Georg Martius. Object-centric learning for real-world videos by predicting temporal feature similarities. In *NeurIPS*, 2023.

Gaoyue Zhou, Hengkai Pan, Yann LeCun, and Lerrel Pinto. Dino-wm: World models on pre-trained visual features enable zero-shot planning. In *ICML*, 2025.

Zhi-Hua Zhou. Abductive learning: Towards bridging machine learning and logical reasoning. *Science China Information Sciences*, 62(7):076101, 2019.

A Details of Method

This appendix provides implementation-level details of the predictive backbone interface and the three Attributors in HASP. The description follows the formulation in the main paper and focuses on the tensor organization and computational flow.

A.1 Predictive Backbone Interface

Given an observed video context $X_{1:t}$, the frozen visual encoder produces

$$Z_c = E(X_{1:t}), \qquad Z_c \in \mathbb{R}^{B\times T_c\times N\times d_v}, \tag{15}$$

where B denotes the batch size, T_c the number of observed temporal positions, N the number of spatial visual tokens, and d_v the backbone feature dimension.

The frozen latent predictor subsequently produces

$$\widehat{Z}_f = P(Z_c), \qquad \widehat{Z}_f \in \mathbb{R}^{B\times T_f\times N\times d_v}, \tag{16}$$

where T_f denotes the number of predicted future positions. No ground-truth future representation is used to construct $\widehat{Z}_f$.

For K entity slots and state dimension d, HASP produces

$$S^E \in \mathbb{R}^{B\times K\times d}, \tag{17}$$

$$S^D \in \mathbb{R}^{B\times K\times T_D\times d}, \tag{18}$$

and

$$S^R \in \mathbb{R}^{B\times |\mathcal{P}|\times T_R\times d}, \tag{19}$$

where

$$\mathcal{P} = \{(i,j) \mid 1 \le i < j \le K\} \tag{20}$$

is the set of unordered candidate entity pairs.

In the Physion++ implementation, $K = 8$ and $d = 256$, yielding

$$|\mathcal{P}| = \binom{8}{2} = 28. \tag{21}$$

A.2 Implementation of HASP

A.2.1 Entity Attributor

The Entity Attributor converts the current patch-organized representation into a fixed set of entity-organized states. We initialize K learnable entity queries

$$Q_E^{(0)} \in \mathbb{R}^{K\times d}. \tag{22}$$

The current representation is flattened and projected to the Attributor dimension:

$$\widetilde{Z}_c = \mathrm{Proj}_E\left(\mathrm{Flatten}(Z_c)\right). \tag{23}$$

The entity queries retrieve evidence from the current representation through cross-attention:

$$A_E = \mathrm{CrossAttn}\left(\mathrm{LN}(Q_E), \mathrm{LN}(\widetilde{Z}_c)\right). \tag{24}$$

After the attribution blocks, the resulting entity states are

$$S^E = F_E(Z_c). \tag{25}$$

Thus, the Entity Attributor operates only on the current representation. The predicted future does not enter the hierarchy at this stage.

A.2.2 DYNAMIC ATTRIBUTOR

The Dynamic Attributor converts each Entity state into a temporally resolved representation, allowing predicted changes to be attributed to individual entities. For entity i at temporal position t, we initialize an entity-conditioned temporal query as

$$q_{i,t}^{(0)} = \phi_E(S_i^E) + e_t + e_{\mathrm{src}(t)}, \tag{26}$$

where $\phi_E(S_i^E)$ carries the identity of entity i, e_t encodes the temporal position, and $e_{\mathrm{src}(t)}$ indicates whether the position corresponds to the observed context or the predicted future. Collecting all queries gives $Q_D^{(0)} \in \mathbb{R}^{K\times T\times d_D}$.

The entity-conditioned queries then retrieve evidence jointly from the observed and predicted-future representations:

$$H_D = \mathrm{CrossAttn}\left(Q_D^{(0)}, \mathrm{Proj}_D\left([Z_c; \widehat{Z}_f]\right)\right). \tag{27}$$

Since each query is tied to a specific Entity state, the retrieved current–future evidence is attributed to that entity rather than pooled into a global temporal representation.

Temporal dependencies are then modeled independently along each entity trajectory:

$$\widetilde{S}_{i,1:T}^D = \mathrm{TemporalTransformer}\left(H_{D,i,1:T}\right). \tag{28}$$

This forms the main attribution path, where $\widetilde{S}^D$ captures entity-specific temporal changes inferred from both the observed state and its predicted future.

In parallel, we introduce a residual information path

$$S_{\mathrm{res}}^D = G_D(S^E, Z_c), \tag{29}$$

where G_D maps the Entity states together with the current visual representation into the same entity–time feature space as $\widetilde{S}^D$. Unlike the main attribution path, G_D does not access $\widehat{Z}_f$. Its role is to preserve entity identity and current-state evidence that may not be expressed as temporal change.

The final Dynamic state combines the two paths:

$$S^D = \widetilde{S}^D + S_{\mathrm{res}}^D. \tag{30}$$

Thus, $S^D \in \mathbb{R}^{K\times T\times d_D}$ retains explicit entity and temporal axes: each $S_{i,t}^D$ contains both the future-conditioned change attributed to entity i and the lower-level current-state information preserved by the residual path.

A.2.3 RELATION ATTRIBUTOR

The Relation Attributor converts entity-specific dynamics into explicit pairwise interaction states. We consider all unordered entity pairs $\mathcal{P} = (i,j) \mid 1 \le i < j \le K$. For each pair (i,j) at temporal position t, we construct a pair representation from their Entity and Dynamic states:

$$r_{ij,t} = \left[S_i^E + S_j^E,\ |S_i^E - S_j^E|,\ S_{i,t}^D + S_{j,t}^D,\ |S_{i,t}^D - S_{j,t}^D|\right] \tag{31}$$

The sum terms capture information shared by the two entities, while the difference terms encode their relative entity and dynamic states. Using symmetric operations also makes $r_{ij,t}$ invariant to the ordering of the pair.

Each pair representation is first projected into the Relation feature space and then aggregated along its temporal trajectory:

$$\widetilde{S}_{ij,1:T_R}^R = \mathrm{TemporalAttention}\left(\mathrm{MLP}_R\left(r_{ij,1:T_R}\right)\right). \tag{32}$$

This forms the main relation-attribution path, where $\widetilde{S}_{ij,t}^R$ represents the interaction evidence associated with entity pair (i,j) at temporal position t.

In parallel, we introduce a residual information path

$$S_{\mathrm{res}}^R = G_R(S^D, Z_c), \tag{33}$$

where G_R maps the Dynamic states together with the current visual representation into the same pair–time feature space as $\widetilde{S}^R$. This path preserves lower-level dynamic and current-state evidence that may not be fully retained by the explicit pairwise transformation.

The final Relation state is obtained by combining the two paths:

$$S^R = \widetilde{S}^R + S^R_{\text{res}}. \tag{34}$$

Importantly, the Relation Attributor does not independently attend to $\widehat{Z}_f$. Future-conditioned evidence has already been attributed to individual entities in S^D and is therefore propagated naturally into pairwise reasoning. The resulting S^R preserves explicit pair and temporal axes, with each $S^R_{ij,t}$ describing the attributed interaction state of entity pair (i, j) at temporal position t.

A.3 Hierarchical State Construction

The complete HASP computation follows the sequential attribution structure:

$$S^E = F_E(Z_c), \tag{35}$$

$$S^D = F_D(S^E, Z_c, \widehat{Z}_f) + G_D(S^E, Z_c), \tag{36}$$

$$S^R = F_R(S^E, S^D) + G_R(S^D, Z_c). \tag{37}$$

Accordingly, the native representation axis changes progressively as

$$(T_c, N) \to (K) \to (K, T_D) \to (|\mathcal{P}|, T_R). \tag{38}$$

This sequential organization distinguishes HASP from independent feature heads: each Attributor consumes the structured state produced by the preceding level.

A.4 Structured State and Downstream Readout

The three levels jointly form the abductive world state

$$S = (S^E, S^D, S^R). \tag{39}$$

Importantly, the levels are complementary rather than interchangeable. S^E provides object-centered identity and spatial evidence, S^D binds predicted temporal change to individual entities, and S^R represents temporally evolving pairwise interactions. Higher levels introduce additional structural organization while residual paths preserve information established at lower levels.

For a downstream task τ with target Y_τ, a lightweight task-specific readout operates on the state:

$$\widehat{Y}_\tau = R_\tau(S^E, S^D, S^R). \tag{40}$$

This separation allows us to evaluate whether a structured abductive state provides a more useful interface for reasoning than the original predictive latent representation without modifying the predictive backbone itself.

A.5 Factor-specific Training Objectives

HASP is optimized with factor-specific objectives defined at the native granularity of each state:

$$\mathcal{L}_{\text{HASP}} = \mathcal{L}_E + \mathcal{L}_D + \mathcal{L}_R + \lambda_{\text{task}}\mathcal{L}_{\text{task}}. \tag{41}$$

Here, $\mathcal{L}_E$ supervises Entity states, $\mathcal{L}_D$ supervises entity-time Dynamic states, and $\mathcal{L}_R$ supervises entity-pair-time Relation states. When a downstream task is jointly optimized, $\mathcal{L}_{\text{task}}$ provides the corresponding task-level supervision.

For the structured probe used in Physion++, the factor-specific objectives are instantiated over the corresponding object, temporal, and pairwise annotations. Invalid object, time, or pair entries are excluded from the respective losses.

Throughout HASP training, the pretrained encoder E and latent predictor P remain frozen. The predictor is trained separately to provide the current-to-future latent prediction and is subsequently fixed during HASP optimization.

Table 5: Additional diagnostics of the Physion++ Entity slots.

Diagnostic	Value
Presence F1	99.77
Slot occupancy	64.52
Identity consistency	70.29
ID switch rate	4.05
Target-object accuracy	95.79
Object-type accuracy	92.86
Color RGB MAE	0.2244
Center MAE	0.0454
Geometry MAE	0.0285
Velocity MAE	0.00629

Table 6: Physion++ OCP confusion counts under probe interventions.

Condition	TP	TN	FP	FN
Visual-only	237	292	129	142
Visual + ShallowProbe	277	297	124	102
Target mask	325	181	240	54
Irrelevant mask	262	305	116	117
Random slot	279	294	127	100

A.6 Complete Inference Procedure

The complete AWM inference procedure can be summarized as follows.

Algorithm AWM (Abductive World Modeling). Given observed video context $X_{1:t}$, first compute $Z_c \leftarrow E(X_{1:t})$ and $\widehat{Z}_f \leftarrow P(Z_c)$. Then construct Entity states $S^E \leftarrow F_E(Z_c)$ from the current representation. Next, use the Entity states together with the current and predicted-future representations to construct the Dynamic states $S^D \leftarrow F_D(S^E, Z_c, \widehat{Z}_f) + G_D(S^E, Z_c)$. Finally, enumerate the unordered entity pairs $\mathcal{P} = \{(i, j) : i < j\}$, construct pair representations from the Entity and Dynamic states, and obtain $S^R \leftarrow F_R(S^E, S^D) + G_R(S^D, Z_c)$. The resulting hierarchical state $S = (S^E, S^D, S^R)$ is passed to the downstream readout $\widehat{Y}_\tau \leftarrow R_\tau(S)$.

B Additional Results

B.1 Object Identity Diagnostics

We further examine whether the Entity state forms an object-centered representation rather than a pooled scene feature. Beyond the native-factor probes reported in the main text, we evaluate slot occupancy, identity consistency, object attributes, and geometric properties on Physion++.

As shown in Table 5, Entity slots exhibit strong object-level semantics, achieving 99.77% presence F1, 95.79% target-object accuracy, and 92.86% object-type accuracy. Identity consistency reaches 70.29%, with an ID switch rate of 4.05%, indicating that object identity is substantially preserved across time. The low center, geometry, and velocity errors further show that individual slots retain spatial and motion information associated with their corresponding objects. These diagnostics complement the native-factor results in the main text by providing a more detailed characterization of the object-centered structure of the Entity state.

B.2 Additional Entity Intervention Diagnostics

We further analyze the Physion++ Entity intervention from Experiment 3 by examining how target-object masking changes the OCP decision pattern. In addition to the aggregate accuracy and AUROC results reported in the main text, we report confusion counts and paired prediction-logit changes.

Table 7: Paired OCP-logit changes under probe-level interventions.

Condition	Mean change	Mean absolute change
Target mask	+0.49195	0.67680
Irrelevant mask	-0.08051	0.17355
Random slot	+0.02437	0.08895

Table 8: Relevant-object intervention on CLEVRER predictive queries. *AWM* uses all Entity evidence; *AWM w/o Relevant Object* removes the object referred to by the query; and *AWM w/o Irrelevant Object* removes an object unrelated to the query.

Condition	Question Correct	Question Acc. ↑	Option Correct	Option Acc. ↑
AWM	1,785/3,557	**50.18%**	5,099/7,114	**71.68%**
AWM w/o Relevant Object	1,546/3,557	43.46%	4,752/7,114	66.80%
AWM w/o Irrelevant Object	1,684/3,557	47.34%	4,948/7,114	69.55%

Table 6 shows that target masking produces a systematic change in the prediction boundary. Relative to the full Visual + ShallowProbe condition, the number of true positives increases from 277 to 325, while true negatives decrease from 297 to 181 and false positives increase from 124 to 240. Thus, removing the target Entity slot does not simply suppress positive evidence; instead, it substantially alters how the readout separates positive and negative outcomes. In comparison, irrelevant-slot masking and random-slot intervention produce much smaller changes in the confusion pattern.

The paired-logit analysis in Table 7 further quantifies this sensitivity. Target masking produces a mean absolute logit change of 0.67680, approximately 3.9 times that of irrelevant masking (0.17355) and 7.6 times that of random-slot intervention (0.08895). These results provide additional evidence that OCP prediction is selectively sensitive to the Entity representation associated with the target object rather than to arbitrary perturbations of the slot representation.

B.3 CLEVRER Relevant-Object Intervention

We additionally evaluate whether predictive reasoning on CLEVRER depends selectively on objects that are relevant to the current query. The evaluation contains 3,557 predictive questions and 7,114 answer options. We compare masking the query-relevant object with masking an irrelevant object while keeping the model and task readout fixed.

As shown in Table 8, masking the relevant object reduces question accuracy from 50.18 to 43.46 and option accuracy from 71.68 to 66.80. In comparison, masking an irrelevant object retains higher performance, with 47.34 question accuracy and 69.55 option accuracy.

The larger degradation under relevant-object masking shows that the performance drop is not caused merely by removing an arbitrary object. Instead, the CLEVRER readout is more sensitive to Entity evidence associated with the current query, providing additional evidence that object-level information is used selectively during downstream reasoning.

C Details of Assets Used in This Paper

In our experiments, we evaluate AWM on three video benchmarks spanning physical prediction, causal reasoning, and action understanding. We follow the corresponding benchmark settings and use the same data splits for all compared methods.

C.1 Physical Prediction Dataset

Physion++ (Tung et al., 2023) is a benchmark for evaluating physical prediction from videos of interacting objects. The scenes contain diverse physical configurations and interactions, requiring models to reason about object dynamics and predict future physical outcomes. In our experiments,

Physion++ is used to evaluate whether the learned representation captures physical information relevant to future evolution.

C.2 Causal Reasoning Dataset

CLEVRER (Yi et al., 2020) is a synthetic video reasoning benchmark designed to evaluate understanding of objects, motion, collisions, and temporal events. It contains questions that require reasoning about observed and future events based on the dynamics and interactions among objects. In our experiments, CLEVRER is used to evaluate whether AWM captures Entity, Dynamic, and Relation structure that supports causal reasoning about video events.

C.3 Action Understanding Dataset

EPIC-KITCHENS-100 (EK100) (Damen et al., 2022) is a large-scale egocentric video benchmark containing unconstrained first-person recordings of everyday activities. Each action segment is annotated with a verb and noun pair, together defining an action class. In our experiments, EK100 is used to evaluate verb, noun, and action recognition, providing a substantially different setting from the synthetic and physics-oriented benchmarks above.

C.4 Data Splits and Evaluation Settings

For each benchmark, we follow its corresponding experimental protocol and use the same training and evaluation splits across all compared methods. Physion++ is evaluated for physical prediction, CLEVRER for causal reasoning, and EK100 for verb, noun, and action recognition. For CLEVRER, we evaluate only on the predictive subset of the test set, which contains questions requiring prediction of future events.

D Details of Methods and Experimental Settings

Table 9: **Main configuration and hyperparameter settings of AWM.** We report the predictive backbone, evaluation configuration, and computational setup used in our experiments.

Parameter	Value
Predictive Backbone	
Backbone	V-JEPA 2 ViT-H
Observed frames	16
Predicted future frames	16
Patch size	16
Tubelet size	2
Predictor depth	12
Predictor embedding dimension	384
Predictor attention heads	12
Evaluation Configuration	
Visual encoder	Frozen
Latent predictor	Frozen
HASP state extractor	Frozen
Trainable modules	Factor probes / task readouts
Readout optimizer	AdamW
Precision	BF16
Random seed	239

D.1 Details of Experimental Setup

In this section, we provide additional details about the implementation of AWM and the baselines, together with the evaluation settings used in our experiments.

Implementation Details of the Baselines. We compare AWM with V-JEPA 2, Orca, and VideoMAE v2. Within each benchmark, we use the same data splits and evaluation metrics whenever the corresponding model supports the required protocol. AWM and V-JEPA 2 share the same predictive backbone and input setting, providing a matched comparison for evaluating the contribution of abductive world modeling.

V-JEPA 2 is the matched predictive baseline of AWM. It uses the same ViT-H encoder and latent predictor as AWM to encode the observed context and predict future latent states. Downstream readouts operate directly on the V-JEPA 2 representations without introducing explicit Entity, Dynamic, or Relation states. This comparison therefore isolates the effect of HASP under the same predictive backbone.

Orca is an external video representation baseline. Since Orca does not provide a future predictor compatible with our evaluation pipeline, it cannot directly generate the predicted future latent representation required by AWM. On Physion++, where ground-truth future frames are available, we therefore encode both the current and future videos with the frozen Orca encoder and use the resulting future representation as a substitute for the predicted future latent. On EK100, the Orca encoder is kept frozen and only a lightweight attentive readout is trained for verb, noun, and action recognition.

VideoMAE v2 is another external video representation baseline. As VideoMAE v2 also does not provide a compatible future predictor, we use the same substitution on Physion++ by encoding the available ground-truth future frames to obtain the future latent representation. On EK100, unlike Orca, the pretrained VideoMAE v2 encoder is fine-tuned together with the task-specific verb, noun, and action classification heads.

Because the predictive subset of CLEVRER requires reasoning about unobserved future events and does not provide the corresponding ground-truth future frames, this substitution is not available for Orca or VideoMAE v2. We therefore do not report CLEVRER results for these two methods rather than approximating the missing predicted future representation.

AWM is our proposed framework and is built on the same V-JEPA 2 predictive backbone used by the matched baseline. Given an observed video context, the frozen encoder produces the current representation Z_c, and the frozen latent predictor generates the predicted future representation $\widehat{Z}_f$ without accessing ground-truth future frames. HASP then performs backward abductive inference over these representations to construct the Entity, Dynamic, and Relation states. During downstream evaluation, the V-JEPA 2 encoder, predictor, and HASP are kept frozen, and only lightweight factor probes or task-specific readouts are trained.

Hyperparameters. Table 9 summarizes the main configuration of AWM used throughout our experiments. Unless otherwise specified, we use the same frozen representation for downstream evaluation and train independent readouts for physical prediction, causal reasoning, and action understanding.

D.2 Algorithm of Abductive World Modeling

We summarize the inference procedure of AWM in Algorithm 1. AWM first predicts a future latent state from the observed video context and then performs backward abductive inference through HASP. The Entity Attributor identifies what exists, the Dynamic Attributor infers how the identified entities change using the predicted future, and the Relation Attributor infers how pairs of entities interact. These factors jointly form the structured causal state used for downstream prediction and reasoning.

Algorithm 1 Abductive World Modeling

Require: Observed video context $X_{1:t}$
Require: Frozen visual encoder E and latent predictor P
Require: Entity, Dynamic, and Relation Attributors F_E, F_D, and F_R
Require: Task-specific readout R_τ

Encode the observed context: $Z_c \leftarrow E(X_{1:t})$
Predict the future latent state: $\widehat{Z}_f \leftarrow P(Z_c)$
Abduce backward through HASP:
Infer the Entity state: $S^E \leftarrow F_E(Z_c)$
Infer the Dynamic state: $S^D \leftarrow F_D(S^E, Z_c, \widehat{Z}_f) + G_D(S^E, Z_c)$
Construct candidate entity pairs: $\mathcal{P} \leftarrow \{(i,j) \mid i < j\}$
for each $(i,j) \in \mathcal{P}$ **do**
    for $t = 1, \ldots, T_R$ **do**
        Construct pairwise Entity and Dynamic features:
$r_{ij,t} \leftarrow \left[S^E_i + S^E_j,\ |S^E_i - S^E_j|,\ S^D_{i,t} + S^D_{j,t},\ |S^D_{i,t} - S^D_{j,t}|\right]$
    end for
end for
Infer the Relation state: $S^R \leftarrow F_R(\{r_{ij,1:T_R}\}_{(i,j)\in\mathcal{P}}) + G_R(S^D, Z_c)$
Construct the unified abductive state: $S \leftarrow (S^E, S^D, S^R)$
Produce the downstream prediction: $\widehat{Y}_\tau \leftarrow R_\tau(S)$
return $S, \widehat{Y}_\tau$