PROBING SPATIAL STATE REPRESENTATIONS IN VISUAL FOUNDATION MODELS

Anonymous authors
Paper under double-blind review

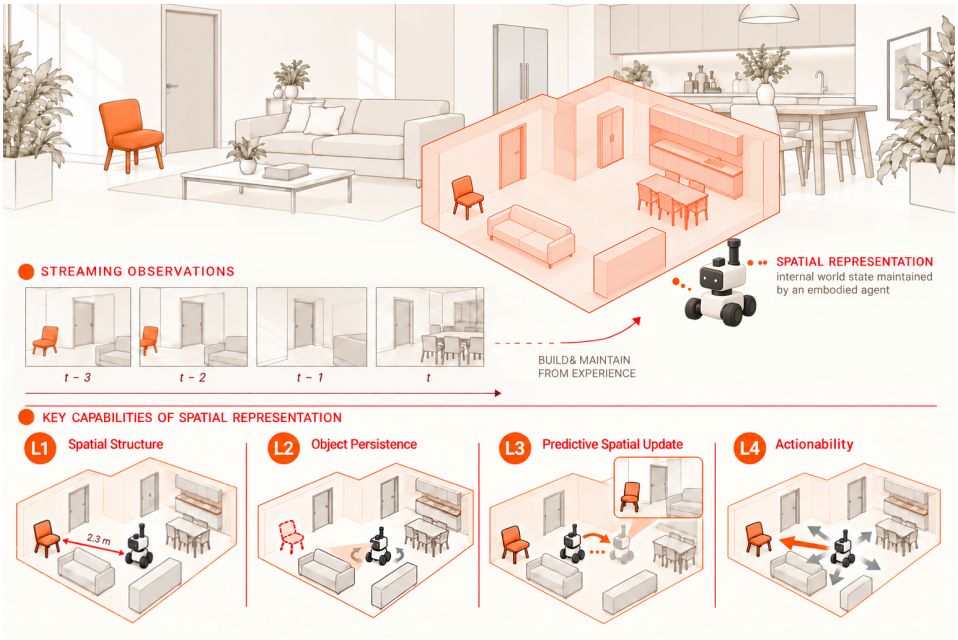


Figure 1: **Spatial state representations from streaming egocentric observations.** A moving observer builds an internal spatial representation from partial views of an environment. **L1** (*spatial structure*) captures scene geometry and object identity across views. **L2** (*object persistence*) retains the observer-relative locations of previously seen objects after they leave the field of view. **L3** (*predictive spatial update*) anticipates what the observer would see after an action or action sequence. **L4** (*actionability*) uses the represented state and a specified goal to select an action that makes progress toward that goal.

ABSTRACT

An observer sees only part of the surrounding world, yet must keep track of what lies beyond the current view. Spatial perception therefore concerns more than recognizing individual images: it calls for a persistent spatial state that can be updated under self-motion and used to guide action. Do visual foundation models provide representations that support such a state? We investigate this question through four levels of spatial capability: spatial structure, object persistence, predictive spatial update, and actionability, using frozen feature probes across ten representations under streaming egocentric observations in static indoor environments. The resulting capability profiles vary across tasks, network layers, and prefix lengths: strong visible scene readout does not ensure accurate object persistence or future feature retrieval. Matched controls further distinguish readout accuracy from reliance on visual history and motion. Native future prediction retains recoverable spatial structure but changes its interface to a fixed spatial readout. Finally, goal-dependent candidate selection does not establish reliable continuous action utility: in the continuous output setting, predicted actions respond to goal changes, yet using the correct goal does not reliably improve progress. Together, these findings motivate evaluating spatial representations not only by what can be decoded, but

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.

by whether that information can be maintained beyond observation, predictively updated, and used for goal-directed action.

Keywords: visual foundation models, spatial state representations, probing, object persistence, actionability.

1 INTRODUCTION

Spatial representation is one of the most fundamental capacities of biological perception. Humans and other animals do not experience the world as a sequence of independent retinal images. Through continuous exploration, they maintain information about surrounding space, keep track of objects that are no longer visible, update their location as the observer moves, and use this information to guide subsequent behavior. These abilities are not inherently linguistic: they arise from perception and sensorimotor interaction with the environment.

Several classical ideas in perception and spatial cognition point to this view. Object file theories describe how object-specific information can remain associated with the same entity across changes in its visual appearance (Kahneman et al., 1992). Spatial updating and path integration describe how an internal estimate of location can be continuously revised using self-motion (McNaughton et al., 2006). Sensorimotor contingency theories further emphasize that perception involves knowledge of how sensory input would change under possible actions (O'Regan & Noë, 2002). Together, these perspectives suggest that a useful representation of space should describe more than what is visible in the current image: it should maintain a state of the surrounding world that persists and evolves as the observer interacts with it.

An observer's **spatial state** describes the configuration of the surrounding environment in relation to the observer, including previously observed objects that are no longer visible. A **spatial state representation** is an internal encoding of this configuration inferred from the observer's visual history. We do not prescribe an explicit 3D map, object-centric scene graph, coordinate system, or neural architecture. Instead, we characterize its functional usefulness through four capabilities:

- **L1 Spatial structure.** Recover the spatial organization of the observed scene through metric geometry, instance identity, and view consistency.
- **L2 Object persistence.** Retain information about a previously observed object's location when the object is no longer directly visible.
- **L3 Predictive spatial update.** Anticipate changes in the observer-relative state under ego-motion, evaluated through one-step prediction and path integration.
- **L4 Actionability.** Use the represented state and a specified goal to predict an action that makes progress toward that goal.

These capabilities, summarized in Fig. 2, describe complementary uses of spatial information: reading the current scene, retaining information beyond observation, predicting motion-conditioned changes, and using this information to act.

We study these functions in a streaming setting: the observer receives past and current visual evidence but no future observations. Representations are extracted from causal visual prefixes of increasing length. Our experimental scope is observer motion through static indoor environments, rather than object dynamics, contact physics, or complete physical world modeling.

This question spans *visual foundation models*, including image and video encoders, 3D models, multimodal models, video generators, and world models. Strong performance on a model's training objective does not establish all four spatial functions, and success at one level need not carry over to the next.

We therefore conduct a **functional probing study** of ten frozen representations. Six main diagnostics provide seven measurements of spatial structure, object persistence, and predictive spatial update: metric geometry contributes separate depth and camera pose metrics. Goal-conditioned action prediction additionally tests actionability in candidate selection and continuous output settings. Beyond a single extraction point, we organize our analyses by

$$\text{network depth} \times \text{streaming context}, \tag{1}$$

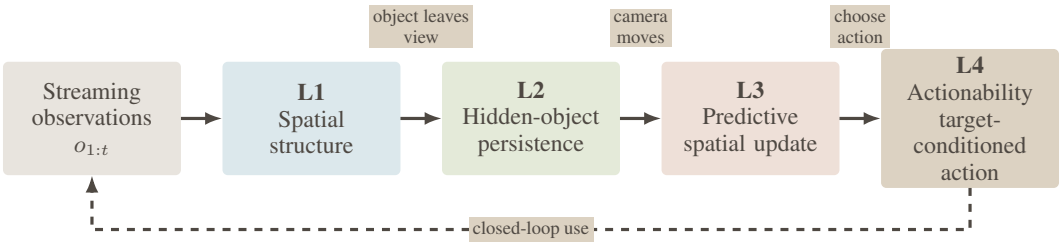


Figure 2: **Functional relationships among the four spatial capabilities.** L1 grounds the spatial state in observed scene structure. L2 extends the demand from representing what is visible to retaining information when objects leave the field of view, while L3 requires anticipating how the observer-relative state changes under ego-motion. L4 connects the represented state to goal-directed action. The dashed return arrow illustrates potential closed-loop use: actions change the observer's viewpoint and thus the subsequent observations, while the environment remains static.

Matched controls alter motion, history, transformation order, or goals while keeping the relevant predictor and target fixed.

Findings. No representation leads on all tasks: Fast3R reads camera pose best, VideoMAE localizes out-of-view objects best despite weaker visible-scene readout, and DINOv2 and Qwen2.5-VL lead future prediction despite weaker persistence (Section 4.1). These profiles shift with layer, and longer prefixes worsen persistence and path integration for every model (Section 4.2). Controls then separate *decodability* from *input use*: copying the last feature rivals learned prediction for V-JEPA2, and visual history helps V-JEPA2 and Wan but not VideoMAE, the most accurate localizer (Section 4.3). Native prediction keeps recoverable geometry but not the original readout interface (Section 4.4), and input use is not *functional utility*: changing the goal rotates predicted actions by 30–46°, yet the correct goal does not reliably improve progress (Section 4.5).

Contributions. We (i) operationally characterize spatial state representations through four functional capabilities; (ii) develop a frozen feature probing framework covering ten representations, with controls for history, motion, predictive readout, and goal use; and (iii) show that readout accuracy, input use, and goal-specific utility dissociate across models, layers, and prefix lengths, motivating multidimensional evaluation of spatial state.

2 Probing Spatial State Representations

We ask what spatial information is accessible in frozen visual features, what inputs support that information, and whether it remains useful for prediction and goal-directed action.

2.1 Four functional capabilities

Figure 3 summarizes the probing tasks used to instantiate the four levels of spatial capability.

L1: Spatial Structure. We first ask what spatial structure is directly readable from the streaming visual representation. The **metric geometry** task takes frozen features extracted from an egocentric video prefix and predicts the current camera pose and scene depth, testing whether metric 3D structure is accessible from the representation. The **instance identity** task uses the same visual features to separate individual object instances, testing whether the representation organizes the visible scene at the object level rather than only encoding global geometry. In addition, the **view consistency** task performs cross-view correspondence matching between features extracted from two perspectives of the same scene, evaluating whether spatial correspondences are preserved across changes in viewpoint.

L2: Object Persistence. We then test whether information about an object remains spatially available after the object is no longer visible. We evaluate this capability through **out-of-view object localization**. A target object is first observed in an earlier frame and represented by an object query q_{obj}. At the end of the video prefix, the target has moved outside the observer's field of view because of camera motion. Given the frozen representation of the whole trajectory together with q_{obj},

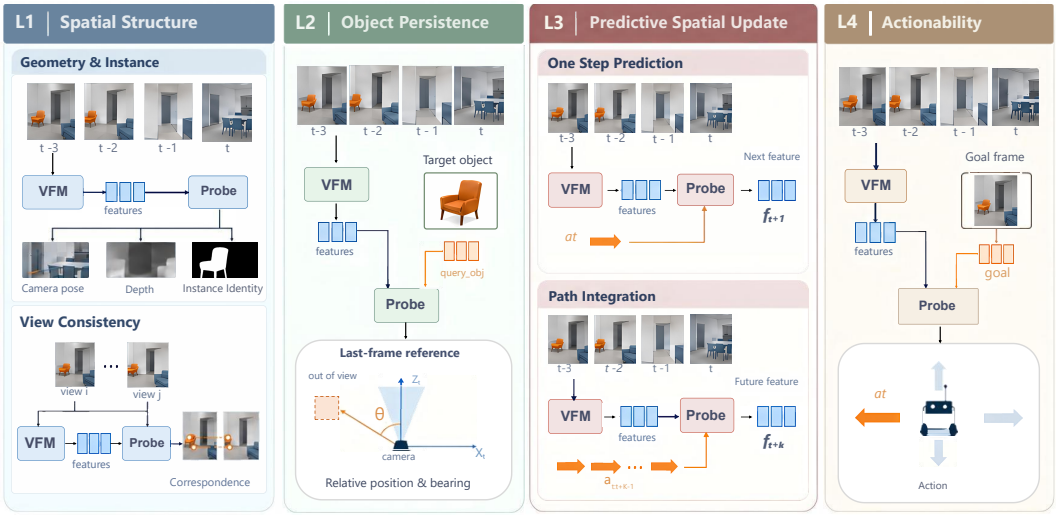


Figure 3: **Probing four spatial capabilities.** Probe inputs and targets for each level: prefix features alone (L1), plus an object query (L2), ego-motions and an isolated future feature (L3), or a goal frame and an action (L4).

the probe predicts the target object's location relative to the final camera, represented by its relative position and bearing. This task therefore asks whether the representation can maintain the spatial state of a previously observed object beyond direct visual evidence, rather than merely detect or recognize objects that remain visible.

L3: Predictive Spatial Update. Next, we ask whether the representation supports motion-conditioned prediction of how the observer-relative spatial state will change. In **one-step prediction**, the probe receives the current visual representation together with the next ego-motion a_t and predicts the feature of the next observation, f_{t+1}. In **path integration**, the probe instead receives a sequence of ego-motions $a_{t:t+K-1}$ and predicts the representation reached after integrating those motions, f_{t+K}. The two tasks test whether spatial state can be updated according to a single movement and over a longer sequence of movements, respectively.

L4: Actionability. Finally, we test whether the spatial representation can be used for goal-directed action through **goal-conditioned action prediction**. The probe receives the representation of the current visual history together with an independently encoded goal frame q_{goal}, and predicts an action a_t intended to move the observer toward that goal. This level asks whether the spatial information exposed by the representation is not only descriptive or predictive, but also usable for selecting behavior conditioned on a desired future state.

2.2 Causal Streaming Formulation

For model m and layer ℓ, frozen encoder E_ℓ^m encodes prefixes that share a start frame and differ only in endpoint p:

$$\mathbf{Z}_p^{m,\ell} = E_\ell^m(\mathbf{I}_{1:p}), \qquad p \in \{8, 12, 16, 24\}. \tag{2}$$

Each prefix is encoded independently, contains no frame after p, and is treated as a *candidate spatial state representation* at every level. For L2, the object query is pooled from frames where the object was visible, and the target is its bearing in the camera frame at p; neither the visibility mask nor the target location is given to the probe. No recurrent state update is required, and since a longer prefix also moves the endpoint, comparisons across p do not isolate added history. L3–L4 additionally use a future frame encoded without access to the prefix or any later frame,

$$\mathbf{Z}_{p+h}^{m,\ell,\text{tar}} = E_\ell^m(\mathbf{I}_{p+h}), \qquad h \in \{1, 2, 4\}, \tag{3}$$

as the retrieval target for one-step prediction (h=1) and path integration ($h \in \{2, 4\}$), and as the goal frame for L4. Matched noise seeds are used for stochastic encoders; further constructions are specified with each experiment.

3 EXPERIMENTAL SETUP

3.1 DATA AND SPLITS

All temporal experiments use posed ScanNet++ trajectories (Yeshwanth et al., 2023) processed by our pipeline. Each sample provides RGB observations, camera poses, instance geometry, and visibility annotations. Scene-level partitions keep related trajectories, objects, and windows together. The historical master test partition overlaps earlier training/development manifests, so the new diagnostics are reported as development validation evidence rather than independent confirmation. Task-specific scene-disjoint splits, sample counts, and seeds are specified with the experiments and in Appendix A.4. Views without a meaningful temporal trajectory are used only for tasks that do not require temporal history or motion.

3.2 MODELS AND EXTRACTION POINTS

We compare **ten frozen representations** spanning image encoders, video encoders, 3D models, multimodal models, video generators, and world models (Table 1) with all backbone parameters remaining frozen. Layer indices are model-specific, not absolute measures of depth across architectures. The designated representative layers are listed in Appendix Table 9.

Within a model and task, readout architecture and training are held fixed across layers and prefix lengths. The main comparison averages the evaluated prefixes at the designated layer; layer and prefix sweeps are analyzed separately rather than using their post-hoc maxima as primary scores.

Family	Model
Image encoder	DINOv2-Large
Video encoder	V-JEPA2 ViT-L VideoMAE-Large
3D encoder	Fast3R ViT-L
Video generator	Wan2.1-T2V-1.3B CogVideoX-5B-I2V MiniMax-H3 LingBot-video MoE-30B-A3B
World model	AetherV1
Multimodal	Qwen2.5-VL-7B-Instruct

Table 1: **Ten representations evaluated.**

Readout metrics. Instance identity uses a cross-entropy loss on per-pixel embeddings, with prototypes computed from ground-truth instance groups within the evaluated sample. It measures identity separation, not mask generation accuracy or segmentation IoU. One-step prediction and path integration retrieve future features within each model's own target space; these scores are not a common scale of physical prediction error.

4 EXPERIMENTS

We first compare spatial readouts across all ten representations and examine their dependence on extraction settings. We then test history and motion use, spatial readout from predicted features, and goal-specific action utility.

4.1 SPATIAL CAPABILITIES DIFFER ACROSS REPRESENTATIONS

The main results appear together in Table 2. For model m and task t, the reported score averages its evaluated prefixes $\mathcal{P}_{m,t}$ at the designated layer ℓ_m:

$$\bar{s}_{m,t} = \frac{1}{|\mathcal{P}_{m,t}|} \sum_{p \in \mathcal{P}_{m,t}} s_{m,t}(\ell_m, p). \tag{4}$$

Spatial structure exposes distinct task strengths. Fast3R and DINOv2 illustrate that metric geometry, view consistency, and instance identity are not interchangeable readouts. Fast3R has the highest camera pose AUC@30 and view accuracy in the main table, but its advantage over DINOv2 is much larger for camera pose (0.3852 versus 0.0386) than for depth loss or view accuracy. DINOv2 instead has lower instance identity loss (1.2563 versus 1.4804). Strong view consistency or instance identity separation therefore does not imply equally strong camera pose performance.

Table 2: **Main results for all ten representations.** Prefix aggregation follows Eq. 4 at each designated layer. The first header row names the diagnostic task and the second row its metric; metric geometry contributes two metrics. **Bold**, colored values indicate the best reported result in each column, with colors matching the corresponding header groups.

	Metric geometry		Instance identity	View consistency	Object persistence	One-step prediction	Path integration
Model (layer)	Depth loss ↓	AUC@30 ↑	Instance loss ↓	Bal. acc. ↑	Ang. err. (°) ↓	Global R@1 ↑	Global R@1 ↑
DINOv2-L (0)	0.1425	0.0386	**1.2563**	0.9597	69.01	0.1592	**0.1470**
FAST3R (24)	**0.1347**	**0.3852**	1.4804	**0.9827**	56.45	0.0263	0.0145
VIDEOMAE-L (23)	0.2888	0.2124	4.2575	0.9210	**55.53**	0.0888	0.0685
V-JEPA2-L (23)	0.2687	0.1849	3.0291	0.9240	59.39	0.1232	0.0839
QWEN2.5-VL (−1)	0.2193	0.0269	2.8522	0.9403	71.36	**0.2375**	0.1374
WAN (20)	0.3599	0.0603	4.8688	0.8175	66.44	0.1150	0.0383
COGVIDEOX (20)	0.4375	0.0302	5.7245	0.7918	70.80	0.0272	0.0112
AETHER (1)	0.5413	0.0198	7.9184	0.7238	62.46	0.0109	0.0015
MINIMAX-H3 (30)	0.4864	0.0893	4.7338	0.5000	79.54	0.0041	0.0014
LINGBOT-VIDEO (30)	0.5683	0.0098	2.9600	0.9164	74.46	0.0851	0.0290

Visible scene accuracy does not determine object persistence. VideoMAE achieves the lowest reported object persistence error (55.53°), compared with 69.01° for DINOv2 and 71.36° for Qwen2.5-VL, although both outperform VideoMAE on depth and view consistency. VideoMAE also outperforms Fast3R on their common prefix settings (53.75° versus 56.45°). Object persistence therefore reveals differences that visible scene readout alone would miss.

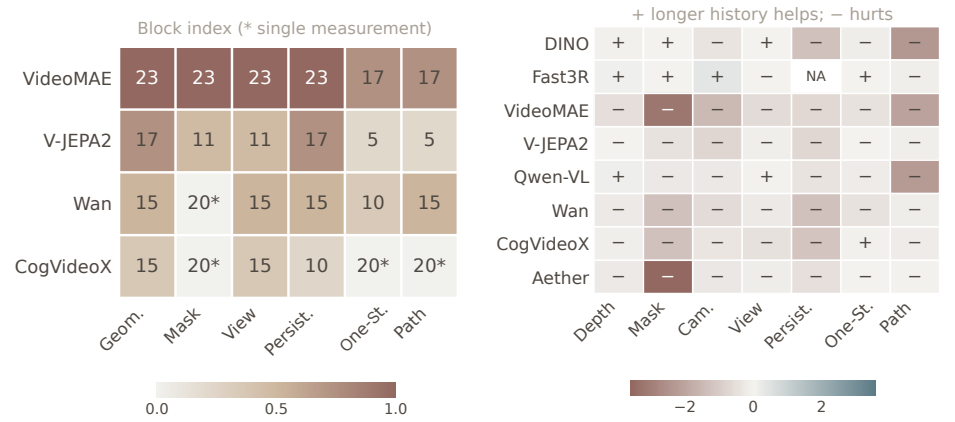
Persistence and prediction yield different model orderings. DINOv2 and Qwen2.5-VL rank first and second in both one-step prediction and path integration, despite their higher object persistence errors than VideoMAE and V-JEPA2. The two prediction tasks also yield different orderings: Wan exceeds VideoMAE on one-step prediction (0.1150 versus 0.0888), whereas VideoMAE exceeds Wan on path integration (0.0685 versus 0.0383). These reversals motivate separate evaluation of object persistence and predictive spatial update, including both prediction tasks.

Representation families do not determine capability profiles. VideoMAE and V-JEPA2 have similar view consistency accuracy but different strengths: VideoMAE performs better on camera pose and object persistence, whereas V-JEPA2 performs better on depth and both prediction tasks. Among the evaluated generative and world model representations, Aether has lower object persistence error than Wan and CogVideoX despite weaker results on the other seven metrics. These comparisons describe the selected representations and extraction points; they do not isolate training objective effects, since architecture and training data also differ.

> **No task-independent model ranking**
>
> The observed advantages depend on the spatial function being tested. Strong visible scene readout does not guarantee accurate object persistence or strong one-step prediction and path integration. A single diagnostic therefore cannot summarize the full spatial capability profile.

4.2 SPATIAL ACCESSIBILITY DEPENDS ON LAYER AND STREAMING PREFIX

Different readouts favor different layers. Figure 4a summarizes the best measured layers for backbones with multi-layer sweeps. V-JEPA2 peaks at block 5 for one-step prediction and path integration, block 17 for object persistence, and block 11 for instance identity. VideoMAE instead peaks at block 23 for localization and instance identity, and block 17 for both prediction tasks. A representative layer selected for one task need not maximize another task's readout.



(a) Best observed layers. (b) Effects of longer prefix.

Figure 4: **Spatial readout depends on both layer and streaming prefix.** (a) Four multi-layer backbones: cells give the best measured block, colored by normalized depth; starred cells have only one measured layer. Metric geometry combines normalized depth and camera metrics. (b) Paired prefix comparisons at the designated layers: direction-corrected, standardized p24–p8 differences (Eq. 5); + favors the longer prefix, − the shorter. Full curves and joint grids are in Appendix A.1.

Longer prefixes do not uniformly improve spatial readout. We compare p24 and p8 using direction-corrected standardized differences:

$$\Delta_{m,t} = \frac{d_t[s_{m,t}(\ell_m, 24) - s_{m,t}(\ell_m, 8)]}{\sigma_t}, \tag{5}$$

where $d_t = +1$ for higher-is-better metrics and -1 otherwise, and σ_t is the corresponding task metric standard deviation used by the visualization. Among the reported endpoint pairs, 82% favor p8 and 18% favor p24. All seven available object persistence pairs and all eight path integration pairs worsen with the longer prefix. For example, VideoMAE's localization error rises from 53.11° to 60.85°, while DINOv2's path integration R@1 falls from 0.1944 to 0.0990. Fast3R nevertheless improves on depth and camera pose, showing that the effect is task-dependent. As longer prefixes also move the endpoint (Section 2.2), these differences do not isolate history alone.

> **Spatial accessibility is setting-dependent**
> Different spatial readouts favor different layers, and longer visual prefixes do not uniformly improve them. The extraction setting is part of the evaluation, not an interchangeable choice of feature interface.

4.3 What do the probes actually use?

We complement the main benchmark with baseline comparisons and input controls to examine the contributions of motion and visual history. For these input controls, the trained probe and prediction targets remain fixed, allowing direct comparison between full and modified conditioning inputs.

Table 3: **Motion and history controls.** Motion gain is correct-minus-wrong R@1; history benefit is the reduction in localization error. Brackets show paired 95% CIs.

Model	Motion gain (%)	History benefit (°)
V-JEPA2	18.84 [13.04, 25.36]	10.26 [3.51, 16.80]
VideoMAE	10.51 [6.16, 15.22]	2.08 [-6.10, 9.97]
Wan	2.90 [-1.45, 7.25]	10.01 [3.42, 16.69]
CogVideoX	4.35 [0.35, 7.97]	-3.41 [-9.39, 2.39]

One-step prediction. We compare learned prediction with copying the last observed feature as a temporal baseline. In global future-feature retrieval, Wan achieves 15.22% with prediction and 11.96% with copying, while V-JEPA2 achieves 13.04% and 19.93%, respectively. We further evaluate motion conditioning using eight pose-displacement-matched candidates. Correct motion improves retrieval over matched wrong motion for V-JEPA2, VideoMAE, and CogVideoX, with gains of 18.84, 10.51, and 4.35 percentage points; Wan's gain is 2.90 points, with an interval span-

ning zero (Table 3). The copying comparison provides a reference for global retrieval, whereas the matched-candidate comparison quantifies the benefit of correct motion.

Object persistence. We compare localization from full visual history with the current frame alone, retaining the same historical object query in both conditions. Full history reduces angular error for V-JEPA2 and Wan by 10.26° and 10.01°, respectively, with both intervals above zero. The corresponding effects for VideoMAE and CogVideoX are smaller and have intervals spanning zero (Table 3). These comparisons supplement the main localization results by quantifying the contribution of historical visual inputs.

Motion-composition consistency. We additionally examine how an external recurrent updater trained on one-step transitions generalizes to two-motion composition. For Wan and V-JEPA2, updating improves endpoint prediction over copying. Correct-order effects remain inconclusive for both the feature-based updaters and a learned geometric control, whereas an analytic SE(3) reference preserves transformation order. The full comparison is provided in Appendix A.3.

> **Benefits from motion and visual history**
>
> Correct motion improves matched-candidate prediction for V-JEPA2, VideoMAE, and CogVideoX, while visual history improves localization for V-JEPA2 and Wan. These controls complement the main capability profiles with measured gains from task-relevant inputs.

4.4 Prediction changes the spatial readout interface

As a further L3 analysis, we ask whether L1 metric geometry accessible in real future features remains accessible after native prediction. A probe trained on real V-JEPA2 future states is applied unchanged to the model's native predictions (Figure 5). Supporting controls first establish real feature readout and motion-conditioned utility for an latent predictor; Details are in Appendix A.5.

At horizon two, normalized-depth MAE rises from 0.182 on real features to 0.246 on native predictions. Retraining an equal-capacity reader on predictions improves MAE to 0.223, better than the 0.241 layout prior but worse than the real feature readout. Horizon four shows the same pattern. Frozen transfer remains better than copied context, but slightly worse than the prior (Table 7). The results distinguish recoverable spatial structure from preservation of the original readout interface.

> **Recoverable structure is not unchanged readability**
>
> Native predicted features retain some recoverable spatial structure, but a reader trained on real features does not retain its accuracy unchanged. Prediction supports neither lossless spatial propagation nor a claim that all spatial information has disappeared.

4.5 Goal-conditioned action prediction

We test actionability through **goal-conditioned action prediction** in two complementary settings. The candidate selection setting chooses between supplied actions, whereas the continuous-output setting predicts an action directly. Both use p8 features and an independently extracted goal feature.

Table 4: **Goal-conditioned action prediction: candidate selection setting.** Accuracy (%), mean±SD over three training seeds; chance is 50%. The endpoint reference observes realized outcomes and is diagnostic.

Representation	Action only	Direct	Predict-then-score	Endpoint ref.
DINOv2-L	67.8±0.4	76.2±0.8	76.2±1.2	85.5
Fast3R	67.8±0.4	79.9±1.3	78.1±0.3	78.0
V-JEPA2	67.8±0.4	78.1±0.7	76.7±0.8	78.0
Wan	67.8±0.4	75.9±0.9	63.6±1.1	68.3

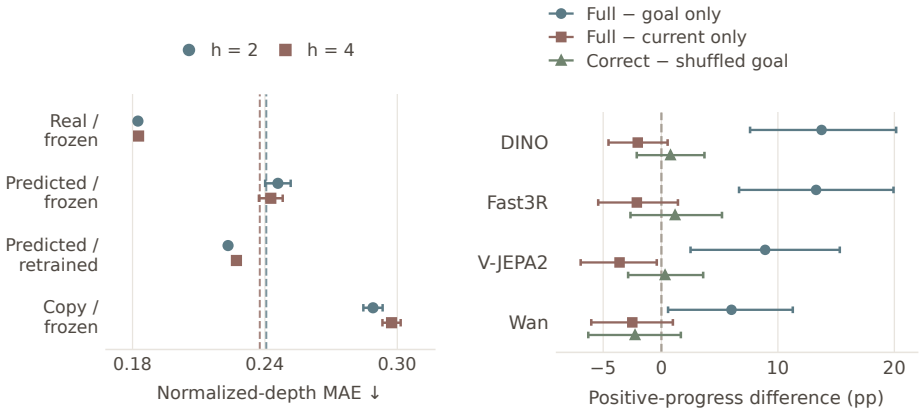


Figure 5: **Testing readable structure and goal-specific utility.** Left: V-JEPA2 normalized-depth MAE at $h = 2, 4$, using the same frozen reader except for the retrained control; dashed lines denote training-layout priors. Right: positive-progress differences under input restriction and fixed-weight goal shuffling. Error bars show three-seed SD (left) and paired 95% scene-bootstrap CIs (right).

Candidate selection setting. A shallow probe chooses between two 9D camera motions by their progress toward the goal. Progress is the decrease in translation distance plus 0.5 times rotation angle in radians to the goal. Direct selection reaches 75.9–79.9% accuracy, exceeding the 67.8% action-only baseline; shuffling goals reduces accuracy by 12.2–20.7 percentage points (Table 4). The readout therefore uses goal information for candidate selection. Predicting each candidate's outcome before scoring it offers no consistent advantage, and reduces Wan's accuracy to 63.6%.

Continuous-output setting. Without candidates, a shallow probe predicts the first observed 9D motion on a four-step trajectory from current p8 and isolated horizon-four goal features. Separate projections c and g are fused as $[c, g, g - c, c \odot g]$ before prediction. The matched-capacity concatenation control are specified in Appendix A.6. FULL achieves positive progress on 50.2–59.8% of trajectories and exceeds GOAL-ONLY by 6.0–13.8 percentage points. However, FULL doesn't improve positive progress over CURRENT-ONLY for any backbone (Table 8). Beating GOAL-ONLY establishes that the current state is useful but not benefit from adding the goal as input.

Changing the goal changes actions, not reliably their utility. Holding the current state and weights fixed while substituting a difficulty-matched, different-scene goal changes the predicted direction by 29.8–45.6°. Yet correct-minus-shuffled-goal effects on positive progress (-2.3 to $+1.2$ percentage points) and goal reduction (-0.006 to $+0.001$) all have 95% intervals containing zero (Figure 5). Thus, goal sensitivity does not translate into reliable goal-specific utility under this readout. The relational head improves some metrics over concatenation, but does not resolve this gap. These are supervised offline diagnostics, not a demonstration of a closed-loop policy.

> **Goal sensitivity is not goal utility**
> Changing the goal changes predicted actions, but the correct goal does not reliably improve progress. Successful prediction in the candidate selection setting therefore does not establish reliable goal-conditioned action prediction in the continuous-output setting.

5 Discussion and Conclusion

We characterize spatial state representations through four levels: spatial structure, object persistence, predictive spatial update, and actionability. Across ten frozen representations, readouts vary across tasks and extraction settings. Complementary analyses of motion, history, prediction, and action distinguish decodability from input use and functional utility, supporting multidimensional evaluation beyond a single spatial task score.

Ethics Statement

This work evaluates ten frozen visual representations using existing ScanNet++ indoor-scene data (Yeshwanth et al., 2023) and pretrained models. The reported experiments involve offline visual observations and camera-motion annotations, without recruiting human participants or deploying physical robots. Our analyses concern scene geometry, object persistence, motion-conditioned prediction, and goal-directed action. Use and redistribution of the datasets and pretrained models are subject to their respective licenses and access conditions. The comparisons characterize spatial capabilities under the specified feature interfaces and probing tasks, rather than general judgments of model quality or assessments of deployment safety.

Reproducibility Statement

Section 2 describes the four levels of spatial capability, their probing tasks, and the causal streaming formulation. Section 3 summarizes the data and model configurations, while Appendix Table 9 specifies model-specific extraction layers and generation coordinates. The experimental sections describe the evaluation metrics and task-specific protocols, including motion and history controls in Section 4.3 and goal-conditioned action evaluation in Section 4.5. Appendix A.1 provides extended layer and history analyses; Appendix A.3 details the motion-composition diagnostic; and Appendix A.7 reports detailed predictive-readout and continuous-action results. Sample counts, training seeds, and uncertainty-estimation procedures are documented with the corresponding experiments and supplementary tables.

AI Use Statement

We used generative AI tools, including large language models, coding assistants, and image-generation tools, to support several aspects of this work. These included drafting, editing, code debugging.

The AI-assisted illustrations are explanatory schematics, not experimental observations. The authors made the final decisions regarding the research scope, experimental design, selection of analyses, and interpretation of results. The authors take full responsibility for the accuracy and integrity of the final manuscript, code, numerical results, figures, and references, including content developed with generative AI assistance.

References

Florian Bordes, Quentin Garrido, Justine T. Kao, Adina Williams, Michael Rabbat, and Emmanuel Dupoux. Intphys 2: Benchmarking intuitive physics understanding in complex synthetic environments. *arXiv preprint arXiv:2506.09849*, 2025.

Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021.

Parsa Esmati, Somjit Nath, Katja Hofmann, Derek Nowrouzezahrai, Samira Ebrahimi Kahou, and Majid Mirmehdi. The invisible hand of physics: When video diffusion models know more than they show. *arXiv preprint arXiv:2606.05328*, 2026.

Quentin Garrido, Nicolas Ballas, Mido Assran, Adrien Bardes, Laurent Najman, Michael Rabbat, Emmanuel Dupoux, and Yann LeCun. Intuitive physics understanding emerges from self-supervised pretraining on natural videos. *arXiv preprint arXiv:2502.11831*, 2025.

Zixuan Huang, Xiang Li, Zhaoyang Lv, and James M. Rehg. How much 3d do video foundation models encode? *arXiv preprint arXiv:2512.19949*, 2025.

Daniel Kahneman, Anne Treisman, and Brian J Gibbs. The reviewing of object files: Object-specific integration of information. *Cognitive Psychology*, 24(2):175–219, 1992. ISSN 0010-0285. doi: https://doi.org/10.1016/0010-0285(92)90007-O. URL https://www.sciencedirect.com/science/article/pii/001002859290007O.

Thomas Kipf, Gamal Elsayed, Aravindh Mahendran, Austin Stone, Sara Sabour, Georg Heigold, Rico Jonschkowski, Alexey Dosovitskiy, and Klaus Greff. Conditional object-centric learning from video. In *International Conference on Learning Representations*, 2022.

Francesco Locatello, Dirk Weissenborn, Thomas Unterthiner, Aravindh Mahendran, Georg Heigold, Jakob Uszkoreit, Alexey Dosovitskiy, and Thomas Kipf. Object-centric learning with slot attention. In *Advances in Neural Information Processing Systems*, 2020.

Grace Luo, Lisa Dunlap, Dong Huk Park, Aleksander Holynski, and Trevor Darrell. Diffusion hyperfeatures: Searching through time and space for semantic correspondence. In *Advances in Neural Information Processing Systems*, 2023.

Bruce L. McNaughton, Francesco P. Battaglia, Ole Jensen, Edvard I. Moser, and May-Britt Moser. Path integration and the neural basis of the 'cognitive map'. *Nature Reviews Neuroscience*, 7 (8):663–678, 2006. ISSN 1471-0048. doi: 10.1038/nrn1932. URL https://doi.org/10.1038/nrn1932.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaa El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *Transactions on Machine Learning Research*, 2024.

J. O'Regan and Alva Noë. A sensorimotor account of vision and visual consciousness. *Behavioral and Brain Sciences*, 24:939–973, 11 2002. doi: 10.1017/S0140525X01000115.

Luming Tang, Menglin Jia, Qianqian Wang, Cheng Perng Phoo, and Bharath Hariharan. Emergent correspondence from image diffusion. In *Advances in Neural Information Processing Systems*, 2023.

Wan Team. Wan: Open and advanced large-scale video generative models. *arXiv preprint arXiv:2503.20314*, 2025.

Jianing Yang, Alexander Sax, Kevin J. Liang, Mikael Henaff, Hao Tang, Ang Cao, Joyce Chai, Franziska Meier, and Matt Feiszli. Fast3r: Towards 3d reconstruction of 1000+ images in one forward pass. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2025.

Zhuoyi Yang, Jiayan Teng, Wendi Zheng, Ming Ding, Shiyu Huang, Jiazheng Xu, Yuanming Yang, Wenyi Hong, Xiaohan Zhang, Guanyu Feng, et al. Cogvideox: Text-to-video diffusion models with an expert transformer. *arXiv preprint arXiv:2408.06072*, 2024.

Chandan Yeshwanth, Yueh-Cheng Liu, Matthias Nießner, and Angela Dai. Scannet++: A high-fidelity dataset of 3d indoor scenes. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023.

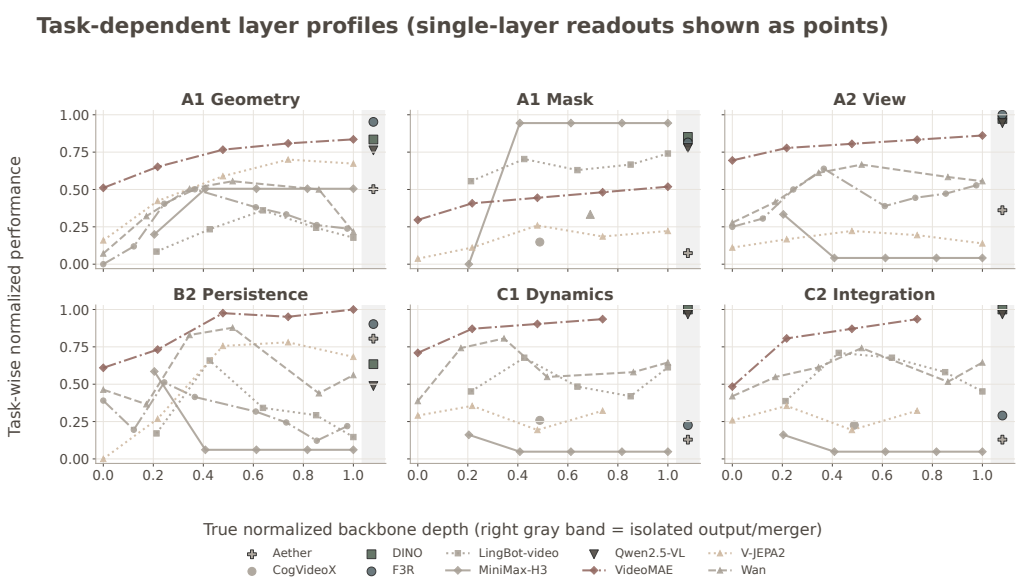


Figure 6: **Capability profiles across measured layers.** Performance is direction-corrected and normalized within each task, with prefix-aggregated readouts shown at each measured layer. The two metric geometry metrics are combined after normalization; instance identity retains its own equally sized panel. A campaign is selected separately for each model and task, preserving instance identity measurements available only at a default layer as individual points. The horizontal axis uses actual backbone depth; isolated output representations occupy a separate band.

A Appendix

A.1 Additional layer, history, and intervention analyses

The main text retains compact summaries. Here we provide the full measured layer profiles, joint instance identity/localization grids, intervention schematic, and scheduler grid. Layer scans complement the representative layer benchmark; they describe extraction depth profiles, not training-time emergence.

A.2 Generation coordinate analysis

This separate ScanNet++ extraction study evaluates layers $\{15, 20, 25\}$ and prefixes $\{8, 12, 16, 24\}$ at scheduler indices $\{499, 999\}$ for Wan and CogVideoX, and $\{500, 1000\}$ for MiniMax-H3. The main comparison uses scheduler step 749 for Wan and CogVideoX and flow coordinate 0.25 for MiniMax-H3; those default coordinates are not part of the two-coordinate sweep. Coordinates are model-specific and are not interpreted as a common physical noise level.

Table 5 reports last-epoch probe validation readouts for block 20, averaged over prefixes. Wan's 12 layer–prefix pairs favor the lower index on depth, view consistency, and object persistence. CogVideoX has zero view pairs and three depth pairs favoring the higher index; its localization error is unchanged at each prefix. MiniMax-H3's reported view and localization values are also unchanged between the two indices. Instance identity has model- and setting-dependent signs, so these results do not establish a universal advantage for one generation coordinate.

The prediction probes reach training-time cosine above 0.99 in this study. Differences of order 10^{-4} in that metric are not treated as retrieval improvements. One-step prediction and path integration R@1 remain the default coordinate measurements in Table 2.

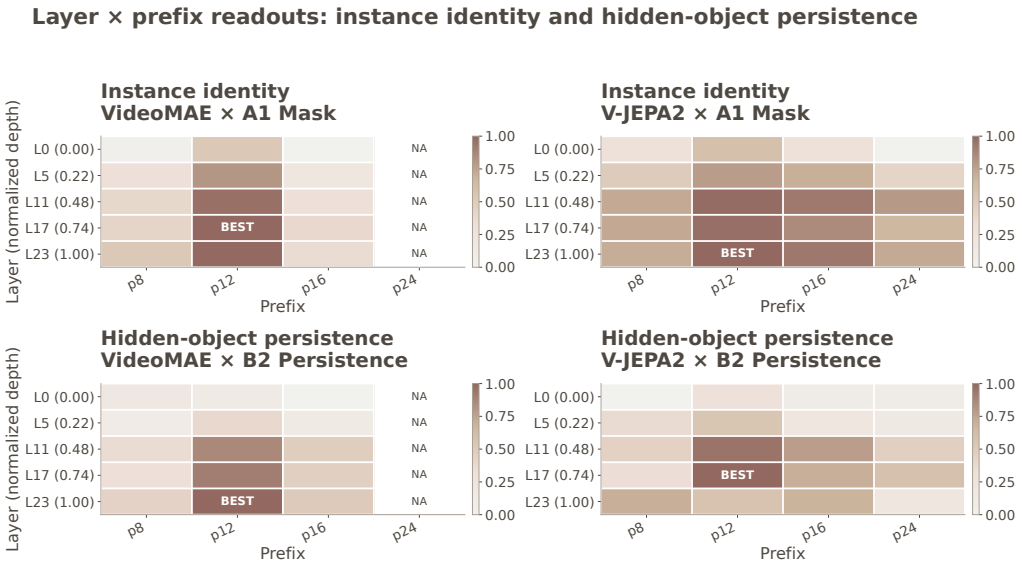


Figure 7: **Layer–prefix structure of instance identity and object persistence.** The four panels visualize instance identity (top) and object persistence (bottom) for VideoMAE (left) and V-JEPA2 (right), across extraction layers and prefix lengths. Colors are normalized within each panel and therefore describe within-model structure rather than absolute cross-model accuracy. The complete model–task atlas also includes all available instance identity tensors; raw values accompany the heatmaps.

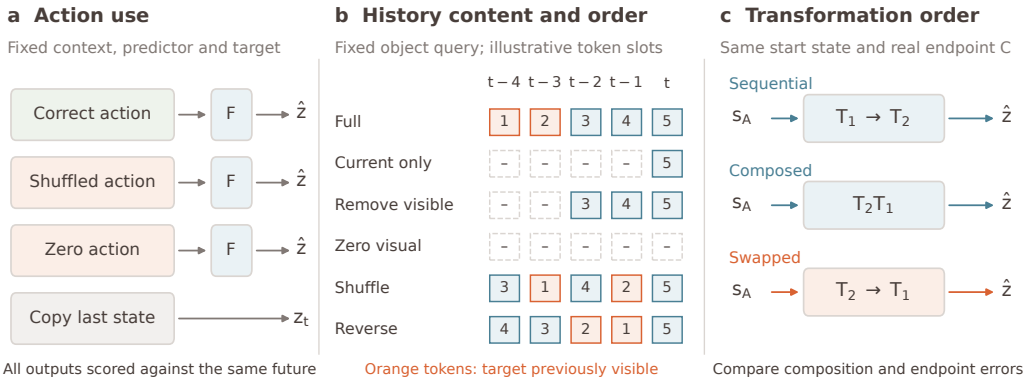


Figure 8: **Interventions separate information use from decodability.** (a) Change the action supplied to a fixed predictor F, or copy the last state. (b) Remove or reorder visual feature tokens while retaining the object query; five illustrative slots are shown, with temporal identifiers and the final frame fixed during order tests. (c) Compare sequential, composed, and swapped updates against the same real endpoint.

	Depth loss ↓		Instance loss ↓		View acc. ↑		Localiz. err. (°) ↓	
Model	lower	higher	lower	higher	lower	higher	lower	higher
WAN (20)	**0.387**	0.430	6.917	**5.695**	**0.834**	0.791	**59.0**	78.0
COGVIDEOX (20)	**0.449**	0.467	**6.559**	7.181	**0.806**	0.805	78.6	78.6
MINIMAX-H3 (20)	—	—	—	—	0.484	0.484	78.6	78.6

Block 20, mean over prefixes $\{8, 12, 16, 24\}$. Lower/higher coordinates are $499/999$ for Wan and CogVideoX, and $500/1000$ for MiniMax-H3. Dashes denote unreported metrics. These are last-epoch validation readouts from the separate coordinate study, not replacements for main table scores.

Table 5: **Spatial readout at two generation coordinates.** Depth, instance identity loss, view consistency, and object persistence at the two scheduler coordinates. Instance identity favors the higher coordinate for Wan and the lower coordinate for CogVideoX at block 20; CogVideoX and MiniMax-H3 localization is unchanged.

Table 6: **One-step-to-two-step structural generalization.** Predeclared representative layers; differences are Wan block 15 minus V-JEPA2 block 11 with paired 95% intervals. The four comparisons form one Holm family. Composition distance and order margin are normalized.

Metric	Wan (15)	V-JEPA2 (11)	Difference [95% CI]	Holm p
Gain vs. copy ↑	0.2051	0.2323	-0.0273 $[-0.0814, 0.0209]$	0.568
Gain vs. no-action ↑	0.0639	0.0452	0.0187 $[-0.0050, 0.0414]$	0.384
Composition distance ↓	0.0377	0.0232	0.0145 $[0.0107, 0.0188]$	< 0.004
Order margin ↑	-0.0001	-0.0007	0.0006 $[-0.0005, 0.0019]$	0.568

A.3 MOTION COMPOSITION CONSISTENCY

We train an external recurrent updater U and readout D only on one-step camera transitions. Sequential, composed, and swapped updates start from the same p12 state s_A:

$$\begin{aligned} \hat{z}_{\text{seq}} &= D(U(U(s_A, T_1), T_2)), \quad \hat{z}_{\text{dir}} = D(U(s_A, T_2 T_1)), \\ \hat{z}_{\text{swap}} &= D(U(U(s_A, T_2), T_1)), \quad T_1 = T_{AB}, \quad T_2 = T_{BC}. \end{aligned} \tag{6}$$

The fixed p12 protocol contains 128 training tuples, 32 validation tuples, and 276 evaluation tuples from 69 scenes; the evaluation frame triples are new, but the scenes were used in the earlier pilot. An external recurrent head is trained only on observed $A \to B$ transitions, with isolated endpoint features as targets, using seeds 42–46. This is not either backbone's native predictor. Cosine endpoint errors are evaluated within each feature space. Relative gains are $(e_{\text{baseline}} - e_{\text{seq}})/e_{\text{baseline}}$; composition distance is the sequential–direct cosine distance normalized by copy-to-endpoint error, using scene-macro means. The normalized order margin compares swapped- and correct-order endpoint errors; a positive value favors the observed order.

All seven evaluated Wan layers and five V-JEPA2 layers improve endpoint prediction over copying the initial state; all but Wan block 0 also improve over the no-action predictor. At the predeclared layers, V-JEPA2 block 11 has a smaller normalized sequential–direct distance than Wan block 15 (0.0232 versus 0.0377).

Table 6 reports the predeclared representative layer comparison, not a best layer search or an overall model ranking. Action-only, matched-shuffle, and additive controls do not reveal a stable order effect, even in high non-commutativity strata. Identity padding is only a stress test because identity lies outside the training action support. The analytic SE(3) oracle has a positive order margin of 0.1107 m, whereas a geometric GRU trained on the same one-step objective has sequential RMSE 0.7010 m and no reliable order margin. This limits attribution of weak latent order effects to the representation itself.

A.4 ADDITIONAL PROTOCOL DETAILS

One-step prediction and history interventions. The correct/copy and matched-wrong-motion tests use 276 samples from 69 scenes per prefix; the restricted retrieval pool contains eight pose-

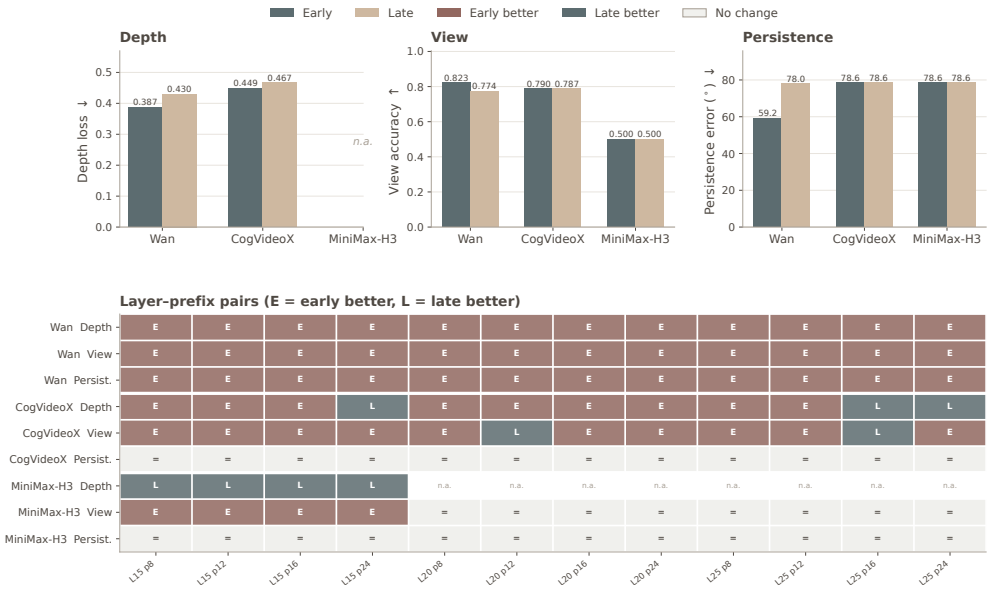


Figure 9: **Spatial readout across generation coordinates.** Top: the same model at the early versus late scheduler index (layer 20, mean over prefixes). Lower is better for depth, instance identity loss, and object persistence; higher is better for view consistency. Bottom: each layer–prefix pair is marked E (early better), L (late better), or = (no change). CogVideoX localization is unchanged.

displacement-matched candidates with candidate seed 42. The p12 object persistence controls use 165 samples from 65 scenes. Both studies use probe training seed 42. Reported intervals are paired 95% scene-clustered bootstrap intervals. The two tested prediction prefixes are not strictly target-paired, so agreement across them does not isolate a history length effect.

For localization, the appearance query is pooled using a historical object mask; neither that mask nor camera pose is supplied to the spatial readout. Content-only shuffling and reversal retain the final frame and temporal identifiers. Moving tokens together with their identifiers is a permutation sanity check, not a distinct temporal order intervention.

A.5 Supporting controls for predictive spatial readout

The fixed ScanNet++ protocol uses 612 training scenes, 69 development scenes, horizons $h \in \{2, 4\}$, and three training seeds. The main question is whether a spatial reader trained on real future states transfers unchanged to native predicted features. Two supporting controls check the specificity and conditional utility of spatial readout.

Target-specific real feature readout. Wan's spatial reader achieves scene-macro normalized-depth MAE of 0.155/0.152 at $h = 2/4$, compared with 0.241/0.238 for a training layout prior. Real target features also outperform within-scene wrong targets and copied context, demonstrating sample-specific information beyond the layout prior.

Motion-conditioned external prediction. For Wan's external predictor, the pooled branch gives lower downstream depth MAE under true motion than identity motion, matched-horizon shuffled motion, or an independently trained context-only predictor. The 0.003–0.007 reductions have paired scene-clustered intervals excluding zero for all three controls at both horizons. This is evidence of conditional prediction utility, not a physical action intervention.

A.6 Action readout protocol and controls

Goal-conditioned action prediction: candidate selection setting. The candidate experiment uses 844 training queries (650 fit and 194 development) from 289 scenes, and 227 validation queries

Table 7: **Spatial readout through V-JEPA2's native predictor.** Scene-macro normalized-depth MAE ↓ on 69 development scenes, spatial branch; mean±SD over three training seeds. The same reader trained on real native target states is used for real, predicted, and copied inputs, except in the explicitly retrained condition. The layout prior uses training labels only.

Input / readout condition	$h = 2$	$h = 4$
Real future / original reader	0.1824 ± 0.0003	0.1827 ± 0.0003
Native prediction / frozen reader	0.2459 ± 0.0058	0.2427 ± 0.0054
Native prediction / retrained reader	0.2233 ± 0.0008	0.2271 ± 0.0005
Copied context / frozen reader	0.2890 ± 0.0044	0.2974 ± 0.0041
Training layout prior	0.2406	0.2377

Table 8: **Goal-conditioned action prediction: continuous-output setting.** FULL uses the relational head; values are mean±SD over seeds 41–45. Progress is a percentage; goal reduction uses the pose distance metric defined in the text. The final column is FULL minus CURRENT-ONLY goal reduction with a paired 95% scene-bootstrap interval (5,000 replicates); positive favors FULL. Seed repeats are averaged within each sample before scene resampling.

Model	Direction (°) ↓	Progress (%) ↑	Goal reduction ↑	FULL − CURRENT-ONLY goal reduction
DINOv2-L	65.3 ± 1.3	53.8 ± 1.1	0.008 ± 0.006	-0.008 $[-0.013, -0.004]$
Fast3R	64.7 ± 1.3	59.8 ± 3.4	0.025 ± 0.007	-0.005 $[-0.012, +0.002]$
V-JEPA2	66.2 ± 1.8	53.8 ± 2.4	0.007 ± 0.007	-0.009 $[-0.015, -0.002]$
Wan	70.0 ± 1.5	50.2 ± 3.5	-0.001 ± 0.009	-0.011 $[-0.017, -0.004]$

from 77 disjoint scenes. A shallow scorer receives frozen current and goal features plus each candidate 9D action. Ground-truth poses, utilities, realized endpoints, and scene identities are not given to the scorer. Accuracy is summarized across three training seeds. The endpoint reference observes realized outcomes and is diagnostic rather than a candidate selection method with the same information.

Goal-conditioned action prediction: continuous-output setting. The continuous experiment uses 1,024 training trajectories from 444 scenes (803 fit and 221 development before refitting), and 256 validation trajectories from 108 disjoint scenes. Epoch selection and normalization use development and fit data, respectively. Features are DINOv2's final project output, Fast3R block 24, V-JEPA2 block 23, and Wan block 29 at scheduler step 749; the latter differs from Wan's main table block 20. Separate current and goal projections are combined as $[c, g, g - c, c \odot g]$. Concatenation uses the same optimization budget.

FULL has positive progress differences of -3.6 to -2.0 percentage points relative to CURRENT-ONLY; only V-JEPA2's interval excludes zero. Mean goal reduction is also lower for all four representations, with intervals excluding zero for DINOv2, V-JEPA2, and Wan. The relational head improves progress over concatenation for DINOv2 and Fast3R, but only Fast3R shows a clear goal reduction gain.

The same-checkpoint intervention substitutes a difficulty-matched goal from another scene while holding the current state and learned weights fixed. Correct-minus-shuffled effects are computed against the original goal. Five seeds (41–45) are averaged within each sample before paired scene-bootstrap resampling with 5,000 replicates. This measures goal sensitivity and utility under the stated input substitution; the candidate selection and continuous-output settings remain separate offline evaluations of the same task.

A.7 Predictive readout and action prediction metrics

Tables 7 and 8 provide the numerical readout and action metrics accompanying Figures 5.

Model	Probed layers	Primary layer	Extraction coordinate
DINOv2-Large	final hidden state	0	–
V-JEPA2 ViT-L	$\{0, 5, 11, 17, 23\}$	23	–
VideoMAE-Large	$\{0, 5, 11, 17, 23\}$	23	–
Fast3R ViT-L	$\{24\}$	24	–
Wan2.1-T2V-1.3B	$\{0, 5, 10, 15, 20, 25, 29\}$	20	scheduler step 749
CogVideoX-5B-I2V	$\{0, 5, 10, 15, 20, 25, 30, 35, 40\}$	20	scheduler step 749
AetherV1	$\{1\}$	1	scheduler step 749
MiniMax-H3	$\{10, 20, 30, 40, 50\}$	30	flow coordinate 0.25
LingBot-video MoE-30B-A3B	$\{10, 20, 30, 40, 48\}$	30	flow coordinate 0.25
Qwen2.5-VL-7B-Instruct	visual merger output	-1	–

Table 9: **Feature extraction details.** Layer indices denote the representations used for layer-wise probing. For generative models, features are extracted at the specified denoising or flow coordinate.

A.8 Feature extraction details

Details are showed at 9.

B Related Work

B.1 Probing spatial and 3D structure in frozen representations

Frozen feature probing asks what information is already accessible in a representation without updating the backbone. Self-supervised image encoders such as DINO and DINOv2 revealed that semantic organization, object boundaries, and strong local correspondences can emerge without dense labels (Caron et al., 2021; Oquab et al., 2024). More recent work extends this methodology to video foundation models. In particular, VidFM3D evaluates whether shallow readouts can recover depth, camera pose, and 3D point structure from frozen video model features, and analyzes how these signals vary with network depth and diffusion timestep (Huang et al., 2025). Feed-forward 3D models such as Fast3R provide a useful geometry-specialized reference point because their training directly targets multiview reconstruction and camera estimation (Yang et al., 2025).

Our work begins from the same evidence-first principle but changes the object of evaluation. Metric geometry is treated as one functional diagnostic, not as a complete definition of spatial state. We ask whether a representation also supports view consistency, object persistence, predictive spatial update, and actionability. Consequently, two models with similar depth readout can receive very different spatial state profiles, and a model with strong future feature retrieval is not credited with predictive spatial update unless it passes matched motion controls.

B.2 Internal representations of diffusion and video generation models

Large video generators such as Wan and CogVideoX are optimized for denoising and visual synthesis rather than for downstream readout (Wan Team, 2025; Yang et al., 2024). Nevertheless, work on image diffusion models has shown that semantic and geometric consistency can be extracted from internal activations, and that useful information is distributed across both network layers and diffusion timesteps (Tang et al., 2023; Luo et al., 2023). Recent probing of video diffusion transformers similarly reports that physical and 3D-related signals can be most accessible at intermediate blocks and nontrivial noise levels (Huang et al., 2025; Esmati et al., 2026).

This motivates our layer $\times$ generation coordinate $\times$ streaming prefix analysis. We compare selected extraction settings rather than assuming that one hidden state characterizes an entire model. The question is whether spatial readouts remain accessible across settings or are concentrated at particular layers and generation coordinates. Such profiles describe accessibility during computation, not when a capability emerges during training.

B.3 Object-centric representations and persistence

Object-centric learning explicitly decomposes scenes into entity-level latent variables. Slot Attention learns exchangeable slots that bind to objects, while sequential extensions such as SAVi propagate and correct object slots through video (Locatello et al., 2020; Kipf et al., 2022). These approaches build an object-centric format into the architecture. In contrast, the foundation models we evaluate generally use distributed patch or video tokens, and we do not assume that their internal state is slot-based.

Our question is operational: can an object's bearing be recovered after it is no longer visible? This differs from violation-of-expectation benchmarks such as IntPhys 2, which assess discrimination between possible and impossible events involving permanence, continuity, solidity, or immutability (Bordes et al., 2025). V-JEPA surprise has shown strong results on several intuitive-physics conditions (Garrido et al., 2025); plausibility discrimination, however, does not directly measure object persistence. We therefore predict an explicit bearing in the current camera frame and use history interventions to test which visual evidence supports that readout.

C Discussion

Our conclusions concern the stated frozen feature interfaces, trained probes, and offline indoor trajectories. Independently encoding causal prefixes does not demonstrate an incremental recurrent state update; retrieval scores in different target feature spaces are not a shared physical error scale; and probe failure does not prove that the backbone lacks a capability. Within this scope, evaluating what is readable, what information is used, and what improves behavior provides a more informative characterization than a single spatial task score.