MOTIONJEPA: PHYSICALLY STRUCTURED RESIDUAL PREDICTIVE REPRESENTATION LEARNING FOR AUTONOMOUS DRIVING

Anonymous authors
Paper under double-blind review

ABSTRACT

Driving world models must predict and represent future scenes relevant to driving decisions. Changes in high-dimensional representations arise both from reference-frame changes caused by ego motion and from temporal evolution of the surrounding environment. When a model directly predicts the complete future, a predictor with limited capacity must fit both known geometric effects and unknown scene changes. This can impair future-scene prediction, produce representations that are less useful for driving decisions, and impose substantial computational costs. We propose MotionJEPA, a physically structured residual predictive representation-learning framework. It uses low-dimensional, interpretable ego-motion variables to construct a physical reference evolution for high-dimensional representations, and then learns the conditional residual not explained by that reference. The method combines explicit SE(2) motion transport, residual prediction around a physical reference, and static–dynamic diagnostics of residual sources. In 10,000 training steps on NAVSIM, model A3, which combines motion transport with an environment-residual predictor, reduces masked Smooth-L1 error by 7.76% relative to A1, which uses motion conditioning alone. A3 also achieves a 7.35% lower best validation Macro Smooth-L1 and a 7.69% lower overall test Smooth-L1 than A1. The additional validation metric, mean Intersection over Union (mIoU), also improves.

Keywords: world models; autonomous driving; JEPA; motion transport; environmental change; residual learning; joint-embedding prediction

1 INTRODUCTION

The rapid development of world models for autonomous driving calls for models that can predict scene changes relevant to driving decisions from limited historical data. Prediction targets may include video, point clouds, occupancy, traffic-participant states, or latent representations; these choices differ in the information they preserve, their computational cost, and how they are evaluated. Generating clear images and providing useful driving predictions are therefore distinct objectives. These research directions differ, but they also show that future modeling must account for both environmental evolution and vehicle motion; generation quality alone cannot establish value for driving. (Tu et al., 2025; Guan et al., 2024)

In the ego-vehicle reference frame, observations change over time for two reasons: the frame translates and rotates, and the environment itself evolves. As the vehicle moves forward, the road surface and stationary buildings move backward in a bird's-eye-view representation. Nearby vehicles and pedestrians also move independently of this coordinate effect. A predictor that directly regresses a future representation from the current one must fit both types of change. DynaDreamer treats the entanglement of ego motion and scene dynamics as a modeling burden and mitigates it with physical conditioning; GAST combines pose-driven transformations with feature refinement for occupancy prediction. These approaches motivate a testable question: when some changes can be approximated from geometric relations, should the learning module focus on predicting what remains after geometric transport? (Z. Wang et al., 2026; Chen et al., 2026)

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.

The Joint-Embedding Predictive Architecture (JEPA) offers a representation-level perspective on this question. Its central idea is to predict target representations from context in a suitable representation space, rather than devoting all learning capacity to reconstructing observational details. LeCun proposed an architecture for autonomous intelligence built around predictive representations. I-JEPA predicts representations of target image regions from image context, while V-JEPA learns appearance and motion information by predicting video features. V-JEPA 2 further distinguishes action-free video pretraining from action-conditioned world-model post-training, underscoring the need for an explicit training connection between representation prediction, dynamics prediction, and planning (LeCun, 2022; Assran et al., 2023; Bardes et al., 2024; Assran et al., 2025). For driving, the question is not merely whether pixels are replaced with latents, but what spatial information the representation preserves and how the predictor uses motion knowledge.

From the perspective of function approximation, a direct predictor must jointly learn the systematic coordinate transformation induced by ego motion in a spatial representation and scene changes such as traffic-participant motion, occlusion, and disocclusion. Supplying motion only as a network condition still requires the network to learn how motion acts on the spatial representation. We encode the known geometric effect in an explicit transport operator and ask the learner to predict the remaining component.

MotionJEPA therefore studies a predictive decomposition into explicit ego-motion transport and a learned residual. It estimates ego-pose change from historical states, transports the current metric BEV representation into the future ego frame, and predicts the difference between this geometric baseline and the future target. Equation (1) in the Methods section gives the overall relation. The residual is defined relative to a geometric reference and is not a generic skip connection in a network. APHYNITY combines physical and data-driven components to model incomplete dynamics; Warp to the Future combines feature transport with direct prediction to address limitations of transformation alone.

The residual should not be interpreted directly as pure environmental motion. With predicted poses, a discrete grid, and a finite field of view, it may also contain traffic-participant motion, occlusion changes, newly visible regions, pose-estimation bias, and interpolation error. Our goal is to allocate prediction work through the model structure to improve efficiency and accuracy. MC-JEPA combines the learning of motion and content features, showing that the two can inform each other (Bardes et al., 2023). We focus on the fact that changes of the ego reference frame have a computable structure in spatial representations. Diagnosing the role of the residual therefore requires analysis by static and dynamic regions, visible support, and prediction horizon; an overall loss reduction alone cannot establish its source.

Our contributions are as follows.

1. We introduce a physically structured predictive formulation that writes a high-dimensional future representation as the sum of an explicit reference transport induced by low-dimensional ego motion and a learned residual, moving physical information from an input condition into the prediction structure.

2. We establish a three-level analysis framework: geometric transport identifies changes explainable by ego motion; residual learning defines what the predictor must add; and static–dynamic grouping diagnoses the limits of that explanation. We compare four models, A0–A3, corresponding to direct prediction without structure, motion as a conditioning input, physical transport alone, and physically structured residual prediction.

3. Using a common field of view and a ground-truth-pose comparison, we quantify the proportion of scene changes explained by motion transport across static and dynamic classes and prediction horizons.

2 Related Work

Detailed related work appears in Appendix A.

3 METHOD

3.1 TASK DEFINITION AND THREE-LEVEL DECOMPOSITION

MotionJEPA organizes future-representation evolution into three levels. The first asks what changes ego motion can explain: low-dimensional relative pose is lifted into a geometric transport applied to the entire metric spatial representation. The second asks what the learner should predict: the transported representation serves as a physical reference, and the model fits only the conditional residual between that reference and the target. The third asks what the residual may contain: ground-truth poses and static–dynamic channels help diagnose its sources, while the residual may still include extrapolation, sampling, finite-field-of-view, and annotation errors.

LEVEL I: REPRESENTATION TRANSPORT INDUCED BY EGO MOTION

Let the current semantic representation be Z_t, the prediction horizon be h, and the ego motion extrapolated only from states available up to the current time be ξ_h.

$$Z_t \qquad h \qquad \widehat{\xi}_h$$

The overall prediction relation is given in Equation (1).

This operator maps low-dimensional motion to a high-dimensional spatial transformation. It represents the nominal future representation under the assumption that the scene is rigidly static while the observation frame moves with the ego vehicle. A2 is therefore a physical-transport baseline: it transports the current representation but does not predict scene evolution.

$$B_h = \mathcal{T}_{\widehat{\xi}_h}(Z_t), \qquad \widehat{Z}_{t+h} = B_h + P_\theta(B_h, \widehat{\xi}_h, h). \tag{1}$$

LEVEL II: CONDITIONAL RESIDUAL PREDICTION AROUND A PHYSICAL REFERENCE

Here, T_{ξ_h}(·) denotes explicit geometric transport, B_h = T_{ξ_h}(Z_t) is the prediction under the assumption that the environment remains static and only the ego vehicle moves, and P_θ(B_h, ξ_h, h) is the signed residual relative to that baseline. A positive residual can increase a semantic response at a location, while a negative residual can remove a response from its original location. A moving object therefore usually requires corrections at both its old and new locations. (Yin et al., 2020; Guen and Thome, 2020)

$$\mathcal{T} \qquad B_h \qquad P_\theta$$

Using physical information as a condition differs from using it as structure. A1 supplies motion to the network but still requires the network to learn how motion acts at each spatial location. A3 applies that effect first and then lets the network correct the physical reference.

The residual-source decomposition, Stage 1 construction, and ego-motion extrapolation are detailed in Appendix B.

3.2 EXPLICIT TRANSPORT IN METRIC BEV

For a static world point, its current coordinate p_t and its coordinate in the future ego frame p_{t+h} satisfy Equation (2).

$$p_t \qquad p_{t+h}$$

$$p_{t+h} = R(\widehat{\psi}_h)^\top (p_t - \widehat{d}_h). \tag{2}$$

Thus, to compute the value at a future output location q, we must query the corresponding current location R(ψ_h)q + d_h using the inverse mapping. The coordinate relation is given in Equation (3).

$$q \qquad R(\widehat{\psi}_h)q + \widehat{d}_h$$

$$B_h(q) = Z_t\left(R(\widehat{\psi}_h)q + \widehat{d}_h\right). \tag{3}$$

This output-to-input sampling arrangement is related to the grid-sampling formulation of spatial transformer networks (Jaderberg et al., 2015). In the current binary-semantic experiments, we use nearest-neighbor interpolation and zero padding outside the boundary with align_corners = False; we do not train a learnable motion module with bilinear sampling. The tensor row axis corresponds to metric x, and the column axis to metric y, so normalized coordinates supplied to the sampler must be ordered as column then row. The validity mask M_h(q) marks query locations that remain within the current BEV field of view. Zero padding cannot recover information outside the finite field of view, and valid overlap does not mean that newly visible regions have been predicted. Prior feature-prediction work combines motion transport of observed features with content prediction to address the limitations of warping alone (Saric et al., 2020).

$$\texttt{align_corners} = \texttt{False} \qquad x \qquad y \qquad M_h(q)$$

3.3 Conditional Residual Predictor

The predictor first uses a convolution whose kernel size and stride are both 4 to convert the 64×64 BEV into 16×16 tokens, each with dimension 128, and adds learnable spatial positional embeddings. The backbone is a three-layer, four-head self-attention Transformer with a feed-forward width of 512 per layer and GELU activations. Self-attention enables information exchange across grid locations without changing the definition of the residual target (Vaswani et al., 2017).

$$64 \times 64 \qquad 16 \times 16 \qquad 128 \qquad 512$$

The motion condition is (d_x/20, d_y/10, ψ/0.5, h/1.5). A two-layer MLP with width 128 and SiLU activations encodes it as a conditioning vector. For each attention and feed-forward sublayer, this vector generates scale and shift parameters for the LayerNorm output. After a final LayerNorm, tokens are reshaped into a spatial grid and passed through a transposed convolution with stride 4, GELU, and a 3×3 output convolution to recover the seven-channel residual. The output has no sigmoid, allowing positive and negative corrections. The three prediction horizons share network parameters and are predicted directly from their respective conditions; no autoregressive rollout is used.

$$(d_x/20, d_y/10, \psi/0.5, h/1.5) \qquad 128 \qquad 4 \qquad 3 \times 3$$

Stage 1 A3 strictly uses $\hat{Z}$_{t+h} = B_h + P_θ(B_h, ξ_h, h): the predictor processes only the transported BEV, with no concatenation of the original BEV and no gating branch. The final output convolution is zero-initialized, so A3 initially equals the parameter-free transport baseline A2; the final conditioning-modulation layer is also zero-initialized. This follows the general motivation for residual connections and residual learning, but the additive reference here has a specific geometric meaning (He et al., 2015).

$$\widehat{Z} = B_h + P_\theta(B_h, \widehat{\xi}_h, h)$$

For a given sample and fixed transport reference B_h, define the true residual R_h* = Z_{t+h} − B_h.

$$B_h \qquad R_h^* = Z_{t+h} - B_h$$

The relationship between prediction error and residual-fitting error is given in Equation (4).

$$\widehat{Z}_{t+h} - Z_{t+h} = P_\theta(B_h, \widehat{\xi}_h, h) - R^*_h. \tag{4}$$

This means that future-scene prediction error equals residual-fitting error as an algebraic identity. It does not guarantee that the true residual has small magnitude, that a finite-parameter network is easier to optimize, or that training will necessarily produce lower error than direct prediction. Zero initialization establishes only a common starting point. Whether the geometric prior improves sample efficiency, optimization stability, or generalization must be determined empirically.

3.4 Four Prediction Structures and Training Objective

The four structures represent different ways of using physical information in prediction. A0 predicts a future representation directly from the current BEV and horizon, without motion conditioning. A1 supplies extrapolated motion as an additional input, while the predictor must still learn the relationship between motion and the spatial representation. A2 applies geometric transport alone as a physical-reference baseline and does not predict scene evolution. A3 adds the predictor's signed residual to the transported representation: future prediction = physical transport reference + learned residual. A1 and A3 receive the same motion information, making them the key comparison between physical information as an input and physical information defining the prediction reference. A2-GT uses the true future pose only for privileged diagnosis.

Table 1 summarizes how the four prediction structures differ in motion conditioning, explicit transport, and learnable residuals, and the role of each comparison.

Model	**Prediction structure**	**Trainable parameters**	**Purpose of comparison**
A0	Directly predicts the future BEV from the current BEV, without motion conditioning	1,127,559	Direct-prediction baseline without an explicit motion prior
A1	Uses historically extrapolated motion as a condition to predict the future BEV	1,127,943	Motion supplied as an input feature
A2	Transports the current BEV using only the extrapolated ego pose	0	Physical-transport reference baseline
A3	Applies the same transport as A2, then predicts and adds a signed residual	1,127,943	Motion defines the reference; the network learns the remaining changes

3.5 Smooth-L1 Loss

Let the target BEV be Z, the model prediction be $\hat{Z}$, and the error at one channel, horizon, and grid location be e = $\hat{Z}$ − Z. We use Smooth-L1 with β = 1 (Girshick, 2015): the pointwise loss is $e^2/2$ when |e| < 1 and |e| − 1/2 when |e| ≥ 1.

$$\mathcal{L}_{S1} = \frac{\sum_{b,h,q} M_{b,h}(q)\, \frac{1}{7} \sum_{c=1}^{7} \updownarrow_1(e_{b,h,c,q})}{\sum_{b,h,q} M_{b,h}(q)}, \qquad \updownarrow_1(e) = \begin{cases} \frac{e^2}{2}, & |e| < 1, \\ |e| - \frac{1}{2}, & |e| \geq 1. \end{cases} \tag{5}$$

The parameter β is the boundary between the quadratic and linear regions. The quadratic region gives small errors a smooth penalty and gradient, supporting stable correction of predictions close to the target. The linear region limits the influence of a few large errors compared with a purely quadratic loss. The function and its first derivative are continuous at |e| = β, combining smooth behavior for small errors with a milder penalty for large ones. With β = 1, errors whose absolute value is at most 1 lie in the quadratic region, and larger errors lie in the linear region.

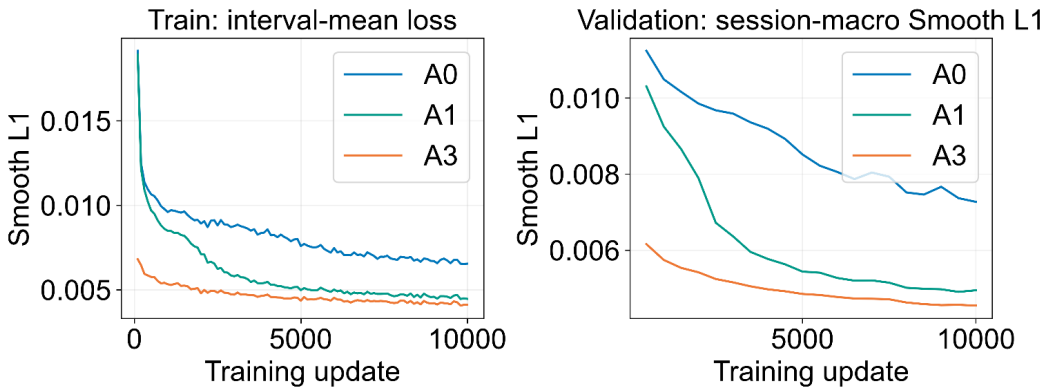

Figure: Training for 10,000 Steps

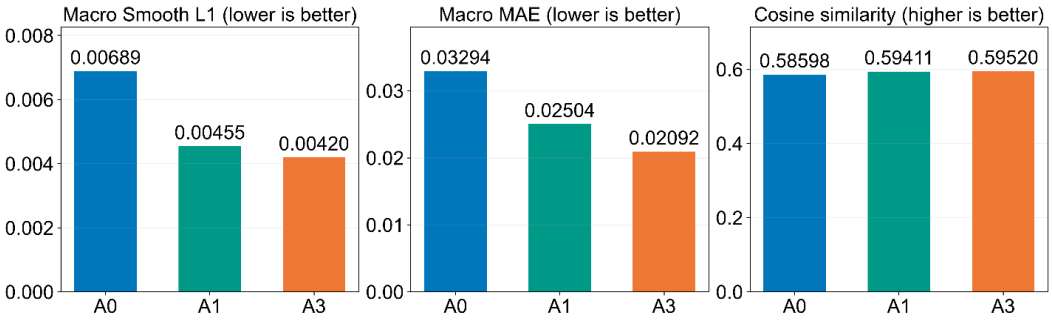

Figure: Test Macro Smooth-L1

4 Experimental Setup and Results

Data Splits and Result Provenance

We construct a seven-channel future semantic-BEV prediction task using the NAVSIM data organization (Dauner et al., 2024). The current input is built from map and scene annotations, and the prediction horizons are 0.5, 1.0, and 1.5 s. This experiment evaluates prediction of fixed semantic scenes. The frozen manifest contains 954 training logs, 119 validation logs, and 119 test logs, with 82,384, 9,887, and 11,017 indexed scenes, respectively. Stage 1 has 75,251 valid training samples and 9,219 valid validation samples. We compare A0 direct prediction, A1 motion-conditioned prediction, A2 parameter-free CTRA transport, and A3 transport plus residual prediction. A1 and A3 each have 1,127,943 trainable parameters; A0 has 384 fewer, and A2 has no trainable parameters. Copy and transport with the true pose, A2-GT, are retained as diagnostics; true future poses are used only in evaluation analyses.

Qualitative predictions appear in Appendix C.

Evaluation Metrics and Statistical Units

Validation Macro Smooth-L1 uses the pointwise loss defined in the Methods section. We first aggregate all samples, prediction horizons, and valid grid cells within each log, then average equally across logs; each log therefore contributes equally to the macro average.

Macro MAE uses the same log-level weighting. Micro-averaged error is computed by pooling all valid grid cells before calculating the metric. Results by horizon use the micro-average convention and represent the pooled error over all valid cells at that horizon; they are not log-level macro averages.

A0–A3 use the same CTRA validity mask, so their results cover only regions that can be sampled from the current field of view. A2-GT and A2 have different valid regions, so their error difference

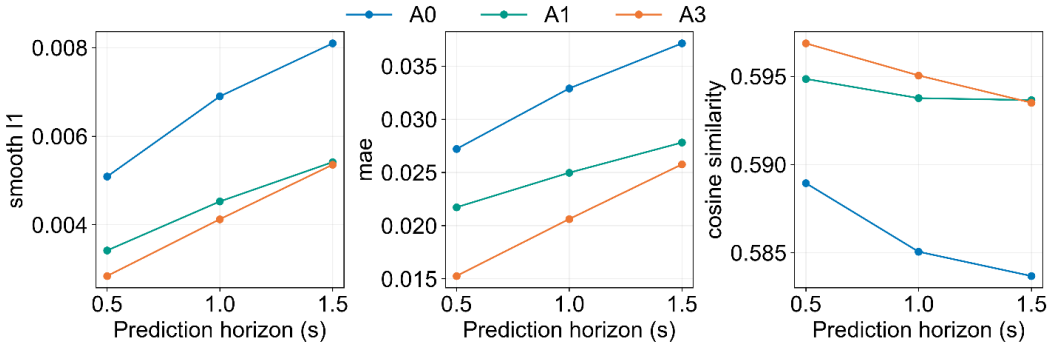

Figure: Test Results by Prediction Horizon

cannot be attributed directly to pose source. The decomposition diagnostic below compares the two transport methods on their common support.

For semantic evaluation, we binarize each channel at a threshold of 0.5 and aggregate true positives (TP), false positives (FP), and false negatives (FN). Per-class IoU is TP/(TP + FP + FN), and mIoU is the unweighted mean over the seven classes. Macro F1 is computed by first calculating F1 for each class and then averaging equally across classes. Classes with a zero denominator are assigned zero by the current implementation.

Table 2: Summary Metrics for Training, Validation, and Test Sets

Model	**Training masked Smooth-L1**	**Validation Macro Smooth-L1**	**Test Macro Smooth-L1 / MAE**	**Validation mIoU**
A0	0.00654368	0.00727024	0.006887 / 0.032939	0.533055
A1	0.00446283	0.00490304	0.004546 / 0.025037	0.590582
A2	No trainable parameters	0.00788889	0.007383 / 0.014766	0.630721
A3	0.00411631	0.00454285	0.004196 / 0.020921	0.658450

Table 3 reports test-set macro averages. A3 achieves a Macro Smooth-L1 of 0.004196, 7.69% below A1's 0.004546. Its Macro MAE is 0.020921, 16.44% below A1's 0.025037. Relative to A2, A3 has 43.16% lower Smooth-L1 but slightly higher MAE.

Table 3: Aggregate Test-Set Continuous Errors

Model	**Selected step**	**Macro Smooth-L1 ↓**	**Macro MAE ↓**
Copy	—	0.014147	0.028294
A0	10,000	0.006887	0.032939
A1	9,500	0.004546	0.025037
A2	—	0.007383	0.014766
A3	10,000	0.004196	0.020921

Micro-averaged errors by horizon are reported in Table 4. A3 achieves Smooth-L1 values of 0.002842, 0.004124, and 0.005356 at 0.5, 1.0, and 1.5 s, respectively, each below A1's 0.003418, 0.004525, and 0.005414. The relative reductions are approximately 16.85%, 8.86%, and 1.07%. The gap between A3 and A1 narrows as the prediction horizon increases.

Table 4: Test-Set Micro-Averaged Smooth-L1 by Prediction Horizon

Model	**Smooth-L1 at 0.5 s** ↓	**Smooth-L1 at 1.0 s** ↓	**Smooth-L1 at 1.5 s** ↓
A0	0.005087	0.006899	0.008097
A1	0.003418	0.004525	0.005414
A2	0.004412	0.007154	0.010141
A3	0.002842	0.004124	0.005356

Validation Continuous Errors and Semantic Quality

Aggregate validation continuous errors and semantic metrics are reported in Table 5. A3 achieves a Macro Smooth-L1 of 0.004543, below A1's 0.004903. Its mIoU is 0.658450, 6.787 percentage points above A1 and 2.773 points above A2. A3's overall MAE is 0.022201, higher than A2's 0.015778, showing that the methods are not ranked consistently across metrics.

Table 5: Aggregate Validation Continuous Errors and Semantic Metrics

Model	**Macro Smooth-L1** ↓	**Overall MAE** ↓	**mIoU** ↑	**Macro F1** ↑
A0	0.007270	0.034637	0.533055	0.629159
A1	0.004903	0.026699	0.590582	0.677209
A2	0.007889	0.015778	0.630721	0.742777
A3	0.004543	0.022201	0.658450	0.762583

Per-class IoU, horizon-wise mIoU, and pose-transport diagnostics appear in Appendix D.

5 Conclusion

We proposed MotionJEPA, which constructs a future-reference representation through BEV geometric transport induced by ego motion and learns the signed residual not explained by that reference. The residual may also contain environmental change, motion-extrapolation error, finite-field-of-view effects, and discretization or sampling error; the current experiments do not causally separate these sources. After 10,000 training steps, A3 achieves a masked Smooth-L1 of 0.00411631, below A1's 0.00446283. The best validation Macro Smooth-L1 values are 0.00454285 for A3 and 0.00490304 for A1, a reduction of about 7.35%. On the overall test set, A3's Macro Smooth-L1 is 0.004196 versus 0.004546 for A1, a reduction of 7.69%; their Macro MAE values are 0.020921 and 0.025037, respectively. On validation, A3 achieves an mIoU of 0.658450, compared with 0.630721 for A2. These results support the effectiveness of constructing a physical reference from ego motion and then predicting a residual for the current fixed-semantic-BEV prediction task.

AI Use Statement

Generative AI tools assisted with manuscript editing. The authors remain responsible for data provenance, computations, and all claims in the submission.

REFERENCES

Assran, Mahmoud, Quentin Duval, Ishan Misra, et al. 2023. *Self-Supervised Learning from Images with a Joint-Embedding Predictive Architecture*. arXiv preprint arXiv:2301.08243. https://arxiv.org/abs/2301.08243.

Assran, Mido, Adrien Bardes, David Fan, et al. 2025. *V-JEPA 2: Self-Supervised Video Models Enable Understanding, Prediction and Planning*. arXiv preprint arXiv:2506.09985. https://arxiv.org/abs/2506.09985.

Bardes, Adrien, Quentin Garrido, Jean Ponce, et al. 2024. *Revisiting Feature Prediction for Learning Visual Representations from Video*. arXiv preprint arXiv:2404.08471. https://arxiv.org/abs/2404.08471.

Bardes, Adrien, Jean Ponce, and Yann LeCun. 2023. *MC-JEPA: A Joint-Embedding Predictive Architecture for Self-Supervised Learning of Motion and Content Features*. arXiv preprint arXiv:2307.12698. https://arxiv.org/abs/2307.12698.

Caesar, Holger, Juraj Kabzan, Kok Seang Tan, et al. 2021. *NuPlan: A closed-loop ML-based planning benchmark for autonomous vehicles*. arXiv preprint arXiv:2106.11810. https://arxiv.org/abs/2106.11810.

Chen, Sitao, Zhuangwei Zhuang, Hui Luo, Qingyao Wu, and Mingkui Tan. 2026. *Geometry-Aware Spatio-Temporal Context Modeling for 4D Occupancy Forecasting*. arXiv preprint arXiv:2608.15279. https://arxiv.org/abs/2608.15279.

Chen, Ting, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. 2020. *A Simple Framework for Contrastive Learning of Visual Representations*. arXiv preprint arXiv:2002.05709. https://arxiv.org/abs/2002.05709.

Dauner, Daniel, Marcel Hallgarten, Tianyu Li, et al. 2024. *NAVSIM: Data-Driven Non-Reactive Autonomous Vehicle Simulation and Benchmarking*. arXiv preprint arXiv:2406.15349. https://arxiv.org/abs/2406.15349.

Girshick, Ross. 2015. *Fast R-CNN*. arXiv preprint arXiv:1504.08083. https://arxiv.org/abs/1504.08083.

Guan, Yanchen, Haicheng Liao, Zhenning Li, et al. 2024. *World Models for Autonomous Driving: An Initial Survey*. arXiv preprint arXiv:2403.02622. https://arxiv.org/abs/2403.02622.

Guen, Vincent Le, and Nicolas Thome. 2020. *Disentangling Physical Dynamics from Unknown Factors for Unsupervised Video Prediction*. arXiv preprint arXiv:2003.01460. https://arxiv.org/abs/2003.01460.

Ha, David, and Jürgen Schmidhuber. 2018. *World Models*. arXiv preprint arXiv:1803.10122. https://arxiv.org/abs/1803.10122.

Hafner, Danijar, Timothy Lillicrap, Ian Fischer, et al. 2018. *Learning Latent Dynamics for Planning from Pixels*. arXiv preprint arXiv:1811.04551. https://arxiv.org/abs/1811.04551.

He, Kaiming, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. 2021. *Masked Autoencoders Are Scalable Vision Learners*. arXiv preprint arXiv:2111.06377. https://arxiv.org/abs/2111.06377.

He, Kaiming, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2015. *Deep Residual Learning for Image Recognition*. arXiv preprint arXiv:1512.03385. https://arxiv.org/abs/1512.03385.

Jaderberg, Max, Karen Simonyan, Andrew Zisserman, and Koray Kavukcuoglu. 2015. *Spatial Transformer Networks*. arXiv preprint arXiv:1506.02025. https://arxiv.org/abs/1506.02025.

Jia, Xiaosong, Zhenjie Yang, Qifeng Li, Zhiyuan Zhang, and Junchi Yan. 2024. *Bench2Drive: Towards Multi-Ability Benchmarking of Closed-Loop End-To-End Autonomous Driving*. arXiv preprint arXiv:2406.03877. https://arxiv.org/abs/2406.03877.

Lang, Alex H., Sourabh Vora, Holger Caesar, Lubing Zhou, Jiong Yang, and Oscar Beijbom. 2019. "PointPillars: Fast Encoders for Object Detection From Point Clouds." *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. https://openaccess.thecvf.com/content_CVPR_2019/html/Lang_PointPillars_Fast_Encoders_for_Object_Detection_From_Point_Clouds_CVPR_2019_paper.html.

LeCun, Yann. 2022. *A Path Towards Autonomous Machine Intelligence*. OpenReview position paper. https://openreview.net/forum?id=BZ5a1r-kVsf.

Li, Yingyan, Yuqi Wang, Yang Liu, Jiawei He, Lue Fan, and Zhaoxiang Zhang. 2025. *End-to-End Driving with Online Trajectory Evaluation via BEV World Model*. arXiv preprint arXiv:2504.01941. https://arxiv.org/abs/2504.01941.

Li, Zhiqi, Wenhai Wang, Hongyang Li, et al. 2022. *BEVFormer: Learning Bird's-Eye-View Representation from Multi-Camera Images via Spatiotemporal Transformers*. arXiv preprint arXiv:2203.17270. https://arxiv.org/abs/2203.17270.

Liu, Zhijian, Haotian Tang, Alexander Amini, et al. 2022. *BEVFusion: Multi-Task Multi-Sensor Fusion with Unified Bird's-Eye View Representation*. arXiv preprint arXiv:2205.13542. https://arxiv.org/abs/2205.13542.

Philion, Jonah, and Sanja Fidler. 2020. *Lift, Splat, Shoot: Encoding Images From Arbitrary Camera Rigs by Implicitly Unprojecting to 3D*. arXiv preprint arXiv:2008.05711. https://arxiv.org/abs/2008.05711.

Saric, Josip, Marin Orsic, Tonci Antunovic, Sacha Vrazic, and Sinisa Segvic. 2020. "Warp to the Future: Joint Forecasting of Features and Feature Motion." *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. https://openaccess.thecvf.com/content_CVPR_2020/html/Saric_Warp_to_the_Future_Joint_Forecasting_of_Features_and_Feature_CVPR_2020_paper.html.

Schubert, Robin, Eric Richter, and Gerd Wanielik. 2008. "Comparison and evaluation of advanced motion models for vehicle tracking." *International Conference on Information Fusion*. https://www.tu-chemnitz.de/etit/nt/publikationen/publikationen.php.en?jahr=2008.

Tu, Sifan, Xin Zhou, Dingkang Liang, et al. 2025. *The Role of World Models in Shaping Autonomous Driving: A Comprehensive Survey*. arXiv preprint arXiv:2502.10498. https://arxiv.org/abs/2502.10498.

Vaswani, Ashish, Noam Shazeer, Niki Parmar, et al. 2017. *Attention Is All You Need*. arXiv preprint arXiv:1706.03762. https://arxiv.org/abs/1706.03762.

Wang, Zhidong, Jingsong Liang, Zirui Li, Zhan Chen, Han Yu, and Chen Lv. 2026. *Ego-Dynamics-Augmented World Model for Autonomous Driving with Zero-Shot Cross-Embodiment Adaptation*. arXiv preprint arXiv:2607.13410. https://arxiv.org/abs/2607.13410.

Yin, Yuan, Vincent Le Guen, Jérémie Dona, et al. 2020. *Augmenting Physical Models with Deep Networks for Complex Dynamics Forecasting*. arXiv preprint arXiv:2010.04456. https://arxiv.org/abs/2010.04456.

Zhang, Yumeng, Shi Gong, Kaixin Xiong, et al. 2024. *BEVWorld: A Multimodal World Simulator for Autonomous Driving via Scene-Level BEV Latents*. arXiv preprint arXiv:2407.05679. https://arxiv.org/abs/2407.05679.

Zheng, Yupeng, Pengxuan Yang, Zebin Xing, et al. 2025. *World4Drive: End-to-End Autonomous Driving via Intention-aware Physical Latent World Model*. arXiv preprint arXiv:2507.00603. https://arxiv.org/abs/2507.00603.

A. Detailed Related Work

Prediction Targets in Driving World Models and the Scope of This Work

World models for autonomous driving can be grouped by prediction target, including latent states, occupancy or semantic scenes, future images, and interactive behavior. The model backbone and prediction target are separate dimensions; Transformers, diffusion models, and JEPA are not mutually exclusive categories. This work focuses on short-horizon, spatialized future representations in latent space. (Ha and Schmidhuber, 2018; Hafner et al., 2018, 2019; Tu et al., 2025; Guan et al., 2024)

From Self-Supervised Representation Learning to JEPA Prediction

Contrastive learning, distillation, and masked reconstruction define self-supervised objectives through similarity, teacher representations, and input reconstruction, respectively. JEPA instead predicts target regions or future features in representation space, avoiding pixel reconstruction as the sole learning signal. I-JEPA, V-JEPA, and V-JEPA 2 demonstrate extensions across image and video representations and action-conditioned settings. They also show that a JEPA architecture depends on how its encoder and prediction target are trained, not simply on whether a network uses latents. (T. Chen et al., 2020; He et al., 2021; Assran et al., 2023; Bardes et al., 2024; Assran et al., 2025)

Metric BEV and Multimodal Scene Representations

BEV organizes information from multiple cameras or sensors in a metric space, enabling spatial alignment and representation of ego motion through view transformation, temporal fusion, and unified multimodal representations. We use a fixed semantic BEV constructed from NAVSIM annotations to isolate the effect of the prediction structure. (Philion and Fidler, 2020; Lang et al., 2019; Z. Li et al., 2022; Liu et al., 2022; Y. Zhang et al., 2024)

Ego Motion, Environmental Change, and Residual Decomposition

Motion transport and residual modeling offer a structured division of labor for scene prediction: first align the reference frame using interpretable motion, then learn changes not explained by that transformation. Physics-informed models and feature-warping studies support this modeling motivation, but they do not establish that a residual has been causally identified as pure environmental dynamics. We use ground-truth poses, static–dynamic channels, and field-of-view boundaries as diagnostic tools, distinguishing structural assumptions from empirical conclusions. (Yin et al., 2020; Guen and Thome, 2020; Saric et al., 2020; Schubert et al., 2008)

Planning Interfaces and Evaluation Protocols

A world model for planning must also be tested for whether its predictions change candidate-trajectory rankings and closed-loop driving performance. Non-reactive evaluations such as NAVSIM, the nuPlan closed-loop benchmark, and interactive scenarios in Bench2Drive address different questions; improvements in prediction error cannot substitute for planning scores. We use NAVSIM scenes for representation prediction, and future work should test planning value under a consistent official protocol. (Caesar et al., 2021; Dauner et al., 2024; Jia et al., 2024; Y. Li et al., 2025; Zheng et al., 2025)

B. Additional Method Details

Level III: Limits on Residual Interpretation

To distinguish environmental evolution from prior error, let S_t denote the ideal continuous semantic field and Q the operator that samples this field onto a common BEV grid; let ξ_h* denote the true future ego pose. Let T^c denote ideal transport in the continuous field, and let Z_{t+h} = Q S_{t+h} + ϵ_{t+h}. The residual target has the identity decomposition in Equation (A1), with its three terms defined in Equations (A2)–(A4).

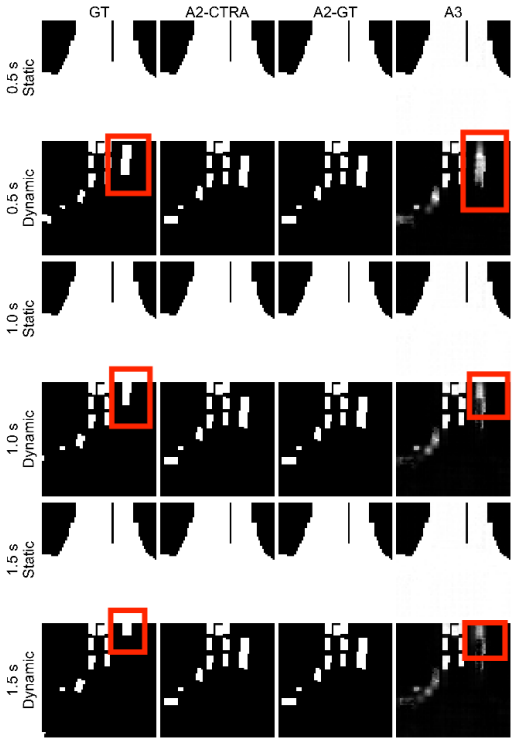

Figure: Static and Dynamic Regions

S_t Q ξ_h^* $Z_{t+h} = QS_{t+h} + \epsilon_{t+h}$ $\mathcal{T}^c$ $\mathcal{T}^d$

$$Z_{t+h} - \mathcal{T}^d_{\widehat{\xi}_h} Z_t = R_h^{env} + R_h^{motion} + R_h^{repr}, \tag{A1}$$

$$R_h^{env} = Q\left(S_{t+h} - \mathcal{T}^c_{\xi_h^*} S_t\right), \tag{A2}$$

$$R_h^{motion} = Q\left(\mathcal{T}^c_{\xi_h^*} S_t - \mathcal{T}^c_{\widehat{\xi}_h} S_t\right), \tag{A3}$$

$$R_h^{repr} = Q\mathcal{T}^c_{\widehat{\xi}_h} S_t - \mathcal{T}^d_{\widehat{\xi}_h} Z_t + \epsilon_{t+h}. \tag{A4}$$

The first term is the environmental change that remains after alignment using true ego motion. The second is scene misalignment caused by motion-extrapolation error. The third includes current- and future-representation errors, discretization, and resampling errors. This equality explains what the residual may contain; it does not show that the three terms are statistically independent, causally disentangled, or uniquely recoverable from the available training labels. The network regresses only their sum. Even with a ground-truth pose, the residual need not equal only the physical motion of other traffic participants: finite sensing range, visibility changes, and label errors can also produce residual changes. We therefore describe this as a structural decomposition of motion transport and remaining environmental change, rather than as an achieved factorization of pure dynamics.

Stage 1 Implementation and Experimental Scope

Our current experiments use a Stage 1 setup: we construct a fixed semantic BEV from NAVSIM map and traffic-scene annotations to control for representation quality when comparing structures. The input Z_t ∈ {0,1}ˆ{7×64×64} has seven channels: road, walkway, lane centerline, static object,

vehicle, pedestrian, and bicycle. Channels may overlap, so the representation is a multilabel grid rather than a mutually exclusive seven-class probability distribution. The forward range is x ∈ [−12, 36) m, the lateral range is y ∈ [−24, 24) m, and the resolution along both axes is 0.75 m. The x-axis points forward, the y-axis points left, and counterclockwise yaw is positive. The prediction targets are semantic BEVs at 0.5, 1.0, and 1.5 s, each constructed in the ego reference frame at its corresponding time. The Stage 1 predictor takes the current semantic BEV as input, and the motion prior is estimated from the two most recent ego states. Ground-truth future poses, future semantics, and expert trajectories are not predictor inputs; future semantics are used only as training and evaluation targets.

$$Z_t \in \{0,1\}^{7\times 64\times 64} \qquad x \in [-12, 36) \qquad y \in [-24, 24)$$
$$0.75 \qquad x \qquad y \qquad 0.5 \qquad 1.0 \qquad 1.5$$

Ego-Motion Extrapolation from Historical States

Let ξ_h = (d_{x,h}, d_{y,h}, ψ_h) denote the translation and yaw of the future ego origin in the current reference frame. We use a constant-turn-rate, constant-acceleration prior, motivated by motion models for vehicle tracking (Schubert et al., 2008). The implementation retains two-dimensional body-frame velocity and acceleration as an engineering extension of this idea. Let τ denote elapsed time; the motion equations are given in Equation (A5).

$$\widehat{\xi}_h = (\widehat{d}_{x,h}, \widehat{d}_{y,h}, \widehat{\psi}_h) \qquad \tau$$

$$\dot{d}(\tau) = R(\psi(\tau))v(\tau), \qquad \dot{v}(\tau) = a_t, \qquad \dot{\psi}(\tau) = \omega_t. \tag{A5}$$

The initial conditions are d(0) = 0 and ψ(0) = 0; v(0) and a_t take the recorded two-dimensional velocity and acceleration. ω_t is obtained by dividing the wrapped yaw difference between the two most recent frames by 0.5 s. Acceleration components remain constant in the vehicle frame, and displacement is integrated in the current reference frame after rotation by the instantaneous yaw. The implementation uses midpoint integration with a maximum step of 0.01 s and extracts poses at each target horizon. This input represents a motion trend extrapolated from historical states; it is not a control command from a future expert.

$$d(0) = 0 \qquad \psi(0) = 0 \qquad v(0) \qquad a_t \qquad \omega_t \qquad 0.5 \qquad 0.01$$

This prior is suitable for short periods of smooth motion, but hard braking, changes in steering intent, and lateral-motion changes increase extrapolation error. We must therefore report ego-translation and yaw-extrapolation errors alongside their relationship to scene-prediction error; otherwise, A3's correction of misalignment could be mistaken for learned traffic interaction. True future poses enter only the separate A2-GT diagnostic path and are not inputs to deployable comparisons.

C. Qualitative Predictions

D. Additional Validation Results

Table 6 reports per-class IoU. Compared with A2, A3 increases vehicle IoU from 0.525118 to 0.626862, a gain of 10.174 percentage points. IoU also improves for road, walkway, lane centerline, and static objects. Pedestrian IoU decreases from 0.390799 to 0.384487, a drop of 0.631 points, while bicycle IoU is 0.237043 for both methods. Gains therefore vary by class.

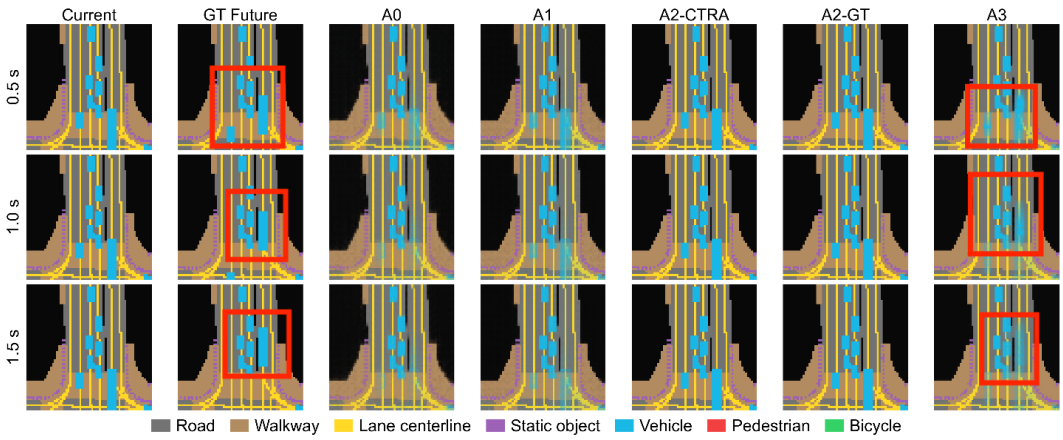

Figure: MotionJEPA Qualitative Results (a)

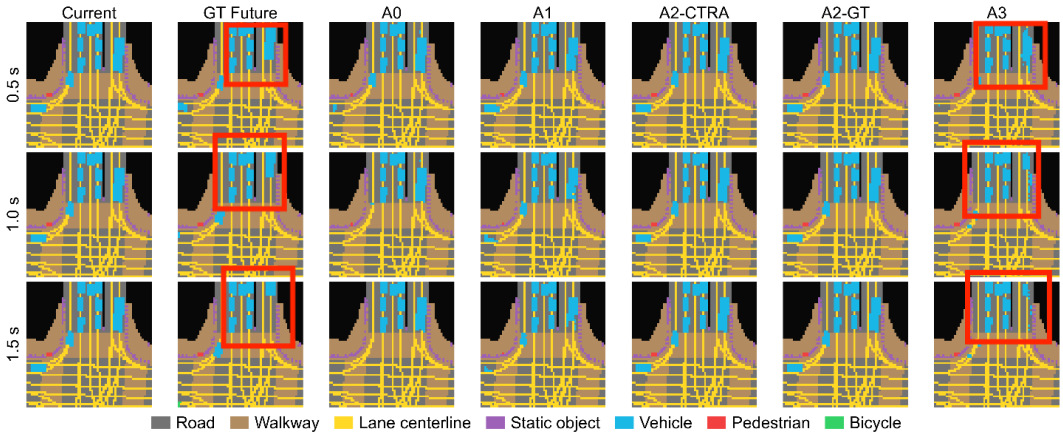

Figure: MotionJEPA Qualitative Results (b)

Table 6: Validation IoU by Semantic Class

Semantic class	**A0 IoU**	**A1 IoU**	**A2 IoU**	**A3 IoU**
Road	0.963152	0.982252	0.976773	0.982987
Walkway	0.836294	0.923925	0.916938	0.931330
Lane centerline	0.714539	0.787327	0.755717	0.809621
Static object	0.465294	0.553122	0.612657	0.636821
Vehicle	0.554450	0.602019	0.525118	0.626862
Pedestrian	0.197657	0.285427	0.390799	0.384487
Bicycle	0.000000	0.000000	0.237043	0.237043

Semantic quality by prediction horizon is reported in Table 7. A3 has the highest mIoU among the four methods at all three horizons: 0.735756, 0.647195, and 0.587833 at 0.5, 1.0, and 1.5 s, respectively. Its advantage over A1 is 10.208, 6.352, and 3.613 percentage points, respectively. A3's own mIoU decreases as the horizon grows.

Table 7: Validation mIoU by Prediction Horizon

Model	**mIoU at 0.5 s ↑**	**mIoU at 1.0 s ↑**	**mIoU at 1.5 s ↑**
A0	0.581029	0.523071	0.491874
A1	0.633674	0.583680	0.551703
A2	0.717438	0.622453	0.557099
A3	0.735756	0.647195	0.587833

To isolate motion-extrapolation error, we compare transport using the CTRA pose and the true future pose over their common support. Table 8 reports macro-average errors and the valid-region fraction for both settings. The true pose is used only for diagnosis.

Table 8: Pose-Transport Diagnostic Errors on the Common Validation Support

Transport pose	**Macro Smooth-L1 ↓**	**Macro MAE ↓**	**Valid-region fraction**
A2-CTRA	0.007833	0.015666	0.906002
A2-GT	0.004980	0.009960	0.906002