TEMPORAL CONTEXT NEEDS THE RIGHT TARGET: PROJECTED LATENT PREDICTION FOR END-TO-END DRIVING

Anonymous authors
Paper under double-blind review

ABSTRACT

Temporal context is essential for motion-aware driving, yet we find that naively introducing temporally fused features into a reconstruction-based latent world model can degrade planning performance. We attribute this failure to a representation–objective mismatch: the predicted latent encodes temporal history and action-conditioned evolution, whereas its supervision target remains a single-frame visual latent whose exact coordinates contain factors unnecessary for planning. To address this mismatch, we propose Temporal-JEPA LAW. Our method first introduces Latent World-Model Temporal Fusion (LWM-TF), which causally aggregates each LAW latent token over time while preserving its spatial identity and maintaining $O(NK^2)$ attention complexity. It then replaces raw latent reconstruction with normalized projected alignment, allowing the world model to match prediction-relevant structure without reproducing every visual-latent coordinate. Following LeWorldModel, a sketched isotropic Gaussian regularizer (SIGReg) prevents collapse in the unnormalized embedding space. Controlled experiments reveal the central effect of this design: temporal fusion alone improves average planning L2 from 0.61 m to 0.53 m, whereas coupling temporal prediction with raw mean-squared error degrades it to 0.66 m. Projected alignment mitigates this negative transfer, achieving 0.51 m average L2 and reducing the average collision rate from 0.30% to 0.09%. These results show that temporal information and predictive supervision must be designed jointly: LWM-TF supplies motion evidence, while JEPA provides a planning-compatible target for latent future prediction.

1 INTRODUCTION

End-to-end autonomous driving increasingly predicts ego trajectories directly from onboard observations, yet planning from a single-timestamp multi-view observation is physically underdetermined. A static snapshot localizes scene elements but discards the history needed to infer velocity, turning intent, occlusion evolution, and interaction trends. Consequently, a single-frame planner may achieve low open-loop displacement error while failing to maintain safe clearance around moving agents.

Latent world models introduce future-state reasoning, but their supervision must match the planning objective. Raw mean-squared reconstruction rewards reproduction of appearance-heavy visual factors such as texture and lighting, even when they are irrelevant to action selection. A planning-oriented target should instead emphasize motion, occupancy, interaction risk, and drivable-space structure.

LAW (Li et al., 2025) encodes multi-view images into compact latents, decodes future waypoints, and reconstructs future visual features with a latent world model. Temporal-JEPA LAW addresses its single-frame input and appearance-faithful target with two changes. LWM-TF applies causal temporal attention over existing LAW tokens without replacing their compact interface, while a JEPA-style objective aligns predicted and target latents after projection rather than coordinate by coordinate.

The main gain is collision reduction rather than marginal displacement improvement: against perception-free LAW, the 1s and average nuScenes collision rates fall from 0.14% and 0.30% to

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.

0.02% and 0.09%, respectively. On the full NAVSIM v2 navtest split (Dauner et al., 2024), the model completes all 12,146 scenarios and achieves high safety and rule-compliance subscores, with EPDMS limited mainly by two-frame extended comfort.

Our contributions are:

- We develop Latent World-Model Temporal Fusion (LWM-TF), which adapts factorized temporal attention to LAW's compact latent tokens through causal, token-wise aggregation and a gated residual path, with $O(K^2N)$ complexity instead of full spatio-temporal $O((KN)^2)$ attention.
- We reformulate LAW-style world-model learning with a Temporal-JEPA projected prediction objective and SIGReg instead of raw next-latent MSE reconstruction.
- We evaluate the resulting planner on nuScenes and full NAVSIM v2 navtest, showing that temporal context and projected world-model supervision improve collision-related behavior while preserving competitive trajectory accuracy.

2 Related Work

End-to-End Autonomous Driving. Perception-based end-to-end methods use intermediate tasks to shape planning representations, as in ST-P3 (Hu et al., 2022), UniAD (Hu et al., 2023), and VAD (Jiang et al., 2023). Recent work explores sparse scene representations, generative trajectory modeling, and language-assisted supervision. SparseDrive (Sun et al., 2025) replaces dense BEV features with sparse instance representations and uses collision-aware planning selection. GenAD (Zheng et al., 2024) formulates end-to-end driving as generative future modeling in a structural latent trajectory space. DiffusionDrive (Liao et al., 2025) introduces a truncated diffusion policy for real-time multi-modal planning, while DiMA (Hegde et al., 2025) distills multi-modal language-model knowledge into a vision-based planner for long-tail robustness. LAW instead uses multi-view features and waypoint supervision without dense perception labels, making image history particularly important for recovering dynamic cues.

Latent World Models for Driving. World models predict future states from current states and actions; LAW predicts latents rather than costly pixels. LeWorldModel (LeWM) (Maes et al., 2026) trains a pixel-based JEPA world model end-to-end using next-embedding prediction and SIGReg, which prevents collapse by matching random one-dimensional latent projections to an isotropic Gaussian. World4Drive (Zheng et al., 2025) also studies annotation-efficient latent world modeling by using vision foundation models to construct intention-aware physical latent states for trajectory generation and selection. We retain latent prediction but supervise it after projection, reducing pressure to reproduce planning-irrelevant visual coordinates.

Temporal Representation Learning. Prior work introduces temporal context through perception backbones, BEV encoders, recurrent memory, or video transformers. TimeSformer and ViViT factorize spatial and temporal attention, including temporal attention among tokens with the same spatial index (Bertasius et al., 2021; Arnab et al., 2021). In autonomous driving, BEVFormer recurrently fuses historical BEV representations through temporal self-attention (Li et al., 2022). Video-transformer and V-JEPA-style approaches can bind time and space early by tokenizing raw video into spatio-temporal patches or tubelets (Bardes et al., 2024; Assran et al., 2025), but this usually requires a substantially larger video encoder and may disrupt the compact planning-token interface used by LAW. LWM-TF applies causal factorization to LAW's compact tokens and uses a gated residual path to share temporal state between the waypoint decoder and world model.

3 Method

Temporal-JEPA LAW follows two design principles. First, trajectory prediction should be conditioned on short-horizon physical context rather than on an isolated visual snapshot, but this temporal context should be injected without replacing LAW's compact planning-token interface with a full video model. Second, the latent world model should be trained with a supervision signal that is closer to planning utility than raw latent reconstruction. Together, these principles define a coupled

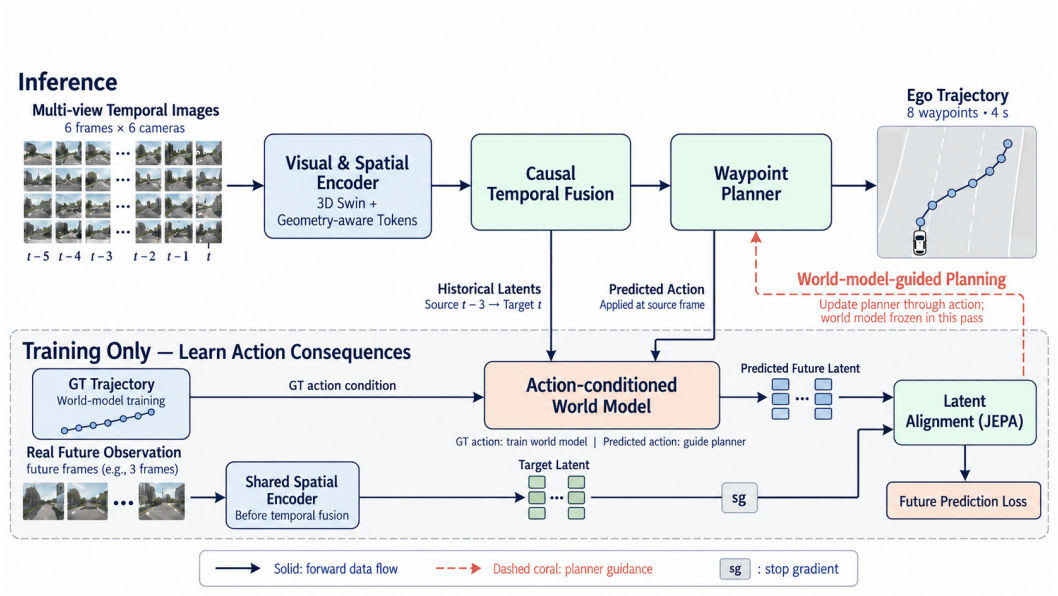

Figure 1: Temporal-JEPA LAW. LWM-TF fuses latent-token histories; the planner predicts waypoints, and the world model aligns its future prediction with a stop-gradient target in projected JEPA space.

representation problem: the planner needs recent motion evidence, while the world-model branch should avoid spending capacity on appearance-level reconstruction.

The overall pipeline proceeds in four stages. First, a shared visual encoder converts each synchronized multi-view observation in a short temporal queue into LAW-style latent tokens. Second, Latent World-Model Temporal Fusion (LWM-TF) fuses tokens across time while preserving the spatial/view identity of each token. Third, the waypoint decoder consumes the current fused latent to predict future ego waypoints; these waypoints also serve as action-conditioned inputs to the temporal world model. Fourth, the world model predicts a future latent from the fused history and predicted waypoints, and the prediction is supervised after JEPA projection instead of by raw latent reconstruction. Training is end-to-end with supervised waypoint imitation, JEPA alignment, and SIGReg; no replay buffer, closed-loop policy optimizer, or external memory system is introduced. Fig. 1 illustrates the resulting architecture.

3.1 Visual Latent Encoding

The visual encoder converts raw camera observations into the latent token sequence shared by the planner and world model. At this stage, we preserve LAW's spatial token structure and defer temporal reasoning to LWM-TF. Let I_t denote the synchronized multi-view camera images at time t. The visual encoder E_θ maps each frame to a set of latent tokens

$$V_t = E_\theta(I_t), \qquad V_t \in \mathbb{R}^{N\times C}, \tag{1}$$

where N is the number of spatial/view tokens and C is the channel dimension. For a queue of K frames, the encoder produces a raw latent sequence

$$\mathbf{V}_{t-K+1:t} = [V_{t-K+1}, \ldots, V_t] \in \mathbb{R}^{K\times N\times C}. \tag{2}$$

These tokens form the common input to the temporal fusion, planning, and world-model modules. Unlike a single-timestamp latent, the sequence $\mathbf{V}_{t-K+1:t}$ retains the recent physical evolution of the scene, including ego motion and the apparent motion of surrounding agents. The resulting sequence is the input to LWM-TF, where temporal evidence is aggregated.

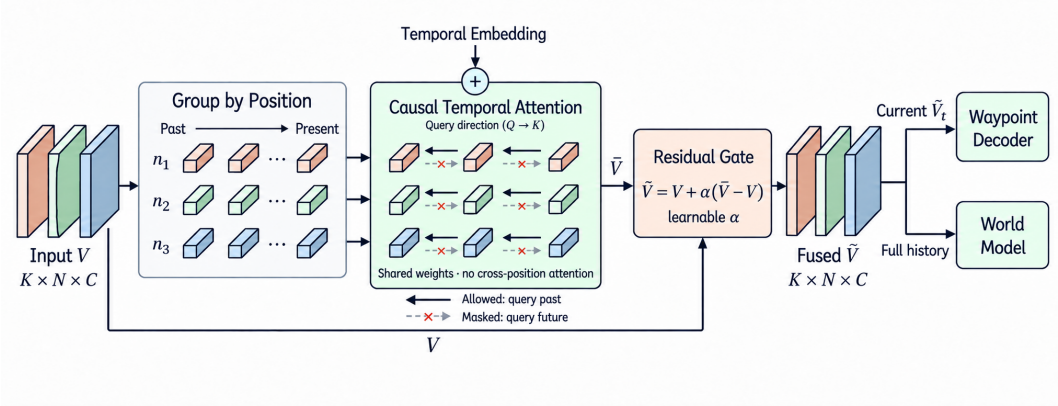

Figure 2: Latent World-Model Temporal Fusion (LWM-TF). Tokens with the same spatial/view identity are grouped across the K-frame queue and processed by shared causal temporal attention. A learnable residual gate preserves the original LAW representation; the current fused latent is sent to the waypoint decoder, while the fused history conditions the world model.

3.2 Latent World-Model Temporal Fusion

Temporal-JEPA LAW uses *Latent World-Model Temporal Fusion* (LWM-TF) to aggregate short-horizon evidence while keeping the latent-token layout compatible with both the original LAW planner and its latent world model. LAW visual tokens already encode spatial and view-specific geometry through position-aware multi-view feature extraction. A direct video-token approach could encode temporal information by processing spatio-temporal tubelets from the raw multi-view image queue, but this would move temporal fusion into a heavier video backbone. A naive spatio-temporal Transformer that mixes all tokens across both time and space would have complexity $O((KN)^2)$ for K frames and N spatial/view tokens, and may also blur the geometric meaning of individual tokens. For planning, this is undesirable: collision avoidance depends on precise spatial localization, so a token representing one image/view location should not arbitrarily mix information with unrelated positions while temporal motion is being estimated.

LWM-TF therefore performs token-wise temporal attention. Following factorized temporal attention in video Transformers (Bertasius et al., 2021; Arnab et al., 2021), each token attends only to its own history across the frame queue, reducing attention complexity to $O(K^2N)$ and preserving spatial token identity. This design addresses the missing-motion limitation of single-frame prediction while avoiding the computational cost of full video input modeling. It respects the geometric structure learned by LAW while still allowing each location to accumulate velocity, occlusion, and interaction cues from previous frames. As illustrated in Fig. 2, tokens are first grouped by their spatial/view position, processed by shared causal temporal attention, and merged with the original sequence through a learnable residual gate.

After adding learnable temporal embeddings P, the token sequence is processed by

$$\bar{\mathbf{V}} = T_\omega(\mathbf{V}_{t-K+1:t} + P; \text{causal mask}), \tag{3}$$

where the causal mask prevents information leakage from future frames. The fused representation is obtained through a learnable residual gate:

$$\tilde{\mathbf{V}}_{t-K+1:t} = \mathbf{V}_{t-K+1:t} + \alpha(\bar{\mathbf{V}} - \mathbf{V}_{t-K+1:t}), \tag{4}$$

where α is a scalar gate initialized to 0.10. The residual gate keeps the raw LAW latent path available while allowing temporal evidence to gradually reshape the planning representation. Because LWM-TF provides the shared input to the planner and temporal world model, both components operate on a consistent temporal latent state. The fused sequence $\tilde{\mathbf{V}}_{t-K+1:t}$ is then passed to the waypoint decoder and the world-model branch.

3.3 WAYPOINT PREDICTION MODULE

The waypoint decoder maps the current fused latent to future ego waypoints:

$$\hat{W}_t = D_\phi(\tilde{V}_t). \tag{5}$$

The waypoint loss is applied to the current prediction and to auxiliary predictions from previous frames in the queue:

$$\mathcal{L}_{traj} = \mathcal{L}^t_{wp} + \lambda_{aux} \sum_{\tau=t-K+1}^{t-1} \mathcal{L}^\tau_{wp}. \tag{6}$$

The auxiliary term keeps the temporal tokens predictive at each queue position instead of only at the final frame. This discourages the temporal module from becoming a passive feature smoother and encourages each history token to preserve planning-relevant physical information. We set $\lambda_{aux} = 0.30$ and distribute it over the previous frames. The predicted waypoints are the planner output and also provide the action condition for the latent world model.

3.4 TEMPORAL LATENT WORLD MODEL

The latent world model uses the fused temporal history and waypoint condition to predict the next latent state:

$$\hat{V}_t = M_\psi(\tilde{V}_{t-K+1:t-1}, \hat{W}_{t-K+1:t-1}). \tag{7}$$

This formulation differs from the original LAW world model, which predicts future latents from a raw single-timestamp latent and a predicted waypoint. Using fused history allows the world model to encode motion cues and temporal context before forecasting the next latent. This makes the prediction more informed by motion history, but it also makes raw reconstruction less appropriate: the predicted latent is produced from history and action evidence, while the target visual latent still contains single-timestamp appearance factors. Instead of applying direct latent reconstruction, Temporal-JEPA supervises the world model with a joint-embedding prediction loss:

$$\mathcal{L}_{jepa} = \mathrm{SmoothL1}\left(\mathrm{LN}(q_\eta(p_\rho(\hat{V}_t))), \mathrm{LN}(\mathrm{sg}(p_\rho(V_t)))\right), \tag{8}$$

where p_ρ is a projector, q_η is a predictor, $\mathrm{LN}(\cdot)$ denotes layer normalization, and $\mathrm{sg}(\cdot)$ denotes stop-gradient. This objective aligns prediction and target after projection rather than forcing element-wise matching in the raw latent space.

3.5 JEPA ALIGNMENT WITH SIGREG

The predicted latent is mapped by a projector p_ρ and predictor q_η, while the target latent is mapped by a stop-gradient projector. This prediction-oriented embedding reduces pressure to reproduce planning-irrelevant appearance coordinates (Assran et al., 2023). The detached target lets the objective update temporal fusion, world model, projector, and predictor without moving the raw target representation toward the prediction.

Following LeWM (Maes et al., 2026), we prevent collapse with the sketched isotropic Gaussian regularizer implemented by `jepa_sigreg_loss`. For flattened embeddings $Z \in \mathbb{R}^{S\times C}$, we sample M unit directions u_m and compare the empirical characteristic function of each projection $h_m = Zu_m$ with that of $\mathcal{N}(0, \sigma_0^2)$:

$$\mathcal{R}_{\mathrm{SIG}}(Z) = \frac{1}{M}\sum_{m=1}^{M}\sum_{\ell} w_\ell \left[\left(\mathbb{E}\cos(s_\ell h_m) - e^{-\sigma_0^2 s_\ell^2/2}\right)^2 + \left(\mathbb{E}\sin(s_\ell h_m)\right)^2\right]. \tag{9}$$

Here s_ℓ are fixed evaluation knots and w_ℓ are trapezoidal Gaussian weights, yielding the Epps–Pulley normality statistic used in the code. We regularize both predicted and target projected embeddings:

$$\mathcal{L}_{\mathrm{sigreg}} = \tfrac{1}{2}\left[\mathcal{R}_{\mathrm{SIG}}(q_\eta(p_\rho(\hat{V}_t))) + \mathcal{R}_{\mathrm{SIG}}(p_\rho(V_t))\right]. \tag{10}$$

3.6 Training Objective

The final training objective combines waypoint imitation with the JEPA-supervised world-model branch:

$$\mathcal{L} = \mathcal{L}_{traj} + \lambda_{jepa}\mathcal{L}_{jepa} + \lambda_{sigreg}\mathcal{L}_{sigreg}. \tag{11}$$

The world model is supervised through $\mathcal{L}_{jepa}$ and $\mathcal{L}_{sigreg}$ rather than direct latent reconstruction. We use $\lambda_{jepa} = 0.20$ and $\lambda_{sigreg} = 0.05$. Trajectory supervision trains the planner to imitate expert waypoints, while the world-model objective regularizes the latent representation through future prediction in the projected space.

4 Experiments

4.1 Experimental Setup

We evaluate on the nuScenes validation set (Caesar et al., 2020). Unless otherwise specified, temporal models use a four-frame queue for training, validation, and testing. We report open-loop planning L2 error at 1s, 2s, and 3s and collision rates; L2 measures imitation fidelity, whereas collisions capture safety-critical failures. Lower is better for all metrics.

We also evaluate the epoch-20 checkpoint on all 12,146 scenarios of the NAVSIM v2 navtest split using the official one-stage protocol (Dauner et al., 2024). We report EPDMS and its safety, compliance, progress, lane-keeping, and comfort subscores.

The implementation uses the perception-free LAW framework. The reference baseline follows its 12-epoch setting; temporal models trained without a LAW checkpoint retain only the external Swin initialization. All models use AdamW, batch size 4, and the same four-frame context during training and evaluation. For LAW-compatible temporal variants, the latent reconstruction weight is set to 0.20. For Temporal-JEPA, direct world-model reconstruction is disabled and the world model is supervised by $\mathcal{L}_{jepa}$ and $\mathcal{L}_{sigreg}$ with weights 0.20 and 0.05, respectively.

Experiments use a single NVIDIA GeForce RTX 4090 GPU with 48GB memory, an Intel Core Ultra 9 285K CPU, and 62GiB system memory; all temporal results follow the same training and validation protocol.

4.2 nuScenes Evaluation

Table 1 provides a landscape comparison with representative nuScenes planners; collision implementations and supervision differ across papers, so Table 3 provides the controlled comparison. Unlike perception-based or multi-task systems such as SparseDrive and DiffusionDrive, Temporal-JEPA targets the perception-free LAW setting. Compared with LAW perception-free, it reduces 1s collision from 0.14% to 0.02% and average collision from 0.30% to 0.09% while keeping the 1s L2 at 0.24m. Thus, the gain appears primarily in occupancy intersections rather than L2.

Table 1: nuScenes planning comparison. "Perception" denotes explicit perception supervision; "Mixed" aggregates earlier methods. Cross-paper collision protocols differ; Table 3 gives the controlled comparison. Lower is better.

Method	Perception	L2 (m) ↓				Collision (%) ↓			
		1s	2s	3s	Avg.	1s	2s	3s	Avg.
Prior Methods	Mixed	0.55	1.20	2.05	1.43	0.04	0.09	0.88	0.33
UniAD	✓	0.48	0.96	1.65	1.03	0.05	0.17	0.71	0.31
VAD	✓	0.41	0.70	1.05	0.72	0.07	0.17	0.41	0.22
BEV-Planner (Li et al., 2024)	–	0.30	0.52	0.83	0.55	0.10‡	0.37‡	1.30‡	0.59‡
SparseDrive-B	✓	0.29	0.55	0.91	0.58	0.01	0.02	0.13	0.06
DiffusionDrive	✓	0.27	0.54	0.90	0.57	0.03	0.05	0.16	0.08
World4Drive	–	0.23	0.47	0.81	0.50	0.02	0.12	0.33	0.16
LAW, perception-free	–	0.26	0.57	1.01	0.61	0.14	0.21	0.54	0.30
LAW, perception-based	✓	0.24	0.46	0.76	0.49	0.08	0.10	0.39	0.19
Ours, Temporal-JEPA LAW	–	0.24	0.47	0.81	0.51	0.02	0.04	0.21	0.09

4.3 NAVSIM FULL NAVTEST EVALUATION

Table 2 reports official one-stage results on full NAVSIM v2 navtest. Our epoch-20 checkpoint completes all 12,146 scenarios and obtains 86.0 EPDMS, with 97.6 NC, 99.3 DDC, 99.9 TLC, 96.9 TTC, and 95.7 LK.

Table 2: Full NAVSIM v2 navtest EPDMS comparison. Our epoch-20 result covers all 12,146 scenarios. Scores are percentages; higher is better.

Method	Backbone	NC	DAC	DDC	TLC	EP	TTC	LK	HC	EC	EPDMS
Ego Status MLP	–	93.1	77.9	92.7	99.6	86.0	91.5	89.4	98.3	85.4	64.0*
TransFuser	ResNet-34	96.9	89.9	97.8	99.7	87.1	95.4	92.7	98.3	87.2	76.7*
Hydra-MDP++	ResNet-34	97.2	97.5	99.4	99.6	83.1	96.5	94.4	98.2	70.9	81.4*
DriveSuprim	ResNet-34	97.5	96.5	99.4	99.6	88.4	96.6	95.5	98.3	77.0	83.1*
ARTEMIS	ResNet-34	98.3	95.1	98.6	99.8	81.5	97.4	96.5	98.3	98.3	83.1*
DiffusionDriveV2	ResNet-34	97.7	96.6	99.2	99.8	88.9	97.2	96.0	97.8	91.0	87.5
SparseDriveV2	ResNet-34	98.1	98.1	99.6	99.8	91.1	97.3	96.9	98.2	78.4	90.1
Temporal-JEPA LAW (Ours)	Swin3D-T	**97.6**	96.2	**99.3**	**99.9**	86.8	**96.9**	**95.7**	98.3	85.6	86.0

* denotes EPDMS reported with the original NAVSIM v2 evaluation code before the human-behavior filtering fix, following prior public comparison tables.

Two-frame extended comfort (55.6), which penalizes acceleration and jerk discontinuities, is the main limitation; all reported safety and rule-compliance scores exceed 94.

4.4 ABLATION STUDY

Table 3 isolates temporal fusion, JEPA supervision, and their combination under the same perception-free LAW setting.

Table 3: Controlled ablation. "WP/WM Temp." denote temporal decoder/world-model inputs; "JEPA Sup." denotes projected supervision. Δ is relative to Official LAW.

Method	Components			L2 (m) ↓				L2 Δ ↓		Collision (%) ↓			
	WP Temp.	WM Temp.	JEPA Sup.	1s	2s	3s	Avg.	1s	Avg.	1s	2s	3s	Avg.
Official LAW	–	–	–	0.26	0.57	1.01	0.61	0.00	0.00	0.14	0.21	0.54	0.30
LAW + JEPA	–	–	✓	0.29	0.58	0.99	0.62	+0.03	+0.01	0.01	0.04	0.27	0.11
Temporal LAW	✓	–	–	0.25	0.50	0.84	0.53	-0.01	-0.08	0.02	0.06	0.24	0.11
Temporal LAW + temporal WM	✓	✓	–	0.28	0.67	1.03	0.66	+0.02	+0.05	0.27	0.22	0.34	0.28
Temporal-JEPA LAW	✓	✓	✓	**0.24**	0.47	0.81	0.51	-0.02	-0.10	0.02	0.04	0.21	0.09

Effect of temporal fusion. LWM-TF reduces average collision from 0.30% to 0.11% and average L2 from 0.61m to 0.53m by exposing short-horizon motion to the decoder while retaining raw-latent world-model supervision. JEPA without LWM-TF also reaches 0.11% collision, but its 1s L2 (0.29m) is worse than Official LAW (0.26m) and the full model (0.24m), showing that projected supervision cannot replace explicit motion evidence.

Effect of JEPA supervision. With raw MSE, the temporal world model worsens average L2 to 0.66m and yields 0.28% collision, exposing the mismatch between motion-conditioned prediction and an appearance-heavy target. Combining temporal evidence with JEPA restores 1s L2 to 0.24m and achieves the best average L2 and collision, 0.51m and 0.09%; the components are therefore complementary.

4.5 QUALITATIVE BEV ANALYSIS

Fig. 3 shows a curved maneuver where single-frame LAW extrapolates tangentially into a collision-risk region, while Temporal-JEPA remains collision-free.



Figure 3: NuScenes camera/BEV comparison. Temporal-JEPA (top) follows the curve outside the highlighted risk region; LAW (bottom) extrapolates tangentially into surrounding occupancy.

By preserving spatial identity while aggregating four-frame history, LWM-TF captures turning and relative-motion cues; JEPA reduces appearance-level reconstruction pressure. The resulting path follows the local geometry, illustrating rather than guaranteeing the measured collision trend.

4.6 Discussion

Collision rates expose safety failures not captured by small L2 differences. The ablations show that temporal fusion supplies motion evidence while JEPA reduces appearance-faithful copying, so world-model supervision should be judged by downstream planning utility as well as reconstruction accuracy.

5 Conclusion

Temporal-JEPA LAW combines temporal fusion over compact planning tokens with projected future-latent supervision. Relative to perception-free LAW, it improves nuScenes average L2 from 0.61m to 0.51m and collision from 0.30% to 0.09%; on full NAVSIM v2 navtest it completes all scenarios and obtains 86.0 EPDMS, primarily limited by two-frame comfort. Together, LWM-TF and JEPA provide complementary motion evidence and planning-compatible supervision without a full video backbone.

AI use statement

Generative AI tools were used to assist with manuscript language editing and LaTeX formatting, including tables and references. All AI-assisted edits were reviewed by the authors. The authors take responsibility for the final content of this work, including all text, claims, and artifacts produced with the aid of generative AI.

Reproducibility statement

Section 3 specifies the model architecture and learning objectives, while Section 4 reports the datasets, training protocol, evaluation metrics, and controlled ablations used in this study. The accompanying code and configurations provide the implementation details needed to reproduce the experiments.

REFERENCES

Anurag Arnab, Mostafa Dehghani, Georg Heigold, Chen Sun, Mario Lučić, and Cordelia Schmid. ViViT: A video vision transformer. In *IEEE/CVF International Conference on Computer Vision*, 2021.

Mahmoud Assran, Quentin Duval, Ishan Misra, Piotr Bojanowski, Pascal Vincent, Michael Rabbat, Yann LeCun, and Nicolas Ballas. Self-supervised learning from images with a joint-embedding predictive architecture. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023.

Mahmoud Assran, Adrien Bardes, David Fan, Quentin Garrido, Russell Howes, Mojtaba Komeili, Matthew Muckley, Ammar Rizvi, Claire Roberts, Koustuv Sinha, Artem Zholus, Sergio Arnaud, Abha Gejji, Ada Martin, Francois Robert Hogan, Daniel Dugas, Piotr Bojanowski, Patrick Labatut, Francisco Massa, Marc Szafraniec, Kapil Krishnakumar, Yong Li, Xiaodong Ma, Sarath Chandar, Franziska Meier, Yann LeCun, and Michael Rabbat. V-JEPA 2: Self-supervised video models enable understanding, prediction and planning. *arXiv preprint arXiv:2506.09985*, 2025.

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

Gedas Bertasius, Heng Wang, and Lorenzo Torresani. Is space-time attention all you need for video understanding? In *International Conference on Machine Learning*, 2021.

Holger Caesar, Varun Bankiti, Alex H. Lang, Sourabh Vora, Venice Erin Liong, Qiang Xu, Anush Krishnan, Yu Pan, Giancarlo Baldan, and Oscar Beijbom. nuScenes: A multimodal dataset for autonomous driving. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020.

Daniel Dauner, Marcel Hallgarten, Tianyu Li, Xinshuo Weng, Zhiyu Huang, Zetong Yang, Hongyang Li, Igor Gilitschenski, Boris Ivanovic, Marco Pavone, Andreas Geiger, and Kashyap Chitta. NAVSIM: Data-driven non-reactive autonomous vehicle simulation and benchmarking. In *Advances in Neural Information Processing Systems, Datasets and Benchmarks Track*, 2024.

Deepti Hegde, Rajeev Yasarla, Hong Cai, Shizhong Han, Apratim Bhattacharyya, Shweta Mahajan, Litian Liu, Risheek Garrepalli, Vishal M. Patel, and Fatih Porikli. Distilling multi-modal large language models for autonomous driving. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2025.

Shengchao Hu, Li Chen, Penghao Wu, Hongyang Li, Junchi Yan, and Dacheng Tao. ST-P3: End-to-end vision-based autonomous driving via spatial-temporal feature learning. In *European Conference on Computer Vision*, 2022.

Yihan Hu, Jiazhi Yang, Li Chen, Keyu Li, Chonghao Sima, Xizhou Zhu, Senyao Chai, Senyao Du, Tianwei Lin, Lewei Lu, Xiaosong Jia, Qiang Liu, Jifeng Dai, Yu Qiao, and Hongyang Li. Planning-oriented autonomous driving. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023.

Bo Jiang, Shaoyu Chen, Qing Xu, Bencheng Liao, Jiajie Chen, Helong Zhou, Qian Zhang, Wenyu Liu, Chang Huang, and Xinggang Wang. VAD: Vectorized scene representation for efficient autonomous driving. In *IEEE/CVF International Conference on Computer Vision*, 2023.

Yingyan Li, Lue Fan, Jiawei He, Yuqi Wang, Yuntao Chen, Zhaoxiang Zhang, and Tieniu Tan. Enhancing end-to-end autonomous driving with latent world model. In *International Conference on Learning Representations*, 2025.

Zhiqi Li, Wenhai Wang, Hongyang Li, Enze Xie, Chonghao Sima, Tong Lu, Yu Qiao, and Jifeng Dai. BEVFormer: Learning bird's-eye-view representation from multi-camera images via spatiotemporal transformers. In *European Conference on Computer Vision*, 2022.

Zhiqi Li, Zhiding Yu, Shiyi Lan, Jiahan Li, Jan Kautz, Tong Lu, and Jose M. Alvarez. Is ego status all you need for open-loop end-to-end autonomous driving? *arXiv preprint arXiv:2312.03031*, 2024.

Bencheng Liao, Shaoyu Chen, Haoran Yin, Bo Jiang, Cheng Wang, Sixu Yan, Xinbang Zhang, Xiangyu Li, Ying Zhang, Qian Zhang, and Xinggang Wang. DiffusionDrive: Truncated diffusion model for end-to-end autonomous driving. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2025.

Lucas Maes, Quentin Le Lidec, Damien Scieur, Yann LeCun, and Randall Balestriero. LeWorld-Model: Stable end-to-end joint-embedding predictive architecture from pixels. *arXiv preprint arXiv:2603.19312*, 2026.

Wenchao Sun, Xuewu Lin, Yining Shi, Chuang Zhang, Haoran Wu, and Sifa Zheng. SparseDrive: End-to-end autonomous driving via sparse scene representation. In *IEEE International Conference on Robotics and Automation*, 2025.

Wenzhao Zheng, Ruiqi Song, Xianda Guo, Chenming Zhang, and Long Chen. GenAD: Generative end-to-end autonomous driving. In *European Conference on Computer Vision*, 2024.

Yupeng Zheng, Pengxuan Yang, Zebin Xing, Qichao Zhang, Yuhang Zheng, Yinfeng Gao, Pengfei Li, Teng Zhang, Zhongpu Xia, Peng Jia, Xianpeng Lang, and Dongbin Zhao. World4Drive: End-to-end autonomous driving via intention-aware physical latent world model. In *IEEE/CVF International Conference on Computer Vision*, 2025.