FROM WORLD STATES TO VIDEOS: V-JEPA FOR DRIVING SCENE GENERATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Predicting a driving video is not only a problem of extending pixels: the generator must decide how traffic, road geometry, and camera motion evolve after the last observed frame. We ask whether a learned world state can make that evolution explicit. A compact Transformer predicts a future-oriented V-JEPA state from observed video, and a latent-video generator uses the state to synthesize the future. Neither future images nor motion metadata are available at inference. Across 130 unseen nuScenes scenes, predicted states improve a fixed semantic–geometric–motion measure by 6.88% over context-only generation in every seed, retaining more than 95% of the benefit of a teacher state extracted from the logged future. A matched predictive pathway built on VideoMAE uses the same predictor and generator budgets, yet predicted V-JEPA states perform 4.24% better. The gain spans all evaluator components and aggregation rules. Motion readouts explain why this is plausible: V-JEPA exposes temporal dynamics more clearly than VideoMAE on two scene cohorts. Oracle-conditioning experiments then show that the generator uses this information and degrades when state and scene no longer correspond. Together, these results support a context-to-state-to-video path in which predicted world states guide future-video synthesis.

1 INTRODUCTION

At the last observed frame of a driving clip, appearance is known but evolution is not. The road must remain geometrically stable while the camera and other traffic continue to move. A pixel-conditioned generator has to recover all of this structure indirectly from its visual context. We explore a more explicit route: first predict a compact state of how the scene will evolve, then let the generator render the future from that state.

V-JEPA offers a natural state space because it learns by predicting video in representation space instead of reconstructing every pixel Bardes et al. (2024); Assran et al. (2025). Yet a representation extracted from future frames is only an oracle; it says little about a system that must act before those frames exist. Our model therefore uses the logged future only to define a training target. At inference, a lightweight predictor maps four observed frames to a V-JEPA state, which is passed to a video generator trained with teacher states and then held fixed.

We follow this hypothesis through three linked empirical questions. First, does V-JEPA make scene motion accessible, and does that signal disappear when temporal order is broken? Second, can a generator turn the state into a better future, with quality falling when the state comes from another scene? Third, does the benefit survive when the oracle state is replaced by one predicted from context? We finally ask whether that benefit is specific to the representation by repeating the predictive path with an equally trained VideoMAE state. This sequence connects representation structure to generative behavior instead of treating either result in isolation.

All three links are observed. V-JEPA yields a clear motion-readout advantage over VideoMAE on both held-out and independent-validation scenes. Its oracle state improves generation over context-only and VideoMAE conditioning, while a wrong-scene state hurts the same generator. Most importantly, the context-predicted state preserves nearly all of the oracle gain without generator retraining. It also outperforms the matched predicted-VideoMAE pathway in every seed. The advantage spans evaluator components and aggregation rules. The result is a complete context-to-state-to-video path:

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.

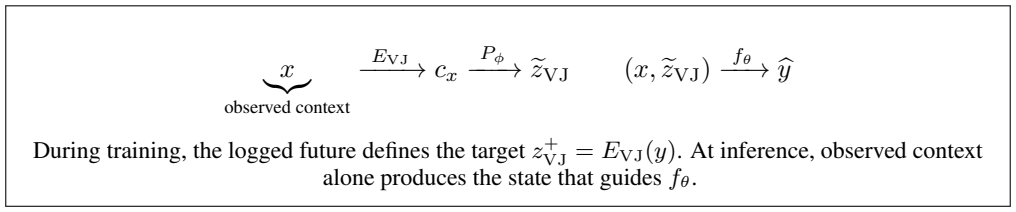


Figure 1: From world states to videos. The logged future supplies a state target during training; at inference, a predictor infers that state from observed frames and uses it to guide the video generator.

a predictive representation becomes useful guidance for future-video synthesis before the future is available.

2 Related Work

Predictive video representations. I-JEPA introduced masked prediction in learned image-representation space, and V-JEPA extended this principle to video Assran et al. (2023); Bardes et al. (2024). V-JEPA 2 scales action-free video pretraining; its separately post-trained V-JEPA 2-AC variant adds action-conditioned latent prediction for planning, whereas V-JEPA 2.1 emphasizes dense spatiotemporal features Assran et al. (2025); Mur-Labadia et al. (2026). VideoMAE V2 is the matched masked-video-modeling comparator in our experiments Wang et al. (2023); InternVideo2 represents broader multimodal video foundation modeling Wang et al. (2024). Earlier work tied visual representations to ego-motion Jayaraman & Grauman (2015); here logged motion is used only as a synchronized readout target. DINOv2 and the Stable Diffusion autoencoder provide frame-wise and reconstruction-latent controls Oquab et al. (2024); Rombach et al. (2022).

Representation-to-generation interfaces. Recent methods connect video representations to synthesis through distinct interfaces. VideoRAE turns frozen video-foundation features into continuous or discrete generation latents Xie et al. (2026); VideoREPA distills token relations from video foundation models during text-to-video fine-tuning Zhang et al. (2025); WMReward uses V-JEPA 2 as an inference-time physics reward Yuan et al. (2026); and PHANTOM jointly models latent physical dynamics and future frames Shen et al. (2026). A matched-capacity study of V-JEPA 2/2.1 and VideoMAE V2 further reports differences in robustness and temporal sensitivity, without testing a generator Alrasheed et al. (2026). These works establish several representation-to-video bridges. We ask a complementary question: can a future-oriented representation be inferred from driving context and still improve a generator trained with oracle states? A shared dense interface lets us trace this path from V-JEPA and VideoMAE representations, through oracle conditioning, and through matched context-to-state prediction for both representation families.

Driving representations and world models. Driving work combines pretrained perception, latent dynamics, controllable generation, and planning. Drive-JEPA adapts JEPA-style pretraining to end-to-end driving; DriveWorld-VLA integrates latent scene evolution with a driving VLA; WorldDrive and DrivingWorld study controllable driving-video generation Wang et al. (2026); Jia et al. (2026); Gui et al. (2026); Hu et al. (2027). These systems motivate our driving-domain evaluation; our focus is the representation-to-generation path: whether a learned summary of driving dynamics remains useful when it must be predicted before the target video.

Evaluator-relative generation. The generation study reuses frozen SegFormer, Depth Anything V2, and RAFT interfaces as semantic, geometry, and motion instruments Xie et al. (2021); Yang et al. (2024); Teed & Deng (2020). Their outputs are pseudo-targets rather than ground truth, and these model citations do not validate our task-specific geometric-mean aggregate as a general-purpose endpoint. We therefore interpret the aggregate only as a paired, evaluator-relative comparison under fixed teachers.

3 World-State Prediction and Evaluation

Let x denote the observed context, y the logged future, and p_h the future ego translation relative to the final context pose. A frozen encoder E_m maps y to a common latent state z_m. We call z_{VJ} a JEPA-derived world state when it exposes synchronized physical motion through a simple readout and improves a matched generator. We mean-pool the state and fit an affine multi-output ridge readout

$$\hat{p}_{1,2,4} = W_m \, \mathrm{mean}(z_m) + b_m \tag{1}$$

on training scenes only. Its error is multihorizon trajectory ADE, aggregated first within each source scene.

The primary representation claim compares 224-equivalent V-JEPA 2.1 Base with the matched VideoMAE V2 Base K710 comparator. The held-out test gain is

$$G_{\mathrm{VJ/VM}} = 1 - \frac{\mathrm{mean}_s \, \mathrm{ADE}_{s,\mathrm{VJ}}}{\mathrm{mean}_s \, \mathrm{ADE}_{s,\mathrm{VM}}}. \tag{2}$$

An independent analysis repeats the exact effect on official-validation scenes with the same train-fitted readouts. Held-out test and independent validation are reported separately and never pooled.

Temporal controls apply the ordered readout without refitting to reversed, shuffled, and repeat-first views. They test dependence on temporal evidence. For generation, a matched decoder produces

$$\hat{y}_m = f_{\theta_m}(x, z_m), \tag{3}$$

where z_m is either zero, the oracle V-JEPA world state, or the matched oracle VideoMAE state. The state-usage control replaces z_{VJ} at test time with a deterministically selected V-JEPA state from another source scene. The context, architecture, training budget, and evaluation noise remain matched.

The predictive experiment removes the future input. A frozen V-JEPA encoder maps the context to c_x, and a predictor P_ϕ estimates

$$\widetilde{z}_{\mathrm{VJ}} = P_\phi(c_x), \qquad \widehat{y} = f_\theta(x, \widetilde{z}_{\mathrm{VJ}}). \tag{4}$$

The target $z^+_{\mathrm{VJ}} = E_{\mathrm{VJ}}(y)$ exists only during predictor training. At inference, neither future RGB nor motion metadata is available. The predicted state is passed to the frozen oracle-state generator without fine-tuning, directly testing whether the learned interface survives context-only prediction. We repeat the complete path with VideoMAE: its context features predict its future state, which conditions the seed-matched generator trained with oracle VideoMAE states. Predictor architecture, update budget, checkpoint rule, generator capacity, and sampling noise are identical across representations.

4 Experimental Setup

We evaluate CAM_FRONT clips from nuScenes Caesar et al. (2020). Each clip contains four context frames followed by sixteen future frames at 4 Hz. Source scenes are disjoint across 569 training scenes, a 130-scene held-out test set, and 149 official-validation scenes. The held-out scenes are excluded from every study-specific parameter update and selection decision; the independent-validation cohort is used only for the representation readout.

All effects are paired at the source-scene level. Clips are averaged within a scene before a paired bootstrap, and stochastic effects are summarized across matched seeds with common scene resamples. Frames, tokens, horizons, clips, and generated samples are never treated as independent observations. Detailed cohort construction and optimization settings are provided in the appendix. Evaluator robustness decomposes semantic, depth, and flow terms and recomputes both contrasts with arithmetic, harmonic, leave-one-component-out geometric, and worst-component reductions, using the same outputs, seed matching, and source-scene bootstrap.

5 A Shared Latent Interface

V-JEPA 2.1 Base and official VideoMAE V2 Base K710 see the same source images and resolution-limited information. We pool their dense outputs to a common spatiotemporal grid before using them

Table 1: Multihorizon translation accessibility. Absolute values are scene-mean ADE in meters with paired 95% intervals. The held-out test set is excluded from training; independent validation reuses both readouts without refitting.

Cohort	V-JEPA ADE	VideoMAE ADE	V-JEPA gain
Held-out test	2.071 [1.919, 2.232]	2.942 [2.727, 3.160]	29.61% [22.92, 35.75]
Independent validation	2.359 [2.181, 2.543]	3.148 [2.917, 3.381]	25.07% [17.15, 32.33]

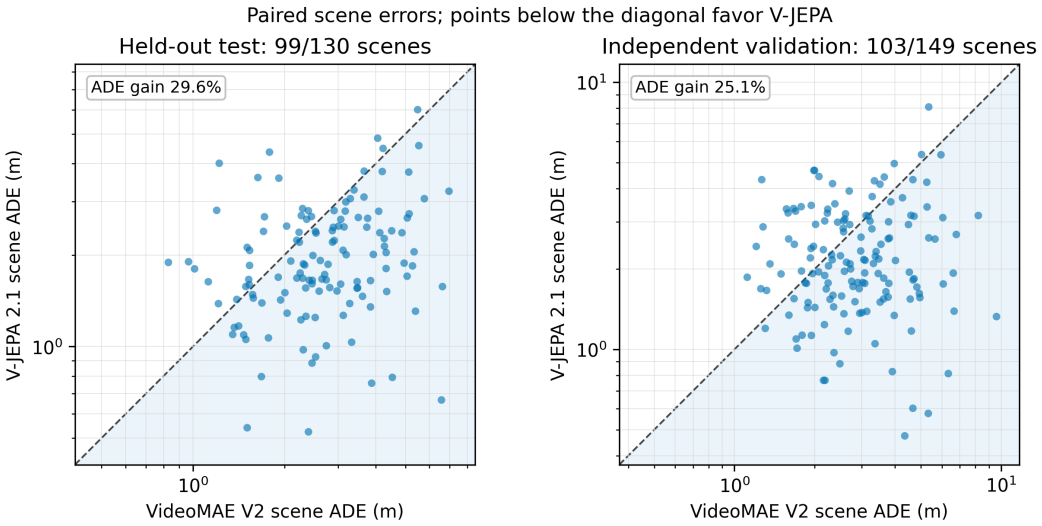


Figure 2: Each point is one source scene; points below the diagonal favor V-JEPA. Most scenes favor V-JEPA in both cohorts, which are shown separately and never pooled.

in either the readout or generator. DINOv2 and a Stable Diffusion VAE provide frame-wise and reconstruction controls for the representation analysis.

For each ordered condition, tokens are mean-pooled and standardized with training-only moments. A single fixed ridge readout predicts logged ego translation; it is reused without refitting for temporal controls and independent validation.

Temporal shuffling, reversal, and repetition alter the input order while leaving the readout fixed, so any degradation reflects loss of the learned feature–target relationship rather than a newly fitted probe.

Context-to-state predictor. The observed frames pass through the same representation interface used to define the oracle target. A compact block-causal Transformer maps context tokens and learned future queries to a predicted state, trained by standardized state regression. Checkpoint selection uses training loss only. Each predictor is paired with the corresponding frozen generator and sampling noise, so the held-out comparison changes only the supplied state. V-JEPA and VideoMAE use separate predictors with identical architectures and optimization budgets; each receives only its own context features and learns only from its own future-state targets.

6 V-JEPA Makes Logged Motion More Accessible

Motion is easier to recover from V-JEPA, twice. The first link in the empirical story asks whether a compact V-JEPA state makes physical change legible. Table 1 answers yes: a simple linear readout recovers logged ego translation more accurately than the same readout on VideoMAE. The advantage appears on the held-out scenes and then reappears on independent official-validation scenes using the original train-fitted readouts. Both intervals exclude zero, so the pattern is not tied to a single scene split.

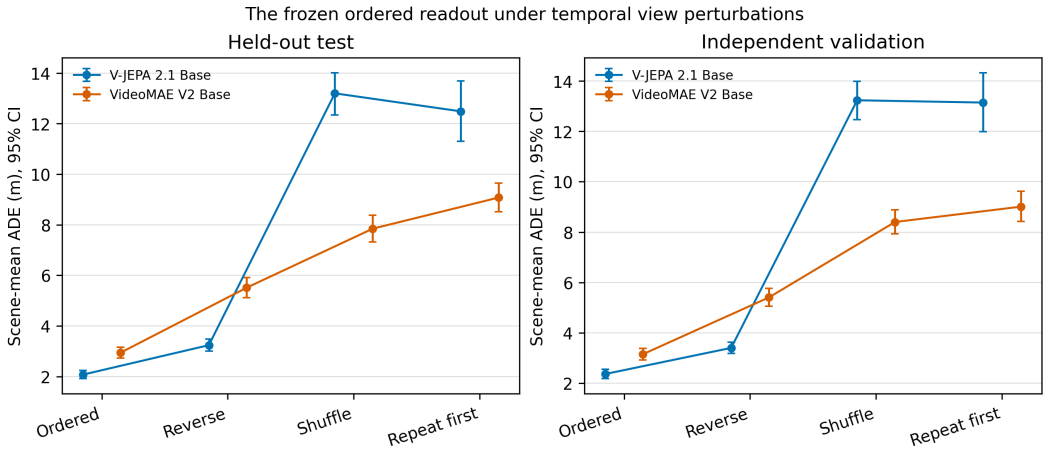


Figure 3: Breaking temporal order weakens the motion signal. The original train-fitted readout is applied without refitting to ordered, reversed, shuffled, and repeat-first clips; error bars are paired source-scene 95% intervals.

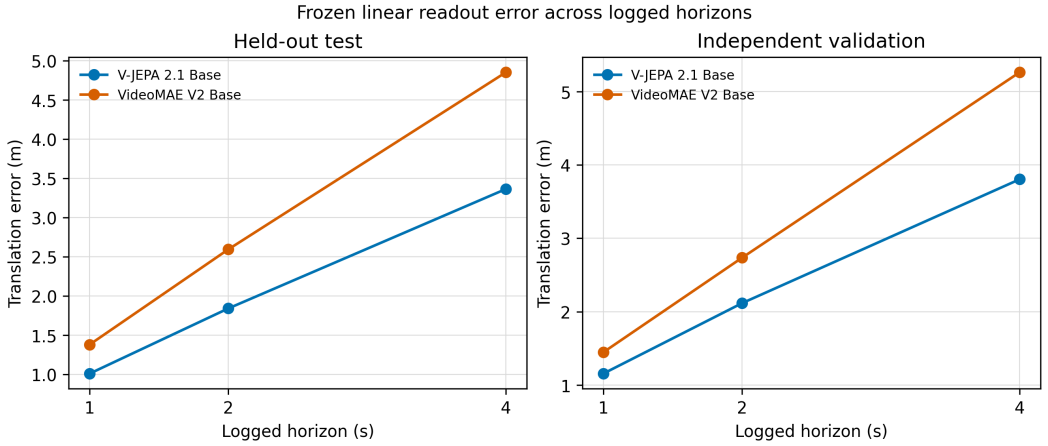


Figure 4: V-JEPA's motion-readout advantage persists from one to four seconds. Values are scene-mean translation errors for the ordered readouts; held-out test and independent validation remain separate.

Lower mean error could still be produced by a few unusually easy clips. Figure 2 shows otherwise: most scenes lie below the diagonal in both cohorts. The state carries a broadly accessible motion signal, although the linear readout itself is only a diagnostic of that signal.

The signal is not static appearance in disguise. When we reverse, shuffle, or repeat frames, the unchanged V-JEPA readout loses much of its accuracy in both cohorts. The disruption is stronger than for VideoMAE, linking the accessibility advantage to temporal organization rather than scene appearance alone. These frame-order changes are diagnostics of the representation, not physical interventions on the driving scene.

The gap also survives as the prediction horizon grows (Figure 4). V-JEPA remains ahead from one to four seconds, when accumulated motion makes the readout progressively harder. The horizons are repeated measurements of the same scenes, not independent replications.

Table 2: The correct V-JEPA state improves generation, beats VideoMAE conditioning, and loses quality when replaced by a state from another scene. Values are median seed effects with paired source-scene 95% intervals; every seed has the favorable sign.

Contrast	Result
V-JEPA state vs context	7.23% [4.64, 9.62]
Scene-swapped state degradation	2.47% [0.29, 5.16]
V-JEPA state vs VideoMAE	4.08% [2.47, 5.38]

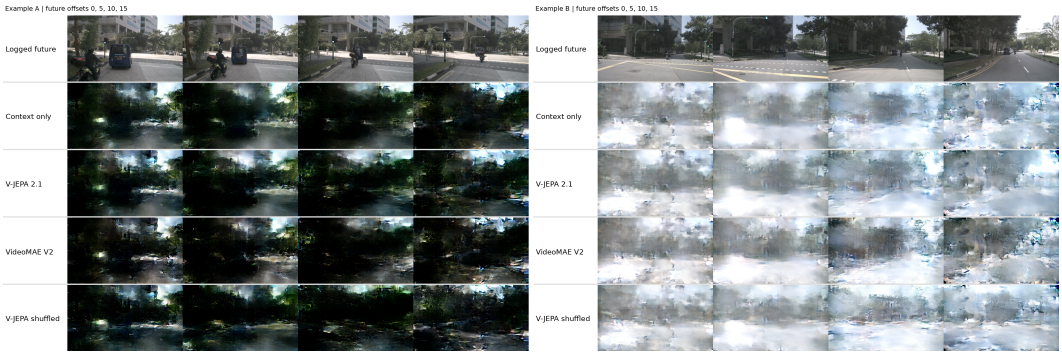


Figure 5: Two driving futures make the conditioning problem concrete. Each column moves deeper into the future; all generated rows share the same observed context but receive context-only, V-JEPA, VideoMAE, or wrong-scene V-JEPA conditioning. Examples were selected by a fixed rule before inspection; quantitative conclusions use every test scene and seed.

7 Future-Derived V-JEPA States Improve Generation

Motion accessibility alone does not mean that a generator can use the state. We therefore turn the representation into an intervention: the observed frames and generator remain the same while the supplied state changes among zero, VideoMAE, and V-JEPA. A state from another scene asks a sharper question: does the generator use scene-specific information or merely respond to any extra conditioning vector? No test clip is used for generator optimization.

Stochastic world-state-conditioned generator. Matched latent-video diffusion models generate the future jointly from the observed context. A LayerNorm projection Ba et al. (2016) injects either no state, the oracle V-JEPA state, or the oracle VideoMAE state. Conditions share architecture, optimization, seed-wise initialization, and sampling noise and use AdamW and deterministic DDIM sampling Loshchilov & Hutter (2019); Song et al. (2021).

The evaluation error vector is one minus SegFormer pseudo-mIoU, median-aligned Depth Anything V2 AbsRel, and RAFT flow EPE normalized by the image diagonal. The aggregate effect is the geometric mean of component error ratios. RGB L1, PSNR, and latent MSE are diagnostics only.

The three rows of Table 2 tell one story. A correct V-JEPA state improves the future over context alone; it also improves over a VideoMAE state with the same downstream capacity. Finally, moving the V-JEPA state to the wrong scene reverses part of that benefit. All three seeds agree, and every interval has the favorable sign. This combination is difficult to explain as an unconditional decoder advantage: generation improves most when the state both comes from V-JEPA and belongs to the scene being rendered.

8 Context-Predicted States Preserve the Generation Gain

The oracle experiment shows that the generator can use a world state, but the state still comes from the future it is meant to generate. The final experiment closes this gap. We replace the oracle with

Table 3: The context-predicted V-JEPA state closes most of the gap to oracle conditioning on the held-out test set. It also outperforms a matched predicted-VideoMAE pathway. Effects are relative reductions in the fixed semantic–geometry–motion error aggregate with paired source-scene bootstrap uncertainty.

Conditioning contrast	Result
Predicted V-JEPA vs context	6.88% [4.32%, 9.10%]
Predicted V-JEPA vs predicted VideoMAE	4.24% [2.44%, 6.23%]
Predicted/oracle gain retention	95.14%

$\widetilde{z}_{\mathrm{VJ}}$, predicted from four observed frames, and pass it directly to the same generator. Future RGB, poses, and labels are unavailable at inference, and the generator is not retrained for predictor outputs.

Prediction preserves the generative value of the state. Relative to context alone, the predicted state improves the fixed evaluator-relative aggregate by 6.88%; every seed improves and the scene-level interval lies above zero. More strikingly, it retains 95.14% of the oracle gain even though the generator has never been adapted to predictor outputs.

The representation choice remains consequential after prediction. Under matched predictor architecture, training budget, generator capacity, and sampling noise, predicted V-JEPA improves over predicted VideoMAE by 4.24% [2.44%, 6.23%]; all three seed effects are positive. Thus the generation gain is not explained merely by adding a learned state or predictor capacity: in this setting, the V-JEPA state is the more useful predictive interface.

The advantage spans the evaluator rather than one favored term. Predicted V-JEPA reduces semantic, depth, and flow errors against both comparators, with favorable signs in every seed and paired source-scene intervals excluding zero (Table 5). Both contrasts also survive arithmetic and harmonic means, every leave-one-component-out geometric mean, and worst-component reduction (Table 6). The gain therefore depends on neither one dominant term nor one aggregation rule.

This closes the empirical loop. The oracle result showed that V-JEPA contains information a generator can use; the context predictor recovers nearly all of that benefit before the target video exists. Additional future pixels or motion metadata cannot explain the gain because neither is available to P_ϕ at inference.

9 Discussion

The three experiments rule out different failure modes. V-JEPA could have been easy to probe but irrelevant to generation; the oracle state could have helped only because it came directly from future frames; or prediction error could have moved the state outside the region the generator understands. Instead, motion is linearly accessible, correct oracle states improve generation, and a context predictor preserves almost all of that gain. The evidence therefore connects representation structure to an observable change in synthesized futures. The matched predicted-VideoMAE comparison further rules out a generic benefit from any equally trained predictive state: the V-JEPA pathway remains ahead after future-derived inputs are removed from both systems.

One plausible reason for the strong oracle-to-predicted transfer is that the state target compresses joint scene evolution without requiring the predictor to reconstruct every future pixel. Errors in texture or other pixel-level details need not destroy the part of the state that guides geometry and motion. Because the generator is unchanged when predicted states are introduced, the result suggests that the predictor lands in a region of representation space that the generator already knows how to interpret.

The controls sharpen this interpretation. Breaking frame order disrupts the motion readout, VideoMAE supplies a strong alternative at both the oracle and predicted interfaces, and the wrong-scene oracle state shows that correspondence matters. Together they make the generator gain more naturally attributable to temporally organized scene information than to generic extra conditioning.

10 Limitations and Outlook

Generation is evaluated on one held-out test split; independent validation covers only representation accessibility. The matched VideoMAE comparison is confined to one dataset, predictor design, and generator family, not universal encoder superiority. Replacing the predicted state with one from another scene changes aggregate error by 1.68% [-0.74%, 4.52%]; the interval includes zero, so scene-specific use of predicted states remains open. The predictor learns from future-derived targets and the endpoint inherits frozen-teacher biases. Broader data, cameras, and generators remain needed. The component and aggregation analyses reuse these generations and test only the paired evaluator-relative comparison, not external teacher validity.

11 Conclusion

A video generator need not infer scene dynamics from pixels alone. On nuScenes, a V-JEPA state predicted from observed frames preserves nearly all of a future-derived teacher's gain and outperforms matched predicted VideoMAE. Motion readouts, representation comparisons, and scene-state swaps trace this benefit to temporally organized information. The resulting context-to-state-to-video path provides concrete guidance for future-video synthesis across all evaluator components and aggregation rules.

Reproducibility Statement

The released protocols and supplement specify cohort construction, model interfaces, optimization, evaluation, uncertainty, and evidence provenance. Tables and figures are exported from verified scene-level aggregates.

Ethics Statement

The study uses front-camera data depicting public-road activity under the non-commercial nuScenes terms. The source dataset applies automated face and license-plate blurring, but the images remain subject to the original access and redistribution conditions. The evaluation is observational and must not be used as evidence that a vehicle-control or deployment system is safe.

AI Use Statement

Generative AI tools assisted with language editing. The authors verified the experiments, claims, citations, and text and take responsibility for the final content.

References

Ali J. Alrasheed, Aryan Yazdan Parast, Basim Azam, James Bailey, and Naveed Akhtar. Latent video prediction learns better world models. *arXiv preprint arXiv:2605.15618*, 2026.

Mahmoud Assran, Quentin Duval, Ishan Misra, Piotr Bojanowski, Pascal Vincent, Michael Rabbat, Yann LeCun, and Nicolas Ballas. Self-supervised learning from images with a joint-embedding predictive architecture. In *CVPR*, pp. 15619–15629, 2023.

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

Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E. Hinton. Layer normalization. *arXiv preprint arXiv:1607.06450*, 2016.

Adrien Bardes, Quentin Garrido, Jean Ponce, Xinlei Chen, Michael Rabbat, Yann LeCun, Mido Assran, and Nicolas Ballas. Revisiting feature prediction for learning visual representations from video. *Transactions on Machine Learning Research*, 2024. URL https://openreview.net/forum?id=QaCCuDfBk2.

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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11621–11631, June 2020.

Xingtai Gui, Meijie Zhang, Tianyi Yan, Wencheng Han, Jiahao Gong, Feiyang Tan, Cheng-zhong Xu, and Jianbing Shen. Bridging scene generation and planning: Driving with world model via unifying vision and motion representation. *arXiv preprint arXiv:2603.14948*, 2026.

Xiaotao Hu, Mingkai Jia, Xiaoyang Guo, Qian Zhang, Xiao-xiao Long, and Wei Yin. DrivingWorld: Constructing world model for autonomous driving via video gpt. In *Pattern Recognition – ICPR 2026*, volume 16813 of *Lecture Notes in Computer Science*, pp. 276–291. Springer, 2027. doi: 10.1007/978-3-032-31583-0_19.

Dinesh Jayaraman and Kristen Grauman. Learning image representations tied to ego-motion. In *ICCV*, pp. 1413–1421, 2015.

Feiyang Jia, Lin Liu, Ziying Song, Caiyan Jia, Hangjun Ye, Xiaoshuai Hao, and Long Chen. DriveWorld-VLA: Unified latent-space world modeling with vision-language-action for autonomous driving. *arXiv preprint arXiv:2602.06521*, 2026.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=Bkg6RiCqY7.

Lorenzo Mur-Labadia, Matthew Muckley, Amir Bar, Mido Assran, Koustuv Sinha, Mike Rabbat, Yann LeCun, Nicolas Ballas, and Adrien Bardes. V-JEPA 2.1: Unlocking dense features in video self-supervised learning. *arXiv preprint arXiv:2603.14482*, 2026.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy V. Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Mido Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Hervé Jégou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. DINOv2: Learning robust visual features without supervision. *Transactions on Machine Learning Research*, 2024. URL https://openreview.net/forum?id=a68SUt6zFt.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10684–10695, June 2022.

Ying Shen, Jerry Xiong, Tianjiao Yu, and Ismini Lourentzou. PHANTOM: Physics-infused video generation via joint modeling of visual and latent physical dynamics. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11185–11194, June 2026.

Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=St1giarCHLP.

Zachary Teed and Jia Deng. RAFT: Recurrent all-pairs field transforms for optical flow. In *Computer Vision – ECCV 2020*, volume 12347 of *Lecture Notes in Computer Science*, pp. 402–419. Springer, 2020. doi: 10.1007/978-3-030-58536-5_24.

Limin Wang, Bingkun Huang, Zhiyu Zhao, Zhan Tong, Yinan He, Yi Wang, Yali Wang, and Yu Qiao. VideoMAE V2: Scaling video masked autoencoders with dual masking. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14549–14560, June 2023.

Linhan Wang, Zichong Yang, Chen Bai, Guoxiang Zhang, Xiaotong Liu, Xiaoyin Zheng, Xiao-Xiao Long, Chang-Tien Lu, and Cheng Lu. Drive-JEPA: Video jepa meets multimodal trajectory distillation for end-to-end driving. *arXiv preprint arXiv:2601.22032*, 2026.

Yi Wang, Kunchang Li, Xinhao Li, Jiashuo Yu, Yinan He, Guo Chen, Baoqi Pei, Rongkun Zheng, Zun Wang, Yansong Shi, Tianxiang Jiang, Songze Li, Jilan Xu, Hongjie Zhang, Yifei Huang, Yu Qiao, Yali Wang, and Limin Wang. InternVideo2: Scaling foundation models for multimodal video understanding. In *Computer Vision – ECCV 2024*, volume 15143 of *Lecture Notes in Computer Science*, pp. 396–416. Springer, 2024. doi: 10.1007/978-3-031-73013-9_23.

Enze Xie, Wenhai Wang, Zhiding Yu, Anima Anandkumar, Jose M. Alvarez, and Ping Luo. SegFormer: Simple and efficient design for semantic segmentation with transformers. In *Advances in Neural Information Processing Systems*, volume 34, pp. 12077–12090, 2021.

Zhihao Xie, Junfeng Wu, Xinting Hu, Junchao Huang, and Li Jiang. VideoRAE: Taming video foundation models for generative modeling via representation autoencoders. *arXiv preprint arXiv:2607.14088*, 2026.

Lihe Yang, Bingyi Kang, Zilong Huang, Zhen Zhao, Xiaogang Xu, Jiashi Feng, and Hengshuang Zhao. Depth anything v2. In *Advances in Neural Information Processing Systems*, volume 37, pp. 21875–21911, 2024. doi: 10.52202/079017-0688.

Jianhao Yuan, Xiaofeng Zhang, Felix Friedrich, Nicolas Beltran-Velez, Melissa Hall, Reyhane Askari-Hemmat, Xiaochuang Han, Nicolas Ballas, Michal Drozdzal, and Adriana Romero-Soriano. Inference-time physics alignment of video generative models with latent world models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 16118–16129, June 2026.

Xiangdong Zhang, Jiaqi Liao, Shaofeng Zhang, Fanqing Meng, Xiangpeng Wan, Junchi Yan, and Yu Cheng. VideoREPA: Learning physics for video generation through relational alignment with foundation models. In *Advances in Neural Information Processing Systems*, volume 38, pp. 135991–136020, 2025. doi: 10.52202/085713-4091.

A Data and cohort details

Training uses 1,474 clips from 569 source scenes; the held-out test set contains 342 clips from 130 disjoint scenes, and independent validation contains 382 clips from 149 official-validation scenes. Clips within a scene are frame-disjoint. Images were obtained through a pinned transport mirror of nuScenes v1.0 trainval and checked against official scene and timestamp metadata. The original licensing and redistribution terms remain in force.

B Representation and readout interface

V-JEPA 2.1 Base uses its public teacher-distilled video model, final transformer block, and a common $4 \times 8 \times 8$ grid of 768-dimensional tokens. Official VideoMAE V2 Base K710 uses the same 224-pixel source-information limit, its learned token normalization, and the same output grid. Both encoders and the training-fitted ridge readouts are frozen before held-out evaluation. Exact public model revisions and machine-readable evidence provenance are supplied outside the manuscript in the anonymous release metadata.

C Absolute and secondary representation endpoints

Table 4: Absolute ADE and four-second FDE in meters with 95% intervals.

Cohort	Representation	ADE	FDE at 4 s
Held-out test	V-JEPA 224-equivalent	2.071 [1.919, 2.232]	3.364 [3.105, 3.640]
Held-out test	VideoMAE V2 Base	2.942 [2.727, 3.160]	4.854 [4.468, 5.238]
Held-out test	DINOv2 Base	5.341 [4.915, 5.799]	9.194 [8.446, 10.012]
Held-out test	SD-VAE	6.869 [6.332, 7.431]	11.865 [10.940, 12.831]
Held-out test	V-JEPA native 384	2.120 [1.976, 2.280]	3.454 [3.205, 3.737]
Independent validation	V-JEPA 224-equivalent	2.359 [2.181, 2.543]	3.804 [3.508, 4.116]
Independent validation	VideoMAE V2 Base	3.148 [2.917, 3.381]	5.261 [4.900, 5.631]

World-state-conditioned generator optimization. The stochastic generator has 15.66M parameters and is trained with AdamW for 8,000 updates. The final exponential-moving-average checkpoint is the only eligible model, and evaluation uses deterministic DDIM sampling. Matched conditions share their architecture, optimization schedule, initialization, and sampling noise; complete hyperparameters are provided in the released configuration.

Context-to-state predictor optimization. Each representation uses a separate but identically configured four-layer block-causal Transformer with learned future queries and a residual output over the final context chunk. Each predictor is trained with AdamW and standardized future-state MSE for 6,000 updates. Training-derived moments standardize both context inputs and future targets, and the checkpoint is selected by training loss without loading held-out outputs. Complete architecture and optimizer settings are provided in the released configuration.

Statistical aggregation. Aggregation first averages clips within source scene, then resamples scenes with all compared conditions paired. Held-out test and independent validation remain separate. Stochastic comparisons preserve scene, seed, and sampling pairing before summarizing effects across seeds.

D Component and aggregation robustness

The robustness analysis reuses the held-out generated clips and changes only how the fixed evaluator outputs are summarized. Effects are relative error reductions; larger values favor predicted V-JEPA.

Table 5: Each evaluator component independently favors predicted V-JEPA. Values are median seed effects with paired source-scene 95% intervals; all three seed effects are positive in every row.

Component	vs context-only	vs predicted VideoMAE
Semantic error (1−pseudo-mIoU)	1.24% [0.98%, 1.56%]	0.85% [0.71%, 1.04%]
Median-aligned depth AbsRel	14.31% [7.87%, 20.46%]	8.92% [3.61%, 14.34%]
Normalized RAFT flow EPE	3.98% [3.51%, 4.53%]	2.61% [2.31%, 3.25%]

Table 6: Both generation comparisons survive every alternative reduction. Values are relative error reductions with paired source-scene 95% intervals; all three seed effects are positive in every row. The original geometric mean is reported in Table 3.

Reduction	vs context-only	vs predicted VideoMAE
Arithmetic mean	6.71% [4.34%, 8.70%]	4.18% [2.41%, 6.03%]
Harmonic mean	7.05% [4.38%, 9.54%]	4.31% [2.44%, 6.46%]
Geometric mean without semantic	9.70% [5.84%, 12.79%]	5.82% [3.21%, 8.71%]
Geometric mean without depth	2.79% [2.46%, 3.10%]	1.82% [1.64%, 2.08%]
Geometric mean without flow	7.87% [4.55%, 11.34%]	5.05% [2.23%, 7.87%]
Worst component	1.24% [0.96%, 1.55%]	0.85% [0.68%, 1.04%]