# LA-CONTROL: A CONTROL-ORIENTED LATENT ACTION MODEL FOR AUTONOMOUS DRIVING

**Anonymous authors**
Paper under double-blind review

## ABSTRACT

Latent Action Models (LAMs) learn representations of visual transitions from video, enabling imitation learning from videos without explicit action labels. For autonomous driving, however, representations learned through visual prediction may not retain the underlying control information. Visually salient states and scene changes can dominate the learned representation, while subtle control differences leave weak visual signatures. We propose LA-CONTROL, a latent action model tailored to driving control that grounds video-learned representations in vehicle state and state changes. Through physical supervision during causal future prediction, LA-CONTROL allocates latent capacity to these complementary quantities while retaining room for other visual dynamics. This organization encourages control-relevant distinctions to remain accessible in the learned latent actions. Once trained, the model produces control-proxy targets for driving videos without action labels. We evaluate how readily small probes recover physical signals from these representations and how effectively the representations support downstream planner pretraining. Across two driving datasets and three planner architectures, LA-CONTROL outperforms alternative latent-action models in physical readout and downstream open-loop trajectory prediction.

## 1 INTRODUCTION

Driving video is abundant. It captures diverse roads, weather, traffic, and rare maneuvers, while synchronized control annotations are available for only a fraction of this footage. Latent Action Models (LAMs) offer a way to learn representations of behavior from visual transitions without explicit action labels. A typical LAM couples an inverse dynamics model (IDM), which encodes consecutive observations into a bottleneck latent action, with a forward dynamics model (FDM), which predicts the next observation representation from the current observation and that latent. The two models are trained jointly through a self-supervised future-prediction objective, using observed video as the training target. The learned latent actions can then serve as targets for policy pretraining on videos without action labels. This recipe underlies action-free policy pretraining in LAPO (Schmidt & Jiang, 2024) and latent-action supervision in LAPA (Ye et al., 2025).

LA-Pose (Wang et al., 2026a) demonstrates that self-supervised latent actions can support metric camera-pose estimation with limited geometric supervision. Inspired by this result, we investigate how latent actions learned from driving video can be grounded in vehicle control signals. This introduces an additional challenge: control signals shape subsequent motion through vehicle dynamics and the current motion state, while images reveal the resulting changes in camera pose. A representation sufficient for pose estimation may consequently underrepresent the distinctions needed for control.

The central difficulty is that future prediction alone does not require a latent interface to preserve these control distinctions. Prior work highlights the importance of action supervision, latent constraints, and the video distribution in shaping learned latent actions (Nikulin et al., 2025; Luo et al., 2026; Garrido et al., 2026).

In driving, this mismatch arises at two levels. First, persistent physical state can be more directly observable than its temporal changes. For example, ego velocity and curvature describe motion state, while acceleration and curvature rate describe how that state changes. Second, even when a latent captures changes in motion, it may represent them unevenly: large steering adjustments produce conspicuous visual changes, whereas subtle corrections can be compressed into nearly



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.

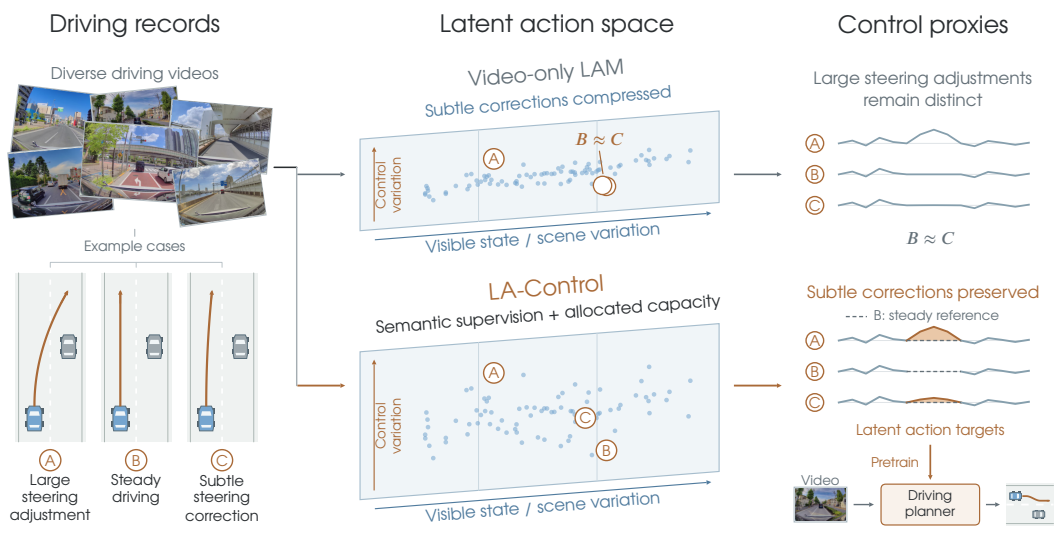


Figure 1: From driving video to control proxies. Driving records feed both latent interfaces; A–C illustrate a large steering adjustment, steady driving, and a subtle steering correction in the same direction. Video-only prediction distinguishes A but can map B and C to nearly identical latent codes. LA-CONTROL preserves the distinction between B and C in its latent actions, providing control proxies for planner pretraining.

indistinguishable codes. A prediction-trained bottleneck can favor visible state and salient motion while underrepresenting subtle control differences. A driving-oriented latent action should therefore preserve both physical state and control-relevant changes, including those with weak visual signatures. Figure 1 illustrates this mismatch and our control-oriented interface.

We introduce LA-CONTROL, a latent action model tailored to driving control, with physically grounded representations designed to preserve both vehicle state and subtle control-relevant changes. It partitions a fixed latent budget into dedicated subspaces for $0^{\text{th}}$-order state anchors, $1^{\text{st}}$-order state-change targets, and residual visual dynamics. Synchronized physical labels orient these subspaces toward their intended semantics during causal future prediction. The IDM uses causal temporal context to infer each transition-level latent action, while the residual region retains capacity for other visual dynamics.

We evaluate the learned representations through action readability, which measures how readily a simple probe recovers physical state and control signals, and planner pretraining, which tests their usefulness as supervision for learning driving behavior. Across these evaluations, LA-CONTROL improves physical readout over alternative LAMs and yields lower open-loop trajectory errors across three representative planner architectures, supporting the use of grounded latent actions as control proxies.

## 2 RELATED WORK

**Learning latent actions from video.** Predictive latent-action models infer transition codes from observations and train them to support future prediction. DynaMo (Cui et al., 2024) uses inverse–forward dynamics to pretrain visual representations for imitation learning. Genie (Bruce et al., 2024) applies inverse–forward learning at scale with discrete codes, whereas Garrido et al. (2026) use constrained continuous latents for in-the-wild actions. AdaWorld (Gao et al., 2025) adapts latents to a known action interface, while other methods explicitly structure them: Olaf-World (Jiang et al., 2026) aligns transitions through their effects, DiLA (Zhang et al., 2026) separates dynamic structure from visual content, and FLAM (Wang et al., 2026b) factorizes independently moving entities' dynamics. AC-LAM (Wei et al., 2026) imposes additive compositional constraints to obtain motion-specific latents with better-calibrated displacement magnitudes.

**Grounding latent actions.** Physical or task supervision can guide which distinctions a predictive latent preserves. LAOM (Nikulin et al., 2025) identifies action-correlated distractors in reconstruction-based latents and uses sparse action supervision during LAM training to improve control and robustness to latent compression. LAOF (Bu et al., 2026) uses optical-flow constraints to encourage action-relevant latent representations. JALA (Luo et al., 2026) jointly aligns a predictive embedding, an inverse-dynamics latent, and available actions. UniVLA (Bu et al., 2025) and Villa-X (Chen et al., 2026) further ground latent actions in task or physical semantics for cross-embodiment policy learning. LA-Pose (Wang et al., 2026a) grounds driving-video latent actions in metric camera pose. Building on these grounding approaches, LA-CONTROL combines physical supervision with dedicated latent regions for vehicle state and state changes. We focus on preserving control-relevant magnitudes and temporal variations within driving transitions, including subtle corrections with weak visual signatures.

**Policy learning with latent actions.** Latent actions also serve as supervision for downstream policies. LAPO (Schmidt & Jiang, 2024) shows that a policy can be pretrained without action labels. LAPA (Ye et al., 2025) pretrains policies on inferred latent-action targets from action-free video before real-action adaptation. Moto (Chen et al., 2025b) similarly bridges video pretraining and robot policy learning with latent motion tokens. UniMimic (Chen et al., 2025a) jointly pretrains policies to predict latent actions and future latent states, combining motion priors with predictive state representations. We adopt latent-action pretraining followed by downstream adaptation to evaluate the usefulness of learned representations across driving planners under shared downstream protocols.

# 3 LA-CONTROL

In this section, we first define the physical quantities and structured latent-action learning, then describe how the resulting control proxies are evaluated through downstream planners.

## 3.1 DRIVING CONTROL FORMULATION

**Video-predictive latent actions.** Let $o_t$ denote the visual observation at time $t$ and $\mathcal{C}_t = (o_{t-H+1}, \ldots, o_t)$ its temporal context. An inverse dynamics model (IDM) $q$ infers the transition ending at $o_{t+1}$, and a forward dynamics model (FDM) $p$ predicts its visual representation:

$$z_t = q(\mathcal{C}_{t+1}) \in \mathbb{R}^D, \qquad \hat{y}_{t+1} = p(\mathcal{C}_t, z_{\leq t}). \tag{1}$$

For a video clip comprising consecutive observations $(o_0, \ldots, o_T)$, these operations produce the latent sequence $z_{0:T-1}$ and predictions $\hat{y}_{1:T}$. At position $t$, the causal FDM uses visual context and latent history through $t$ to predict the target at $t+1$. We optimize the video-prediction objective $\mathcal{L}_{\mathrm{FDM}}(\hat{y}_{t+1}, y_{t+1})$ over all clips in the video collection $\mathcal{D}$. We follow prior work and call $z_t$ a *latent action* because it represents an observation transition. Each $z_t$ represents one transition over the sampled interval $\Delta t$. Increasing $H$ adds temporal context without changing the interval represented by each latent or the output frequency.

**State anchors and state changes.** We use synchronized vehicle-motion signals to supervise state anchors and their first-order changes. Lateral motion is expressed in linearized curvature coordinates under the small-angle kinematic approximation. The state-anchor vector $\mathbf{s}_t = (v_t, \kappa_t)$ contains ego velocity and curvature. The state-change vector $\mathbf{u}_t = (a_t, \dot{\kappa}_t)$ contains acceleration and curvature rate. Over a sampling interval $\Delta t$, their semantics correspond to $a_t \approx (v_{t+1} - v_t)/\Delta t$ and $\dot{\kappa}_t \approx (\kappa_{t+1} - \kappa_t)/\Delta t$. These relations specify physical semantics; label construction follows each dataset's signal definitions and preprocessing. Here $0^{\text{th}}$-order and $1^{\text{st}}$-order refer to the motion state $\mathbf{s}_t$ and its temporal changes, independently of global vehicle pose. A labeled subset $\mathcal{D}_\ell \subset \mathcal{D}$ contains synchronized $(\mathbf{s}_t, \mathbf{u}_t)$. The remaining video supplies transition supervision.

Our goal is to learn latent actions that retain the state and state-change information underlying each transition while remaining predictive for the FDM. Figure 2 summarizes the LAM and its use in downstream evaluation.

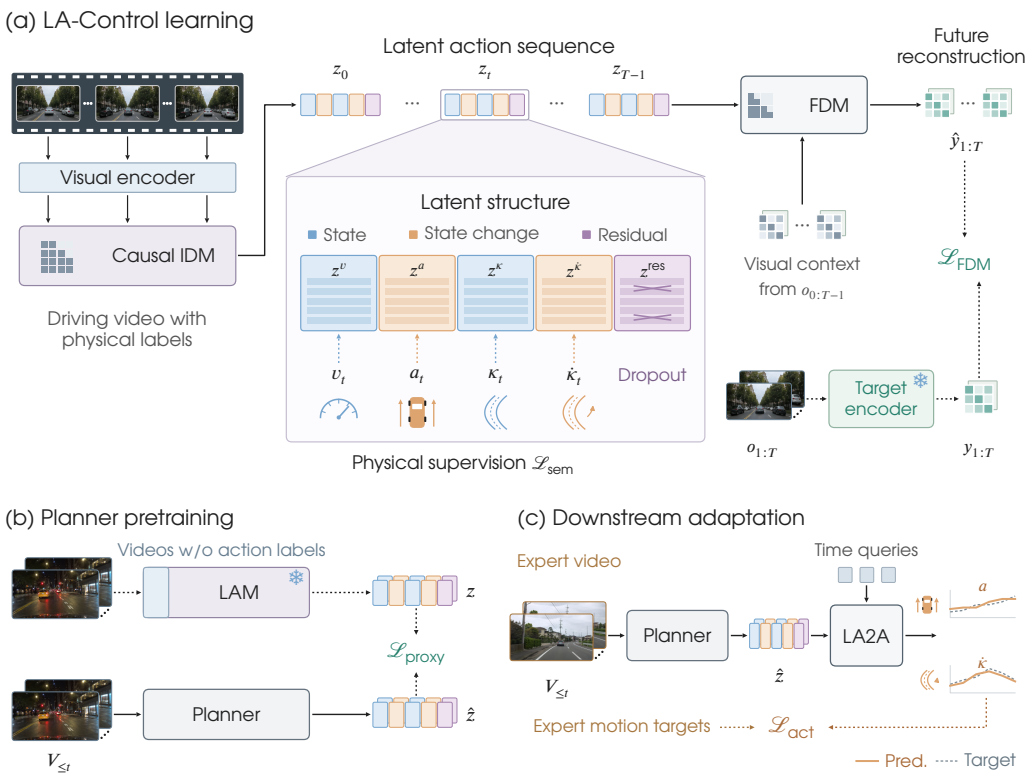


Figure 2: Overview of LA-CONTROL. (a) A causal IDM produces structured latent actions with physically supervised regions and a free residual; the FDM predicts future visual representations from visual context and latent history. (b) The frozen LAM supplies targets for planner pretraining on videos without action labels. (c) Expert demonstrations adapt the pretrained planner to physical motion targets.

### 3.2 STRUCTURED LATENT-ACTION REPRESENTATION

**Video-predictive backbone.** Given an RGB clip, an online image encoder produces visual tokens. For every sampled transition, a causal IDM uses $\mathcal{C}_{t+1}$ to produce a query and maps it through a continuous $D = 50$ bottleneck to $z_t$. A shared expansion maps each latent into the FDM token space, where the FDM predicts discrete visual codes produced by the frozen Cosmos tokenizer (NVIDIA et al., 2025). The targets provide a dense video objective without defining the semantics of the latent interface.

**Capacity for state and state changes.** Inspired by the representation design of Concept Embedding Models (Espinosa Zarlenga et al., 2022), which represent supervised concepts with multidimensional embeddings, we assign dedicated latent regions to physical factors while keeping the total latent size fixed:

$$z_t = \left[ z_t^v \,\middle|\, z_t^a \,\middle|\, z_t^\kappa \,\middle|\, z_t^{\dot\kappa} \,\middle|\, z_t^{\text{res}} \right] \in \mathbb{R}^D. \tag{2}$$

The velocity and curvature regions anchor visually dominant state. The acceleration and curvature-rate regions explicitly reserve capacity for 1st-order state changes whose visual signatures may be weak. The free residual provides capacity for visual change not captured by the supervised physical quantities. All regions jointly condition the FDM, coupling physical supervision with future prediction.

Our $D = 50$ model assigns 10 coordinates to each region. Each supervised region is a multidimensional embedding whose assigned physical quantity is read out by a small head. Physical labels

specify what should be recoverable from the region, while future prediction also shapes its coordinates to support visual dynamics. This leaves flexibility to encode transition-specific information beyond the scalar supervision target.

### 3.3 Physical Supervision and Training

**Semantic anchoring.** For labeled clips, let $\xi_t = (v_t, a_t, \kappa_t, \dot{\kappa}_t)$ collect the two state anchors and two state-change quantities, indexed by $k \in \{v, a, \kappa, \dot{\kappa}\}$. A factor-specific head $h_k$ reads only its assigned region and predicts a categorical distribution $\pi_t^k = h_k(z_t^k)$ over target bins. The corresponding continuous estimate is the expectation over bin centers $c_b$:

$$\hat{\xi}_t^k = \sum_{b=1}^{B_k} \pi_{t,b}^k c_b, \qquad \mathcal{L}_{\text{sem}}^k = \text{CE}(\pi_t^k, b_k(\xi_t^k)) + \lambda_{\text{reg}} \rho_k(\hat{\xi}_t^k - \xi_t^k), \tag{3}$$

where $B_k$ is the number of bins for factor $k$, $b_k(\cdot)$ maps a target value to its bin index, and $\rho_k$ is a robust regression loss with fixed factor-specific scaling. The classification loss supervises the target bin, while the regression loss penalizes the difference between the expected value and the continuous target. Targets are aggregated over the physical duration of each sampled transition so their horizon remains aligned when the temporal stride changes. Semantic losses apply only to $\mathcal{D}_\ell$, while the video-prediction loss is applied to all clips.

**Allocation enforcement and training objective.** To encourage the FDM to use the supervised regions, we apply residual dropout, randomly masking residual coordinates before the shared FDM expansion during training. This mechanism supports the prescribed allocation and is disabled at evaluation; its ablation is reported in Appendix D.2. We average factor losses within each motion axis and give longitudinal and lateral axes equal aggregate weight:

$$\begin{aligned} \mathcal{L}_{\text{long}} &= \tfrac{1}{2}(\mathcal{L}_{\text{sem}}^v + \mathcal{L}_{\text{sem}}^a)\,, & \mathcal{L}_{\text{lat}} &= \tfrac{1}{2}\big(\mathcal{L}_{\text{sem}}^\kappa + \mathcal{L}_{\text{sem}}^{\dot{\kappa}}\big)\,, \\ \mathcal{L}_{\text{LAM}} &= \mathcal{L}_{\text{FDM}} + \tfrac{\lambda_{\text{sem}}}{2}\,(\mathcal{L}_{\text{long}} + \mathcal{L}_{\text{lat}})\,. \end{aligned} \tag{4}$$

We use $\lambda_{\text{sem}} = 1$.

### 3.4 Transfer to Downstream Planners

Following the recipe of LAPA (Ye et al., 2025), we evaluate the learned representations through latent-action pretraining followed by downstream adaptation. The adapted planners predict acceleration and curvature rate, which are converted into trajectories for evaluation. After LAM training, the frozen image encoder and IDM construct latent-action targets offline from observation transitions, while the downstream planner predicts them from causal visual context. For planners that decode predicted latents into physical motion targets (Planners A and B in our experiments), the workflow is

$$\begin{aligned} \mathcal{C}_{t+1} &\xrightarrow{q} z_t, & &\text{offline proxy construction,} \\ V_{\le t} &\xrightarrow{g} \hat{z}_{t:t+K_z-1} \xrightarrow{d} \hat{\mathbf{u}}_{t:t+K_a-1}, & &\text{causal planner inference.} \end{aligned} \tag{5}$$

Here $V_{\le t}$ is the visual context available to the planner and $K_z$ and $K_a$ denote the latent-action and motion-target sequence lengths. The offline IDM may observe the target frame to construct each label; the planner observes no future frames.

**Latent-action pretraining.** The planner learns to predict latent-action sequences by minimizing

$$\hat{z}_{t:t+K_z-1} = g(V_{\le t}), \qquad \mathcal{L}_{\text{proxy}} = \sum_{k=0}^{K_z-1} \ell_{\text{LA}}(\hat{z}_{t+k}, z_{t+k}), \tag{6}$$

where $\ell_{\text{LA}}$ is the planner's latent-sequence prediction loss.

**Downstream adaptation.** For planners with an action decoder, we attach a time-query latent-action-to-action (LA2A) decoder $d$ to the pretrained planner and fine-tune on expert demonstrations. For

details of architecture-specific adaptation, please refer to Appendix A.2. On expert demonstrations, the composed predictor is trained against sequences of physical motion targets $\mathbf{u}_t = (a_t, \dot{\kappa}_t)$:

$$\hat{\mathbf{u}}_{t:t+K_a-1} = d(\hat{z}_{t:t+K_z-1}), \qquad \mathcal{L}_{\text{act}} = \sum_{k=0}^{K_a-1} \ell_{\text{act}}(\hat{\mathbf{u}}_{t+k}, \mathbf{u}_{t+k}). \tag{7}$$

This adaptation jointly updates the visual encoder, planner, and LA2A decoder using only $\mathcal{L}_{\text{act}}$. For these architectures, we refer to $g$ and $d$ jointly as the planner.

# 4 EXPERIMENTS

We evaluate latent-action representations along two axes: how readily physical information can be recovered from them, and how effectively they transfer to downstream driving planners. Controlled ablations then examine the roles of physical supervision and latent organization.

## 4.1 EXPERIMENTAL SETUP

**Data.** Public representation diagnostics use nuScenes (Caesar et al., 2020). We use a proprietary driving dataset, termed *Internal*, for LAM and planner training and evaluation. A recording is a source video sequence; a clip is a continuous segment sampled from it. Both Internal and nuScenes use disjoint recordings for training and evaluation. We train LA-CONTROL on 800k Internal video clips, each with labels for velocity, acceleration, curvature, and curvature rate. These clips do not overlap with either the planner pretraining or downstream adaptation data. Training uses eight NVIDIA H100 GPUs; model architecture and temporal sampling are described in Appendix A.

**Downstream planners.** We evaluate transfer using three planner architectures representing established policy families: **Planner A**, a Transformer-based action-sequence predictor (e.g., ACT (Zhao et al., 2023)); **Planner B**, a world–action planner (e.g., GR-1 (Wu et al., 2024)); and **Planner C**, a diffusion-based sequence planner (e.g., Diffusion Policy (Chi et al., 2023)). All three follow an imitation-learning framework, with latent-action pretraining followed by behavior cloning on expert motion-target sequences. Planner B jointly predicts future visual features and latent actions during pretraining. During downstream adaptation, Planners A and B use time-query LA2A decoding, while Planner C directly denoises physical motion sequences. Each planner is pretrained on latent-action targets from 200k video clips and adapted using the same 10k motion-labeled recordings across LAM interfaces. The transfer procedure is described in Section 3.4.

**Compared methods.** For benchmarking, we compare LA-CONTROL with external LAMs in action readability (Table 1). For planner transfer, we use AdaWorld (Gao et al., 2025), Olaf-World (Jiang et al., 2026), and Olaf-World +AR as baselines (Table 2). Here, +AR denotes supervised action retargeting using physical labels. Within each planner, all interfaces use the same training data, pretraining and adaptation budgets, adaptation recipe, and evaluation protocol.

For controlled ablations, we compare variants that share the LAM backbone, video data, and latent dimension ($D = 50$) to isolate our design choices. *Video-only* uses only future prediction. *Naive auxiliary* and LA-CONTROL use the same physical labels; Naive auxiliary predicts state anchors and state changes from the full latent, without assigning dedicated regions to individual physical quantities. LA-CONTROL uses the five-way partition in Equation (2) and residual dropout. Notably, physical supervision on only 5% of training recordings can already shape a latent space with readability close to full-label training; see Appendix D.1.

**Evaluation metrics.** We assess action readability by fitting a small MLP to frozen latent actions, following Villa-X (Chen et al., 2026) and CoMo (Yang et al., 2026). It jointly predicts velocity, acceleration, curvature, and curvature rate; macro $R^2$ and nRMSE summarize readability across the four targets. Probe architectures and data splits are described in Appendix B.1.

We convert predicted motion quantities into trajectories using a shared kinematic bicycle model and evaluate them against recorded odometry using ADE and FDE at 1 s and 3 s. Both metrics measure position error in meters: ADE averages errors over the evaluation horizon, while FDE measures endpoint error.

Table 1: Latent-action readability with a joint probe. $R^2$ and nRMSE average four physical targets. Subtle uses the shared subset defined in Section 4.1.

| Representation | Internal | | | | nuScenes | |
|---|---|---|---|---|---|---|
| | Overall | | Subtle | | Overall | |
| | $R^2 \uparrow$ | nRMSE $\downarrow$ | $R^2 \uparrow$ | nRMSE $\downarrow$ | $R^2 \uparrow$ | nRMSE $\downarrow$ |
| AdaWorld (Gao et al., 2025) | 0.304 | 0.804 | -1.004 | 1.388 | 0.297 | 0.822 |
| DiLA (Zhang et al., 2026) | 0.266 | 0.830 | -0.876 | 1.345 | 0.329 | 0.799 |
| Olaf-World (Jiang et al., 2026) | 0.315 | 0.787 | -0.119 | 0.976 | 0.361 | 0.769 |
| Olaf-World +AR (Jiang et al., 2026) | 0.443 | 0.687 | -0.537 | 1.091 | 0.401 | 0.737 |
| LAPA (Ye et al., 2025) | 0.219 | 0.869 | -6.020 | 2.228 | 0.257 | 0.850 |
| LAOM (Nikulin et al., 2025) | -0.010 | 0.913 | -7.200 | 2.621 | 0.264 | 0.823 |
| Villa-X (Chen et al., 2026) | 0.187 | 0.892 | -0.921 | 1.359 | 0.222 | 0.873 |
| Villa-X +AR (Chen et al., 2026) | 0.339 | 0.768 | -0.403 | 1.126 | 0.343 | 0.781 |
| LA-CONTROL | **0.918** | **0.271** | **0.570** | **0.605** | **0.593** | **0.602** |

For planner evaluation, Overall evaluates unique timestamps in the held-out Internal recordings. *Subtle* comprises two-second clips with deceleration and light-to-moderate steering adjustments (Appendix B.4) and is shared by the readability and planner evaluations.

### 4.2 LATENT-ACTION READABILITY

LA-CONTROL achieves the highest macro $R^2$ and lowest nRMSE on both datasets and the shared Subtle subset (Table 1), indicating that physical information is more readily recovered from its latents by a small probe.

Separating state anchors from state changes reveals where these gains arise (Appendix C). Across the external LAMs, velocity and curvature are more readable than acceleration and curvature rate, consistent with the distinction between persistent state and state changes that motivates our design. For the two state-change targets, LA-CONTROL achieves a mean $R^2$ of $0.881$ on Internal and $0.356$ on nuScenes, compared with $0.112$ and $0.060$ for the strongest external baseline, Olaf-World +AR. The gains therefore extend to the first-order transition quantities used in downstream evaluation.

The reduction in readability gains from Internal to nuScenes is larger for acceleration and curvature rate than for state anchors (Table 4). This suggests that recovering state changes is more sensitive to domain differences, although these quantities remain the main source of improvement over external LAMs on both datasets.

**Qualitative latent-action readouts.** Figure 3 complements the aggregate metrics with six selected Internal clips. In these examples, compact probes recover acceleration build-up and release as well as signed curvature-rate changes from LA-CONTROL latents, whereas several alternative interfaces produce attenuated or inconsistent responses. Timing and amplitude errors remain visible.

### 4.3 DOWNSTREAM PLANNER TRANSFER

**Transfer across planners.** LA-CONTROL achieves the lowest Overall ADE and FDE for all three planners at both horizons (Table 2). The magnitude of improvement varies across architectures: Planner A benefits at both horizons, Planner B shows larger relative gains at 3 s, and Planner C shows larger relative gains at 1 s. One possible explanation for Planner B's smaller Overall gains is that jointly predicting future visual features and latent actions already encourages learning action-relevant information from visual transitions. This additional visual objective may make the planner less dependent on the control information supplied by its latent-action targets. These results support the utility of LA-CONTROL latents for planner pretraining across different architectures and adaptation schemes.

**Subtle.** On this shared subset, LA-CONTROL achieves the highest action readability and the lowest trajectory errors across all three planners at both horizons (Tables 1 and 2). Relative trajectory

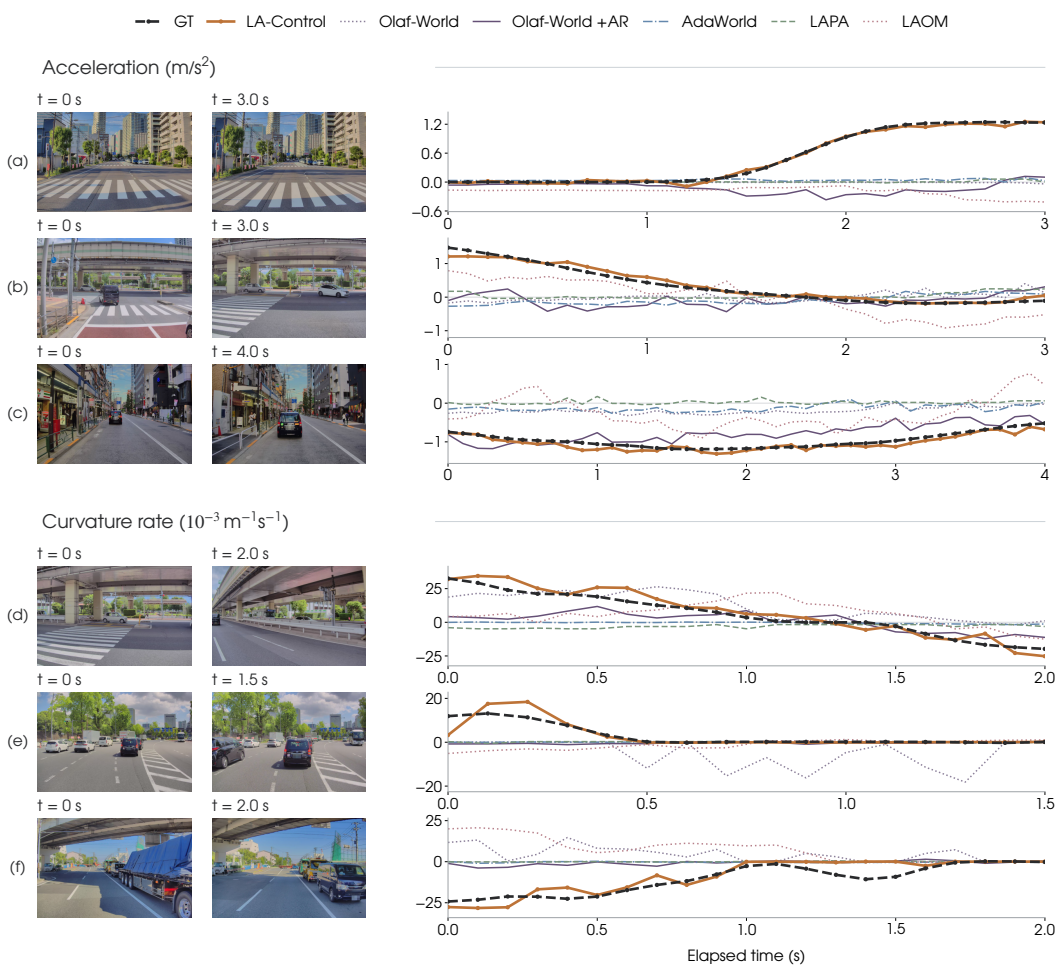


Figure 3: Acceleration (a–c) and curvature-rate (d–f) readouts on selected held-out Internal clips. Images show the first and last frames of each plotted interval; curves compare GT with the joint-probe readouts from Table 1. Acceleration is in $\mathrm{m/s^2}$ and curvature rate in $10^{-3}\,\mathrm{m^{-1}s^{-1}}$.

improvements are larger than on Overall for every planner. The subset rule was selected after inspecting planner results, so these findings are exploratory.

### 4.4 Ablation: Control-Oriented Supervision

Under the same physical supervision and latent dimensionality, LA-Control improves both action readability and downstream planner transfer over unstructured auxiliary supervision (Table 3). The gains are particularly pronounced on Subtle: macro $R^2$ increases from $0.440$ to $0.570$, while Planner A's $3\,\mathrm{s}$ ADE decreases by $12.63\%$. These results suggest that organizing the latent interface helps preserve control-relevant distinctions beyond adding physical supervision alone.

LA-Control also retains more recoverable physical information after Internal-to-nuScenes adaptation than Naive auxiliary (Appendix E).

## 5 Discussion and Limitations

Our results suggest that learning control-oriented latent actions involves both specifying what information should be recoverable and allocating capacity to preserve it. Physical supervision provides the semantic targets, while explicit allocation reserves space for quantities whose importance to control may exceed their visual salience. The additional gains on Subtle are consistent with this role of

Table 2: Open-loop rollout ADE/FDE ($10^{-2}$ m) across three planner architectures. Bold and underlining mark the best and second-best values. Baselines are AdaWorld (Gao et al., 2025) and Olaf-World (Jiang et al., 2026); +AR denotes action retargeting.

| Planner | Interface | @1 s | | | | @3 s | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | Overall | | Subtle | | Overall | | Subtle | |
| | | ADE ↓ | FDE ↓ | ADE ↓ | FDE ↓ | ADE ↓ | FDE ↓ | ADE ↓ | FDE ↓ |
| A | AdaWorld | 5.006 | 12.424 | 7.640 | 19.150 | 43.468 | 114.076 | 68.550 | 180.836 |
| | Olaf-World | 4.923 | 12.185 | 6.635 | 16.511 | 41.937 | 109.381 | 57.600 | 151.039 |
| | Olaf-World +AR | <u>4.375</u> | <u>10.838</u> | <u>5.988</u> | <u>15.164</u> | <u>37.983</u> | <u>99.775</u> | <u>54.929</u> | <u>145.745</u> |
| | LA-CONTROL | **4.075** | **10.026** | **5.260** | **13.119** | **35.564** | **93.875** | **48.489** | **129.709** |
| B | AdaWorld | 4.776 | 11.763 | 7.608 | 18.768 | 40.000 | 103.790 | 64.348 | 167.240 |
| | Olaf-World | 4.308 | 10.611 | 6.196 | 15.335 | 36.785 | 96.185 | 53.484 | 140.088 |
| | Olaf-World +AR | <u>3.997</u> | <u>9.830</u> | <u>5.712</u> | <u>14.257</u> | <u>34.600</u> | <u>91.145</u> | <u>50.516</u> | <u>133.066</u> |
| | LA-CONTROL | **3.986** | **9.806** | **5.046** | **12.607** | **34.302** | **89.955** | **45.718** | **120.948** |
| C | AdaWorld | <u>5.495</u> | <u>13.538</u> | 9.305 | 22.874 | <u>45.569</u> | <u>117.791</u> | 72.719 | 182.174 |
| | Olaf-World | 5.544 | 13.725 | <u>7.975</u> | <u>19.380</u> | 47.234 | 123.362 | <u>60.799</u> | <u>151.861</u> |
| | Olaf-World +AR | 6.449 | 15.950 | 10.367 | 25.537 | 53.968 | 140.100 | 84.089 | 215.298 |
| | LA-CONTROL | **4.671** | **11.691** | **5.582** | **13.543** | **42.643** | **113.793** | **43.593** | **110.707** |

Table 3: Controlled ablation of physical supervision and latent organization. Readability is measured by macro $R^2$ and nRMSE; planner columns report Planner A's 3 s ADE ($10^{-2}$ m).

| Representation | Readability | | | | | | Planner transfer | |
|---|---|---|---|---|---|---|---|---|
| | Internal | | | | nuScenes | | Planner A: ADE ↓ @3 s | |
| | Overall | | Subtle | | Overall | | Overall | Subtle |
| | $R^2$ ↑ | nRMSE ↓ | $R^2$ ↑ | nRMSE ↓ | $R^2$ ↑ | nRMSE ↓ | | |
| Video-only | 0.183 | 0.882 | -0.691 | 1.271 | 0.153 | 0.909 | 59.935 | 85.329 |
| Naive auxiliary | 0.903 | 0.289 | 0.440 | 0.681 | 0.584 | 0.607 | 35.858 | 55.496 |
| LA-CONTROL | **0.918** | **0.271** | **0.570** | **0.605** | **0.593** | **0.602** | **35.564** | **48.489** |

allocation. Meanwhile, future prediction shapes the supervised regions into transition-dependent embeddings that can retain visual context beyond their scalar physical labels.

**Limitations and future work.** The latent allocation is tailored to driving-specific physical variables; extending it to other physical motion parameterizations remains future work. Planner experiments assess early adaptation through open-loop trajectory prediction, leaving closed-loop interaction untested. The Subtle subset was selected after inspecting validation results, so its gains require confirmation on independent data. Broader domain-shift evaluations would further establish the scope of transfer.

## 6 CONCLUSION

We introduced LA-CONTROL, a latent action model tailored to driving control that combines physical supervision with structured semantic allocation during causal future prediction. Its design reserves capacity for vehicle state and state changes, including control-relevant distinctions with weak visual signatures. Across two driving datasets and three planner architectures, LA-CONTROL improves physical readout and downstream planner transfer. During continued training on a new driving domain, the structured representation also better retains recoverable physical information from the original domain. These results show how physical labels can shape a latent action space that supplies control-oriented supervision to videos without action annotations while retaining its physical grounding through further adaptation.

## Ethics Statement

This work evaluates driving planners through open-loop trajectory evaluation. These evaluations do not establish real-world driving safety, which requires separate closed-loop validation.

## AI Use Statement

We used AI agents to assist with literature review, development of evaluation and visualization code, and manuscript revision. The authors take responsibility for the final content of this work.

## References


Jake Bruce, Michael D. Dennis, Ashley Edwards, Jack Parker-Holder, Yuge Shi, Edward Hughes, Matthew Lai, Aditi Mavalankar, Richie Steigerwald, Chris Apps, Yusuf Aytar, Sarah Bechtle, Feryal Behbahani, Stephanie C. Y. Chan, Nicolas Heess, Lucy Gonzalez, Simon Osindero, Sherjil Ozair, Scott Reed, Jingwei Zhang, Konrad Zolna, Jeff Clune, Nando de Freitas, Satinder Singh, and Tim Rocktäschel. Genie: Generative interactive environments. In *Proceedings of the 41st International Conference on Machine Learning*, pp. 4603–4623, 2024.

Qingwen Bu, Yanting Yang, Jisong Cai, Shenyuan Gao, Guanghui Ren, Maoqing Yao, Ping Luo, and Hongyang Li. Learning to act anywhere with task-centric latent actions. In *Proceedings of Robotics: Science and Systems*, Los Angeles, California, USA, 2025.

Xizhou Bu, Jiexi Lyu, Fulei Sun, Ruichen Yang, Zhiqiang Ma, and Wei Li. LAOF: Robust latent action learning with optical flow constraints. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 27334–27344, 2026.

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, 2020.

Guangyan Chen, Meiling Wang, Te Cui, Luojie Yang, Qi Shao, Lin Zhao, Tianle Zhang, Yihang Li, Yi Yang, and Yufeng Yue. Unifying latent action and latent state pre-training for policy learning from videos. In *Proceedings of the SIGGRAPH Asia 2025 Conference Papers*, pp. 1–11, 2025a.

Xiaoyu Chen, Hangxing Wei, Pushi Zhang, Chuheng Zhang, Kaixin Wang, Yanjiang Guo, Rushuai Yang, Yucen Wang, Xinquan Xiao, Li Zhao, Jianyu Chen, and Jiang Bian. Villa-X: Enhancing latent action modeling in vision-language-action models. In *International Conference on Learning Representations*, 2026.

Yi Chen, Yuying Ge, Weiliang Tang, Yizhuo Li, Yixiao Ge, Mingyu Ding, Ying Shan, and Xihui Liu. Moto: Latent motion token as the bridging language for learning robot manipulation from videos. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 19752–19763, 2025b.

Cheng Chi, Siyuan Feng, Yilun Du, Zhenjia Xu, Eric Cousineau, Benjamin CM Burchfiel, and Shuran Song. Diffusion policy: Visuomotor policy learning via action diffusion. In *Proceedings of Robotics: Science and Systems*, 2023.

Zichen Jeff Cui, Hengkai Pan, Aadhithya Iyer, Siddhant Haldar, and Lerrel Pinto. DynaMo: In-domain dynamics pretraining for visuo-motor control. In *Advances in Neural Information Processing Systems*, volume 37, pp. 33933–33961, 2024.

Mateo Espinosa Zarlenga, Pietro Barbiero, Gabriele Ciravegna, Giuseppe Marra, Francesco Giannini, Michelangelo Diligenti, Zohreh Shams, Frederic Precioso, Stefano Melacci, Adrian Weller, Pietro Lió, and Mateja Jamnik. Concept embedding models: Beyond the accuracy-explainability trade-off. In *Advances in Neural Information Processing Systems*, volume 35, 2022.

Shenyuan Gao, Siyuan Zhou, Yilun Du, Jun Zhang, and Chuang Gan. AdaWorld: Learning adaptable world models with latent actions. In *Proceedings of the 42nd International Conference on Machine Learning*, pp. 18744–18771, 2025.

Quentin Garrido, Tushar Nagarajan, Basile Terver, Nicolas Ballas, Yann LeCun, and Michael Rabbat. Learning latent action world models in the wild. In *Proceedings of the 43rd International Conference on Machine Learning*, 2026.

Yuxin Jiang, Yuchao Gu, Ivor W. Tsang, and Mike Zheng Shou. Olaf-World: Orienting latent actions for video world modeling. In *Proceedings of the 43rd International Conference on Machine Learning*, 2026.

Jason Kong, Mark Pfeiffer, Georg Schildbach, and Francesco Borrelli. Kinematic and dynamic vehicle models for autonomous driving control design. In *Proceedings of the IEEE Intelligent Vehicles Symposium*, pp. 1094–1099, 2015.

Hao Luo, Ye Wang, Wanpeng Zhang, Haoqi Yuan, Yicheng Feng, Haiweng Xu, Sipeng Zheng, and Zongqing Lu. Joint-aligned latent action: Towards scalable VLA pretraining in the wild. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 35047–35058, 2026.

Fabian Mentzer, David Minnen, Eirikur Agustsson, and Michael Tschannen. Finite scalar quantization: VQ-VAE made simple. In *International Conference on Learning Representations*, 2024.

Alexander Nikulin, Ilya Zisman, Denis Tarasov, Nikita Lyubaykin, Andrei Polubarov, Igor Kiselev, and Vladislav Kurenkov. Latent action learning requires supervision in the presence of distractors. In *Proceedings of the 42nd International Conference on Machine Learning*, pp. 46427–46447, 2025.

NVIDIA et al. Cosmos world foundation model platform for physical AI. *arXiv preprint arXiv:2501.03575*, 2025.

Dominik Schmidt and Minqi Jiang. Learning to act without actions. In *International Conference on Learning Representations*, 2024.

Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. In *International Conference on Learning Representations*, 2021.

Zhengqing Wang, Saurabh Nair, Prajwal Chidananda, Pujith Kachana, Samuel Li, Matthew Brown, and Yasutaka Furukawa. LA-Pose: Latent action pretraining meets pose estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 34460–34469, 2026a.

Zizhao Wang, Chang Shi, Jiaheng Hu, Kevin Rohling, Roberto Martín-Martín, Amy Zhang, and Peter Stone. Factored latent action world models. In *Proceedings of the 43rd International Conference on Machine Learning*, 2026b.

Hangxing Wei, Xiaoyu Chen, Chuheng Zhang, Tim Pearce, Jianyu Chen, Alex Lamb, Li Zhao, and Jiang Bian. Learning additively compositional latent actions for embodied AI. *arXiv preprint arXiv:2604.03340*, 2026.

Hongtao Wu, Ya Jing, Chilam Cheang, Guangzeng Chen, Jiafeng Xu, Xinghang Li, Minghuan Liu, Hang Li, and Tao Kong. Unleashing large-scale video generative pre-training for visual robot manipulation. In *International Conference on Learning Representations*, 2024.

Jiange Yang, Yansong Shi, Haoyi Zhu, Mingyu Liu, Kaijing Ma, Yating Wang, Gangshan Wu, Tong He, and Limin Wang. CoMo: Learning continuous latent motion from internet videos for scalable robot learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 42352–42363, 2026.

Seonghyeon Ye, Joel Jang, Byeongguk Jeon, Se June Joo, Jianwei Yang, Baolin Peng, Ajay Mandlekar, Reuben Tan, Yu-Wei Chao, Bill Yuchen Lin, Lars Liden, Kimin Lee, Jianfeng Gao, Luke Zettlemoyer, Dieter Fox, and Minjoon Seo. Latent action pretraining from videos. In *International Conference on Learning Representations*, 2025.

Tianqiu Zhang, Muyang Lyu, Yufan Zhang, Fang Fang, and Si Wu. DiLA: Disentangled latent action world models. In *Proceedings of the 43rd International Conference on Machine Learning*, 2026.

Tony Z. Zhao, Vikash Kumar, Sergey Levine, and Chelsea Finn. Learning fine-grained bimanual manipulation with low-cost hardware. In *Proceedings of Robotics: Science and Systems*, 2023.

# A Implementation Details

## A.1 Model Architecture and Temporal Sampling

All model-training runs use eight NVIDIA H100 GPUs. A 12-block per-frame image tokenizer produces visual tokens of width 1,536. Four causal spatiotemporal IDM blocks produce 20 future-aware transition queries; a three-layer MLP maps each query from 1,536 to the continuous 50-dimensional LA. A second three-layer MLP expands each LA to one 1,536-dimensional action token. The FDM has four causal spatiotemporal blocks followed by two next-token-prediction refiners and predicts next-frame targets produced by the Cosmos-DI8x8 tokenizer (NVIDIA et al., 2025), which uses finite scalar quantization (FSQ) (Mentzer et al., 2024). Each spatial location predicts four $2 \times 2$ subpositions and six FSQ channels, yielding 24 categorical heads. The frozen Cosmos cache serves exclusively as a prediction target; RGB tokenization and LA inference are performed online.

The canonical regions have dimensions $10/10/10/10/10$. Each semantic head is $10 \to 128 \to B_k$. The Smooth $L_1$ regression loss on the probability-weighted bin centers uses $\lambda_{\text{reg}} = 0.1$, with residuals expressed in fixed signal scales. Residual dropout is applied before the shared $50 \to 1536$ expansion. Its probability gradually increases from zero to $0.5$ during training, and dropout is disabled at evaluation.

**Temporal sampling.** Each item supplies 101 consecutive frames from a 10 Hz source. One stride is drawn per batch and a valid random offset selects 21 frames; RGB and vehicle-motion tensors use the same indices. The sampling distribution and resulting span are:

| Stride | 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|---|
| Probability | 0.40 | 0.30 | 0.15 | 0.10 | 0.05 |
| Clip span (s) | 2.0 | 4.0 | 6.0 | 8.0 | 10.0 |

Both real timestamps and ordinal frame positions are passed through frequency embeddings to the tokenizer and IDM. Transition labels average the valid physical annotations within each sampled transition.

## A.2 Planner Adaptation

For downstream adaptation, we attach a time-query LA2A decoder and prediction head to Planners A and B, mapping their pretrained latent-action representations to physical motion targets at the required timestamps. The visual encoder, planner backbone, LA2A decoder, and prediction head are updated jointly using only the motion-target prediction loss; Planner B's future-feature objective is not retained during adaptation.

Planner C transfers its pretrained diffusion backbone and replaces the target input and output projections and normalization for physical motion prediction. It directly denoises physical motion sequences using the diffusion objective, without an LA2A decoder.

# B Evaluation Protocols

## B.1 Probe Protocol

The joint probe is $50 \to 8 \to 4$ and has 444 parameters for the controlled $D = 50$ models. For external LAMs, the input width matches the native latent dimension $D$, giving a $D \to 8 \to 4$ probe. Each frozen representation receives a freshly fitted probe, with common fit/evaluation splits across methods in each domain. The Internal probe-fitting and evaluation transitions come from disjoint recordings. Subtle reuses the fitted Internal probes without refitting. For the readability tables, macro $R^2$ and nRMSE equally average the four targets. Each target's RMSE is normalized by its population standard deviation within the corresponding evaluation set.

### B.2 External LAMs

AdaWorld, Olaf-World, DiLA, and Villa-X use their publicly released checkpoints; LAPA uses the released Open X-Embodiment checkpoint. For Olaf-World and Villa-X, +AR denotes additional driving-domain supervision on velocity, acceleration, curvature, and curvature rate. LAOM is reimplemented and trained from scratch on Internal with the same four physical targets. The external representations differ in upstream training data and objectives.

AdaWorld, Olaf-World, and LAPA use $D = 32$; DiLA uses $D = 256$; Villa-X uses $D = 64$; and LAOM uses $D = 8192$. The +AR variants retain their original dimensions.

### B.3 Downstream Evaluation

**Open-loop trajectory metrics.** We convert the predicted motion quantities to trajectories using a shared kinematic bicycle model (Kong et al., 2015). Each rollout starts from the recorded motion state, with position and heading set to zero in the current ego frame. We linearly interpolate the predicted motion coordinates between prediction anchors and integrate their state changes. The model's fixed geometric mapping converts lateral motion coordinates for bicycle integration, using midpoint steps of $0.005$ s and clamping speed to be nonnegative. The lateral sign follows the dataset's coordinate convention.

Ground-truth positions come from recorded odometry expressed in the current ego frame and are linearly interpolated to the evaluation times. For horizon $T \in \{1, 3\}$ s and $t_k = kT/10$, $k = 1, \ldots, 10$, ADE is the mean of $\|\hat{\mathbf{p}}(t_k) - \mathbf{p}(t_k)\|_2$, and FDE is the corresponding distance at $t_{10} = T$.

**Diffusion inference.** Planner C jointly denoises the motion-target sequence using ten deterministic DDIM steps (Song et al., 2021), with the same initial noise across interfaces.

### B.4 Subtle Corrections

Subtle is characterized by a median observed speed above 20 km/h and a mean signed acceleration of at most $-0.2\,\mathrm{m/s^2}$, with acceleration below $-0.2\,\mathrm{m/s^2}$ for at least $70\%$ of the time. The mean absolute curvature rate lies in approximately $[0.16, 1.56] \times 10^{-3}\,\mathrm{m^{-1}s^{-1}}$.

## C Additional Readability Results

Table 4 separates state anchors $(v, \kappa)$ from state-change quantities $(a, \dot{\kappa})$ to show which physical information is recovered from each representation. Group $R^2$ averages the two corresponding per-target scores from the probes used in Table 1. LAMs generally make state anchors more readable than state changes, and LA-Control improves both groups over these interfaces.

Table 4: State and state-change group readability ($R^2$). State averages $(v, \kappa)$ and State change averages $(a, \dot{\kappa})$. Subtle denotes the shared subset defined in Appendix B.4.

| Representation | Internal | | | | nuScenes | |
|---|---|---|---|---|---|---|
| | Overall | | Subtle | | Overall | |
| | State ↑ | State change ↑ | State ↑ | State change ↑ | State ↑ | State change ↑ |
| AdaWorld (Gao et al., 2025) | 0.636 | -0.028 | -1.026 | -0.981 | 0.570 | 0.024 |
| DiLA (Zhang et al., 2026) | 0.574 | -0.041 | -0.361 | -1.390 | 0.613 | 0.045 |
| Olaf-World (Jiang et al., 2026) | 0.674 | -0.044 | 0.619 | -0.858 | 0.697 | 0.024 |
| Olaf-World +AR (Jiang et al., 2026) | 0.775 | 0.112 | 0.738 | -1.812 | 0.742 | 0.060 |
| LAPA (Ye et al., 2025) | 0.438 | 0.001 | -10.702 | -1.337 | 0.506 | 0.008 |
| LAOM (Nikulin et al., 2025) | 0.718 | -0.738 | -3.252 | -11.149 | 0.635 | -0.107 |
| Villa-X (Chen et al., 2026) | 0.372 | 0.002 | -0.724 | -1.119 | 0.437 | 0.007 |
| Villa-X +AR (Chen et al., 2026) | 0.679 | -0.001 | 0.264 | -1.070 | 0.672 | 0.014 |
| LA-Control | 0.955 | 0.881 | 0.766 | 0.374 | 0.830 | 0.356 |

# D Additional Ablations

## D.1 Label Efficiency

The primary LA-CONTROL model uses all available semantic labels in the labeled source. Motivated by the finding of LAOM (Nikulin et al., 2025) that sparse action supervision is most effective during LAM training, we additionally evaluate a 5%-label variant as a supplementary diagnostic. All recordings remain available for IDM/FDM learning, while the 5%-label variant restricts semantic supervision to a fixed, randomly selected 5% of recordings. This variant alternates ordinary video batches with batches resampled from the fixed labeled subset (label replay); the full-label model uses ordinary sampling.

Table 5: Label efficiency of LA-CONTROL on Internal.

| Labeled recordings (%) | Macro $R^2 \uparrow$ | nRMSE $\downarrow$ |
|---|---|---|
| 5 | 0.913 | 0.279 |
| 100 | 0.918 | 0.271 |

With label replay, the 5%-label variant remains close to the full-label model in joint readability: macro $R^2$ decreases by 0.005 and nRMSE increases by 0.008 (Table 5).

## D.2 Counterfactual Latent Interventions

We use counterfactual latent interventions to test how residual dropout changes the FDM's reliance on different latent regions. For each validation window, we replace one region throughout its latent sequence with that from another recording, keeping the recipient's visual context, other latent regions, and prediction targets fixed. Table 6 specifies the physical quantities matched or varied when choosing each donor. Both models are evaluated with dropout disabled.

Pairs are chosen across recordings using normalized, window-averaged physical labels. For longitudinal and lateral interventions, donors are selected to match the preserved physical quantities while differing in those being changed (Table 6); residual interventions use the nearest donor across all four labels.

Table 6: Counterfactual latent-region interventions. Relative FDM loss increases over unmodified predictions are averaged across validation sources; larger values indicate greater sensitivity to replacement.

| Replaced region | Donor quantities | | FDM loss increase (%) | |
|---|---|---|---|---|
| | Matched | Changed | Trained w/o dropout | Trained w/ dropout |
| Longitudinal $(v, a)$ | $(\kappa, \dot{\kappa})$ | $(v, a)$ | 9.99 | 11.22 |
| Lateral $(\kappa, \dot{\kappa})$ | $(v, a)$ | $(\kappa, \dot{\kappa})$ | 11.54 | 17.48 |
| Residual | $(v, a, \kappa, \dot{\kappa})$ | — | 11.85 | 0.07 |

Without residual dropout, residual replacement disrupts prediction about as much as replacing a physical region, even when the donor has similar measured motion. With residual dropout, sensitivity to lateral replacement increases, while sensitivity to residual replacement nearly vanishes under this matching rule. This shift supports greater use of the designated lateral region for future prediction. We also provide a supplementary video of counterfactual rollouts, comparing autoregressive FDM predictions from the same initial scene under its original latent-action sequence and a sequence transferred from another driving clip. Video pairs contrast stationary/moving clips with similar curvature, or straight/turning clips with comparable speeds, across different recordings with the same vehicle and camera configuration. Displayed examples are selected after visual review and include rollouts using the recipient's original latent sequence for comparison.

## E CROSS-DOMAIN ADAPTATION

Explicit organization can also benefit a representation as it continues to learn from a new domain. We compare Video-only, Naive auxiliary, and LA-CONTROL after adaptation from Internal to nuScenes. Table 7 measures how much original-domain physical information remains accessible to a newly fitted compact probe.

Table 7: Semantic retention on Internal after adaptation to nuScenes. Readability drop is defined as $R^2_{\text{before}} - R^2_{\text{after}}$; negative values indicate improved readout. Frozen-FDM damage follows Equation (8).

| Representation | Internal $R^2_{\text{after}} \uparrow$ | Probe $R^2$ drop $\downarrow$ | Frozen-FDM damage $\downarrow$ |
|---|---|---|---|
| Video-only | 0.197 | $-0.029$ | 2.6% |
| Naive auxiliary | 0.860 | 0.060 | 16.5% |
| LA-CONTROL | 0.906 | 0.010 | 13.7% |

We continue training the Internal-pretrained models used in the readability evaluation for 2,000 steps on nuScenes. All trainable LAM parameters, including the FDM, are updated on the same nuScenes windows using each variant's original objective: future prediction alone for Video-only, with physical supervision added for Naive auxiliary and LA-CONTROL.

For the compatibility test, we hold the pre-adaptation FDM and target tokenizer fixed and compare original versus updated latents under the same visual context. Let $\bar{\mathcal{L}}_{\text{old}}$, $\bar{\mathcal{L}}_{\text{new}}$, and $\bar{\mathcal{L}}_0$ denote this frozen FDM's mean cross-entropy with original, updated, and zero latents, respectively:

$$\text{FDM damage} = 100\,\frac{\bar{\mathcal{L}}_{\text{new}} - \bar{\mathcal{L}}_{\text{old}}}{\bar{\mathcal{L}}_0 - \bar{\mathcal{L}}_{\text{old}}}\ \%. \tag{8}$$

This normalizes the adaptation-induced loss increase by the penalty for removing the original latent.

Video-only has the lowest frozen-FDM damage ($2.6\%$), but retains little readable physical information ($R^2 = 0.197$). Compared with Naive auxiliary, LA-CONTROL reduces the readability drop from $0.060$ to $0.010$ and frozen-FDM damage from $16.5\%$ to $13.7\%$. The latter measures compatibility with the original visual predictor. Together, these results suggest that structured physical grounding helps maintain a readable and reusable latent interface during cross-domain adaptation.

## F SUPPLEMENTARY VIDEO

We provide a supplementary video combining the counterfactual rollouts in Appendix D.2 with visualizations of latent principal components (PCs) and their alignment with acceleration and curvature rate. Synchronized driving footage and curves show how latent variation follows physical changes in selected clips, comparing LA-CONTROL with other representations. For each method and physical target, we select the most correlated of the first eight PCs by absolute Pearson correlation on a separate fitting set; the PCA basis, component choice, and sign are then fixed for visualization. PC and ground-truth curves are standardized within each recording to compare their temporal patterns. The video complements the probe-based readability results with a direct view of how individual latent directions relate to control signals.