DriveWAM: Video Generative Priors Enable Scalable World-Action Modeling for Autonomous Driving

Chen Shi[1*] Jinrui Xu[1*] Shaoshuai Shi[2] Kehua Sheng[2] Bo Zhang[2] Li Jiang[1,3†]

[1]The Chinese University of Hong Kong, Shenzhen [2]Voyager Research, Didi Chuxing [3]Shenzhen Loop Area Institute

*Equal contribution †Corresponding author Project page: **chenshi3.github.io/drivewam.github.io**

Project Page GitHub Code

NeurIPS SYDNEY

Introduction

Motivation

VLA driving policies inherit VLMs pretrained on **static image–text pairs**; temporal dynamics must come from driving data. **Video generative models** already carry motion and scene-evolution priors, but lack **semantic planning** and **bounded memory**. **DriveWAM** turns a pretrained video diffusion transformer into an **autoregressive video-action driving policy**.

Contributions

- **Unified world-action policy.** A video DiT predicts future video latents and decodes ego actions by **inverse dynamics** under one **joint flow-matching** loss.
- **Scene-evolving driving guidance.** A **frozen VLM** supplies causal, chunk-specific intent, injected through **temporally localized cross-attention**.
- **Selective KV memory.** Training-free **relevance–redundancy** selection keeps bounded, modality-aware video and action pools for long-horizon rollout.

90.1	0.83 m	16%	>12×
NAVSIM v1 PDMS	PhysicalAI-AV ADE@4s	close encounters	lower rollout cost
single front camera	best baseline: 1.22 m	AlpaSim · prior best 28%	KV memory & FLOPs

Experiments

NAVSIM v1

SV / MV: single- / multi-view camera; L: LiDAR

Method	Sensors	NC ↑	DAC ↑	TTC ↑	C. ↑	EP ↑	PDMS ↑
Human	–	100	100	100	99.9	87.5	94.8
UniAD CVPR'23	MV	97.8	91.9	92.9	**100.0**	78.8	83.4
LAW ICLR'25	SV	96.4	95.4	88.7	99.9	81.7	84.6
DiffusionDrive CVPR'25	MV & L	98.2	96.2	94.7	**100.0**	82.2	88.1
WoTE ICCV'25	MV & L	98.5	96.8	94.4	99.9	81.9	88.3
VLA-based methods							
ReCogDrive* ICLR'26	MV	98.1	94.7	94.2	**100.0**	80.9	86.5
DriveVLA-W0 ICLR'26	SV	**98.7**	96.2	95.5	**100.0**	82.2	88.4
AutoVLA NeurIPS'25	MV	98.4	95.6	**98.0**	99.9	81.9	89.1
DriveDreamer-Policy arXiv'26	MV	98.4	97.1	95.1	**100.0**	83.5	89.2
DriveVLA-W0† ICLR'26	SV	98.7	99.1	95.3	99.3	83.3	90.2
World-action (WA) methods							
Epona ICCV'25	SV	97.9	95.1	93.8	99.9	80.4	86.2
WorldDrive arXiv'26	SV	98.4	95.8	95.2	99.8	83.3	89.0
DriveWAM (ours)	**SV**	**98.3**	**98.1**	**95.2**	**100.0**	**84.3**	**90.1**

* imitation learning only † trained with multiple trajectory anchors

PhysicalAI-AV

curated 1,000-clip test subset · single front camera

Method	Params	PhysicalAI data	ADE@3s ↓	FDE@3s ↓	ADE@4s ↓	FDE@4s ↓
VaVAM*	1.3B	✗	2.31	4.32	–	–
VaVAM	1.3B	✓	1.38	3.29	2.15	6.67
Alpamayo-1.5	10B	✓	0.80	2.31	1.44	4.18
DriveVLA-W0	8.5B	✓	0.80	2.06	1.22	3.24
DriveWAM (ours)	**5B + 8B**	✓	**0.47**	**1.35**	**0.83**	**2.47**

* released checkpoint, not trained on PhysicalAI data (3 s horizon only)

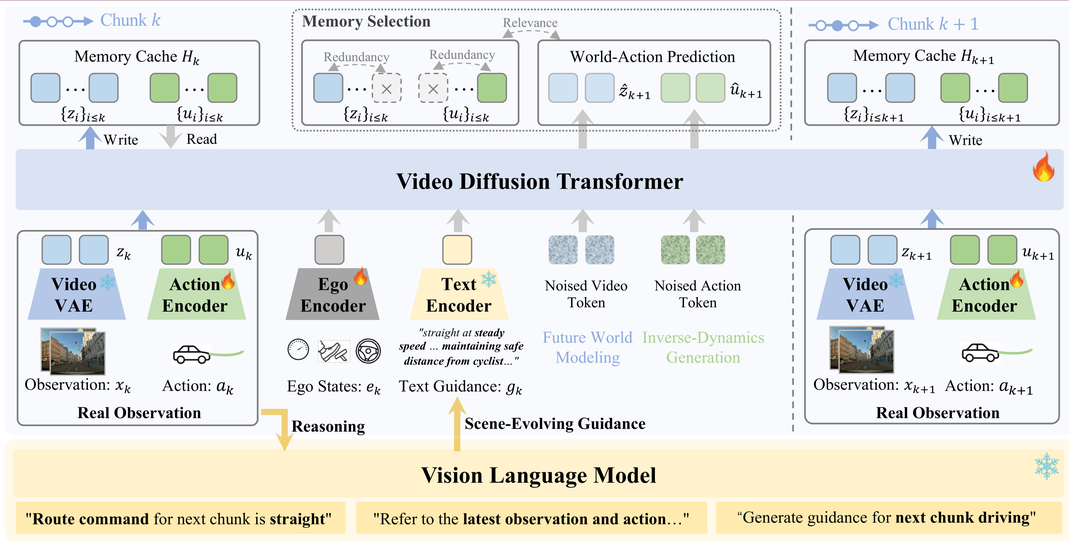
Closed-loop AlpaSim

50 reconstructed PhysicalAI scenes

Method	Close enc. all ↓	Close enc. at-fault ↓	Off-road ↓	Score all ↑	Score at-fault ↑
VaVAM	46.0%	16.0%	28.0%	0.22	0.30
Alpamayo-1.5	28.0%	10.0%	20.0%	0.39	0.51
DriveWAM (ours)	**16.0%**	**8.0%**	**14.0%**	**0.47**	**0.55**

More: NAVSIM v2 EPDMS **87.2** (navhard **35.6**) · full PhysicalAI test, 61.6k clips: ADE@4s **0.91** vs. 1.58 (Alpamayo-1.5)

DriveWAM Architecture



(a) Autoregressive Video-Action Generation

Video and action chunks form one causal token sequence. A shared flow-matching video DiT T_ω first predicts the **future video latent**, then decodes the **ego action conditioned on that imagined future** — inverse dynamics, not a separate trajectory head.

$$\hat{v}^z_{k+1,\tau} = T_\omega(z_{k+1,\tau}; H_k, e_k, g_k, \tau)$$
$$\hat{v}^a_{k+1,\tau} = D_a(T_\omega(u_{k+1,\tau}; \tilde{z}_{k+1}, H_k, e_k, g_k, \tau))$$
$$\mathcal{L} = \mathbb{E}_{k,\tau}\left[\left\|\hat{v}^z_{k+1,\tau} - v^z_{k+1,\tau}\right\|_2^2 + \beta_a \left\|\hat{v}^a_{k+1,\tau} - v^a_{k+1,\tau}\right\|_2^2\right]$$

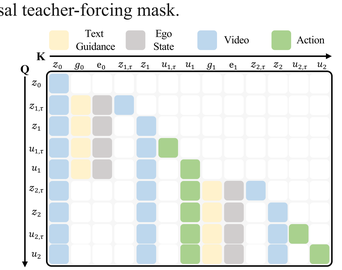
- **Future world modeling** under history H_k, ego state e_k and guidance g_k preserves the pretrained spatio-temporal prior.
- **Inverse-dynamics action flow:** clean future latent in training, generated latent at inference.
- **Full-clip training** under a causal teacher-forcing mask.

(b) Scene-Evolving Driving Guidance

Video priors model near-future dynamics, not **route intent, right-of-way or traffic signals**. At every decision step a **frozen Qwen3-VL-8B** reads only causally available context — latest observation x_k, recent ego motion a_k, route command c_k — and writes fresh guidance for the next chunk.

$$g_k = \Phi_{\text{VLM}}(x_k, a_k, c_k)$$ *queried once per 4 s chunk, reused across all denoising steps*

- **Temporally localized injection:** a block-diagonal text mask lets chunk k+1 attend only to g_k, so no guidance leaks across chunks during parallel training.
- **Causal by construction:** g_k never sees the target chunk — no future information is used.

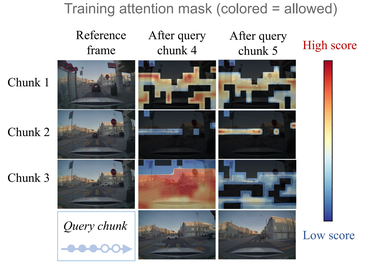


(c) Selective KV Memory

Full KV caching grows linearly with the rollout; FIFO evicts old but decision-critical cues. DriveWAM keeps **bounded, modality-aware** video and action pools, pruned at inference by a **training-free relevance–redundancy score**.

$$\rho_j^m = \frac{1}{|Q_k^m|}\sum_{\mathbf{q}\in Q_k^m}\left[\text{softmax}_{\ell\in H_k^m}\left(\frac{\mathbf{q}^\top \mathbf{k}_\ell^m}{\sqrt{d}}\right)\right]_j \qquad \eta_j^m = \text{mean}_{\ell\neq j}\cos(\mathbf{k}_j^m, \mathbf{k}_\ell^m)$$
$$s_j^m = \lambda\rho_j^m - (1-\lambda)\eta_j^m \qquad H_{k+1}^m \leftarrow \text{Top}_{B^m-|\Delta H_{k+1}^m|}(H_k^m)\cup\Delta H_{k+1}^m$$

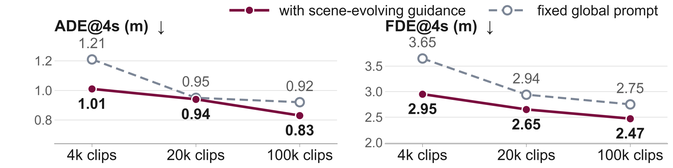
- ρ: relevance to the current queries · η: redundancy with other cached keys



Retained tokens: vehicles, stop signs and lanes stay; repeated sky and road are dropped

Ablations & Analysis

Data scaling × scene-evolving guidance



No saturation from 4k to 100k clips; guidance helps at **every** scale.

Video prior

Pretrained init.	Video sup.	ADE@4s ↓	FDE@4s ↓
✗	✗	1.40	4.21
✓	✗	1.23	3.79
✗	✓	1.10	3.26
✓	✓	**0.83**	**2.47**

KV memory strategy

Cache	ADE@4s ↓	FDE@4s ↓	Mem. (GB) ↓	GFLOPs ↓
Full	0.83	2.47	3.07	17.37
FIFO	1.40	3.47	0.25	1.05
Selective	**0.89**	**2.52**	**0.25**	**1.44**

accuracy: 20 s clips · cost: 300 s rollout

Pretrained init. **and** video supervision are both needed. Selective KV stays near full-cache accuracy at **>12× lower** memory and FLOPs.

Latency: 871 ms per 4 s chunk (one H20 GPU, 5 action steps) vs. 900 ms for Alpamayo-1.5.

Qualitative Results

Joint future video & trajectory generation

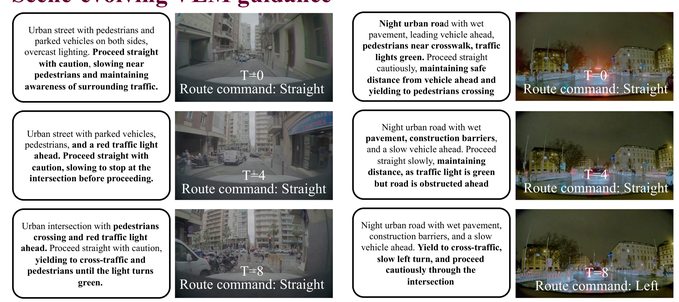


NAVSIM BEV: predicted vs. ground-truth trajectory · generated frames T = 1–4



PhysicalAI-AV trajectory on the current frame · generated frames T = 1–4

Scene-evolving VLM guidance



Guidance adapts to pedestrians, traffic lights and barriers as the scene and route change (T = 0, 4, 8).