A Comprehensive Survey on World Models for Embodied AI

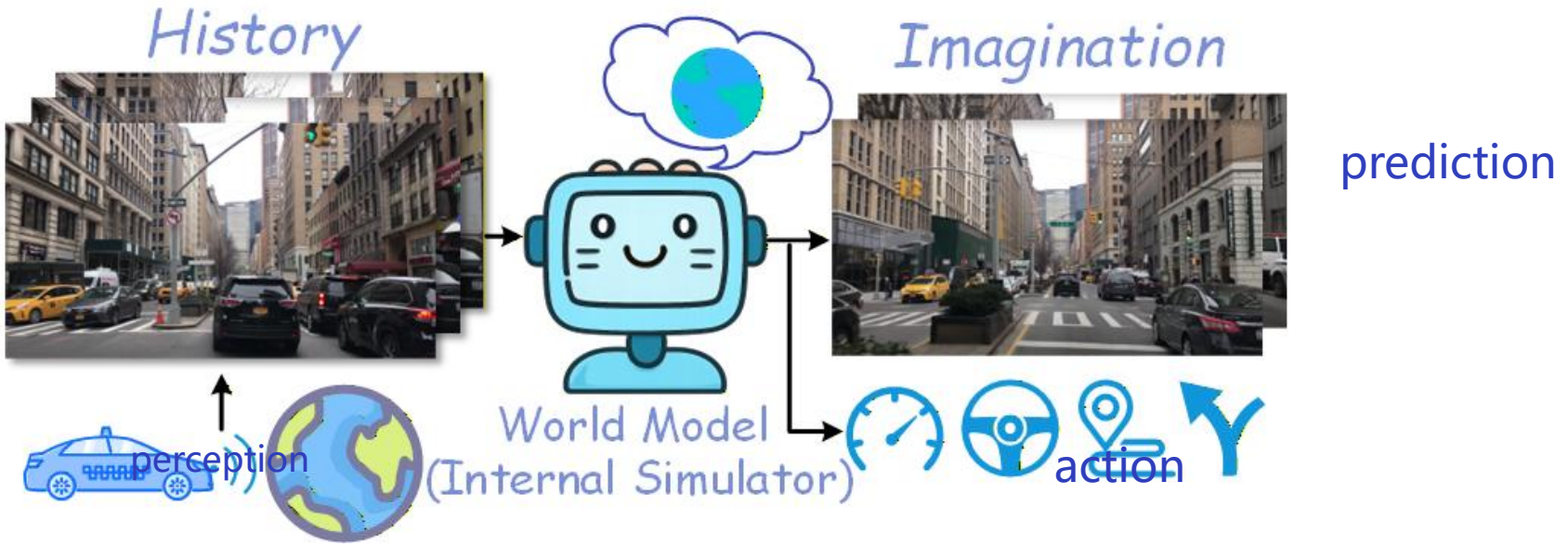
A three-axis taxonomy across functionality, temporality, and spatiality

Xinqing Li

arXive: 2510.16732 | GitHub: Li-Zn-H/AwesomeWorldModels

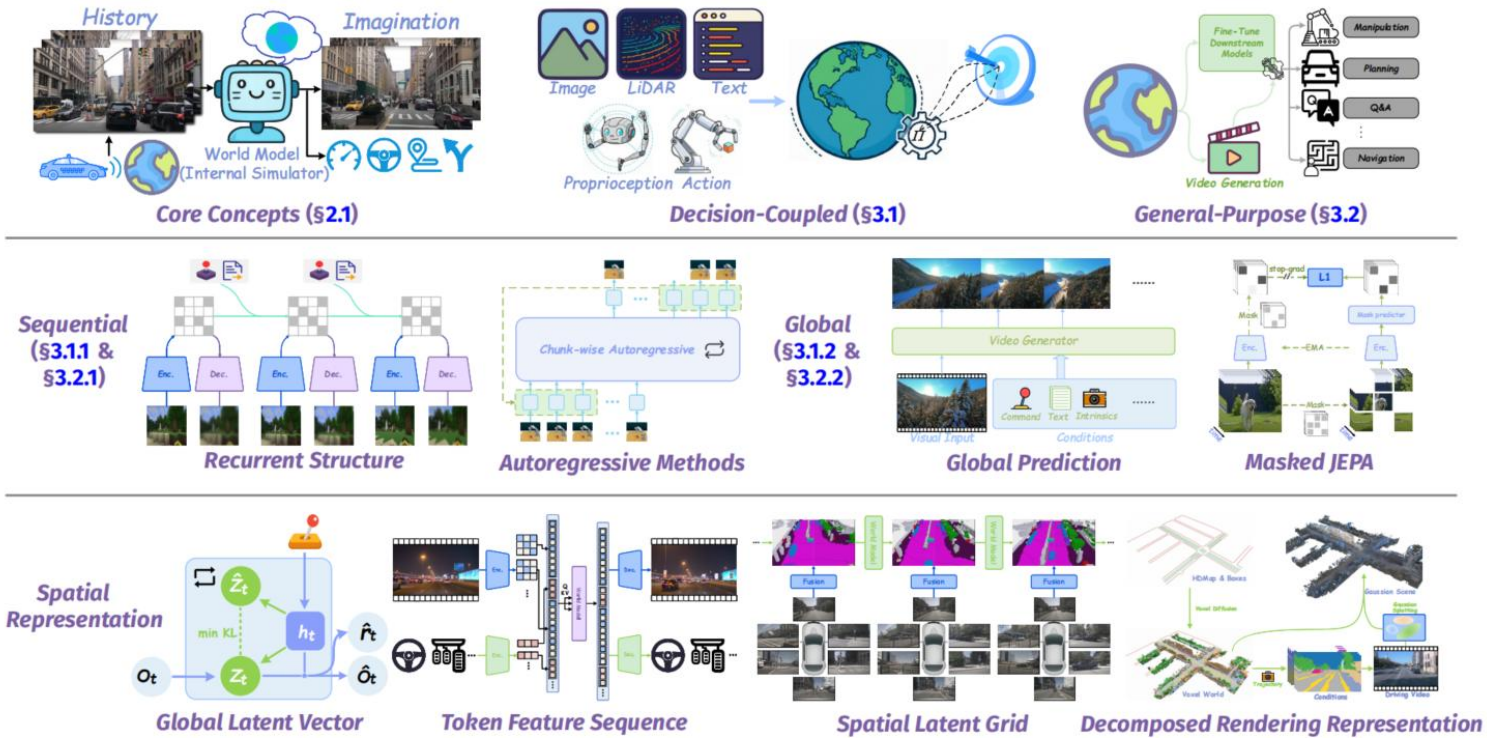
World Model:

- **World models**: perception→action→prediction (closed loop for embodied intelligence)
- **Origins**: Model-based RL (Ha & Schmidhuber, 2018), used RNNs to learn environment dynamics for internal simulation.
- **Status & gap**: generative models have enabled diverse architectures, but a unified survey and taxonomy remain lacking.



A Unified Three-Axis Framework

- Functionality: Decision-Coupled ↔ General-Purpose (defines optimization objectives).
- Temporality: Sequential Simulation ↔ Global Prediction (defines temporal modeling).
- Spatiality: Global Latent Vector / Token Feature Sequence / Spatial Latent Grid / Decomposed Rendering (defines state representation).



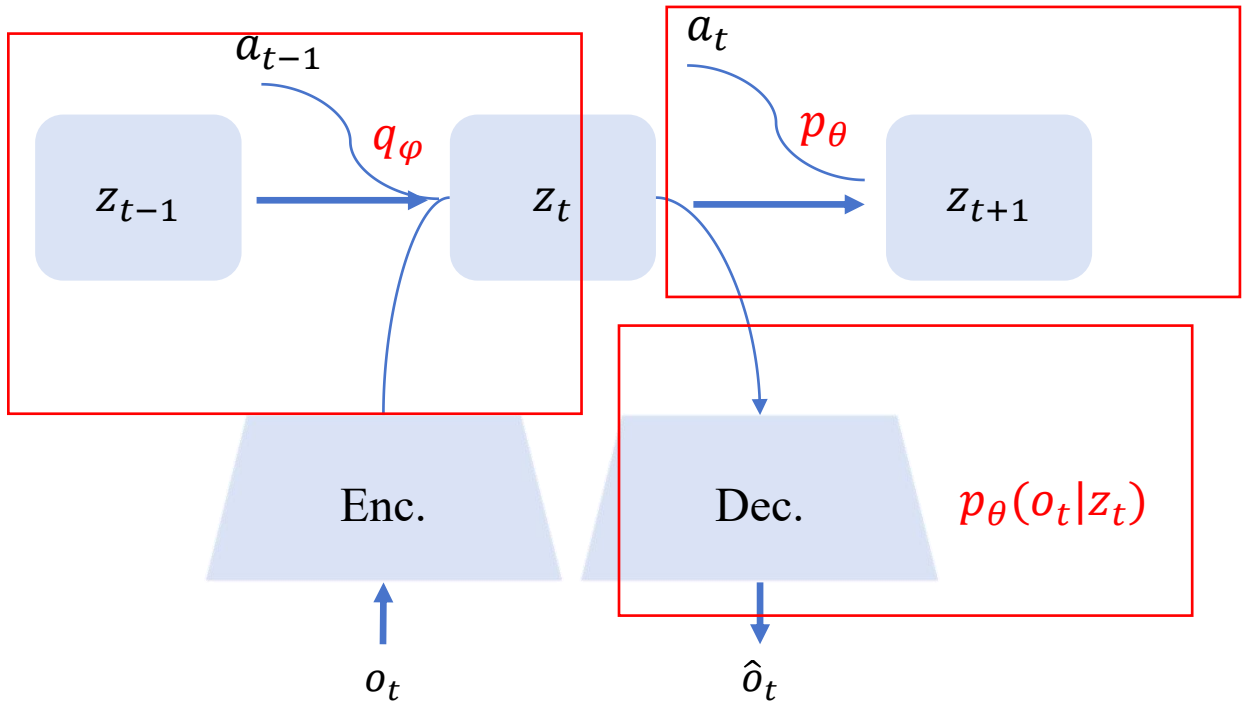
Mathematical Framework & Training Objective

- Task(POMDP): Learn the latent state z_t, the dynamics $p_\theta(z_t|z_{t-1}, a_{t-1})$, and the observation model $p_\theta(o_t|z_t)$ under partial observability.

- Inference: Use an approximate posterior $q_\varphi(z_t|z_{t-1}, a_{t-1}, o_t)$ to estimate the latent state—the belief over z_t given the current observation and history.

- Objective (ELBO):

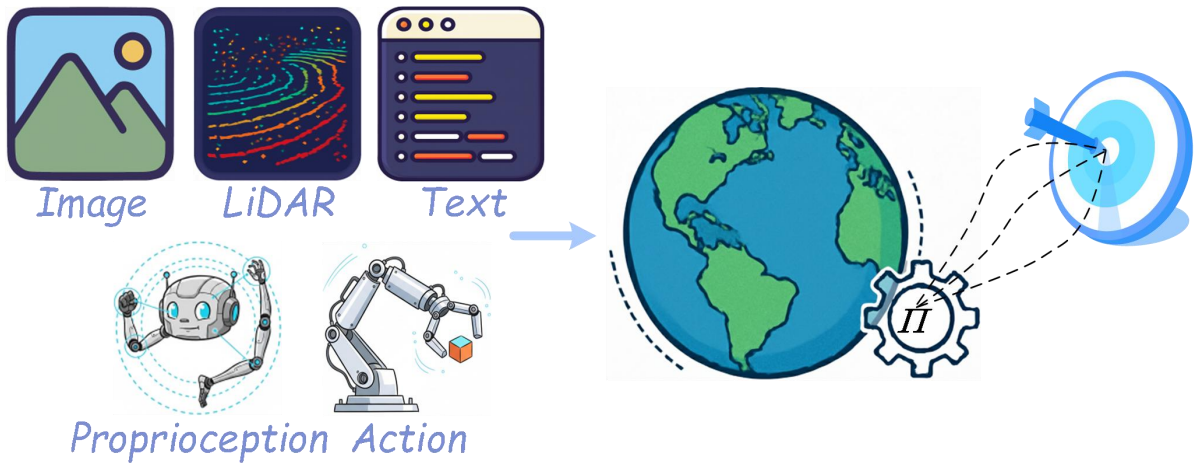
$$\mathcal{L}(\theta, \phi) = \sum_{t=1}^{T} \underbrace{\mathbb{E}_{q_\phi}\big[\log p_\theta(o_t \mid z_t)\big]}_{\text{reconstruction}} - \underbrace{D_{\mathrm{KL}}\big(q_\phi(z_t \mid z_{t-1}, a_{t-1}, o_t) \,\big\|\, p_\theta(z_t \mid z_{t-1}, a_{t-1})\big)}_{\text{KL regularization}}$$

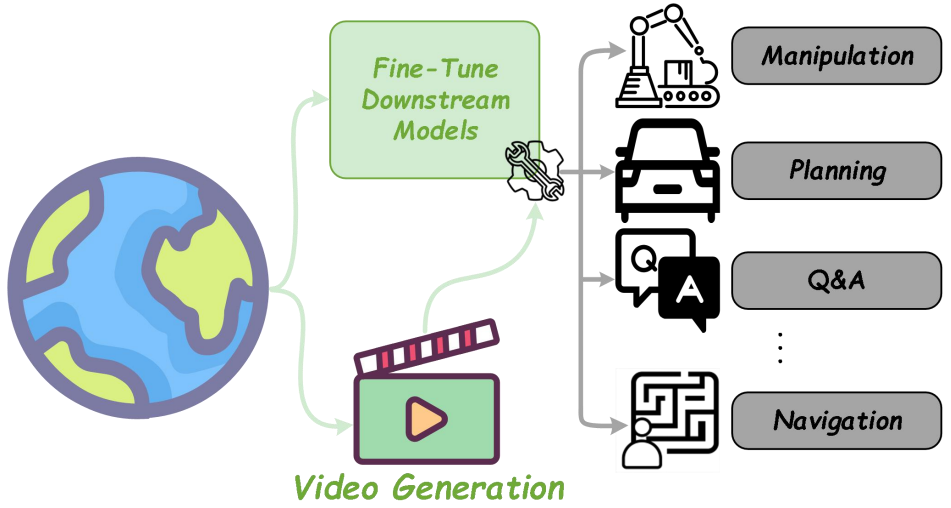


During rollout/deployment there is no o_t. the agent must rely on the prior dynamics p_θ to predict z_t. The KL term aligns the prior with the posterior, ensuring the prior can substitute the posterior at test time.

Decision-Coupled vs. General-Purpose

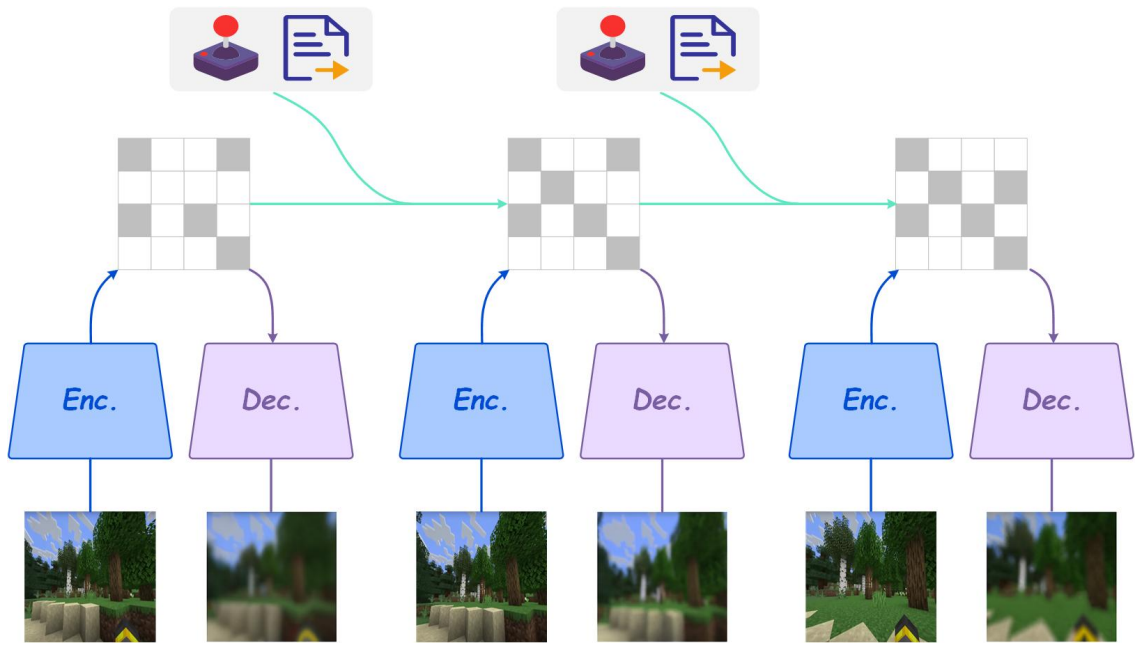
	Decision-Coupled	General-Purpose
Positioning	Task-specific, policy-tied	Task-agnostic environment modeling
Strengths	Strong plug-in performance	Broad transfer across tasks
Limitations	Overfits task; weak transfer	Heavier training; needs adaptation
Examples	MPC-style models; task-specialized WM	Video/general-purpose WM; fine-tune for downstream

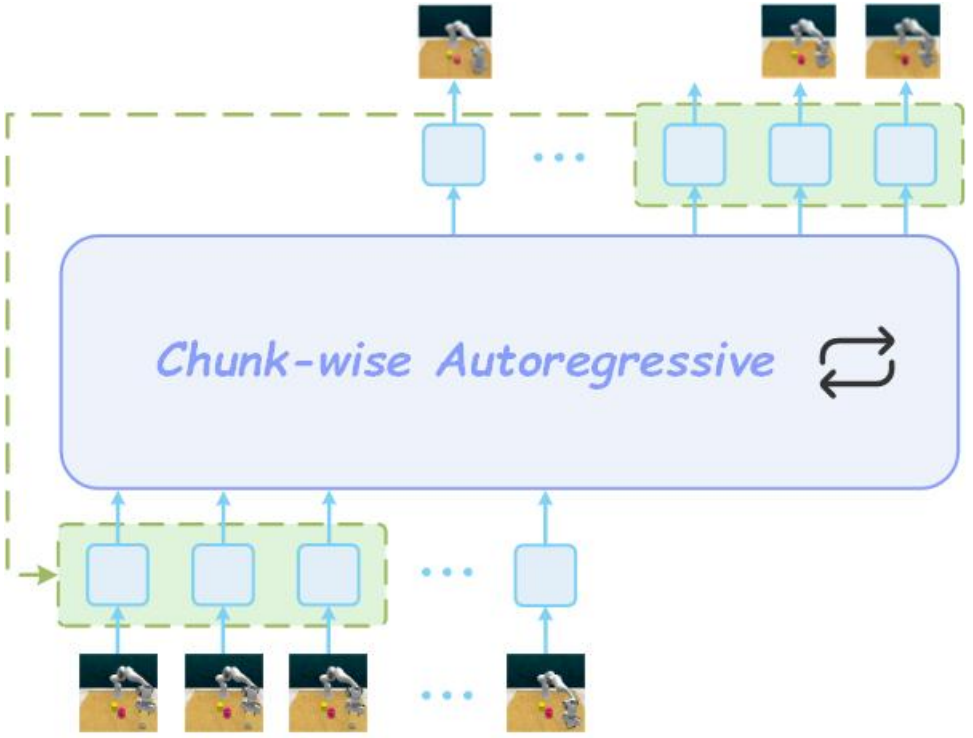




Sequential

- Definition: autoregressively simulates future over time (RNNs/SSMs such as Mamba, or autoregressive Transformers, combine with goal decomposition / CoT).
- Pros： causal consistency; simple interfaces; easy temporal alignment.
- Cons： long-horizon error accumulation; limited parallelism.

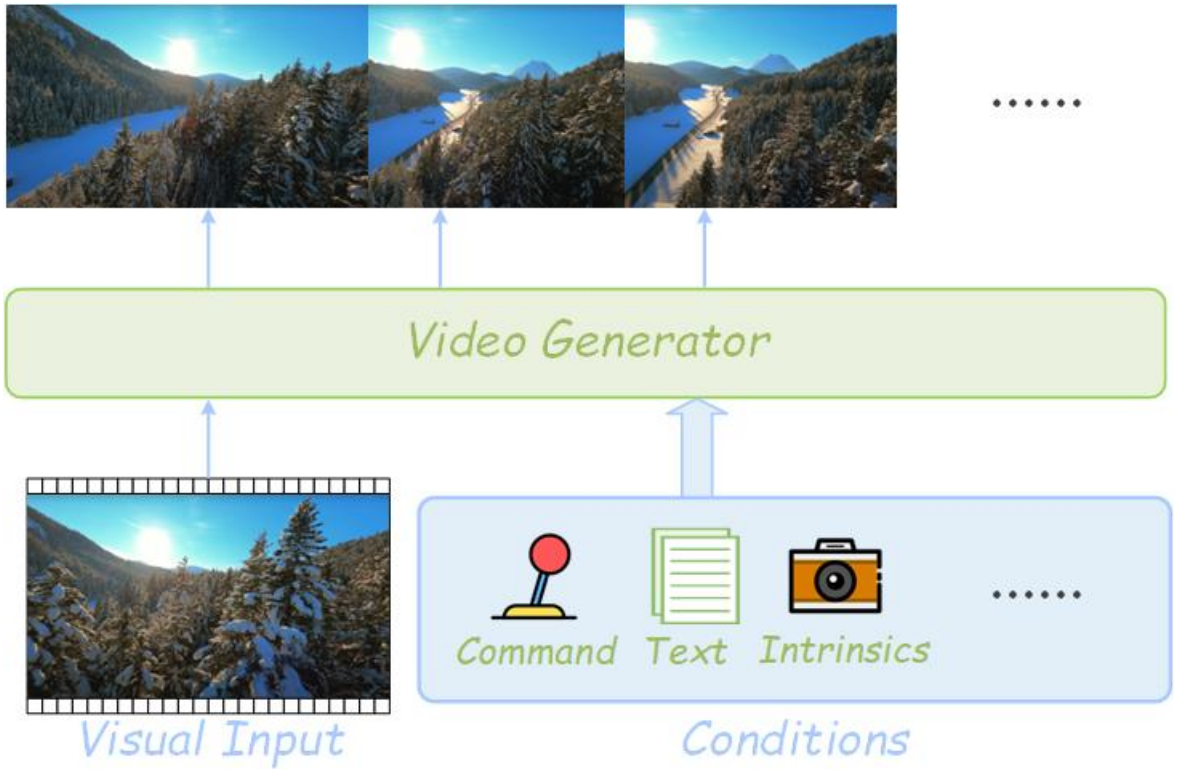


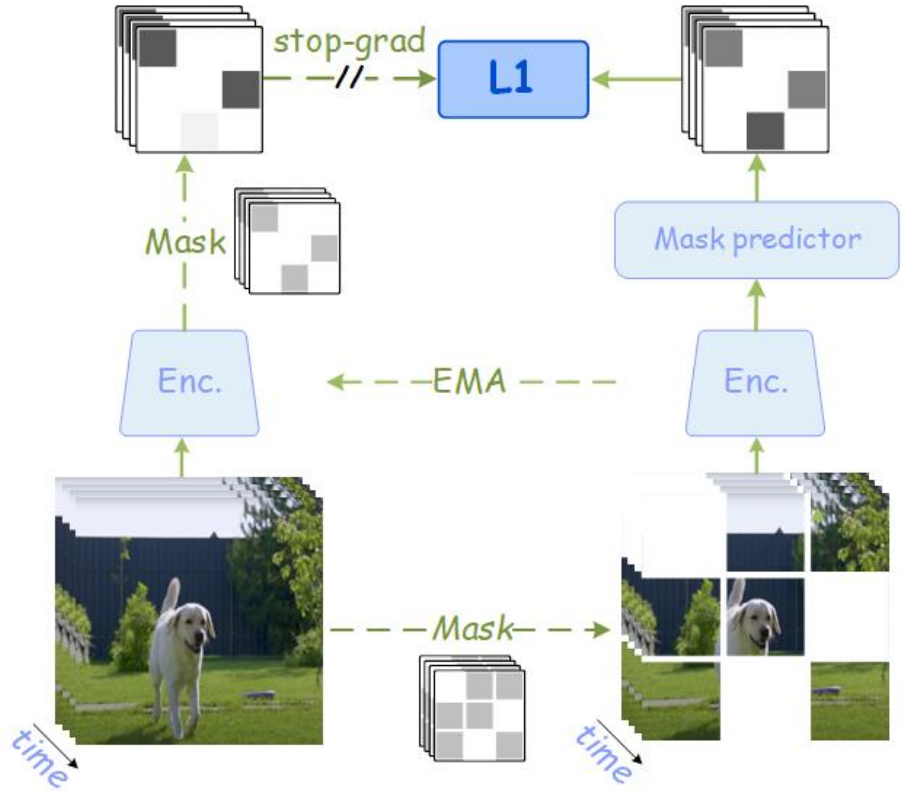


chunked autoregression improves throughput.

Global

- Definition: predicts an entire future segment in one shot (parallel diffusion/flow video, masked prediction, JEPA-style feature prediction).
- Pros: less error accumulation; higher parallel efficiency; easy to add global constraints/conditions.
- Cons: weaker closed-loop interactivity (no explicit action feedback); may miss fine-grained dynamics.

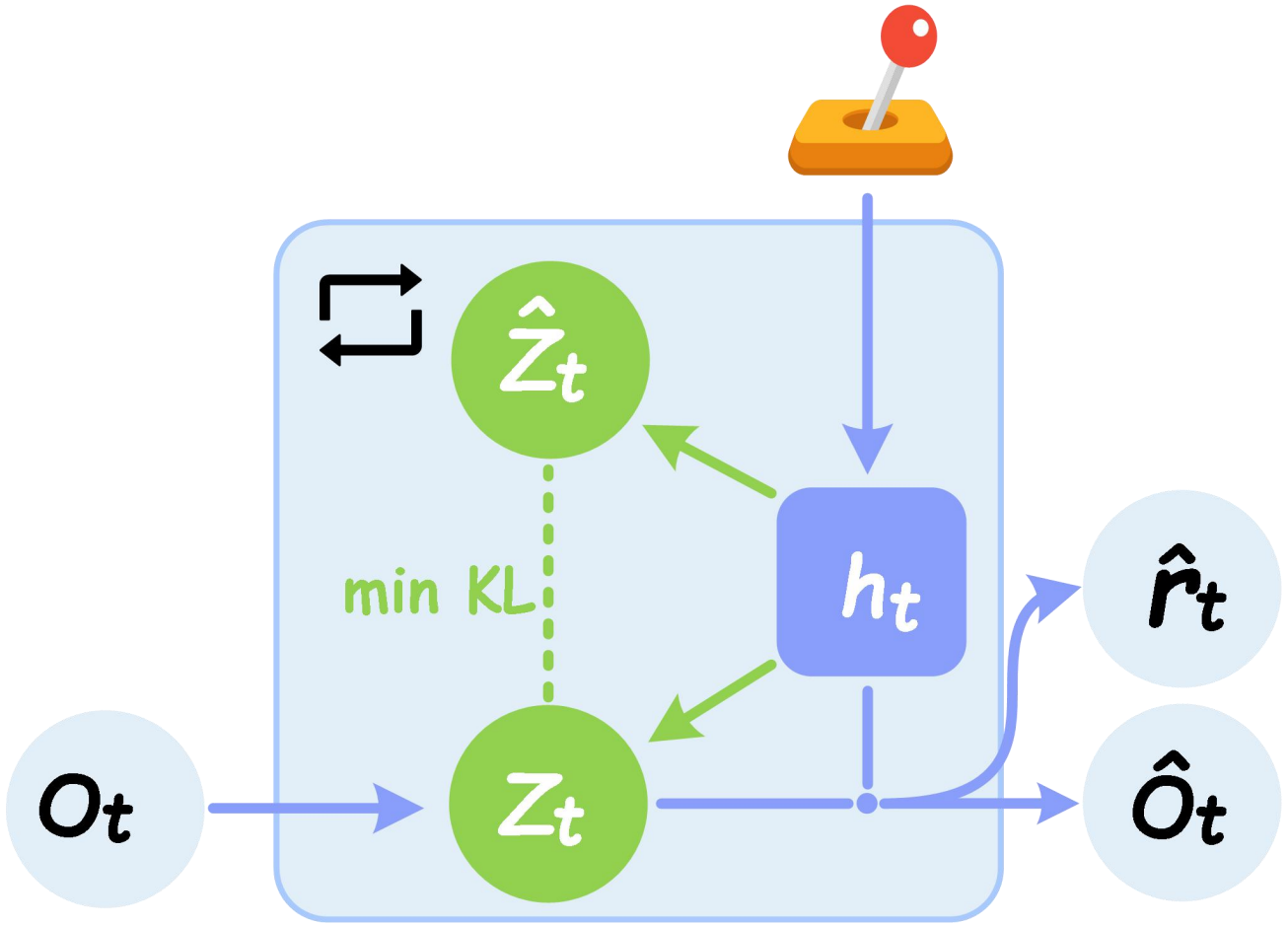




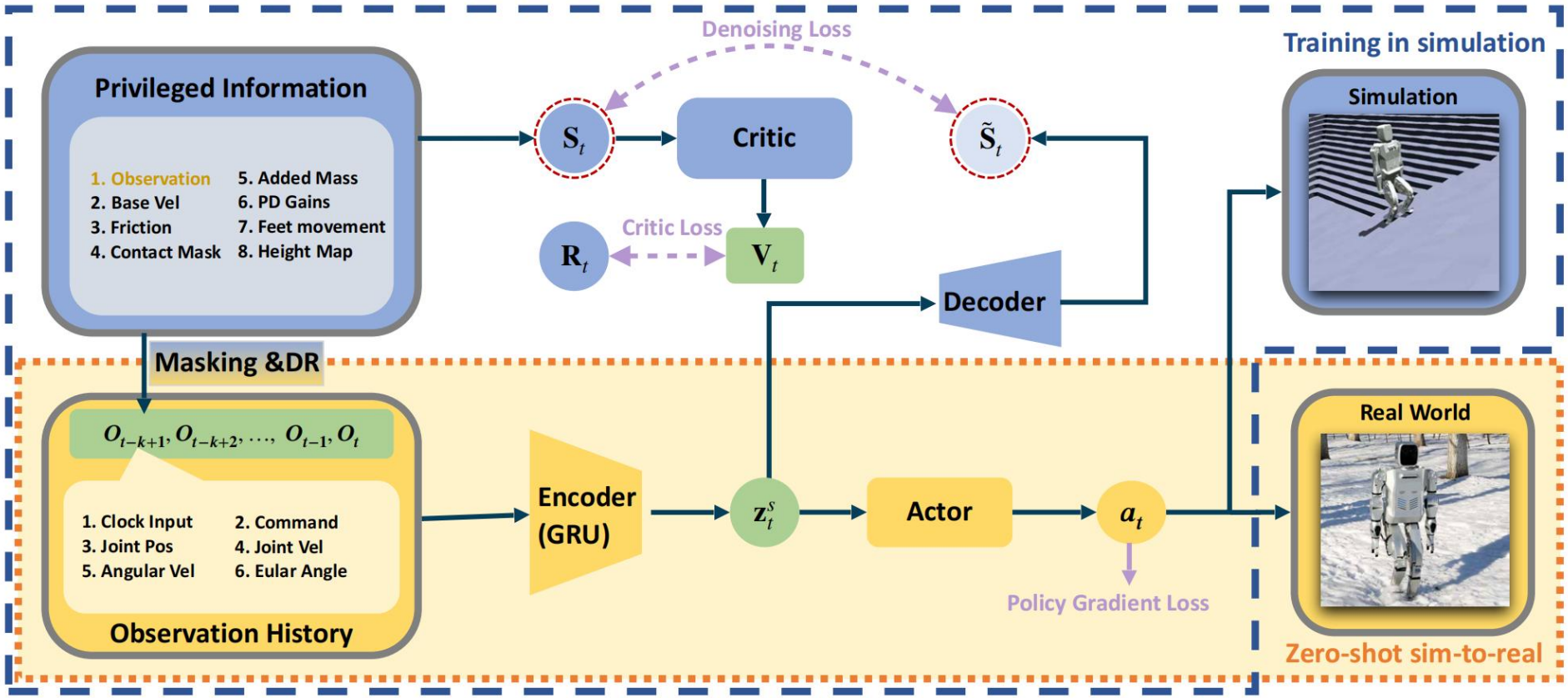
JEPA (LeCun) learns temporal causality by predicting mutually exclusive tokens in embedding space.

Global Latent Vector

- Definition: Compress the scene state into a single low-dimensional vector z_t, and build dynamics on this compact representation.
- Pros: Compute- and latency-friendly; lightweight and deployment-oriented.
- Cons: Loses fine-grained spatiotemporal details.



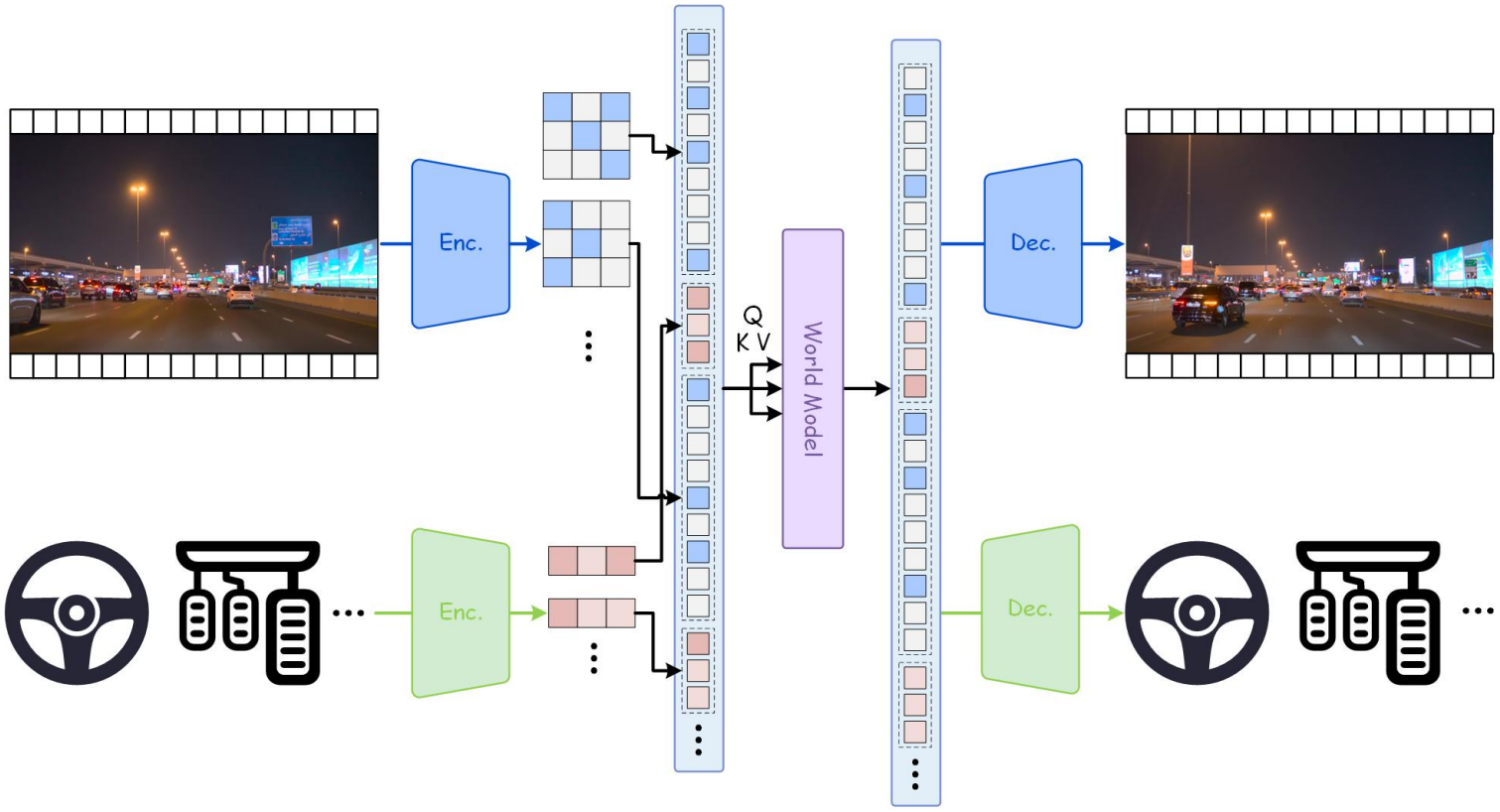
Global Latent Vector



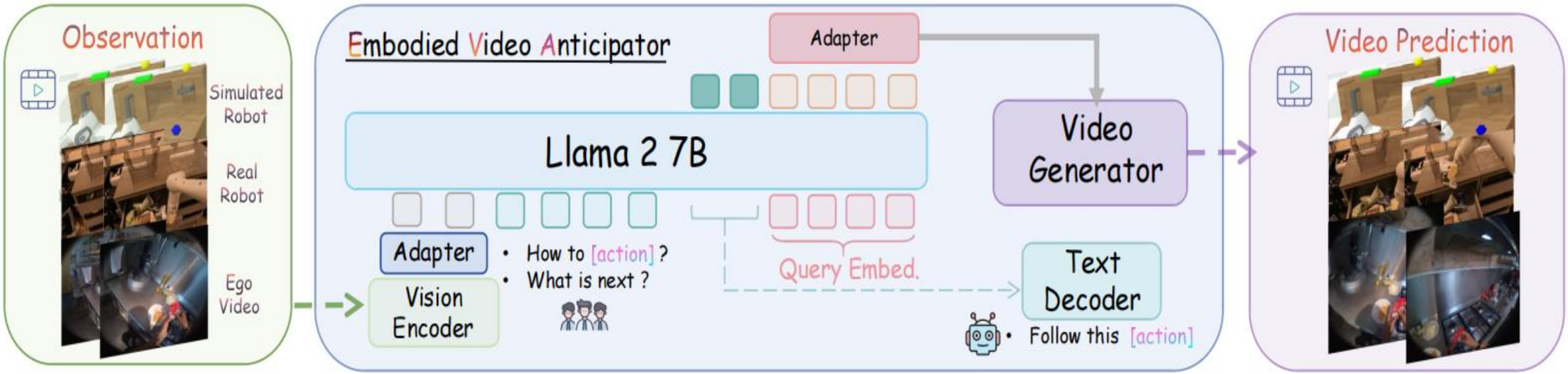
Example (RSS'24 Best Paper, DWL)： During training in simulation, DWL used privileged information to train a critic that predicts V_t， the value signal guided the actor via policy-gradient updates. At deployment, privileged inputs are unavailable, so the agent encodes the observation history with a GRU to infer a short-term latent state Z_t^s directly to actions, enabling zero-shot sim-to-real (S2R) transfer.

Token Feature Sequence

- Definition: discretizes state into a sequence of tokens, enabling dependency modeling among tokens.
- Pros: pairs naturally with attention; fine-grained for complex, multimodal scenes.
- Cons: data-hungry; large models; expensive inference.



Token Feature Sequence

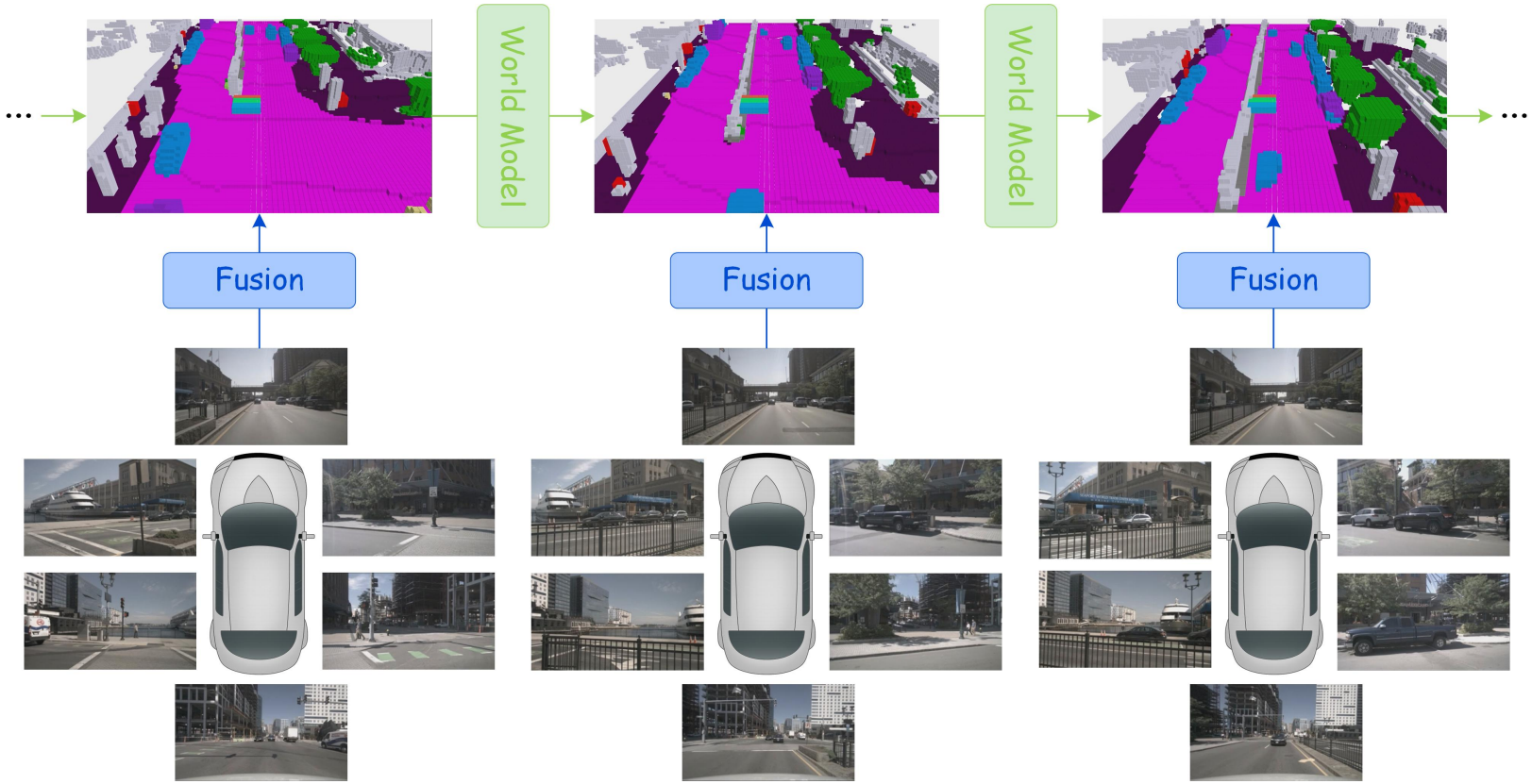


Example (ICML'25, EVA):

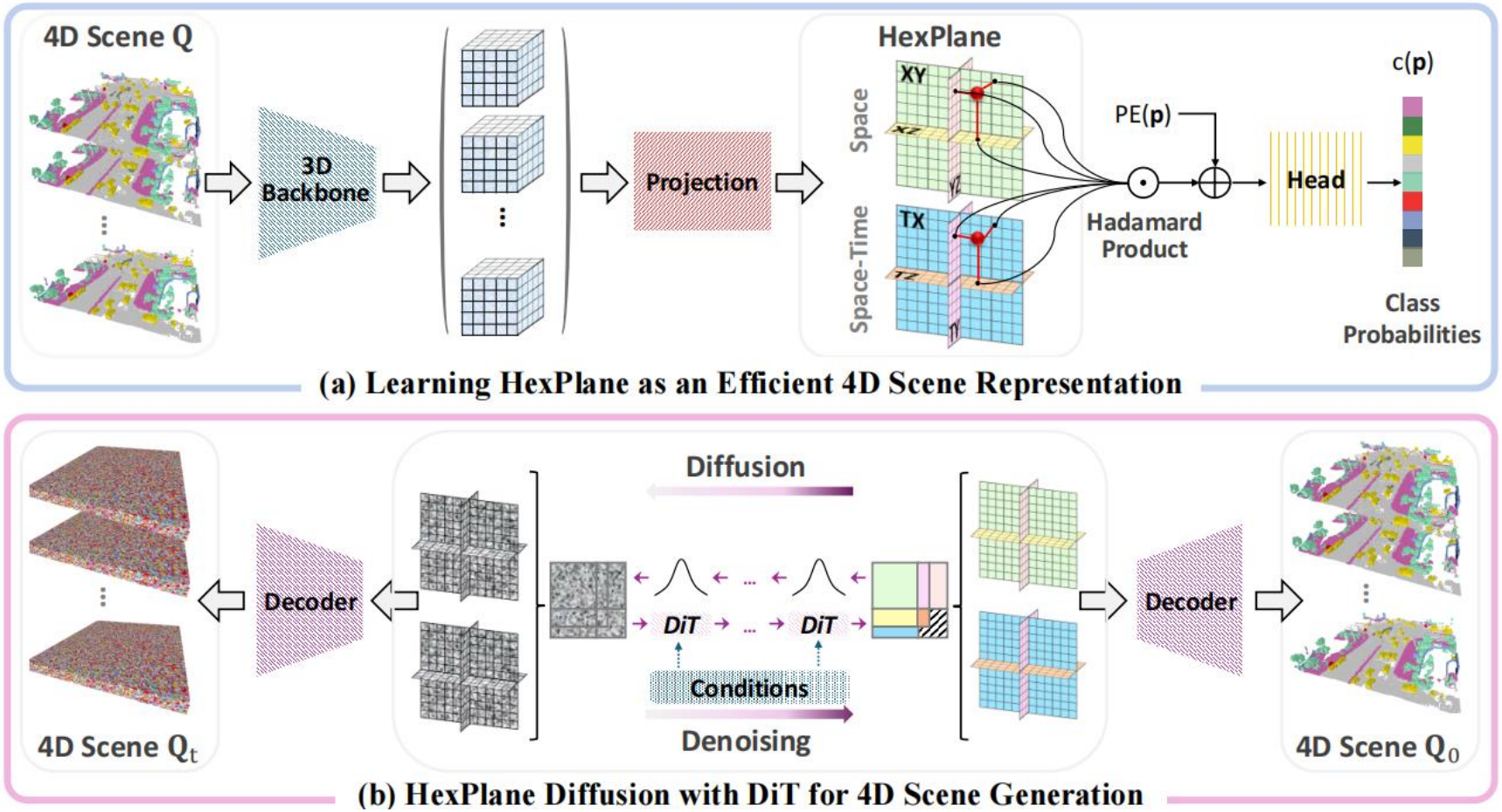
- first-person videos (sim/real) → visual tokens;
- user text → query;
- LLaMA fuses modalities on a unified token sequence;
- dual decoders output action text and future video.

Spatial Latent Grid

- Definition: injects spatial priors or encodes scenes on grids; a mainstream choice in autonomous driving.
- Pros： preserves local topology; well-suited to multi-view fusion and map generation.
- Cons： large representations; resolution constraints; weaker for unstructured environments.



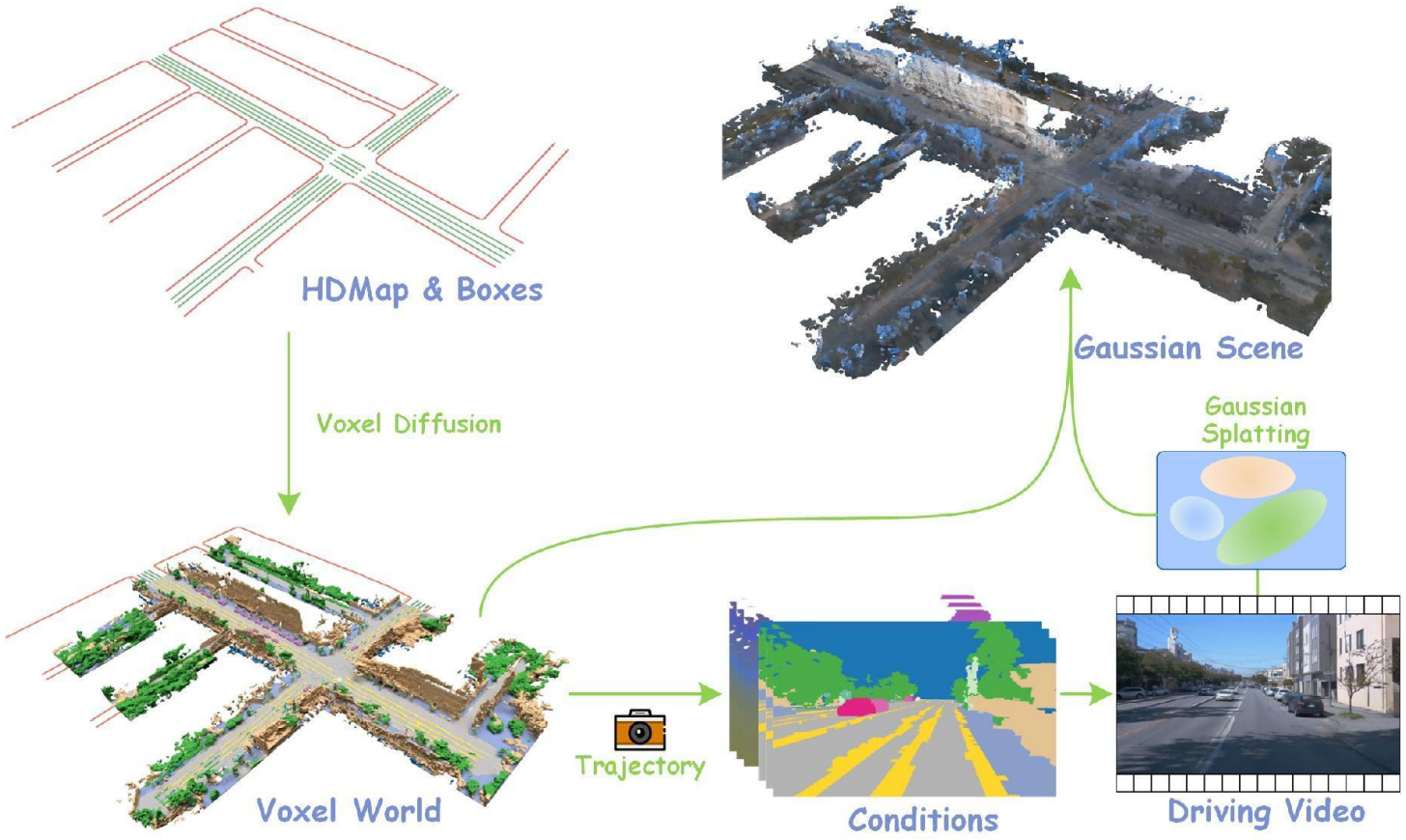
Spatial Latent Grid



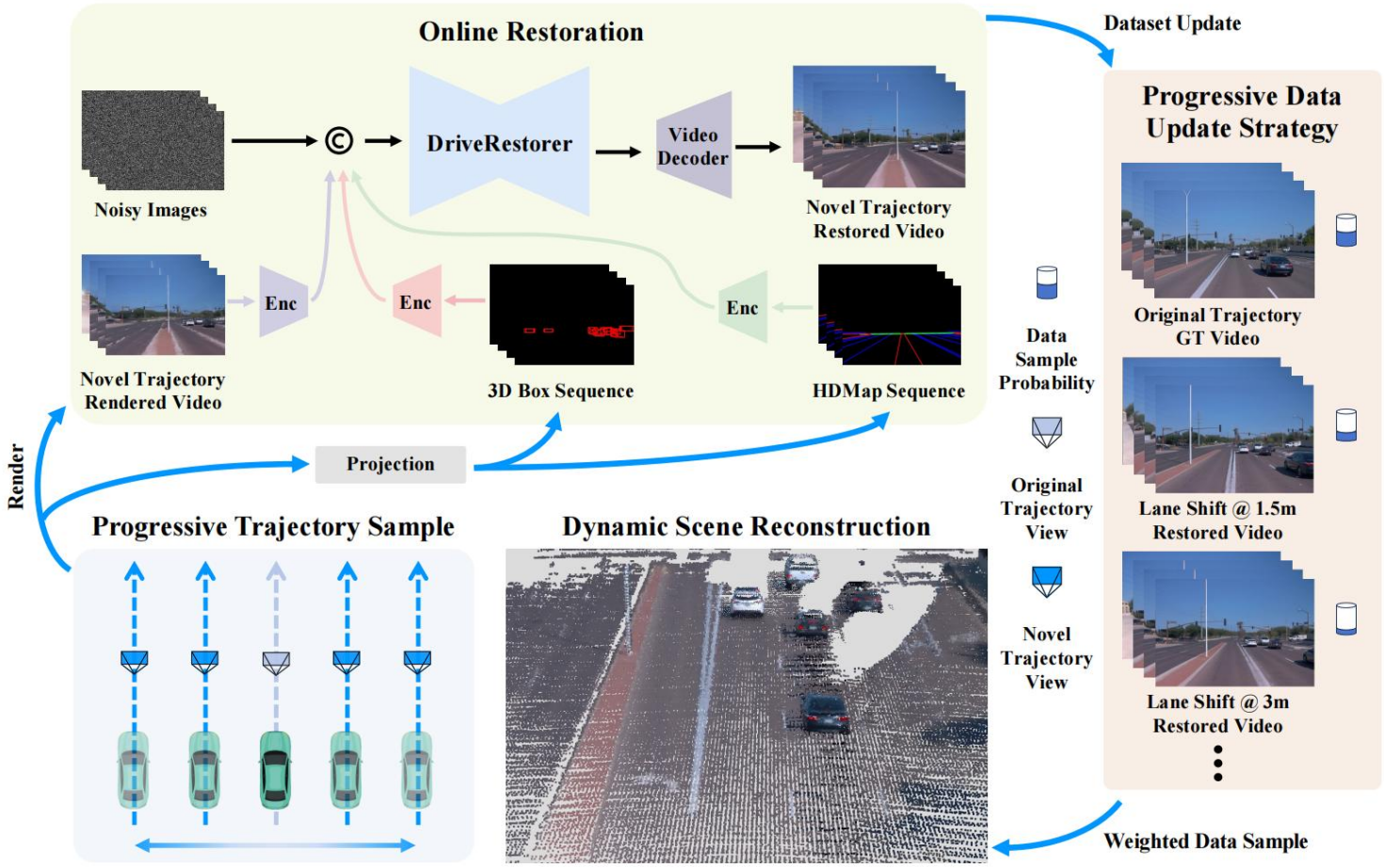
Example (ICLR'25 Spotlight, DynamicCity) encodes a 4D dynamic scene as six 2D HexPlanes (xy, yz, zx, tx, ty, tz). For each query, features from the six planes are combined element-wise, fused with positional encoding, and passed to a small head to predict occupancy probabilities. For generation, noise is added to the six subspaces and a Diffusion Transformer denoises before a decoder reconstructs the scene. This design preserves spatial topology with compact memory, enabling efficient 4D inference and synthesis.

Decomposed Rendering Representation

- Definition: decomposes scenes into renderable primitives, then uses rendering or generative pipelines to form observations.
- Pros: geometric consistency; high fidelity; supports object-level operations.
- Cons: heavy training cost; dynamic/topological changes are hard to update in real time.



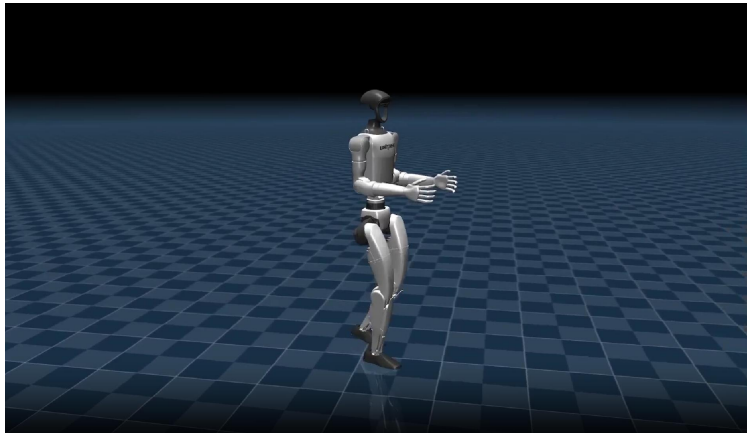
Decomposed Rendering Representation



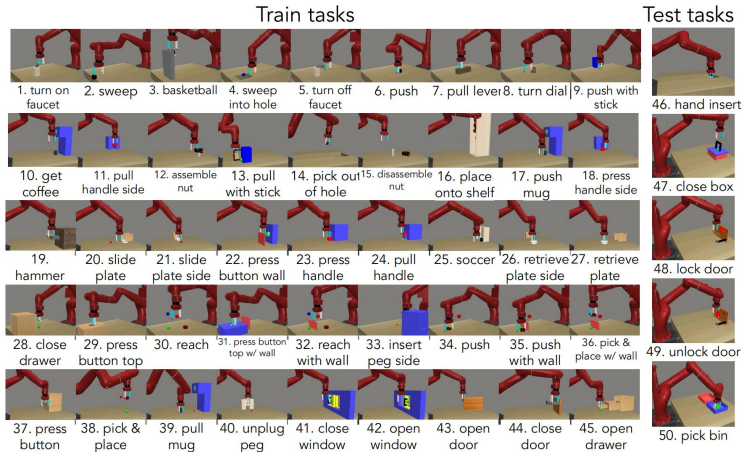
Example (CVPR'25, ReconDreamer): progressive side-offset "new trajectories" → render → online repair (DriveRestorer) → add original + repaired videos to a data pool → weighted sampling from small to large offsets → closed-loop training that stabilizes off-trajectory views.

Data Resources

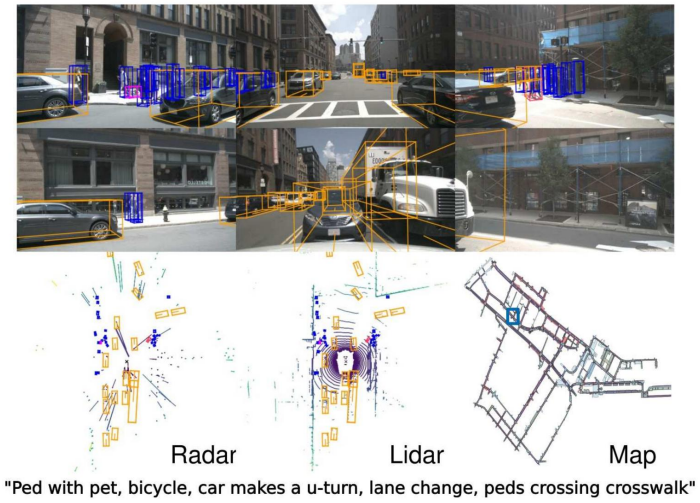
Four categories for embodied AI data: Simulation Platforms, Interactive Benchmarks, Offline Datasets, Real-world Robot Platforms.



MuJoco Simulation



Meta-world Benchmarks



nuScenes Datasets

Unitree G1 Robot

Data Resources

TABLE 3
An overview of data resources for training and evaluating embodied world models.

Category	Name	Year	Task	Input	Domain	Scale	Protocol[1]
Platform	MuJoCo [218]	2012	Continuous control	Proprio.	Sim	-	-
	CARLA [219]	2017	Driving simulation	RGB-D/Seg/LiDAR/Radar/GPS/IMU	Sim	-	✓
	Habitat [220]	2019	Embodied navigation	RGB-D/Seg/GPS/Compass	Sim	-	✓
	Isaac Gym [221]	2021	continuous control	Proprio.	Sim	-	-
	Isaac Lab [222]	2023	Robot learning suites	RGB-D/Seg/LiDAR/Proprio.	Sim	-	-
Benchmark	Atari [223]	2013	Discrete-action game	RGB/State	Sim	55+ Games	✓
	DMC [224]	2018	Continuous control	RGB/Proprio.	Sim	30+ Tasks	✓
	Meta-World [225]	2019	Multi-task manipulation	RGB/Proprio.	Sim	50 tasks	✓
	RLBench [226]	2020	Robotic manipulation	RGB-D/Seg/Proprio.	Sim	100 tasks	✓
	nuPlan [227]	2021	Driving planning	RGB/LiDAR/Map/Proprio.	Real	1.5k hours	✓
	LIBERO [228]	2023	Lifelong manipulation	RGB/Text/Proprio.	Sim	130 tasks	✓
Dataset	SSv2 [229]	2018	Video-action understanding	RGB/Text	Real	220k videos	169k/24k/27k
	nuScenes [230]	2020	Driving perception	RGB/LiDAR/Radar/GPS/IMU	Real	1k scenes	700/150/150
	Waymo [231]	2020	Driving perception	RGB/LiDAR	Real	1.15k scenes	798/202/150
	HM3D [232]	2021	Indoor navigation	RGB-D	Real	1k scenes	800/100/100
	RT-1 [233]	2022	Real-robot manipulation	RGB/Text	Real	130k+ trajectories	-
	Occ3D [234]	2023	3D occupancy	RGB/LiDAR	Real	1.9k scenes	600/150/150; 798/202/-
	OXE [235]	2024	Cross-embodiment pretraining	RGB-D/LiDAR/Text	Real	1M+ trajectories	-
	OpenDV [90]	2024	Driving video pretraining	RGB/Text	Real	2k+ hours	-
	VideoMix22M [14]	2025	Video pretraining	RGB	Real	22M+ samples	-
Robot	Franka Emika [236]	2022	Manipulation	Proprio.	Real	-	-
	Unitree Go1 [237]	2021	Quadruped locomotion	RGB-D/LiDAR/Proprio.	Real	-	-
	Unitree G1 [238]	2024	Humanoid manipulation	RGB-D/LiDAR/Proprio./Audio	Real	-	-

[1] **Protocol**: For interactive benchmarks, a check mark (✓) indicates available evaluation protocols. For datasets, it indicates official data splits are provided.

Evaluation Metrics

Metrics span multiple abstraction levels:

Level	Meaning	Metrics	Use
pixel	Visual realism / temporal coherence	FID, FVD, SSIM, PSNR	Video generation, reconstruction perceptual quality
State	Geometric / semantic / trajectory consistency	mIoU, mAP, ADE/FDE, BEV-IoU	Occupancy & semantics, planning consistency
Task	Ability to achieve the goal	Success, Return, Collision	Closed-loop control, end-to-end evaluation

Recent evaluations should include physical compliance and causal consistency to complement traditional metrics.

Challenges & Future Directions

Challenges:

1. Data & Evaluation: Lack of unified, large-scale multimodal datasets and metrics that assess physical compliance and causal consistency.
2. Real-time Performance: SOTA models (e.g., DiT, VLMs) are computationally heavy, making real-time control difficult; speeding up without sacrificing accuracy is critical.
3. Modeling Trade-offs: Finding the right balance between sequential simulation vs. global prediction and among different spatial representations.

Future Directions:

1. Unified Benchmarks: Build cross-domain datasets that explicitly evaluate physical consistency.
2. Efficient Inference: Apply model compression (quantization, pruning, distillation, sparsity) and adopt more efficient temporal models (e.g., Mamba).
3. Strategy & Balance: Use hybrid schemes—sequential + global modeling, explicit memory, and CoT-style task decomposition—to support long-horizon reasoning with coarse-to-fine refinement, reducing error accumulation and feedback loops.

Thanks