Opinion: A Unified World Model is the cornerstone for integrating perception, reasoning, and decision-making in embodied AI

Anonymous Authors
Affiliation withheld for double-blind review

Abstract

We argue that a unified world model is a foundational mechanism for integrating perception, reasoning, and decision-making in embodied agents. Concretely, we define a visuo-conceptual, reconstructive latent state learned jointly with dynamics and policy that connects pixel-grounded 2D/3D scene understanding to language and action. By enabling internal simulation with decodable futures, such a model supports long-horizon planning, cross-modal knowledge transfer from multimodal LLMs, and end-to-end optimization in closed-loop settings. We synthesize converging evidence from world-model reinforcement learning, vision–language–action systems, diffusion-based control, and applications in robotics, autonomous driving, and open-ended environments. We outline a concrete research agenda: (i) a bidirectional scene memory that decodes to images, video, and affordance fields; (ii) differentiable imagination for evaluating and selecting actions; (iii) grounding language priors in latent 3D and temporal structure; and (iv) rigorous sim-to-real evaluation with uncertainty. We distill design patterns, failure modes, and actionable benchmarks to accelerate progress.

1 Why a unified world model is the cornerstone

Classical modular pipelines place vision, mapping, prediction, and planning behind brittle interfaces. They under-specify what must be preserved for control and make credit assignment across modules difficult. In contrast, a *unified world model* learns predictive latent states that support imagination and control, with strong evidence for sample efficiency and long-horizon competence in diverse domains (e.g., Dreamer-style agents) [1]. The crucial step is a visuo-conceptual latent that (1) is optimized for decision-making and (2) remains *decodable* to pixels, depth, flow, occupancy, or video, so the agent can both *think ahead* and *show its work*. Reconstruction—including action-conditioned video synthesis—acts as a powerful self-supervision signal that shapes temporally consistent, controllable latents.

Design principle. Treat representation $\mathbf{z}_t$ as an *actionable scene memory*: it fuses multi-view 2D/3D cues, task context, and language priors; it can be rolled forward by a dynamics model $p(\mathbf{z}_{t+1}|\mathbf{z}_t, \mathbf{a}_t)$ and decoded to images/video $p(\mathbf{x}_t|\mathbf{z}_t)$, affordance/goal fields, and safety costs. This collapses "see" and "simulate" into the same substrate, enabling tight perception–control coupling.

2 Unified world model meets MLLMs and VLAs: knowledge meets grounding

Recent embodied MLLM/VLA systems demonstrate that injecting web-scale knowledge and reasoning into grounded policies improves generalization and instruction following. PaLM-E streams

Preprint. Under review.

visual tokens into a pretrained language model to reason over scenes and plan robot behaviors [2]; RT-2 tokenizes actions and co-trains with internet V+L tasks to yield emergent semantic control [3]; open-source VLAs (e.g., OpenVLA) leverage Open X-Embodiment for cross-robot transfer [4, 5]. These models benefit from language priors yet still lack *explicit* internal simulation. Our thesis: **marry VLA semantics with reconstructive world models**. Language supplies abstract goals and commonsense; the world model supplies temporally coherent, physically plausible futures on which to plan.

3 From observation to simulation to decision: an integrated loop

End-to-end optimization aligns the encoder, dynamics, and policy with downstream reward/costs, avoiding error cascades and letting intermediate features emerge task-adaptively [6]. Concretely:

1. **Encode.** Multi-camera video (and proprioception) $\rightarrow$ latent scene $\mathbf{z}_t$ with 3D inductive biases (BEV, slots, Gaussian splats, or neural radiance).
2. **Imagine.** Roll out imagined futures $\{\mathbf{z}_{t:t+H}\}$ under candidate actions; decode to video for self-supervision and to occupancy/flow/contact to query constraints.
3. **Evaluate.** Score imagined futures with differentiable value/cost heads (task, safety, comfort), optionally guided by LLM-inferred subgoals or constraints.
4. **Act.** Optimize actions (gradient-based planning or actor) and execute; update memory and uncertainty; learn from both real and imagined data.

Diffusion policies are a natural fit for multi-modal action distributions and can be nested in this loop as the action sampler or as the video/trajectory generator [7]. For driving and robotics, unified transformers with streaming histories (e.g., DriveTransformer/UniAD line) show how to couple perception, prediction, and planning around shared features [8, 9].

4 Applications

Robotics. Household and mobile manipulation benefit from VLA priors (object affordances, tool use) combined with reconstructive latents to plan contact-rich behaviors. RT-2-style action tokenization can coexist with latent MPC over imagined futures. Open X-Embodiment enables cross-robot pretraining; the world model adds temporally grounded control [3, 4].

Autonomous driving. Planning-oriented stacks increasingly unify tasks; adding action-conditioned video generation and 3D occupancy decoding provides closed-loop counterfactuals for safe exploration and contingency planning [8–10]. Generative simulators learned from data bridge sim$\leftrightarrow$real when paired with uncertainty and causal interventions [11].

Open-ended environments (e.g., Minecraft). LLM-powered agents like Voyager show the value of curriculum, skill libraries, and code synthesis [12]. Adding a reconstructive world model yields more stable long-horizon execution and safety validation before acting.

5 Practical implementation guidelines and ablation protocol

We operationalize the proposed approach with five tightly coupled components that can be implemented incrementally and ablated systematically. First, learn a spatiotemporal latent with explicit 3D structure (e.g., BEV or object/slot-centric) that serves as a visuo–conceptual scene memory, jointly optimized with reconstruction losses (RGB, depth, optical flow) and control-aware auxiliaries (affordance/goal heatmaps). Second, enable differentiable imagination by rolling out H-step latent futures under candidate actions and decoding action-conditioned video; regularize with physics priors and differentiable consistency checks to suppress implausible trajectories. Third, couple the world model with a VLA/MLLM planner by summarizing the scene into compact tokens, using the LLM to propose subgoals and constraints, and grounding them back into differentiable masks or penalties so that credit propagates to perception and dynamics. Fourth, use diffusion both for multi-modal action proposals and for controllable future-frame generation, conditioning on the latent state and textual

Lane	Inputs	State update	Decoders/Artifacts	Decision signal
Perception	multi-cam video, proprio	$\mathbf{z}_t \leftarrow E_\theta(\mathbf{x}_t)$	depth/flow/occupancy	$\mathcal{L}_{\text{rec}} + \lambda_{\text{aff}}\,\mathcal{L}_{\text{aff}}$
World model	$\mathbf{z}_t, \mathbf{a}_t$	$\mathbf{z}_{t+1} \sim f_\theta(\mathbf{z}_t, \mathbf{a}_t)$	video $\hat{\mathbf{x}}$, affordances	$\mathcal{L}_{\text{dyn}} + \mathcal{L}_{\text{roll}}$
Reasoning	goal text g, summary s_t	$(\tilde{g}, \phi) \leftarrow \text{MLLM}(s_t, g)$	constraints/subgoals	feasibility, consistency
Decision	$\mathbf{z}_t, \phi$	$\mathbf{a}_{t:t+H-1} \leftarrow \pi(\cdot \mid \mathbf{z}_t, \phi)$	—	value V_ψ, risk c

Table 1: End-to-end lanes and their interfaces. ϕ denotes differentiable constraints grounded from language; $\mathcal{L}_\cdot$ are training losses.

Head	Symbol	Training signal	Used for	Notes
Reward/Value	V_ψ, Q_ψ	TD/λ-returns	plan/policy	end-to-end credit [6]
Safety cost	c_s	counterfact. risk	veto/penalty	closed-loop eval
Comfort	c_c	kinematic priors	penalty	jerk/lat-acc bounds
Feasibility	ϕ	LLM-grounded labels	constraint	mask latent/action [3, 4]
Uncertainty	σ	ensembles/variances	risk-aware	exploration/safety

Table 2: Scoring/constraint heads supervising imagination and decision.

goals to capture real-world multi-modality. Fifth, estimate epistemic uncertainty with ensembles or variational methods, veto actions whose imagined rollouts exceed risk thresholds, and close the sim-to-real gap via small-scale real-data finetuning and domain-randomized generative augmentation. For transparency and reproducibility, report ablations over: (a) reconstruction vs. control auxiliaries, (b) imagination horizon H and presence of video decoding, (c) LLM-guided constraints on/off, (d) diffusion vs. non-diffusive policies/decoders, and (e) uncertainty heads and safety veto mechanisms.

6 Failure modes and evaluation methodology

Robustness hinges on three recurrent failure modes—hallucinated futures, language–vision misalignment, and cross-loop credit assignment—and we evaluate progress with targeted, comparable metrics. To detect hallucinated futures, compare imagined rollouts to realized trajectories using per-step reconstruction and dynamics consistency (PSNR/SSIM/FVD for decoded video; occupancy/flow errors; constraint-violation rates) and report control impact via planning regret and the open-loop vs. closed-loop gap. To quantify language–vision alignment, measure referential grounding accuracy between latent slots and text, instruction-conditioned success and latency, and calibration (e.g., ECE) of feasibility predictions; stress-test with disambiguation and counterfactual phrasing. To probe credit assignment across perception–dynamics–policy, track value/policy consistency (TD error, value calibration), gradient-through-imagination effectiveness (improvement of actor–critic MPC over behavior cloning), and data efficiency as the planning horizon scales. We standardize reporting across domains: robots (mobile manipulation with language goals and safety constraints; success rate, constraint satisfaction, episode length), driving (Bench2Drive and nuPlan; infraction rate, comfort, counterfactual risk), and Minecraft-like open-ended tasks (skill acquisition rate, safety-gate pass rate, ablations over imagination horizon). All metrics are logged on both real data and model-generated rollouts to attribute gains to representation, imagination, or decision components.

7 Tabular summaries and formalization

System components and interfaces. We summarize the proposed end-to-end loop using a structured table instead of figures. It captures the inputs/outputs, latent interactions, and decision signals of each lane.

Rollout scoring heads and constraints. We list the heads used to score imagined futures and how they contribute to optimization.

Model	V	L	A	Internal sim.	E2E dec.	Notes
PaLM-E [2]	✓	✓	✓	○	○	Visual tokens into LLM; positive transfer
RT-2 [3]	✓	✓	✓	○	✓	Actions as tokens; emergent semantic control
OpenVLA [5]	✓	✓	✓	○	✓	Open-source; Open-X pretraining
Gato [13]	✓	✓	✓	○	○	Generalist sequence policy, limited planning
Dreamer-v3 [1]	✓	○	✓	✓	✓	Latent imagination; diverse domains incl. Minecraft
DriveTransformer/UniAD [8, 9]	✓	○	✓	○ → ✓	✓	Unified stack; trending to generative planning

Table 3: Support for end-to-end decision optimization. V/L/A mark vision/language/action modalities. ✓=fully supported; ○=limited/implicit; ○ → ✓=evolving capability.

Formal objective and planning. We formalize training and control with reconstructive and decision-aware terms:

$$\min_{\theta,\phi,\psi,\eta} \sum_t \Big[\underbrace{\mathcal{L}_{\text{rec}}(\mathbf{x}_t, D_\phi(\mathbf{z}_t))}_{\text{reconstruction}} + \lambda_{\text{aff}}\, \mathcal{L}_{\text{aff}}(\mathbf{z}_t) + \lambda_{\text{dyn}}\, \mathcal{L}_{\text{dyn}}\big(\mathbf{z}_{t+1}, f_\theta(\mathbf{z}_t, \mathbf{a}_t)\big) \Big] + \lambda_{\text{val}}\, \mathcal{L}_{\text{val}}(\psi) + \lambda_{\text{align}}\, \mathcal{L}_{\text{align}}(\mathbf{z}_t; g, \eta). \tag{1}$$

Rollout-based control optimizes actions against value and costs under constraints:

$$\mathbf{a}^*_{t:t+H-1} \;=\; \arg\max_{\mathbf{a}_{t:t+H-1}} \; \mathbb{E}\Big[\sum_{k=0}^{H-1} \gamma^k \big(r(\mathbf{z}_{t+k}, \mathbf{a}_{t+k}) - \beta_s\, c_s - \beta_c\, c_c \big) \Big] \quad \text{s.t. } \phi(\mathbf{z}_{t+k}, \mathbf{a}_{t+k}) \le 0,\ \forall k. \tag{2}$$

Language-grounded constraints are compiled into differentiable masks or penalties:

$$\phi(\mathbf{z}, \mathbf{a}) \;:=\; g_\eta(\mathbf{z};\; g_{\text{text}}). \tag{3}$$

Diffusion policies [7] can parameterize π or serve as controllable decoders for $\hat{\mathbf{x}}_{t:t+H}$.

8 Position and outlook

Position. The most promising path to robust, general embodied agents is a *unified world model*—the cornerstone that integrates perception, reasoning, and decision-making—trained end to end with reconstructive self-supervision, LLM-guided reasoning, and decision-aware objectives. This closes the loop between seeing, imagining, and acting.

Immediate opportunities. (1) *Video-grounded planning*: train value heads on decoded futures instead of raw features; (2) *Language-to-latent constraints*: compile textual rules/subgoals into differentiable costs or masks; (3) *Diffusion everywhere*: unify scene and action generation under shared conditioning; (4) *Risk-sensitive imagination*: uncertainty-aware veto and scenario stress-testing before execution.

Limitations. Training stability, compute, and safety validation remain challenging; however, hybrid actor–critic MPC, curriculum from LLM planners, and uncertainty-aware decoders provide practical levers.

Reproducibility notes for tables. Tab. 1 can be instantiated from a trained model by enumerating module IOs and recording which losses supervise each artifact. Tab. 2 is derived by exporting head definitions and their training targets; the objective and constraints are directly computed from logged rollouts and labels.

References

[1] D. Hafner, J. Pasukonis, J. Ba, and T. Lillicrap, "Mastering diverse domains through world models," *arXiv*, 2023. arXiv:2301.04104.

[2] D. Driess, F. Xia, M. S. M. Sajjadi, C. Lynch, A. Chowdhery, B. Ichter, A. Wahid, J. Tompson, Q. Vuong, T. Yu, W. Huang, Y. Chebotar, P. Sermanet, D. Duckworth, S. Levine, V. Vanhoucke, K. Hausman, A. Zeng, I. Mordatch, P. Florence, M. Toussaint, and K. Greff, "Palm-e: An embodied multimodal language model," *arXiv*, 2023. arXiv:2303.03378.

[3] A. Brohan, N. Brown, J. Carbajal, Y. Chebotar, X. Chen, K. Choromanski, T. Ding, D. Driess, A. Dubey, C. Finn, P. Florence, C. Fu, M. Gonzalez Arenas, K. Gopalakrishnan, K. Han, K. Hausman, A. Herzog, J. Hsu, B. Ichter, A. Irpan, N. Joshi, R. Julian, D. Kalashnikov, Y. Kuang, I. Leal, L. Lee, T.-W. E. Lee, S. Levine, Y. Lu, H. Michalewski, I. Mordatch, K. Pertsch, K. Rao, K. Reymann, M. Ryoo, G. Salazar, P. Sanketi, P. Sermanet, J. Singh, A. Singh, R. Soricut, H. Tran, V. Vanhoucke, Q. Vuong, A. Wahid, S. Welker, P. Wohlhart, J. Wu, F. Xia, T. Xiao, P. Xu, S. Xu, T. Yu, and B. Zitkovich, "Rt-2: Vision-language-action models transfer web knowledge to robotic control," *arXiv*, 2023. arXiv:2307.15818.

[4] Open X-Embodiment Collaboration, "Open x-embodiment: Robotic learning datasets and rt-x models," *arXiv*, 2023. arXiv:2310.08864.

[5] M. J. Kim, K. Pertsch, S. Karamcheti, T. Xiao, A. Balakrishna, S. Nair, R. Rafailov, E. P. Foster, P. R. Sanketi, Q. Vuong, T. Kollar, B. Burchfiel, R. Tedrake, D. Sadigh, S. Levine, P. Liang, and C. Finn, "Openvla: An open-source vision-language-action model," *arXiv*, 2024. arXiv:2406.09246.

[6] Y. LeCun, "A path towards autonomous machine intelligence." OpenReview preprint, June 2022.

[7] C. Chi, Z. Xu, S. Feng, E. Cousineau, Y. Du, B. Burchfiel, R. Tedrake, and S. Song, "Diffusion policy: Visuomotor policy learning via action diffusion," *arXiv preprint arXiv:2303.04137*, 2023.

[8] X. Jia, J. You, Z. Zhang, and J. Yan, "Drivetransformer: Unified transformer for scalable end-to-end autonomous driving," in *International Conference on Learning Representations (ICLR)*, 2025.

[9] Y. Hu, J. Yang, L. Chen, K. Li, C. Sima, X. Zhu, S. Chai, S. Du, T. Lin, W. Wang, L. Lu, X. Jia, Q. Liu, J. Dai, Y. Qiao, and H. Li, "Planning-oriented autonomous driving," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, IEEE/CVF, 2023.

[10] D. Chen, V. Koltun, and P. Krähenbühl, "Learning to drive from a world on rails," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 15590–15599, IEEE/CVF, 2021.

[11] Y. Hu, S. Chai, Z. Yang, J. Qian, K. Li, W. Shao, H. Zhang, W. Xu, and Q. Liu, "Solving motion planning tasks with a scalable generative model," in *Computer Vision – ECCV 2024*, Lecture Notes in Computer Science, pp. 386–404, Springer Nature Switzerland, 2024.

[12] G. Wang, Y. Xie, Y. Jiang, A. Mandlekar, C. Xiao, Y. Zhu, L. Fan, and A. Anandkumar, "Voyager: An open-ended embodied agent with large language models," *arXiv preprint arXiv:2305.16291*, 2023.

[13] S. Reed, K. Zolna, E. Parisotto, S. Gomez Colmenarejo, A. Novikov, G. Barth-Maron, M. Gimenez, Y. Sulsky, J. Kay, J. T. Springenberg, T. Eccles, J. Bruce, A. Razavi, A. Edwards, N. Heess, Y. Chen, R. Hadsell, and O. Vinyals, "A generalist agent," *arXiv*, 2022. arXiv:2205.06175.