CLAW: Learning Continuous Latent Action World Models via Adversarial Latent Regularization

Tewodros W. Ayalew[1] Matthew Jeung[1] Samuel Wheeler[3] Xiao Zhang[1] Andre de la Cruz Arce[1] Kaylene Stocking[2] Michael Maire[1] Matthew R. Walter[2]

[1]University of Chicago [2]Toyota Technological Institute at Chicago [3]Argonne National Laboratory

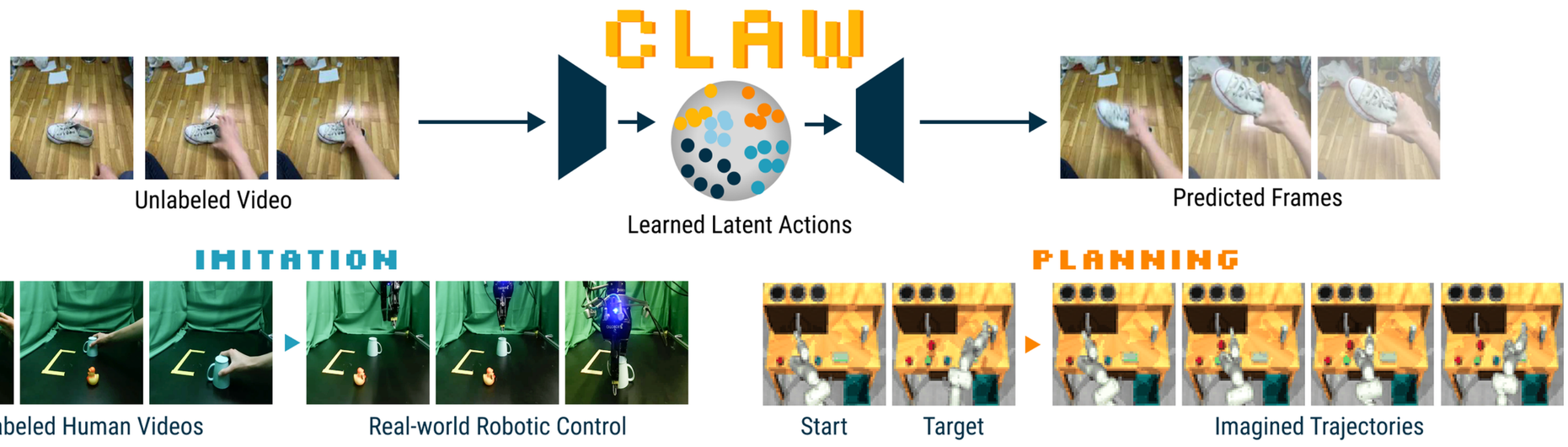

Figure 1. From unlabeled video, CLAW learns continuous latents + a world model for imitation and goal-directed planning.

TL;DR

CLAW jointly learns continuous latent actions and a diffusion world model from **action-free video** using adversarial latent regularization—enabling imitation from observation [2], goal-directed planning [5], and real-robot policy learning [13] *without* retraining the world model.

Motivation

Action-conditioned world models [8] struggle to scale to unlabeled internet video and to transfer across embodiments when actions are agent-specific [13, 7]. Latent action world models (LAWMs) infer actions from observation alone [2, 10], but continuous latents often **leak** future pixels (so the dynamics model "cheats") or **collapse** to uninformative codes. CLAW addresses both via end-to-end joint training with gradient-reversal regularization [6], yielding reusable latents for ILfO [12], planning, and latent-action policy pretraining.

Method

We learn a world model $P(o_{i+1} \mid o_i, z_i)$ from action-free video $\mathcal{V} = \{o_{1:T}\}^N$ by jointly training a continuous Latent Action Model (LAM) and a diffusion forward model—no action labels, no VQ codebook.

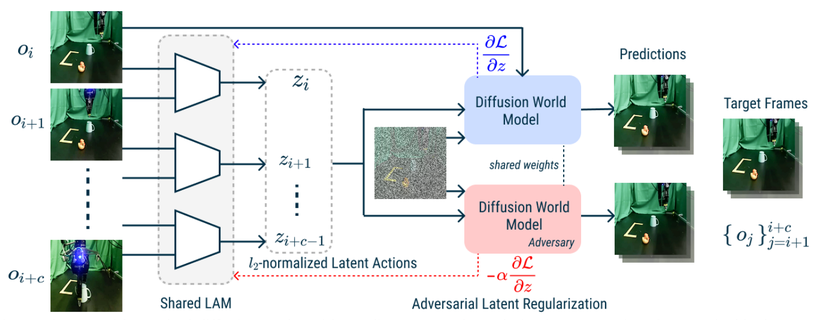

Figure 2. Shared LAM encodes transitions; dual diffusion paths with gradient reversal regularize latents against future-pixel leakage while preserving controllable dynamics.

Latent Action Model (LAM)

Similar to an inverse dynamics model [10], a ViT [4] reads consecutive frames (o_i, o_{i+1}) (patch tokens + spatial/temporal embeddings + an action token) and maps them to $z_i \in \mathbb{R}^n$, then ℓ_2-normalized onto the unit sphere.

Diffusion world model

A UViT [1] denoises a horizon of future frames given the initial observation o_i and latent sequence $\{z_j\}$. Training uses a variance-preserving diffusion process on image patches; a *causal* attention mask lets each noised target attend only to its own frame, the start frame, and actions up to its timestep—blocking peeking at future actions or later observations so sequential transfer stays well-defined.

Adversarial latent regularization

Pixel predictors can "cheat" by packing future appearance into z_i. CLAW doubles each batch into two shared-weight paths [6]: (i) standard $(o_i, z_i) \to$ future frames and (ii) z_i alone through a **gradient-reversal layer** (GRL). The GRL path is trained to predict futures from z only, but gradients reverse into the LAM—so latents that leak future pixels are penalized, while the standard path still rewards transition-informative codes. Joint end-to-end optimization ties the LAM and world model together under this adversarial constraint.

Latent Action Retrieval

We probe the latent space with L_2 nearest-neighbor retrieval: encode a query transition and recover the closest dataset neighbors.

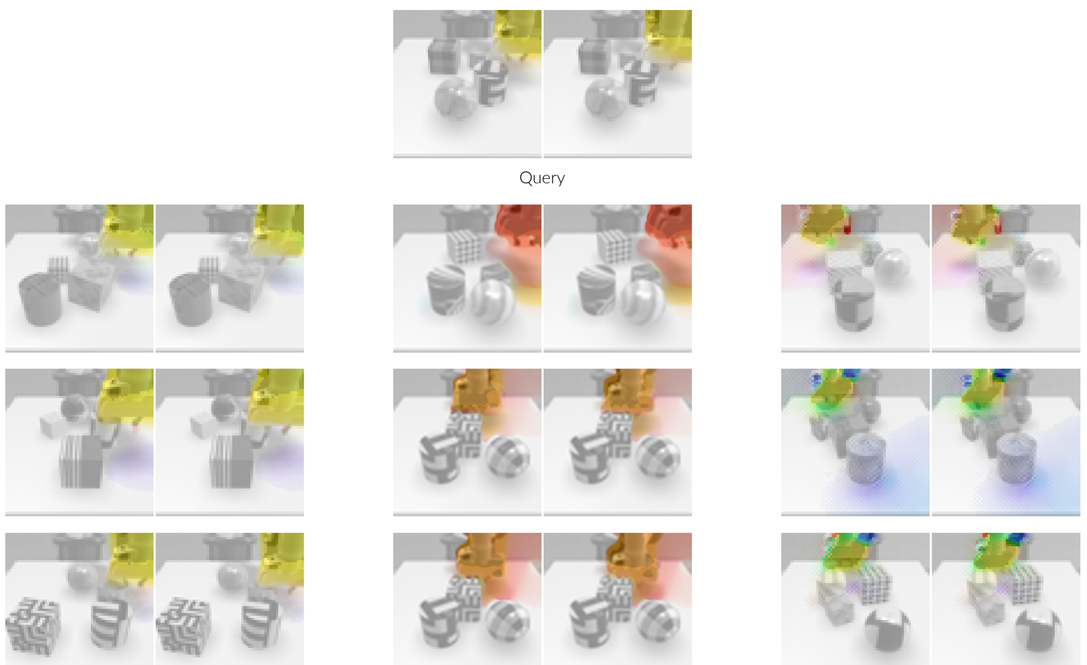

Figure 3. CLAW neighbors match **end-effector motion** with varying objects; AdaWorld / LAPO [7, 10] more often retrieve by scene appearance (color / layout).

Latent Action Policy Learning

In the imitation-from-observation setting [2, 12], we extract latents z_i from expert videos and train $\pi(z_i \mid o_i)$ by behavior cloning. At test time each predicted latent is mapped to a real action via nearest-neighbor lookup in a small labeled reference set—no separate action adapter.

Table 1. Despite task-agnostic pretraining (no wood reward), CLAW leads Crafter harvest / acquire and most OGBench [9] drawer tasks; AdaWorld edges Push Button.

	Crafter			OGBench		
Method	Harvest	Avg. W	Acquire	Open	Close	Push
Random	25.0	0.28	27.5	2.5	14.5	8.0
LAPO	35.5	1.12	38.0	4.5	39.5	9.5
AdaWorld	43.5	1.60	46.5	7.5	29.0	**16.5**
CLAW	**86.0**	**4.46**	**84.0**	**13.5**	**42.0**	15.0

Visual Planning

We plan latent-action sequences with MPC [5] (CEM on Procgen) toward a goal image, then execute via nearest-neighbor retrieval from a labeled reference set on VP² [11] (Robosuite / RoboDesk) and Procgen [3].

Quantitative planning. CLAW leads on four of six VP² tasks (esp. Open Slide, Red Button, Upright Block); Blue/Green Button remain hard under partial occlusion. On Crafter + Procgen, best on three of four.

Table 2. VP² success (%; mean over 3 runs). *AdaWorld paper.

Method	Push	Open	Blue	Green	Red	Upright
LAPO	18.3	11.7	8.9	6.7	3.3	15.6
AdaWorld*	**64.6**	5.8	**29.2**	10.8	10.0	5.0
CLAW	63.7	**23.3**	12.2	**11.1**	**25.6**	**43.3**

Table 3. Crafter / Procgen planning success (%).

Method	Craft.	Jump.	Climb.	Ninja
LAPO	46.7	58.3	53.3	48.3
AdaWorld	53.3	56.7	**66.7**	46.7
CLAW	**68.3**	**63.3**	60.0	**56.7**

Visual Planning (qualitative)

Qualitative plans. MPC rollouts executed with retrieved low-level actions (see tables at left for success rates).

Figure 4. Successful VP² plan on RoboDesk (upright block): start → goal via imagined latents.

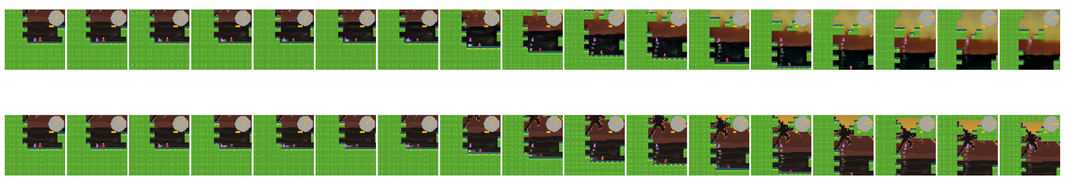

Figure 5. Successful Procgen Jumper plan: agent clears the gap toward the goal.

Latent Action Transfer

Action-swap protocol [7]: latents from a *source* video condition rollouts from a different *target* initial frame—testing whether actions are reusable across scenes.

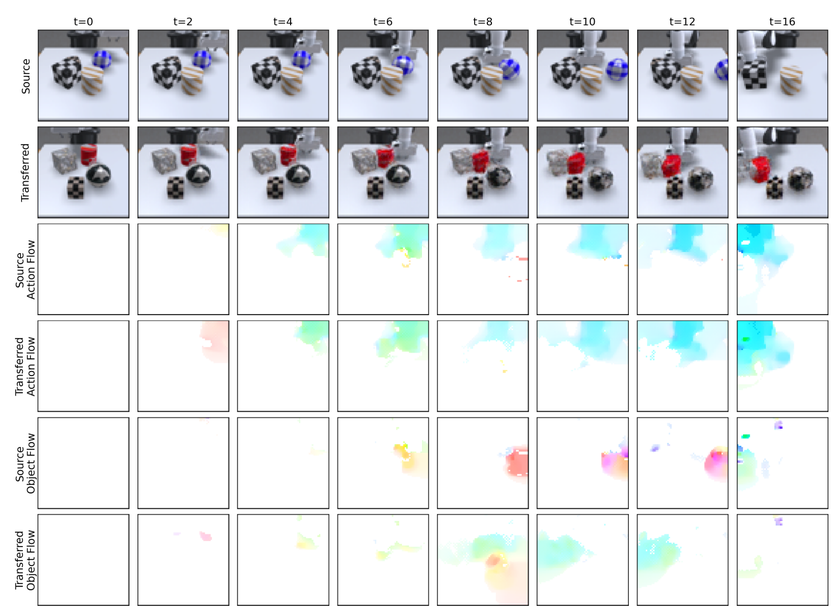

Figure 6. Robosuite: by t=4, transferred *action* flow matches the source gripper motion; *object* flow adapts to the new layout (reusable control, not appearance copy).

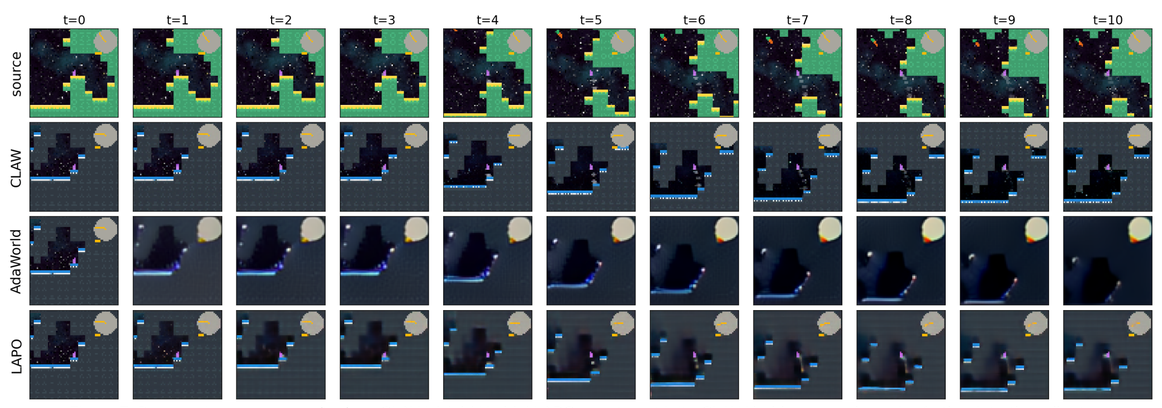

Figure 7. Procgen Jumper: source (top) vs. CLAW / AdaWorld / LAPO on a target init—CLAW holds source dynamics farther into the rollout.

Emergent Structure in Latent Action Space

To test whether continuous latents capture action semantics—without a VQ codebook or action labels at training—we encode Crafter transitions $z_i = \mathrm{LAM}_\theta(o_i, o_{i+1})$ and visualize them with t-SNE. Points are colored by ground-truth actions *for visualization only*. Locomotion (Move Left/Right/Up/Down), placement, and crafting form coherent clusters, showing that meaningful action structure emerges from self-supervised adversarial training alone.

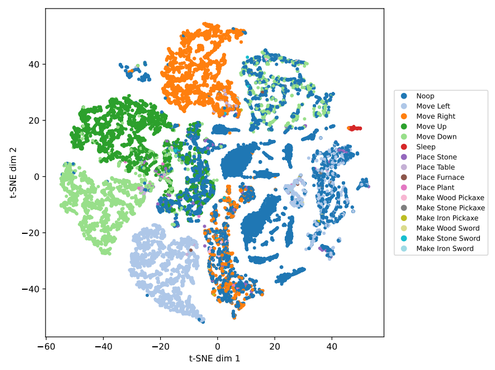

Figure 8. t-SNE of CLAW latents on Crafter: behaviorally related transitions occupy distinct regions of the continuous action space (no action supervision during training).

References

[1] Fan Bao, Shen Nie, Kaiwen Xue, Yue Cao, Chongxuan Li, Hang Su, and Jun Zhu. All are worth words: A vit backbone for diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 22669–22679, 2023.

[2] Jake Bruce, Michael D Dennis, Ashley Edwards, Jack Parker-Holder, Yuge Shi, Edward Hughes, Matthew Lai, Aditi Mavalankar, Richie Steigerwald, Chris Apps, et al. Genie: Generative interactive environments. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2024.

[3] Karl Cobbe, Christopher Hesse, Jacob Hilton, and John Schulman. Leveraging procedural generation to benchmark reinforcement learning. *arXiv preprint arXiv:1912.01588*, 2019.

[4] Alexey Dosovitskiy. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

[5] Frederik Ebert, Chelsea Finn, Sudeep Dasari, Annie Xie, Alex Lee, and Sergey Levine. Visual foresight: Model-based deep reinforcement learning for vision-based robotic control. *arXiv preprint arXiv:1812.00568*, 2018.

[6] Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario March, and Victor Lempitsky. Domain-adversarial training of neural networks. *Journal of Machine Learning Research*, 17(59):1–35, 2016.

[7] Shenyuan Gao, Siyuan Zhou, Yilun Du, Jun Zhang, and Chuang Gan. AdaWorld: Learning adaptable world models with latent actions. *arXiv preprint arXiv:2503.18938*, 2025.

[8] Danijar Hafner, Timothy Lillicrap, Ian Fischer, Ruben Villegas, David Ha, Honglak Lee, and James Davidson. Learning latent dynamics for planning from pixels. *arXiv preprint arXiv:1811.04551*, 2019.

[9] Seohong Park, Kevin Frans, Benjamin Eysenbach, and Sergey Levine. Ogbench: Benchmarking offline goal-conditioned rl. In *International Conference on Learning Representations*, 2025.

[10] Dominik Schmidt and Minqi Jiang. Learning to act without actions. *arXiv preprint arXiv:2312.10812*, 2024.

[11] Stephen Tian, Chelsea Finn, and Jiajun Wu. A control-centric benchmark for video prediction. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2023.

[12] Faraz Torabi, Garrett Warnell, and Peter Stone. Recent advances in imitation learning from observation. *arXiv preprint arXiv:1905.13566*, 2019.

[13] Seonghyeon Ye, Joel Jang, Byeongguk Jeon, Sejune 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. *arXiv preprint arXiv:2410.11758*, 2025.

Scan for project website
ripl.github.io/claw