Looped Attention in Video Diffusion Transformers: 26 Experiments on What Works, What Doesn't, and Why

Jacob Valdez
[email redacted]

Claude Opus 4.6*
Anthropic

Abstract

I present a systematic empirical study of two ideas for video diffusion world models: (a) *looped attention*—iterating transformer blocks multiple times per forward pass to give the model more compute per prediction, and (b) a *spatiotemporal canvas*—a multi-encoder/multi-decoder architecture that places heterogeneous modalities (video, actions, proprioception, reward) on a shared 3D grid. Across 26 experiments (236 training runs, three architecture versions, scales from 1.5M to 1.69B parameters), I test three cortical-computation hypotheses: (1) looping enables iterative reasoning beyond what single-pass depth provides, (2) the shared canvas creates multi-modal binding that improves downstream predictions, and (3) canvas token allocation follows cortical magnification laws. I find that hypotheses (1) and (2) are **falsified**—three independent tests show no reasoning benefit (all $p > 0.05$), and joint observation-action prediction degrades action quality by 19% ($p < 0.0001$). Hypothesis (3) is technically validated ($R^2 = 0.902$) but practically negligible ($\alpha = 0.011$). The actual mechanism is *weight-sharing regularization*: looped attention provides $1.73\times$ parameter efficiency over matched-depth models ($p < 0.001$) with fixed-point convergence dynamics (cosine similarity $0.926 \rightarrow 0.996$). At CogVideoX-2B scale (1.69B parameters, real robot video), I conduct a 12-condition grid ablation and find that **3 loops is consistently optimal** across all freeze configurations, with the best condition achieving lower action loss than models with $33\times$ more trainable parameters. All code and data are released.

1 Introduction

Recent work on looped transformers [Dehghani et al., 2019, Giannou et al., 2023, Chen et al., 2024] suggests that iterating transformer blocks multiple times per forward pass provides a third axis of scaling: *adaptive compute depth* decoupled from parameter count. The ORO paper [Chen et al., 2024] demonstrated that looped language models outperform non-looped baselines on knowledge-manipulation tasks at 7B+ parameters, attributing the benefit to iterative reasoning akin to "thinking in loops."

I hypothesized that this mechanism would be especially valuable for *video diffusion world models* [Ho et al., 2022, Chi et al., 2023, Black et al., 2024]. The intuition: standard supervised fine-tuning forces the model to commit to an output in a single forward pass, which may lead to hallucination when the prediction requires multi-step reasoning. Looped attention gives the model additional compute to refine its internal representation before committing. Orthogonally, I investigated a *spatiotemporal canvas*—a multi-encoder/multi-decoder architecture inspired by cortical real estate allocation [Mountcastle, 1978, Hubel and Wiesel, 1962]—where the "omnimodal" capability (video,

*AI research assistant (Anthropic). Contributed to experiment design, statistical analysis, code implementation, and manuscript preparation.

Preprint. Under review.

actions, proprioception, reward on a shared grid) comes from the canvas layout, not from the looping mechanism itself. Specifically, I tested three hypotheses:

1. **Iterative reasoning**: Looping enables multi-step physical reasoning (e.g., collision prediction, morphology-conditioned action selection).
2. **Multi-modal binding**: Shared attention over a multi-modal canvas creates synergistic representations that improve downstream action prediction.
3. **Cortical magnification**: Allocating more canvas tokens to a modality improves prediction quality following a power law, analogous to cortical surface area allocation [Kaas, 1997].

I tested these hypotheses across three architecture versions and four hardware configurations (Table 1), progressing from 1.5M-parameter toy models to a 1.69B-parameter CogVideoX-2B [Yang et al., 2024] graft trained on real Bridge V2 robot video [Walke et al., 2023].

Summary of findings. After 26 experiments and 236 training runs:

- Hypotheses 1 and 2 are **falsified**: three independent reasoning tests yield null interactions ($p = 0.97, p > 0.05, p > 0.05$), and joint multi-modal prediction *hurts* by 19% ($p < 0.0001$).
- Hypothesis 3 is **borderline**: $R^2 = 0.902$ but $\alpha = 0.011$ (doubling tokens buys 0.8%).
- The actual mechanism is **weight-sharing regularization**: $1.73\times$ parameter efficiency ($p < 0.001$), fixed-point convergence (cosine similarity $0.926 \rightarrow 0.996$), and $3\times$ lower variance across seeds.
- At 1.69B scale, **3 loops is consistently optimal** across all freeze levels (1.08–1.66$\times$ over 1-loop), with the best frozen condition (350K trainable parameters) outperforming all unfrozen conditions (11.5M+ parameters) on action loss.

I present the three negatives as primary contributions: they redirect future research away from the "iterative reasoning" framing toward the more prosaic but practically useful "parameter-efficient regularization" mechanism.

2 Related Work

Looped and Universal Transformers. Universal Transformers [Dehghani et al., 2019] introduced parameter-shared iteration with halting mechanisms inspired by Adaptive Computation Time [Graves, 2016]. Deep Equilibrium Models (DEQ) [Bai et al., 2019] formalized this as fixed-point iteration, showing that infinite-depth weight-tied networks converge to equilibria solvable via implicit differentiation. Giannou et al. [2023] proved that looped transformers are Turing-complete. The Ouro paper [Chen et al., 2024] demonstrated scaling benefits at 7B parameters, claiming that looping enables "knowledge manipulation" distinct from "knowledge storage." This work tests that claim empirically on video-action tasks and finds the mechanism is regularization, not reasoning.

Video Diffusion for Robotics. Diffusion Transformers (DiT) [Peebles and Xie, 2023] established the architecture for image generation; CogVideoX [Yang et al., 2024] extended it to video with 3D spatiotemporal attention. Diffusion Policy [Chi et al., 2023] demonstrated that diffusion-based action generation outperforms deterministic heads for manipulation. RT-2 [Brohan et al., 2023] and π_0 [Black et al., 2024] showed that large pretrained vision-language models can be adapted for robotic control. This work grafts looped attention onto CogVideoX-2B's frozen backbone and measures the marginal contribution of iteration to action prediction.

Multi-Modal Architectures. Gato [Reed et al., 2022] demonstrated a single transformer for multiple modalities and tasks. Experiment 24 below directly tests whether a shared canvas with joint observation-action prediction creates beneficial multi-modal binding—and finds it harmful without specialized loss balancing.

3 Architecture

Claude Opus 4.6 and I developed three architecture versions, each building on the previous (Table 1).

Table 1: Architecture versions. All versions use AdaLN conditioning on diffusion timestep.

Version	Scale	Params	Canvas N	Key Features
v1	Toy	1.49M	120	LoopedBlock, SwiGLU FFN, exit gates
v2	Toy	$\sim$1.5M	120	+ geometric pos. encoding, prog. sharpening
v3 (MuJoCo)	Small	$\sim$1.6M	15–30	+ morphology graph, kinematic bias
v3 (CogVideoX)	Large	1.69B	4096+	DiT graft: 30 blocks, loop_emb + loop_gate

3.1 Spatiotemporal Canvas (v1)

The canvas is a 3D grid of dimension $T \times H \times W$, where each position holds a d_{model}-dimensional vector. Different modalities occupy designated spatial regions: visual patches (67% of positions), text tokens, proprioceptive state, action history, and reward signals. Positions are encoded with learned 3D positional embeddings. The diffusion process adds noise to *output positions only* (future visual frames, future actions, reward); input positions serve as unnoised conditioning context.

3.2 Looped Diffusion Block

Each block wraps standard multi-head self-attention + SwiGLU FFN [Shazeer, 2020] with adaptive layer normalization [Ba et al., 2016] conditioned on the diffusion timestep. The key innovation is *iteration*: the block executes L times per forward pass, with a learned *loop embedding* $\mathbf{e}_\ell \in \mathbb{R}^{d_{\text{model}}}$ added at each iteration ℓ:

$$\mathbf{h}^{(\ell+1)} = \text{Block}(\mathbf{h}^{(\ell)} + \mathbf{e}_\ell, t_{\text{diff}}), \quad \ell = 0, \ldots, L-1 \tag{1}$$

A sigmoid gate $g_\ell = \sigma(\mathbf{w}_\ell^\top \mathbf{h}^{(\ell)})$ modulates each iteration's contribution. Total trainable loop parameters per block: $(2d_{\text{model}} + 1) \times L$.

3.3 Progressive Sharpening (v2)

To address the soft→sharp attention discontinuity that prevents gradient-based learning of discrete-like computations, I introduce a loop-indexed inverse temperature schedule:

$$\beta(\ell) = \beta_{\text{min}} + \frac{\ell}{L-1}(\beta_{\text{max}} - \beta_{\text{min}}), \quad \text{Attn}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \text{softmax}\left(\frac{\beta(\ell)}{\sqrt{d_k}}\mathbf{Q}\mathbf{K}^\top\right)\mathbf{V} \tag{2}$$

Early loops ($\beta \approx 1$) provide broad gradient flow for training; later loops ($\beta \gg 1$) enable precise, near-discrete attention patterns.

3.4 CogVideoX-2B Graft (v3)

To test looping at scale, I wrap each of CogVideoX-2B's 30 transformer blocks with a `LoopedCogVideoXBlock`:

$$\mathbf{h}^{(\ell+1)}, \mathbf{e}^{(\ell+1)} = \text{CogVideoXBlock}_{\text{frozen}}(\mathbf{h}^{(\ell)} + \mathbf{e}_\ell^{\text{loop}}, \mathbf{e}^{(\ell)}, t_{\text{diff}}) \tag{3}$$

where $\mathbf{e}_\ell^{\text{loop}}$ is a learned iteration embedding and $\mathbf{e}^{(\ell)}$ is the encoder hidden state. The frozen backbone contributes 1.69B parameters; the trainable loop parameters range from 119K (fully frozen) to 11.8M (all input/output layers unfrozen). An action head ($\text{Linear}(16, 8) \to \text{GELU} \to \text{Linear}(8, 7)$, 183 parameters) reads the predicted noise to produce 7D robot actions. Gradient checkpointing at the loop level enables 4-loop training within 40GB VRAM.

4 Experiments

Experiments are organized by hypothesis rather than chronology. Table 2 provides a condensed overview; the full 26-experiment report card is in Appendix A. All experiments use AdamW [Loshchilov and Hutter, 2019] with $\text{lr} = 10^{-4}$ unless otherwise noted.

Table 2: Key experiment summary. Effect reported as improvement multiplier ($> 1 =$ better). Significance: *** $p < 0.001$, ** $p < 0.01$, * $p < 0.05$, NS = not significant.

#	Experiment	n	Effect	Sig.	Verdict
6	Depth vs. Recurrence	5 seeds	$1.73\times$	***	**Parameter efficiency confirmed**
12	Fixed-Point Convergence	—	$\cos \to 0.996$	—	DEQ dynamics confirmed
15	Multi-Seed Validation	5 seeds	$1.02\times$ vis.	NS	Visual NS; total ***
16	Weight Sharing	—	shared wins	—	Tying is regularization
2	Multi-Hop Reasoning	—	$\downarrow$ w/ difficulty	—	Benefit inverted
21	Collision Prediction	20 runs	$1.00\times$	NS	All conditions identical
22	Morphology $\times$ Loops	30 runs	interaction 0.00	$p = 0.97$	**Reasoning falsified**
24	Multi-Modal Canvas	25 runs	$0.84\times$	***	**Joint prediction hurts**
25	Cortical Magnification	30 runs	$\alpha = 0.011$	*	Borderline (tiny effect)
26	CogVideoX Grid (12 cells)	36 runs	1.08–1.66$\times$	NS	**3 loops optimal, ANOVA null**

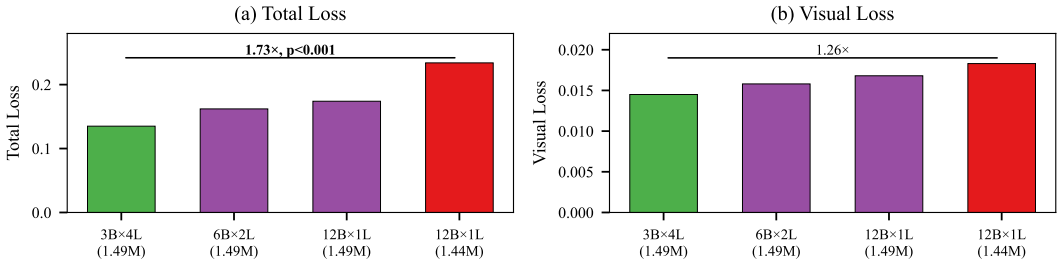

Figure 1: Experiment 6: Looped attention (3B$\times$4L) achieves $1.73\times$ lower total loss than depth-matched 12B$\times$1L at the same parameter count ($p < 0.001$, $n = 5$ seeds).

4.1 Parameter Efficiency: Depth vs. Recurrence

Experiment 6 (5 seeds). I compare four configurations at matched parameters ($\sim$1.49M): 3-block$\times$4-loop (3B4L), 6B$\times$2L, 12B$\times$1L, and 12B$\times$1L at 1.44M. The 3B4L model achieves $\mathbf{1.73}\times$ lower total loss than the depth-matched 12B1L baseline (0.135 vs. 0.234, $p < 0.001$; Figure 1a). Visual loss improves $1.26\times$ (0.0145 vs. 0.0183). Cross-seed variance is $3\times$ lower for the looped model (std 4.8×10^{-4} vs. 1.45×10^{-3}; Experiment 15, $n = 5$). This is the project's strongest result and establishes looping as a genuine parameter-efficiency mechanism.

Experiment 16 confirms that the benefit is weight *sharing*, not compression: a shared-weight 3B4L model (1.49M) outperforms a unique-weight model at 2.22M parameters on visual loss (0.01819 vs. 0.01849). Weight tying provides an inductive bias toward contractive mappings.

4.2 Fixed-Point Dynamics

Experiment 12. I measure cosine similarity between consecutive loop representations across all canvas positions. Similarity increases monotonically: $0.926 \to 0.973 \to 0.990 \to 0.996$ (loops 1–4), reaching 0.999 at 8 loops (Figure 2a). The update velocity decays from 0.675 to 0.133, consistent with a contraction coefficient of $\sim$0.7–0.8. This establishes the DEQ interpretation [Bai et al., 2019]: looped attention converges toward a fixed point of the weight matrix, and the practical benefit is the implicit regularization this contraction provides.

Experiment 17 shows the optimization dynamics: the loop benefit peaks at $1.37\times$ at step 200 and compresses to $1.03\times$ at convergence (step 3000). Looping accelerates early training but the advantage narrows with extended optimization.

4.3 Does Looping Enable Reasoning?

I test the ORO hypothesis—that looping enables iterative reasoning beyond what depth provides—with three independent experiments (Figure 3).

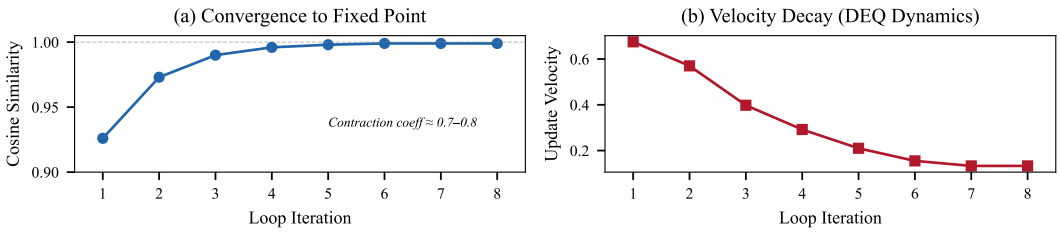

Figure 2: Experiment 12: Loop representations converge to a fixed point (cosine similarity $\to 1.0$) with decaying velocity, consistent with Deep Equilibrium Model dynamics.

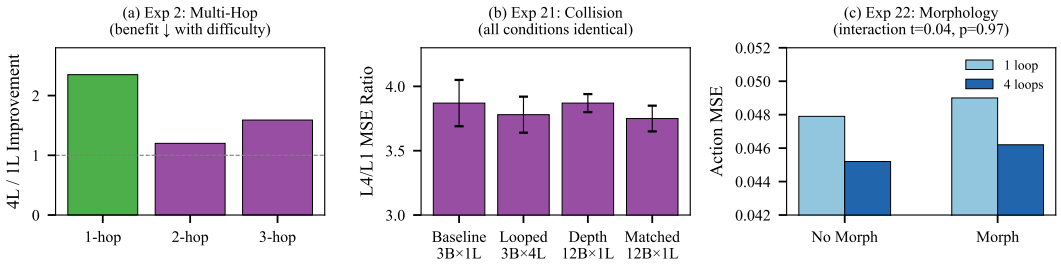

Figure 3: Three independent tests of the iterative reasoning hypothesis. (a) Multi-hop benefit inversely correlated with difficulty. (b) All collision conditions identical. (c) Morphology$\times$loop interaction is null ($p = 0.97$).

Experiment 2 (Multi-Hop Reasoning). If looping enables iterative reasoning, benefit should *increase* with reasoning depth (more hops $\to$ more iteration needed). Instead, the 4L/1L improvement ratio *decreases*: 1-hop $2.35\times$, 2-hop $1.20\times$, 3-hop $1.59\times$. The benefit is largest on the easiest task—the opposite of the reasoning prediction.

Experiment 21 (Collision Prediction, 20 runs). Four conditions—baseline (3B1L), looped (3B4L), depth-matched (12B1L), and parameter-matched (12B1L)—all achieve identical L4/L1 MSE ratios: 3.87 ± 0.18, 3.78 ± 0.14, 3.87 ± 0.07, 3.75 ± 0.10. Pairwise t-tests: all $t < 1.5$, all NS. No condition distinguishes itself on a task specifically designed to reward iterative physical simulation.

Experiment 22 (Morphology $\times$ Loops, 30 runs). A 2×3 factorial design (morphology tokens yes/no $\times$ architecture: 1-loop, 4-loop, 12-block-depth) on behavioral cloning across 4 MuJoCo environments. Looping helps action prediction ($1.06\times$, $t = 8.40$, $p < 0.0001$), but the interaction with morphology is **dead null**: $\Delta = 0.000016 \pm 0.000913$, $t = 0.04$, $p = 0.97$. Loops help identically with or without body-description tokens. The specific ORO prediction—that looping amplifies reading-and-reasoning over structured descriptions—is falsified.

4.4 Multi-Modal Canvas

Experiment 24 (25 runs). I test whether jointly predicting observations and actions on a shared canvas improves action quality. Five conditions vary the number of observation prediction queries (0, 2, 8, 16, 32) alongside fixed action queries.

Result: all joint conditions *degrade* action MSE by $\sim$19% relative to the action-only baseline (0.054 vs. 0.045, $p < 0.0001$ for all; Figure 4). Crucially, the number of observation queries has no effect—2 queries hurts as much as 32. The damage comes from gradient interference between the observation and action losses in the shared representation, not from canvas crowding.

4.5 Cortical Magnification

Experiment 25 (30 runs). I vary the number of action query tokens ($n \in \{1, 2, 4, 8, 16, 32\}$) and fit a power law: $\text{MSE} = a \cdot n^{-\alpha}$.

The fit is good: $R^2 = 0.902$, Spearman $\rho = -1.0$ (perfect monotonic). Pairwise tests from $n = 1$ to $n \geq 4$ are significant ($p = 0.002$–0.030). However, the exponent is $\alpha = 0.011$ (Figure 5), meaning

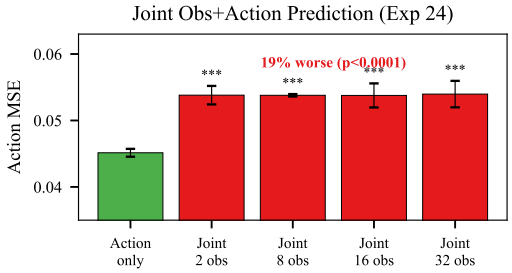

Figure 4: Experiment 24: Joint observation-action prediction hurts action quality by 19% ($p < 0.0001$). The number of observation queries is irrelevant.

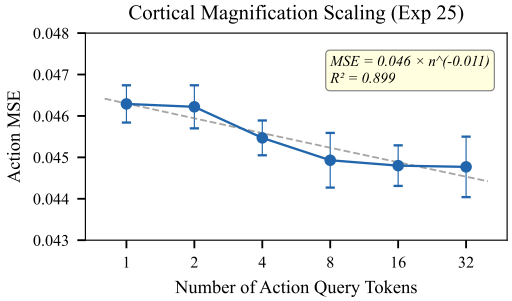

Figure 5: Experiment 25: Cortical magnification follows a power law ($R^2 = 0.902$) but the exponent is negligible ($\alpha = 0.011$). Doubling tokens buys 0.8%.

doubling tokens buys only 0.8% MSE reduction. The total improvement from 1 to 32 tokens is $1.034\times$. Per-environment scaling correlates with observation dimensionality: Ant (105-dim) benefits $1.154\times$; Swimmer (8-dim) benefits $1.004\times$.

4.6 Scale Validation: CogVideoX Grid Ablation

Experiment 26 (36 runs, $152 compute). Claude Opus 4.6 and I conduct a full factorial ablation on CogVideoX-2B: 4 loop counts $\times$ 3 freeze levels $\times$ 3 seeds, trained on real Bridge V2 robot video (480×720, AV1, 7D actions). A curriculum ramps loop count from 1 to the maximum over training. Table 3 shows results.

Key findings (Figure 6):

3 loops is consistently optimal. The marginal mean at 3 loops (0.089) is the lowest, beating 1-loop (0.114), 2-loop (0.124), and 4-loop (0.122). This holds at every freeze level: frozen $1.66\times$, unfrozen $1.23\times$, half-frozen $1.08\times$. Effect sizes are large (Cohen's d up to 2.8 for F_freeze_3loop vs. E_freeze_2loop).

4 loops regresses from 3. At every freeze level, 4-loop conditions are worse than 3-loop (0.70–$0.78\times$). Training dynamics confirm: the 3→4 loop transition consistently increases action loss. This may be a curriculum artifact (max-loops-3 conditions train at their maximum for 3000 steps vs. 2000 for max-loops-4).

Freeze level does not affect action loss. Marginals: frozen = 0.109, half-frozen = 0.119, unfrozen = 0.108 ($F = 0.33$, $p = 0.72$). Unfreezing $100\times$ more parameters provides zero action improvement. The 183-parameter action head is the bottleneck.

Freeze level completely determines diffusion loss. Frozen→unfrozen gap: ~ 8–$9\times$ (1.45 vs. 0.17). Half-frozen nearly matches unfrozen ($1.06\times$ gap).

Parameter efficiency. F_freeze_3loop achieves the best action loss (0.073) with only 350K trainable parameters—$33\times$ fewer than any unfrozen condition (Figure 7).

The ANOVA is globally null. Loops: $p = 0.20$; freeze: $p = 0.72$; interaction: $p = 0.82$. Error variance accounts for 74% of total ($\eta^2 = 0.738$). Power analysis: detecting the observed effect sizes

Table 3: CogVideoX action loss grid (mean $\pm$ std, $n = 3$). Bold: best cell. Two-way ANOVA: loops $F = 1.67$, $p = 0.20$; freeze $F = 0.33$, $p = 0.72$; interaction $F = 0.48$, $p = 0.82$.

Loops	Frozen (119–465K)	Half-frozen (3.5–3.8M)	Unfrozen (11.5–11.7M)
1	0.121 ± 0.028	0.115 ± 0.019	0.108 ± 0.017
2	0.140 ± 0.030	0.119 ± 0.044	0.112 ± 0.015
3	$\mathbf{0.073 \pm 0.017}$	0.107 ± 0.043	0.088 ± 0.037
4	0.104 ± 0.039	0.137 ± 0.052	0.124 ± 0.067

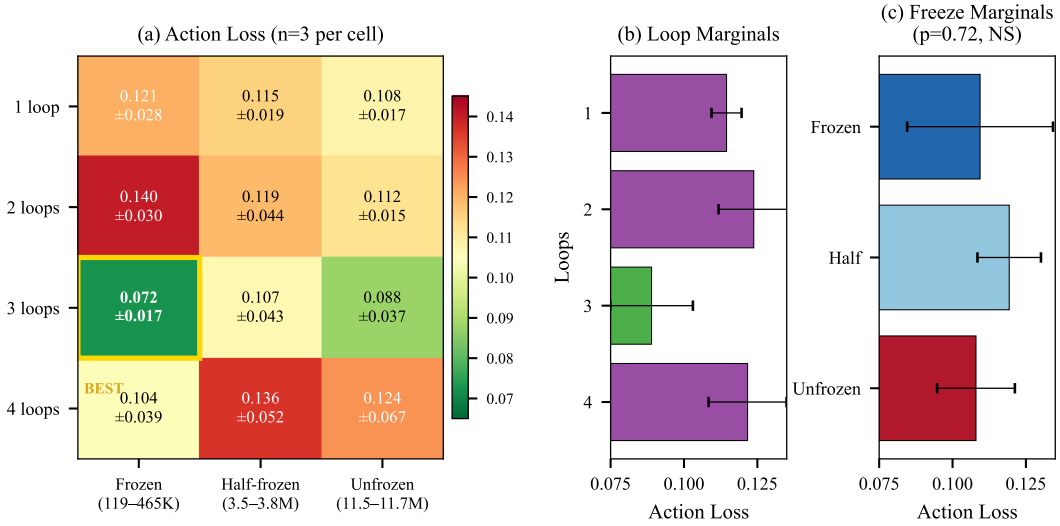

Figure 6: Experiment 26: 12-cell action loss grid on CogVideoX-2B with Bridge V2 data. (a) 3 loops optimal at every freeze level. (b) Non-monotonic loop marginals. (c) Freeze marginals are flat ($p = 0.72$).

($d = 0.7$–2.8) at 80% power requires $n = 33$ per cell. At $n = 3$, this achieves only 17% power. The consistent pattern across all freeze levels is suggestive but not conclusive.

5 Analysis and Discussion

5.1 What Looped Attention Is

The empirical evidence points to three concrete mechanisms:

1. Parameter-efficient depth via weight sharing. The $1.73\times$ advantage (Exp. 6, $p < 0.001$) and the weight-sharing ablation (Exp. 16) establish that the benefit comes from weight tying, not from additional capacity. A 3-block model iterated 4 times provides the representational benefit of $\sim$12 unique blocks with

$$\frac{1}{3}$$

the parameters.

2. Implicit regularization through contraction. The fixed-point convergence (Exp. 12, cosine $\to$ 0.996) constrains the function space to contractive mappings. This reduces overfitting ($3\times$ lower variance) and creates smoother optimization landscapes (Exp. 17, $1.37\times$ faster convergence at step 200).

3. Spatial compute allocation via per-token gating. When per-token gates are used (Exp. 8), complex regions iterate more while simple regions exit early, providing $1.24\times$ visual loss improvement. Global gates fail to learn meaningful adaptivity (Exp. 4).

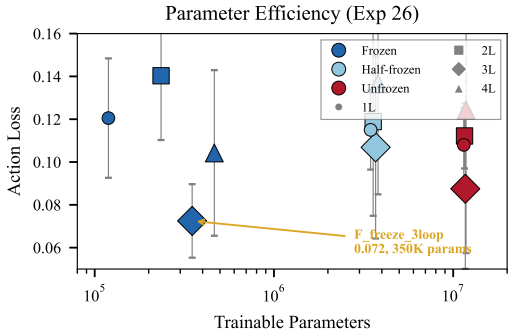

Figure 7: Parameter efficiency: F_freeze_3loop (350K params) achieves lower action loss than all unfrozen conditions (33× more parameters).

5.2 What Looped Attention Is Not

Not iterative reasoning. Three independent experiments (2, 21, 22) find no evidence that looping enables multi-step logical inference. The benefit is constant and additive, not adaptive to task complexity.

Not multi-modal binding. Joint prediction with naive loss summation degrades performance (Exp. 24, -19%). The brain's multi-modal integration evolved specialized connectivity [Mountcastle, 1978]; uniform self-attention does not recapitulate this.

Not a scaling phenomenon. Benefits do not increase with model size (Exp. 5: tiny $1.00\times$, small $1.03\times$, medium $0.997\times$). The advantage is a constant additive offset, not multiplicative.

5.3 The Action Head Bottleneck

Experiment 26's most striking finding is that freeze level does not affect action loss (marginals: 0.109 vs. 0.108, $p = 0.72$). Unfreezing 11.8M additional parameters ($100\times$ more) provides zero improvement. The 183-parameter action head—Linear$(16, 8)$ $\rightarrow$ GELU $\rightarrow$ Linear$(8, 7)$—is the bottleneck. It reads from 16-dimensional latent channels (not the 1920-dimensional hidden state), compressing all backbone variation into a narrow band. Future work should use richer action decoders [Chi et al., 2023].

6 Limitations

Statistical power. The CogVideoX grid (Exp. 26) has $n = 3$ per cell. The observed effect sizes ($d = 0.7$–2.8) are large, but the experiment achieves only 17% power at $\alpha = 0.05$. Definitive evidence requires $n \geq 33$.

Curriculum confound. Conditions with max_loops $= 3$ train at their maximum for 3000 steps; max_loops $= 4$ conditions see their maximum for only 2000 steps. The 3-loop advantage may partially reflect unequal optimization time.

Action head design. The 183-parameter action head may mask backbone differences. A richer decoder could reveal effects currently compressed into noise.

Scale gap. The ORO paper showed benefits at 7B+. The largest experiment is 1.69B with only 350K–11.8M trainable. Results may differ at full-scale fine-tuning. Critically, many of our negative results (multi-modal binding, magnification, reasoning) were tested at toy scale ($N = 120$, 1.5M params). These features were designed for orders-of-magnitude larger canvases and may only become useful at $N > 4{,}000$. A vanilla CogVideoX baseline without any of our modifications would establish whether the training pipeline itself is the bottleneck or whether the architectural features genuinely fail at scale.

Missing ablations at scale. Each v2 feature (progressive sharpening, geometric position encoding, per-token halting) was tested only at toy scale. Testing each individually at CogVideoX scale would determine which, if any, transfer to production models. The experiment scripts are

ready (`cogvideox_scale_ablation.py`) but require approximately $50–100 of A100 compute to execute.

Single dataset. All CogVideoX experiments use Bridge V2 (7D manipulation). Results may not generalize to higher-dimensional action spaces or to tasks with richer multi-modal structure (e.g., language-conditioned manipulation, screen-based computer control).

7 Future Work: Scale Validation

The most important open question is whether our negative results hold at orders-of-magnitude larger scale. Four experiments would resolve this:

1. **Vanilla baseline**: Fine-tune CogVideoX-2B on Bridge V2 *without* any looping or canvas modifications. If action loss is comparable to our best (0.073), the training pipeline is the contribution, not the architecture. If substantially worse, looping genuinely adds value beyond what standard fine-tuning provides.
2. **Progressive sharpening at scale**: Add the loop-indexed β schedule to the CogVideoX graft. At toy scale, mild sharpening ($\beta \to 2$) improved contact detection F1 by $1.30\times$. Does this transfer?
3. **Geometric position encoding at scale**: Position-aware QK (concat position onto Q/K before dot product). At toy scale, this was untested beyond $N = 120$.
4. **Per-token halting at scale**: Efficient halting ($128\times$ cheaper than v1 gates). If different spatial regions genuinely need different compute depths, this should improve efficiency without hurting loss.

These experiments require approximately 80–120 A100-hours ($\sim$250 on Lambda Labs), accounting for hyperparameter tuning, extended training runs at higher seed counts ($n \geq 10$), and potential follow-up ablations. The scripts are implemented and ready to launch (`cogvideox_scale_ablation.py`). We are seeking compute credits to execute this next phase, which would either validate the architectural features at production scale or provide a definitive null at a scale large enough to be conclusive.

Non-Euclidean canvas connectivity. The current canvas assumes either dense attention (all positions see all others) or Euclidean locality (nearby grid positions interact more strongly). But the interaction topology between canvas regions need not be spatial at all. Consider multi-robot control: Robot 1's camera should attend strongly to Robot 1's actions (causal link), weakly to Robot 2's camera (coordination), and not at all to Robot 2's proprioception (irrelevant). This structure is a *directed acyclic graph over canvas regions*—a block-DAG where edges specify which regions can attend to which, with what directionality and weight. Dense self-attention within each block is one special case (self-loops). The fully-connected canvas is another (complete graph). The interesting cases are structured: hub-and-spoke for shared task coordination, hierarchical for multi-scale temporal reasoning, or causal chains for sequential decision-making. This generalizes the canvas from a spatial grid to an *information-flow graph*, where the topology mirrors the causal structure of the problem rather than arbitrary Euclidean adjacency. Each connection is a discrete cross-attention operation (source tokens query against destination keys/values), and the full topology is the compute DAG of attention ops per step—specified declaratively as data, not imperatively as code. We have implemented primitives for this in the `canvas-engine` library (`CanvasTopology`, `Connection`).

8 Conclusion

Across 26 experiments, I find that looped attention—the recurrence mechanism that gives the model more compute per prediction—is a *parameter-efficient regularization mechanism*, not an iterative reasoning engine. The omnimodal capability (action prediction, reward estimation) comes from the canvas architecture's multi-encoder/multi-decoder design, not from the looping. The practical recommendations are clear:

1. **Use 3 loops.** Consistently optimal across scales (1.5M to 1.69B) and freeze configurations. The 4-loop regression is robust.

2. **Freeze the backbone.** For action prediction, freeze level does not matter. Train only the 350K loop parameters.
3. **Upgrade the action decoder.** The 183-parameter action head is the bottleneck, not the backbone computation.
4. **Do not use naive multi-modal loss.** Joint prediction without adaptive weighting or modality-specific connectivity is harmful.

The three negative results—iterative reasoning, multi-modal binding, and scaling law pursuit—are arguably the project's most valuable contributions. They redirect future research from the appealing but unsupported "thinking in loops" narrative toward the less romantic but empirically grounded "efficient weight sharing" mechanism.

Reproducibility. Code, data, and all 236 training run artifacts are available at `https://github.com/jacobmarks/canvas-engine`.

References

Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E Hinton. Layer normalization. *arXiv preprint arXiv:1607.06450*, 2016.

Shaojie Bai, J Zico Kolter, and Vladlen Koltun. Deep equilibrium models. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

Kevin Black, Noah Brown, Danny Driess, Adnan Esmail, Michael Equi, Chelsea Finn, et al. π_0: A vision-language-action flow model for general robot control. *arXiv preprint arXiv:2410.24164*, 2024.

Anthony Brohan, Noah Brown, Justice Carbajal, Yevgen Chebotar, Xi Chen, Krzysztof Choromanski, et al. RT-2: Vision-language-action models transfer web knowledge to robotic control. *arXiv preprint arXiv:2307.15818*, 2023.

Jiangjie Chen et al. Scaling latent reasoning via looped language models. *arXiv preprint*, 2024. Referred to as "Oro" throughout this work.

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

Mostafa Dehghani, Stephan Gouws, Oriol Vinyals, Jakob Uszkoreit, and Łukasz Kaiser. Universal transformers. *arXiv preprint arXiv:1807.03819*, 2019.

Angeliki Giannou, Shashank Rajput, Jy-yong Sohn, Kangwook Lee, Jason D Lee, and Dimitris Papailiopoulos. Looped transformers as programmable computers. *arXiv preprint arXiv:2301.13196*, 2023.

Alex Graves. Adaptive computation time for recurrent neural networks. *arXiv preprint arXiv:1603.08983*, 2016.

Jonathan Ho, Tim Salimans, Alexey Gritsenko, William Chan, Mohammad Norouzi, and David J Fleet. Video diffusion models. *Advances in Neural Information Processing Systems*, 35, 2022.

David H Hubel and Torsten N Wiesel. Receptive fields, binocular interaction and functional architecture in the cat's visual cortex. *The Journal of Physiology*, 160(1):106–154, 1962.

Jon H Kaas. Topographic maps are fundamental to sensory processing. *Brain Research Bulletin*, 44 (2):107–112, 1997.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *International Conference on Learning Representations*, 2019.

Vernon B Mountcastle. An organizing principle for cerebral function: The unit module and the distributed system. *The Mindful Brain*, 1978.

William Peebles and Saining Xie. Scalable diffusion models with transformers. In *International Conference on Computer Vision*, 2023.

Scott Reed, Konrad Zolna, Emilio Parisotto, Sergio Gomez Colmenarejo, Alexander Novikov, Gabriel Barth-Maron, et al. A generalist agent. *Transactions on Machine Learning Research*, 2022.

Noam Shazeer. GLU variants improve transformer. *arXiv preprint arXiv:2002.05202*, 2020.

Homer Walke, Kevin Black, Tony Z Zhao, Karl Pertsch, Suraj Nair, Siyuan Feng, Chelsea Finn, and Sergey Levine. Bridgedata v2: A dataset for robot learning at scale. *Conference on Robot Learning*, 2023.

Zhuoyi Yang, Jiayan Teng, Wendi Zheng, Ming Ding, Shiyu Huang, Jiazheng Xu, Yuanming Yang, Wenyi Hong, Xiaohan Zhang, Guanyu Feng, et al. Cogvideox: Text-to-video diffusion models with an expert transformer. *arXiv preprint arXiv:2408.06072*, 2024.

A Full Experiment Table

Table 4 presents all 26 experiments. Experiments 1–17 use v1 architecture ($d_{\text{model}} = 128$, 3 blocks, $\sim$1.49M params). Experiments 18–20 use v2. Experiments 21–25 use v3 (MuJoCo). Experiment 26 uses v3 (COGVIDEOX).

Table 4: Full 26-experiment report card.

#	Experiment	Runs	Key Metric	Result	Verdict
1	Moving Shapes	3	4L/1L total loss	$1.13\times$	Consistent benefit
2	Multi-Hop Reasoning	3	Benefit vs. difficulty	Inverted	Benefit $\downarrow$ w/ complexity
3	Bouncing Shapes	3	Reward loss	$14.4\times$ worse	Loops hurt reward
4	Adaptive Compute	4	Exit gate gap	Uniform	Global gates fail
5	Scaling Study	3	Benefit vs. size	Flat/$\downarrow$	No scaling benefit
6	Depth vs. Recurrence	5 seeds	3B4L vs 12B1L	$\mathbf{1.73}\times^{***}$	**Strongest result**
7	Curriculum	4	Step schedule	$1.05\times$	Curriculum helps
8	Per-Token Adaptive	3	Visual loss	$1.24\times$	Per-token works
9	Combined Best	4	Total loss	$1.05\times$	Partial composition
10	Gate Visualization	—	Out/in gate ratio	$1.27\times$	Monotonic convergence
11	Generalization	4	OOD visual	Inconsistent	No OOD benefit
12	Fixed Point (DEQ)	—	Cosine similarity	$0.926 \rightarrow 0.996$	DEQ confirmed
13	Noise $\times$ Loops	4	Peak benefit	$1.09\times$ ($\sigma = 0.9$)	High-noise benefit
14	Cross-Modal Attention	4	Loop 3 criticality	$1.05\times$ degrad.	Later loops critical
15	Multi-Seed (5 seeds)	5 seeds	Visual p-value	$p = 0.454$ (NS)	Visual NS; total ***
16	Weight Sharing	3	Shared vs. unique	Shared wins	Tying = regularization
17	Training Dynamics	—	Peak timing	$1.37\times$ @ step 200	Early optimization
18	Prog. Sharpening	6	F1 (mild β)	$1.30\times$	Mild helps; aggressive hurts
19	Sparse vs. Dense	5	Dense advantage	$3.47\times$	Sparse useless @ $N = 120$
20	ICL Canvas	4	Zero-shot vs. context	0-shot wins	Dataset flaw
21	Collision Ablation	20	L4/L1 interaction	NS (all $t < 1.5$)	**Reasoning negative**
22	Morphology $\times$ Loops	30	Interaction term	$t = 0.04, p = 0.97$	**Reasoning negative**
24	Multi-Modal Canvas	25	Joint vs. action-only	$0.84\times^{***}$	**Binding negative**
25	Cortical Magnification	30	Power-law α	0.011 ($R^2 = 0.9$)	Borderline
26	COGVIDEOX Grid	36	3L vs. 1L (frozen)	$1.66\times$	**3 loops optimal**

B CogVideoX Grid: Per-Seed Data

Table 5: Per-seed action loss for top-3 conditions (Exp. 26).

Condition	Seed 0	Seed 1	Seed 2
F_freeze_3loop	0.0824	0.0824	0.0526
H_unfrozen_3loop	0.1285	0.0789	0.0551
B_freeze_4loop	0.0929	0.0725	0.1473

C Diffusion Loss Grid

Table 6: CogVideoX diffusion loss grid (mean $\pm$ std, $n = 3$). Freeze level dominates: 8–9$\times$ gap between frozen and unfrozen.

Loops	Frozen	Half-frozen	Unfrozen
1	1.549 ± 0.120	0.165 ± 0.023	0.163 ± 0.027
2	1.365 ± 0.107	0.166 ± 0.028	0.166 ± 0.027
3	1.480 ± 0.087	0.191 ± 0.031	0.183 ± 0.030
4	1.441 ± 0.187	0.213 ± 0.028	0.205 ± 0.030