FORK-READY WORLD MODELS: TRAINING DISTILLED WORLD MODELS TO BE BRANCHABLE

Anonymous authors
Paper under double-blind review

ABSTRACT

Tree-structured imagination relies on continuations sampled from a shared prefix, but a shared noisy cache can correlate those continuations and reduce the benefit of best-of-k selection. We characterize this effect in a linear-Gaussian model and measure it in a 1.4B distilled autoregressive video model built on Self-Forcing and Wan2.1. The shipped clean-cache recipe has zero cache-noise variance in our protocol, whereas increasing cache noise reduces the measured variance ratio from 1.0 to 0.25. We introduce *Fork-Invariance*, a training objective that compares noised-cache rollouts with clean-cache references using a shared sampling tape. Its strongest variance-ratio result (0.964) is misleading: full-weight training damages decoded video and reduces sensitivity to the prefix. Capping the objective's gradient norm at 0.2 of the distillation gradient instead yields a variance ratio of 0.888, retains prefix sensitivity, and shows no detected quality degradation under two frozen visual judges. At matched measured variance ratio, the retained best-of-k search gain increases from 52% to 84%. These results distinguish cache-noise suppression from preservation of useful conditioning and motivate evaluating branch groups through variance, selection performance, decoded quality, and prefix sensitivity together. The gradient-matched result uses one training run, and judge scores alone do not establish perceptual equivalence. Code, configurations, and recorded measurements are available in an anonymous repository.

1 INTRODUCTION

A learned world model is increasingly used not just to roll out a single imagined future but to *branch*: to fork the current state and grow many continuations, then select or search among them—best-of-k sampling, Monte-Carlo tree search, and branching rollouts for planning and RL in imagination (Ha & Schmidhuber, 2018; Schrittwieser et al., 2020; Hafner et al., 2023; Gao et al., 2023). On modern autoregressive-diffusion video world models, the state that must be forked is large—the per-layer key/value (KV) cache over a long context window—so implementations increasingly borrow the copy-on-write (CoW) machinery from LLM serving (Kwon et al., 2023; Wang et al., 2026): a fork shares the parent's cache pages, and each branch only appends its own.

CoW forking is usually framed as an engineering question of *completeness*: is the snapshot sufficient, and will a branch that replays from it reach the same state? Execution-state capsules study complete snapshots and byte-exact restoration in their evaluated setting (Su, 2026); ForkKV targets efficient multi-LoRA cache sharing and includes a lossy approximation (Wang et al., 2026). We identify a second, orthogonal question of *validity—exchangeability*: are the k continuations grown from a shared forked prefix valid, interchangeable draws from the model's conditional law $P_\theta(\cdot \mid s)$?

These are different properties, and a fork can satisfy the first while violating the second. **A fork can be byte-exact and still be an invalid branch.** Best-of-k, MCTS, and every downstream selection rule assume exchangeability, not just completeness.

A concurrent line of work identifies the rollout cache as a *diversity* bottleneck, analyzes it through a total-variance decomposition, and injects stochasticity during training and inference (Duan et al., 2026). We examine the distinct question of how sharing a realized cache affects a group of sibling continuations and their subsequent selection. Preserving marginal diversity does not make a forked prefix a valid branch: each sibling can still be distributed as the model's conditional, but best-of-k and MCTS consume the *joint*, and it is the joint that a shared prefix corrupts. We measure the quantity a

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

marginal test cannot see, and find the two can decouple completely: a forked cache can be byte-exact, marginally diverse, and still not an exchangeable sample.

We show that exchangeability fails in a concrete, characterizable, and increasingly common setting. In a distilled autoregressive-diffusion world model whose rollout cache is written at a non-zero noise level, the shared forked prefix embeds a single, *branch-invariant* noise realization. Every sibling conditions on the same nuisance draw. The siblings are therefore *under-dispersed*: their spread is smaller than that of independent draws from $P_\theta(\cdot \mid s)$, and choosing the best of them over-selects on a shared error. The pathology is silent—each individual branch looks like a perfectly valid rollout—and it degrades exactly the decision that motivated branching in the first place.

Contributions.

1. **A pathology, formalized (§3).** We define branch exchangeability and prove an exact linear-Gaussian law for the sibling variance ratio, $\mathrm{VR} = \Sigma/(\Sigma + \sigma^2 W W^\top)$, with $\mathrm{VR} = 1$ iff the shared cache is noise-free. We connect $\mathrm{VR} < 1$ to a best-of-k excess regret that grows with k and σ; at $\sigma = 2$, $k = 64$ the realized search gain is more than halved (Fig. 1).
2. **Real-model confirmation (§4).** On a 1.4B distilled video world model we verify (i) that the distilled student is genuinely stochastic (so the phenomenon has a subject), (ii) that the shipped clean-cache recipe is *exactly* exchangeable ($\mathrm{VR} = 1.0000$, for every prompt, at every horizon tested), (iii) that injected cache noise drives VR from 1.0 to 0.25 along a curve matched by a one-parameter σ^2 law, and (iv) that the damage is decision-level: under best-of-k on the real model, a shared cache retains only $\approx 1/4$ of the search gain of independent branches, on top of a bias floor no selector removes.
3. **A repair that no baseline achieves, and a failure mode the repair can hide (§5).** We introduce *Fork-Invariance*, a training objective that pairs the noised rollout against a clean-cache reference on a shared sampling tape, so the sampling noise cancels exactly in the difference of sibling means and the branch-invariant component—invisible to every per-sample marginal loss—becomes the quantity minimized. The natural alternatives fail: serving-side re-keying moves neither VR nor the offset, a tree rollout restores VR only to the untrained level, and naive consistency fine-tuning *damages* it ($0.802 \to 0.730$). The objective itself then splits into two regimes that VR cannot tell apart: at full weight it attains $\mathrm{VR} = 0.964$ by converging to a solution that ignores the cache (pixels collapse under two independent VLM judges, $t = -8.2/-8.5$; continuations depend on the sampling seed more than on the prefix), and because Adam normalises away the loss scale, ten times less weight reproduces the damage. The working lever is the gradient *share*: capped at 0.2 of the DMD gradient, the same objective yields the only configuration we found that is above untrained on VR (0.888), pixel-intact, and still cache-reading, lifting the retained best-of-k search gain from 52% to 84% at matched VR.
4. **A statistic that cannot certify a fork (§4, §5).** Because a fork corrupts the siblings' *joint* law while the variance ratio is marginal, VR bounds neither the damage nor the repair: a model can reach its best VR by ceasing to condition on the cache, and matching VR across arms does not match their decision-layer performance. We treat VR as necessary but not sufficient, and pair it with pixel and prefix-sensitivity checks throughout.
5. **A practitioner's audit (§6).** We trace the cache-writing noise level through shipped pipelines and give a decision procedure—readable from a config—for whether a given model may be safely forked.

2 Setup and problem statement

Distilled autoregressive-diffusion world models. We study chunked autoregressive-diffusion video world models in the Self-Forcing / CausVid lineage (Huang et al., 2025; Yin et al., 2025), built on a Wan2.1 DiT backbone (Team Wan et al., 2025). Generation proceeds in blocks of a few latent frames; each block is produced by a few-step distilled sampler (DMD-style; Yin et al., 2024), and a rolling per-layer KV cache summarizes the context so far. After a block is denoised, the pipeline performs a *cache-write pass*: it re-runs the model on the finished block at a designated timestep and stores the resulting keys/values. The timestep of that write, which we call the *context noise* level, is a

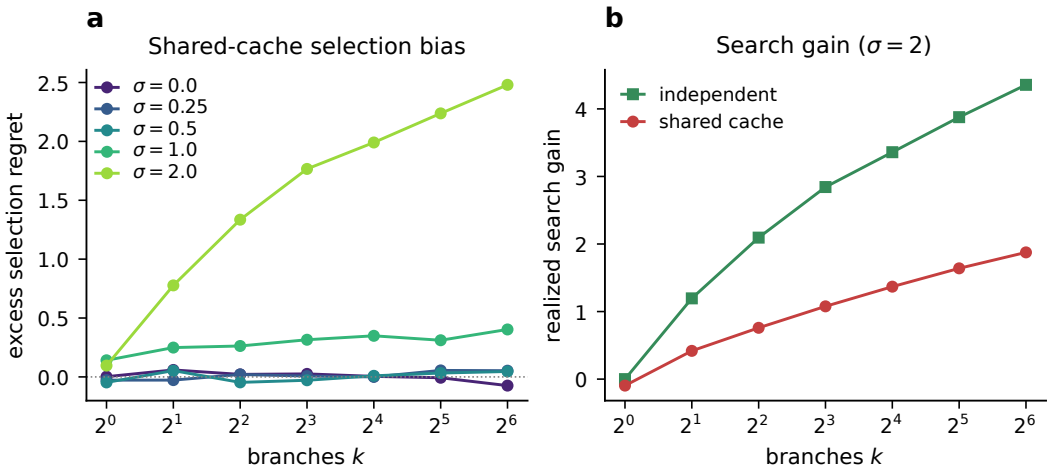

Figure 1: **The pathology is a decision-level failure.** Left: best-of-k excess regret from shared-cache branching grows with k and cache noise σ; the $\sigma = 0$ curve stays within noise of zero (clean-cache immunity). Right: at $\sigma = 2$ the realized search gain is more than halved relative to independent branches. Linear-Gaussian analytic model, scorer SNR $= 4$.

configuration knob, and recent members of this family expose it as a training-time hyperparameter set above zero (Liu et al., 2026).

Forking. A fork at state s snapshots the KV cache; k branches then each append their own generated blocks. Under CoW the branches physically share the parent's cache bytes. Branch identity is realized through the exogenous randomness a branch consumes: the initial latent it draws and the per-step re-noising it applies.

The question. Let x be a continuation grown from the forked prefix. Best-of-k and its relatives assume the k siblings $\{x^{(1)}, \ldots, x^{(k)}\}$ are *exchangeable*: their joint law is invariant to permutation, and each marginal equals the model conditional $P_\theta(\cdot \mid s)$ that an independently re-materialized rollout would sample. We ask when this holds, and quantify the damage when it does not.

3 The Pathology, Exactly

A tractable model. Let the true state be s, and let the context written into the cache be $c = As$. If the cache is written at noise level σ, the value physically stored is $c' = c + \varepsilon$, $\varepsilon \sim \mathcal{N}(0, \sigma^2 I)$. A continuation reads the cache and samples $x \mid c' \sim \mathcal{N}(Wc', \Sigma)$, where Σ is the legitimate conditional variance the model would express and W maps the cached context to the continuation mean. Two forking protocols differ only in how ε is shared:

- **shared** (copy-on-write): all k siblings reuse one cache, hence one ε;
- **fresh** (re-materialize per branch): each sibling draws its own ε_i.

Theorem 1 (Under-dispersion law)**.** *Under the model above,*

$$\mathrm{Var}[x \mid c, \varepsilon] = \Sigma, \qquad \mathrm{Var}_{\mathrm{fresh}}[x \mid c] = \Sigma + \sigma^2 W W^\top,$$

so the variance ratio between shared and fresh siblings is

$$\mathrm{VR} \;:=\; \frac{\mathrm{Var}_{\mathrm{shared}}}{\mathrm{Var}_{\mathrm{fresh}}} \;=\; \frac{\Sigma}{\Sigma + \sigma^2 W W^\top} \;\leq\; 1,$$

with equality iff $\sigma = 0$. Moreover the shared siblings share a mean offset with squared magnitude $\sigma^2 \,\mathrm{tr}(WW^\top)/P$ (for P output dimensions).

Table 1: **Real-model variance ratio vs. cache noise** (Self-Forcing, 1.4B; 5×5 grid, one continuation block). The shipped clean-cache recipe ($\sigma = 0$) is exactly exchangeable. A one-parameter law $\mathrm{VR} = \Sigma/(\Sigma + c\sigma^2)$ with $\Sigma = 0.00753$, $c = 0.01854$ (single fit) tracks the measured curve; nuisance grows as σ^2. Beyond $\sigma \approx 1.2$ the perturbation leaves the linear-response regime.

σ (frac. of cache RMS)	VR measured	VR predicted	nuisance measured
0.00	**1.0000**	1.0000	0.00000
0.15	0.9631	0.9475	0.00028
0.30	0.8493	0.8186	0.00119
0.60	0.4584	0.5301	0.00680
1.20	0.2501	0.2200	0.02776
2.40 (excl. from fit)	0.3949	0.0659	0.07847

The proof is a direct second-moment computation (Appendix C). Two consequences matter. First, *clean-cache pipelines are exactly immune*: $\sigma = 0 \Rightarrow \mathrm{VR} = 1$ and zero mean offset. Second, the failure is *dose-dependent* and continuous in σ, so it can be characterized rather than merely flagged. A Monte-Carlo instantiation ($P = D = 64$) matches the theorem to 1–3% across the curve ($\mathrm{VR} = 1.00/0.80/0.50/0.20$ measured at $\sigma = 0/0.5/1/2$, against identical theoretical values).

From variance ratio to decision damage. Under-dispersion is a symptom; the consequence is biased selection. We measure *excess regret*: the extra distance-to-oracle incurred when best-of-k selects among shared-cache siblings rather than independent ones, holding the scorer's signal-to-noise ratio fixed. Figure 1 shows the excess regret is negligible at $\sigma = 0$ for all k (immunity again), and grows *monotonically in both k and σ*: more branches cannot average away an error every branch shares. At $\sigma = 2$, $k = 64$ the realized best-of-k search gain falls from 4.36 (independent branches) to 1.88 (shared cache)—more than halved.

4 Confirmation on a real 1.4B world model

The theory is exact but linear-Gaussian. We now show the same mechanism governs a real 1.4B distilled video world model (Self-Forcing on Wan2.1-T2V-1.3B), running its shipped few-step sampler and rolling KV cache.

The student is genuinely stochastic. If a few-step distilled student were effectively deterministic (ODE-like), forking would have no subject. It is not: the pipeline is bit-deterministic given identical inputs (mean $|a-b| = 0$ across repeated runs), isolating measured spread from kernel nondeterminism, yet across initial draws the continuation spread is $2.13\times$ the natural frame-to-frame variation of the signal, growing with horizon ($0.26 \to 0.48$ over 9 frames)—exactly the Σ of Theorem 1.

Measuring VR by variance decomposition. Estimating VR as a ratio of two separately measured k-sample variances is statistically noisy. Instead we evaluate an $n_s \times n_c$ grid of continuations from one clean prefix, where row i fixes the per-sibling *sampling* seed (initial latent and re-noising tape) and column j fixes an independent cache *noise realization* at level σ. This yields a clean decomposition,

$$\text{signal} = \mathbb{E}_j \operatorname{Var}_i x(s_i, r_j), \qquad \text{nuisance} = \mathbb{E}_i \operatorname{Var}_j x(s_i, r_j), \qquad \mathrm{VR} = \frac{\text{signal}}{\text{signal} + \text{nuisance}},$$

where "signal" is the legitimate sampling spread at a fixed cache (the shared protocol) and "nuisance" is the extra dispersion an independent cache injects. At $\sigma = 0$ every cache is identical, so nuisance $= 0$ and $\mathrm{VR} = 1$ by construction. Because the shipped recipe writes a clean cache ($\sigma = 0$), we emulate the noised-cache lineage by injecting Gaussian noise of scale $\sigma \cdot \mathrm{RMS}$ into the cached keys/values, holding everything else fixed.

Result. Table 1 and Figure 4 (Appendix F) report the outcome. (i) *Immunity is exact*: at $\sigma = 0$, $\mathrm{VR} = 1.0000$ and the nuisance variance is identically zero. This turns "clean-cache pipelines are safe to fork" from a paper inference into a reproducible fact on the real network. (ii) *Contamination monotonically destroys exchangeability*: VR falls $1.00 \to 0.96 \to 0.85 \to 0.25$ as σ rises. (iii) *The*

same σ^2 law from Theorem 1 governs the deep model: a single-coefficient fit predicts the entire curve through $\sigma = 1.2$ (measured 0.25 vs. predicted 0.22), breaking only under extreme corruption ($\sigma = 2.4$, where the model's response saturates). The linear-Gaussian theory is not a toy analogy; it is quantitatively predictive of a 1.4B network. (iv) *The magnitude is honest*: at small σ the effect is mild (VR ≈ 0.96–0.85) because the student's own sampling spread still dominates the nuisance—the severity of the branching bias is exactly the ratio of cache-noise variance to the student's sampling variance, an interpretable quantity a practitioner can reason about.

Robustness across horizons and prompts. Both properties survive the two natural stress tests. Repeating the sweep at imagination horizon 3 (9 generated frames), immunity remains exact (VR $= 1.0000$ at $\sigma = 0$) and the under-dispersion is horizon-stable (VR $= 0.96, 0.86, 0.55, 0.27$ vs. $0.96, 0.85, 0.46, 0.25$ at horizon 1; Appendix F)—the shared cache-noise realization keeps biasing every sibling for the full imagined trajectory, which is exactly the regime tree search operates in. Repeating it over four diverse prompts, each with its own prefix noise, gives exact immunity and a monotone dose response for all four, with content-dependent magnitudes that track the student's own sampling spread Σ as Theorem 1 predicts (Table 4).

Anchoring σ to the shipped configuration knob. The noised-cache lineage does not literally add Gaussian noise of scale $\sigma \cdot \mathrm{RMS}$: in its training pipeline the finished block is re-noised by the diffusion forward process at timestep $t =$ `context_noise` before the cache-write pass—and the noise drawn there is precisely the branch-invariant state of Theorem 1. We implemented faithful noised-cache serving (re-noise at t, then write) and repeated the ANOVA measurement natively, rebuilding the entire prefix per cache realization with identical trajectory randomness, so columns differ *only* in the write noise. On the same 1.4B model: VR $= 1.0000, 0.984, 0.938, 0.785, 0.554$ at `context_noise` $= 0, 25, 50, 100, 200$. Immunity is again exact at the shipped value 0; at `context_noise` $= 100$ a CoW fork is already substantially under-dispersed (VR ≈ 0.78, corresponding to $\sigma \approx 0.42$ on the emulated axis). The question "may I fork this model?" is thereby answered in the pipeline's own configuration units.

Decision damage on the real model. Finally we convert VR into what a planner actually loses, on the real network. From one clean prefix we grow k siblings under three protocols that share the *same* per-sibling sampling seeds, so differences isolate the cache treatment: **clean** ($\sigma = 0$), **shared** (one cache-noise realization read by all k siblings—a CoW fork of a noised-cache pipeline), and **independent** (a fresh realization per sibling—what re-materialization provides, and the dispersion that path-keying restores). The scorer is transparent: distance to a target defined as the mean of $M = 8$ held-out clean-cache continuations (a proxy for the noise-marginalized conditional mean; its estimation error is a constant shared by all protocols), with both an oracle selector and a fixed-scale noisy selector (Appendix E). Figure 2 (Appendix B) shows two separate injuries. First, a *bias floor*: at $\sigma = 1.2$ every noised-cache protocol selects $\approx 15\times$ worse than clean (0.115 vs. 0.0075)—most of it the injected noise itself—and no selector can remove it, because it is the shared mean offset (§5). Second, *search-gain collapse*: growing k from 1 to 8 cuts the independent protocol's error by 0.0042 but the shared protocol's by only 0.0010—a CoW fork retains about a *quarter* of the search gain that valid branches would provide, and its best-of-8 barely beats its $k = 1$. The collapse is a decision-layer quantity that a serving-side re-keying cannot repair, but training largely restores it: at matched variance ratio, the retained search gain rises from 52% (untrained) to 84% (trained; §5). One caution about reading VR as a certificate: the same VR $= 0.554$ is reached here by two models that retain 52% and 84% of the gain respectively; a broken variant with *higher* VR retains more than either because it no longer reads its cache at all (§5); and the matched-VR difference does not replicate on our second prompt. We therefore treat VR as necessary but not sufficient, and pair it with pixel and prefix-sensitivity checks throughout.

5 Two failure modes, one repairable

The pathology of §4 decomposes into two quantities that earlier work—and our own first draft—conflated. Distinguishing them decides who can fix what.

Mode 1: the mean offset is *correct* behaviour, and it is invisible to marginal training. When k siblings share a cache written with noise ε, every sibling's conditional mean is shifted by $W\varepsilon$. This

Table 2: **Training arms at context noise 100.** The first five rows use four paired seed offsets, each measured on a 6×6 grid; FI-gm is a single 8×8 measurement. Parentheses in the first column give training-loop steps. Naive ctx100 uses marginal training on noised paths; fork100 uses marginal training on tree rollouts. Both FI arms use paired clean references, with a gradient-share cap for FI-gm. The full-weight FI arm has the highest VR but fails the pixel and cache-sensitivity checks. Pixel failures show the Qwen2.5-VL paired t statistic (Qwen3-VL: $t = -8.5$); cache checks show prefix-over-sampling sensitivity where available.

training arm (steps)	VR	mean offset	pixels	cache check
base (untrained)	0.802 ± 0.021	0.00269	✓	✓
naive ctx100 (300)	0.730 ± 0.029	0.00314	✓	✓
fork100 (300)	0.800 ± 0.022	0.00250	✓	✓
fork100 (500)	0.784 ± 0.024	0.00269	✓	✓
FI (full weight)	0.964 ± 0.025	0.00129	$\times$ (-8.2)	$\times$ ($0.30\times$)
FI-gm (100)	**0.888**	0.00366	✓	✓($1.8\times$)

shift is not an error the model makes: *given the contaminated context it was handed*, $W(c + \varepsilon)$ is the correct conditional mean, and a model that suppressed it would be ignoring its own context. It is therefore invisible to the objective training normally minimizes: a per-sample marginal loss sees only "this contaminated cache produced this continuation," which a model that treats the cache noise as signal and a model that discards it fit equally well. Consistently, neither naive consistency fine-tuning at the noised write nor a tree rollout removes it (*ctx100* and *fork100*, Table 2), and the one arm in which it does move has also stopped reading its cache (below), so that movement is a withdrawal rather than a repair.

Serving-side keying cannot reach it. In principle a serving layer could: if *every* draw a branch consumes—including the noise committed when the shared prefix's cache was written—were keyed by a pseudorandom function of the branch's full (action, outcome-index) path, siblings from a shared fork would be computationally indistinguishable from independent draws, with no change to the model (Proposition 1, stated and proved in Appendix B). A real CoW deployment, however, violates the premise: the prefix is generated *once*, and its noise realisation is committed to the shared cache pages before any branch-local draw happens, so the natural implementation re-keys only noise that was never shared in the first place. We verified this directly on the 1.4B model: a serving-side path-keyed arm moves *neither* VR nor the mean offset at any cache-noise level (`context_noise` 25–400; paired $|t| < 1.3$, all n.s.), while a prefix-re-derivation arm that *does* re-draw the committed write noise (*replay-fork*) cuts the offset by 1.9–$2.5\times$ ($|t| = 14$–27) but is capped by the model's own learned dispersion. The full protocol comparison, and a third violation (shared re-noising of emitted chunks, which no inference-side keying repairs), are in Appendix B.

Mode 2: under-dispersion is a property of training, and training *can* move it. The variance ratio, unlike the offset, is not pinned by the cache: it measures how correlated the siblings' sampling noise is, which is what the student learned. Here training is the right lever—and, surprisingly, naive training is the wrong direction.

Naive consistency fine-tuning *damages* exchangeability. We fine-tuned the same 1.4B student on the noised-cache training semantics at `context_noise` $= 100$ for 300 steps (*ctx100*; everything else unchanged), and measured VR on four independent seed offsets: it *falls* from the untrained 0.802 to 0.730 (paired $t = 9.1$; Welch $t = 4.0$). Training the model to be accurate under its own noised cache makes its siblings *more* correlated, not less: the student learns to reproduce the very noise realisation its context carries, which is exactly the correlation a fork must not inherit.

Forked-prefix self-forcing restores it. The fix is to change the *sampling process*, not the loss. In *forked-prefix self-forcing* each gradient step rolls out a tree rather than a path: a shared prefix is generated once, its KV cache is physically replicated k times (a literal copy-on-write fork, in-graph), and the k siblings continue with independent initial latents, re-noising tape, and cache-write noise. The distillation loss is untouched; only what the model sees during training changes. With this tree rollout (*fork100*)—run under a lighter per-step supervision footprint (one block per step, to fit the

forked graph)—the same measurement gives $\mathrm{VR} = 0.800$—statistically indistinguishable from the untrained 0.802 (paired $t = -0.3$) and clearly above ctx100 (paired $t = 7.3$). Forked-prefix training thus *removes the damage* that naive fine-tuning inflicts, though at this budget it does not exceed the untrained model.

Fork-Invariance: making the shared component visible. Can training reach what a marginal objective cannot even see? The diagnosis above names the missing ingredient. A marginal loss is *unidentified*: it sees only "this contaminated cache produced this continuation," which both a model that treats the cache noise as signal and a model that discards it fit equally well. The comparison that would separate them—*the same continuation under a clean cache*—is never presented. *Fork-Invariance* supplies it. Each step runs two arms over the same prefix content and the same pre-drawn sampling tape, differing in exactly one respect: whether the cache write is noised. Because the arms share the tape, the sampling noise cancels exactly in the difference of their sibling means, leaving the branch-invariant component, and the objective is that difference normalised by the clean arm's own sibling spread:

$$\mathcal{L}_{\mathrm{FI}} \;=\; \frac{\|\overline{x}^{\,\mathrm{noised}} - \overline{x}^{\,\mathrm{clean}}\|^2}{\mathrm{spread}_{\mathrm{clean}}} \;=\; \frac{k-1}{k}\Big(\frac{1}{\mathrm{VR}} - 1\Big),$$

i.e. Theorem 1 turned into a loss: it vanishes exactly when the fork is exchangeable, and it costs one extra no-gradient arm. Unlike a serving-side keying scheme—which cannot reach noise already committed to the shared prefix—Fork-Invariance acts where the contamination is produced.

The objective has two regimes, and the variance ratio cannot tell them apart. Our first implementation added $\mathcal{L}_{\mathrm{FI}}$ to the DMD loss with a weight of 1.0 (*w1.0*). It is the strongest result on the variance-ratio axis we have measured ($\mathrm{VR} = 0.964 \pm 0.025$, four seed offsets), and the model is unusable: scoring its decoded frames against four prompts with two independent frozen VLM judges, the city, ocean and fire videos are severely corrupted (paired $t = -8.2$ and -8.5 on a 0-to-10 quality scale, n=16), and only the lowest-frequency prompt (a sunset lake) survives. A direct prefix-sensitivity probe—two different clean prefixes, same continuation seed—explains why: on the city prompt the trained model's continuation depends on the prefix content *less* than on its own sampling seed ($0.30\times$; untrained $3.3\times$), i.e. it has stopped reading its cache. The "cure" was the model ceasing to condition on the contaminated cache at all.

This reading also disposes of a claim we had previously made on w1.0's authority: that the paired objective removes the mean offset. A model that has stopped reading its cache has no branch-invariant offset to inherit, so w1.0's offset reduction is the same artifact as its VR—a withdrawal, not a repair. Consistently, our healthy model does not move the offset (raw mean offset 0.0037–0.0041 against the untrained 0.0028); on the evidence we have, the offset remains training-invisible for any model that still conditions on its cache.

The mechanism is structural, and it matters beyond this paper. The loss numerator is $\|W\varepsilon\|^2$ with ε the freshly drawn write noise, which the model cannot predict; the only legitimate way to shrink it is Bayesian denoising of the cache read, and that route has a floor. Past the floor, the gradient has two degenerate routes left—shrink W (stop reading the cache), or inflate the sampling spread in the detached denominator—and the marginal DMD objective, which asks only that the video look like *some* video for this prompt, does not defend against either. We then tested the obvious control, ten times less weight (*w0.1*). It reproduced w1.0's damage at the same step count ($\mathrm{VR} = 0.888$ vs. 0.853 at step 50; the city frame near-black in both), which identifies the loss scale as the wrong variable: $\mathcal{L}_{\mathrm{FI}}$'s gradient is coherent across batches and *eight times* the DMD gradient at initialisation, so under Adam (which normalises each coordinate, β_1=0) the optimiser is insensitive to the loss *scale*. At w=1.0 the paired term supplies 89% of the update direction; at w=0.1 it still supplies 45%. Scaling cannot fix a direction.

What fixes it is constraining the *gradient share*. We cap the paired term's gradient at the rollout latent to a fixed fraction (0.2) of the DMD gradient norm, evaluated online (an all-reduce keeps the multiplier identical on every FSDP rank); the same configuration can be run at a fixed equivalent weight once the online measurement has been made (0.025 reproduces it, and avoids an extra autograd pass). This model (*FI-gm*, 100 steps) is the one the rest of this section reports, and it is the only configuration we found that satisfies all three of: VR well above untrained, pixel-level generation indistinguishable from base under both judges, and a model that still reads its cache ($1.8\times$ prefix-

over-sampling sensitivity, versus 3.3 untrained and 0.30 for w1.0). Table 2 collects all training arms.

What this means. Serving-side re-keying cannot repair a CoW fork after the fact, and even the best serving-side protocol applied to the untrained model (replay-fork, $\mathrm{VR} = 0.872$) does not reach a *plain shared* fork of the gradient-matched trained model (0.888; Appendix B). Fork-Invariance with a capped gradient share repairs the dispersion once, at training time, with no change to serving; neither training regime moves the mean offset, consistent with Mode 1. The full-weight case, finally, is the sharpest illustration of this paper's own thesis—a VR value can look cured while the thing it measures has been destroyed—and it is why our evaluation protocol decodes to pixels and probes prefix sensitivity before accepting any VR.

6 Which pipelines are affected

Theorem 1 makes "may I fork this model?" answerable from a single configuration value: the cache-writing noise level. **Immune (clean cache, $\sigma = 0$):** the Self-Forcing / CausVid recipe writes the KV cache from the finished, fully denoised block at timestep zero—we verified in the shipped code that the cache-write pass reruns "with timestep zero to update KV cache using clean context" with `context_noise: 0` in the config (Huang et al., 2025; Yin et al., 2025), and §4 confirms the consequence ($\mathrm{VR} = 1.0000$, nuisance identically zero), so forks of any such pipeline are exactly exchangeable. **Vulnerable (noised cache, $\sigma > 0$):** the lineage that exposes context noise as a training-time hyperparameter and sets it above zero (Liu et al., 2026) writes a cache that carries a branch-invariant noise state and inherits $\mathrm{VR} < 1$; our native measurement quantifies it in the pipeline's own units, $\mathrm{VR} = 0.98/0.94/0.78$ at `context_noise` $= 25/50/100$ (§4). **Vulnerable via re-noising:** pipelines that re-noise and recompute emitted chunks (overlap/backward-noising variants) incur the shared re-noising violation of Appendix B, which no serving-side keying repairs. For the models most people ship today, then, CoW forking is safe, and we can say so with a reproducible $\mathrm{VR} = 1.0000$.

7 Related work

World-model serving and forking. PagedAttention manages KV memory through paging (Kwon et al., 2023), and ForkKV supports copy-on-write sharing of disaggregated caches across LoRA agents with a quality–efficiency tradeoff (Wang et al., 2026). Execution-state capsules instead snapshot and restore complete execution boundaries over contiguous static buffers, reporting byte-exact stored-state restoration in their evaluated setting (Su, 2026). These address memory reuse and execution-state recovery. We study the statistical dependence among continuations that share a noisy cache. Diffusion Forcing (Chen et al., 2024) provides a related setting for generating multiple futures under per-token noise schedules.

Branching, diversity, and dependence. Uncertainty DMD identifies collapsed first-chunk uncertainty and its propagation through an autoregressive cache, uses a total-variance decomposition, and injects stochasticity during training and inference (Duan et al., 2026). BranchGRPO explicitly combines shared and independent Gaussian noise to control branch correlation, studies that correlation through ablations, and reports marginal-distribution comparisons using MMD^2/KID (Li et al., 2025). Thus, prior work already studies cache uncertainty and branch dependence. Our focus is the effect of a committed noisy prefix on sibling dispersion and downstream selection, evaluated together with decoded quality and prefix sensitivity. Matching individual-branch marginals alone does not determine the joint law of the branch group.

Invariance training and paired estimators. Sharing a pre-drawn randomness tape across two arms is an instance of common random numbers (Owen, 2013). DMD provides the underlying distribution-matching objective (Yin et al., 2024). Invariant risk minimization and risk extrapolation address stability across training environments (Arjovsky et al., 2019; Krueger et al., 2021), whereas consistency models learn consistency across noise levels along a generative trajectory (Song et al., 2023). Our paired term compares clean and noised cache writes under the same prefix content and

random tape. Exact cancellation of the sampling term requires the additive model assumptions; in the nonlinear generator the paired loss is a surrogate, evaluated alongside the empirical variance ratio.

Model-based search. Best-of-k, MCTS, and value-guided search in imagination assume the branched samples are valid draws (Schrittwieser et al., 2020; Hafner et al., 2023); our result identifies a regime where that assumption fails silently and quantifies the selection bias. The failure is a cousin of reward-model overoptimization (Gao et al., 2023): best-of-k over-selects on a shared error, except here the shared error is injected by the serving-time cache rather than by the scorer.

8 Conclusion

Branchability is a property of the world model and how its exogenous noise is threaded through shared state, not merely of the serving layer that snapshots it: a forked cache can be byte-exact and still yield under-dispersed, non-exchangeable siblings that bias best-of-k search. We characterized this exactly, confirmed it on a real 1.4B distilled world model—exactly exchangeable under the shipped clean-cache recipe, decaying along the predicted σ^2 law under a noised cache—and separated it into two failure modes of unequal tractability: the mean offset is committed to the shared prefix and survives every method we tried that leaves the model intact, while the loss of dispersion is repairable, though naive fine-tuning *damages* it and a tree rollout merely restores the untrained level.

The training story turned out to have a second act, and it is the one we would want a reader to carry away. The objective implied by the theory—written down correctly, with the branch-invariant component made visible—*works* on the statistic it targets and destroys the model that produced it, because the same gradient that removes the shared component can also remove the model's dependence on its cache; the failure is invisible to VR and we found it only by decoding to pixels. What separates the repair from the ruin is where the gradient is allowed to act: matched to one-fifth of the distillation gradient rather than dominating it, the identical objective yields a model above untrained on VR, intact under both pixel judges, and still conditioned on its prefix. Whether a branch is valid is a property of the siblings' *joint* law; a marginal statistic can be matched while that law is corrupted, so the test for when imagination-time search over a forked world model is trustworthy must ultimately be behavioural, not scalar. Limitations of our study—the single-run status of the gradient-matched model, the single-prompt seed replication, the supervision-budget asymmetry between arms, and the reliance on frozen VLM judges rather than human evaluation—are discussed in Appendix A.

Reproducibility Statement

Our experiments run a single publicly released model, Self-Forcing on the Wan2.1-T2V-1.3B backbone (Huang et al., 2025; Team Wan et al., 2025), using its shipped few-step sampler and rolling KV cache; all fine-tuning starts from the released checkpoint and changes only the training objective described in §5 and Appendix E. Every measurement is bit-deterministic given its inputs (we verify this directly in §4: repeated runs with identical inputs give a mean absolute deviation of 0), so all reported quantities are functions of seeds we fix explicitly and list in the appendix. The variance decomposition of §4 is an $n_s \times n_c$ grid with the sampling seed on one axis and the cache-noise realization on the other; the choice $n_s = n_c = 6$ for the seed-replicated arms and $n_s = n_c = 8$ for the single-run arms is stated with each result. Statistical tests are paired over seed offsets wherever the same offsets are reused across arms (Table 2); elsewhere we state the test used and report t and n. Code, experiment configurations, and recorded result JSON files are available in our anonymous repository: `https://anonymous.4open.science/r/Forkready-1E88/`. The repository includes installation instructions, evaluation commands, and an index of the recorded results; model weights and full training datasets must be obtained separately as described in the repository. We flag in §5 and the Limitations that our pixel-level and prefix-sensitivity checks rely on frozen VLM judges rather than human evaluation, and we report both judges' verdicts rather than a single score.

AI USE STATEMENT

We used AI tools to polish the language and improve the readability of the manuscript.

ETHICS STATEMENT

Our work studies the statistical validity of branching in learned world models and proposes a training method to repair it; we do not develop or release any capability intended for harmful use, and the video model we experiment on is a publicly released open-weights text-to-video system. The failure mode we identify—a forked world model whose branches are under-dispersed and therefore mislead best-of-k selection or tree search—is a reliability concern for planning systems that act on imagined rollouts; we regard characterizing and repairing it as a safety-relevant contribution rather than a risk-introducing one, and we report the mode of failure that our own repair can hide (§5) precisely so that the field does not certify such systems on a single scalar. Our VLM judges are used only to score the quality and prompt-alignment of generated video and are not trained or fine-tuned on our data.

REFERENCES

Martin Arjovsky, Léon Bottou, Ishaan Gulrajani, and David Lopez-Paz. Invariant risk minimization. *arXiv preprint arXiv:1907.02893*, 2019. URL https://arxiv.org/abs/1907.02893.

Boyuan Chen, Diego Martí Monsó, Yilun Du, Max Simchowitz, Russ Tedrake, and Vincent Sitzmann. Diffusion forcing: Next-token prediction meets full-sequence diffusion. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024. URL https://arxiv.org/abs/2407.01392. arXiv:2407.01392.

Zixuan Duan, Xunzhi Xiang, Yabo Chen, Xin Zhang, Changhan Liu, Haibin Huang, Chi Zhang, Qi Fan, and Xuelong Li. Uncertainty DMD: Restoring diversity in few-step autoregressive video distillation. *arXiv preprint arXiv:2609.11265*, 2026. URL https://arxiv.org/abs/2609.11265.

Leo Gao, John Schulman, and Jacob Hilton. Scaling laws for reward model overoptimization. In *International Conference on Machine Learning (ICML)*, 2023. URL https://proceedings.mlr.press/v202/gao23h.html. arXiv:2210.10760.

David Ha and Jürgen Schmidhuber. Recurrent world models facilitate policy evolution. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2018. URL https://arxiv.org/abs/1809.01999. arXiv:1809.01999.

Danijar Hafner, Jurgis Pasukonis, Jimmy Ba, and Timothy Lillicrap. Mastering diverse domains through world models. *arXiv preprint arXiv:2301.04104*, 2023. URL https://arxiv.org/abs/2301.04104.

Xun Huang, Zhengqi Li, Guande He, Mingyuan Zhou, and Eli Shechtman. Self forcing: Bridging the train-test gap in autoregressive video diffusion. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2025. URL https://arxiv.org/abs/2506.08009. arXiv:2506.08009.

David Krueger, Ethan Caballero, Joern-Henrik Jacobsen, Amy Zhang, Jonathan Binas, Dinghuai Zhang, Remi Le Priol, and Aaron Courville. Out-of-distribution generalization via risk extrapolation (REx). In *International Conference on Machine Learning (ICML)*, 2021. URL https://proceedings.mlr.press/v139/krueger21a.html.

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph E. Gonzalez, Hao Zhang, and Ion Stoica. Efficient memory management for large language model serving with PagedAttention. In *ACM Symposium on Operating Systems Principles (SOSP)*, 2023. URL https://arxiv.org/abs/2309.06180. arXiv:2309.06180.

Yuming Li, Yikai Wang, Yuying Zhu, Zhongyu Zhao, Ming Lu, Qi She, and Shanghang Zhang. BranchGRPO: Stable and efficient GRPO with structured branching in diffusion models. *arXiv preprint arXiv:2509.06040*, 2025. URL https://arxiv.org/abs/2509.06040.

Jinxiu Liu, Xuanming Liu, Kangfu Mei, Yandong Wen, Ming-Hsuan Yang, and Weiyang Liu. Streaming autoregressive video generation via diagonal distillation. In *International Conference on Learning Representations (ICLR)*, 2026. URL https://openreview.net/forum?id=X7YW6STzeL. arXiv:2603.09488.

Art B. Owen. *Monte Carlo Theory, Methods and Examples*. https://artowen.su.domains/mc/, 2013.

Julian Schrittwieser, Ioannis Antonoglou, Thomas Hubert, Karen Simonyan, Laurent Sifre, Simon Schmitt, Arthur Guez, Edward Lockhart, Demis Hassabis, Thore Graepel, Timothy Lillicrap, and David Silver. Mastering atari, go, chess and shogi by planning with a learned model. *Nature*, 588 (7839):604–609, 2020. URL https://doi.org/10.1038/s41586-020-03051-4.

Yang Song, Prafulla Dhariwal, Mark Chen, and Ilya Sutskever. Consistency models. In *International Conference on Machine Learning (ICML)*, 2023. URL https://proceedings.mlr.press/v202/song23a.html.

Liang Su. Execution-state capsules: Graph-bound execution-state checkpoint and restore for low-latency, small-batch, on-device physical-AI serving. *arXiv preprint arXiv:2606.20537*, 2026. URL https://arxiv.org/abs/2606.20537.

Team Wan et al. Wan: Open and advanced large-scale video generative models. *arXiv preprint arXiv:2503.20314*, 2025. URL https://arxiv.org/abs/2503.20314.

Shao Wang, Rui Ren, and Lin Gui. ForkKV: Scaling multi-LoRA agent serving via copy-on-write disaggregated KV cache. *arXiv preprint arXiv:2604.06370*, 2026. URL https://arxiv.org/abs/2604.06370.

Tianwei Yin, Michaël Gharbi, Richard Zhang, Eli Shechtman, Frédó Durand, William T. Freeman, and Taesung Park. One-step diffusion with distribution matching distillation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024. URL https://arxiv.org/abs/2311.18828. arXiv:2311.18828.

Tianwei Yin, Qiang Zhang, Richard Zhang, William T. Freeman, Frédó Durand, Eli Shechtman, and Xun Huang. From slow bidirectional to fast autoregressive video diffusion models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025. URL https://arxiv.org/abs/2412.07772. arXiv:2412.07772.

A LIMITATIONS

Our native `context_noise` measurements run the noised-cache write semantics on the clean-cache-trained base model, and our fine-tuning study compares one budget (300 steps, plus a 500-step fork arm) at 1.4B scale. The gradient-share-matched model is reported at 100 steps as a *single* run: its VR $= 0.888$ carries no error bar of its own (a four-seed replication at the Table 2 protocol is in progress). The seed replication uses a single content prompt, and per-prompt effect sizes vary widely, so reported VR shifts are within-content comparisons; the claim that training does not move the mean offset rests on non-significance at $n{=}4$ rather than an equivalence test. Two limitations are structural. First, VR *is not sufficient*: matched-VR arms differ by $2.2\times$ in retained search gain, and any VR-only evaluation—including several in this paper's own earlier drafts—can be satisfied by a model that has stopped reading its cache, which is precisely how our full-weight variant scored 0.964 while being unusable. Second, the training budget asymmetry: forked-prefix arms supervise one block per step (the memory-feasible configuration) whereas the naive arm supervises all blocks, so "same steps" is not "same supervision". Finally, our best-of-k uses a transparent distance-to-conditional-mean scorer rather than a task reward (a binary VLM task probe saturates: 6 of 8 prompts fix the queried attribute, so all siblings agree—a resolution limit consistent with the theory, not counter-evidence), and our pixel and prefix-sensitivity checks use frozen VLM judges and a two-prefix probe rather than human evaluation: sufficient to detect damage this severe (both judges agree, $t \approx -8$), but not calibrated to certify a repaired model as *undamaged*.

B SERVING-SIDE PROTOCOLS: PATH-KEYING, REPLAY-FORK, AND THE COMMITTED PREFIX

Path-keyed noise. If the exogenous noise a branch consumes is a deterministic function of that branch's identity—keyed by a pseudorandom function (PRF) on the full *(action, outcome-index)* path from the fork point—then siblings that took different paths draw independent noise, while a branch replayed along the same path is reproducible. Crucially the key must chain the entire outcome-index history: keying on the action sequence alone couples two branches that diverged at an earlier stochastic node but later chose the same actions.

Proposition 1 (Path-keying restores exchangeability)**.** *If* all *exogenous noise a branch consumes—including the noise committed when the shared prefix's cache was written—is drawn as* $\varepsilon_t = \mathrm{PRF}(\textit{root}, a_{1:t}, u_{1:t})$ *keyed on the full path, sibling continuations from a shared forked prefix are exchangeable and computationally indistinguishable from independent draws from* $P_\theta(\cdot \mid s)$*, without any change to the model weights or loss.*

Why the obvious serving-side fix does not work in practice. Proposition 1 requires that the prefix's own cache-write noise be a function of the branch's identity. In a real CoW deployment it is not: the prefix is generated *once*, and its noise realisation is already committed to the shared cache pages before any branch-local draw happens. Re-keying only the sibling's downstream draws—the natural implementation—therefore re-keys noise that was never shared in the first place. We verified this directly on the 1.4B model: a serving-side path-keyed arm moves *neither* VR nor the mean offset at any cache-noise level (`context_noise` 25–400; paired $|t| < 1.3$, all n.s.), while a prefix-re-derivation arm that *does* re-draw the committed write noise cuts the offset by 1.9–$2.5\times$ ($|t| = 14$–27, $p < 10^{-3}$). The distinction is the whole point: the branch-invariant state is the *shared prefix's* write noise, and it must be addressed where it enters.

A third, serving-invisible violation. When emitted chunks are re-noised and recomputed—the practice of overlap and backward-noising pipelines—path-keying overshoots: at a clean cache but with shared re-noising, path-keyed VR $= 2.00$ (over-dispersed), while clean stays at 1.00. A shared re-noising injection both offsets the siblings and inflates their variance, and no inference-side keying repairs it; this one additionally wants the training objective.

Table 3: **No serving-side protocol recovers what training provides** (Self-Forcing 1.4B; VR under each serving protocol, eight independent seed bases, 6×6; `context_noise` $= 100$ and 400). *Path-keyed* re-keys branch-local draws only—the natural serving fix—and moves neither quantity. *Replay-fork* re-derives the prefix's committed write noise and helps, but its ceiling is the model's own dispersion. The trained rows are the *full-weight* variant; its VR and offset are the strongest we obtain, but it is the model that fails pixel-level inspection (§5), so this table should be read as a comparison of *serving protocols applied to the untrained model*—where replay-fork is the best available and still falls short of a plain shared fork of even the gradient-share-matched model (VR $= 0.888$ at ctx 100, pixels intact)—rather than as an endorsement of the full-weight trained rows.

		VR (@ctx 100)		VR (@ctx 400)	
model	serving protocol	VR	mo_{raw}	VR	mo_{raw}
untrained	shared (CoW)	0.774	3.2e-3	0.400	1.8e-2
	path-keyed	0.775	3.3e-3	0.410	1.8e-2
	replay-fork	0.872	1.7e-3	0.716	6.5e-3
	independent	0.508	—	0.504	—
Fork-Invariance (full weight)	shared (CoW)	0.952	1.7e-3	0.778	1.0e-2
	path-keyed	0.952	1.9e-3	0.797	9.3e-3
	replay-fork	0.964	1.2e-3	0.842	6.7e-3
	independent	0.515	—	0.512	—

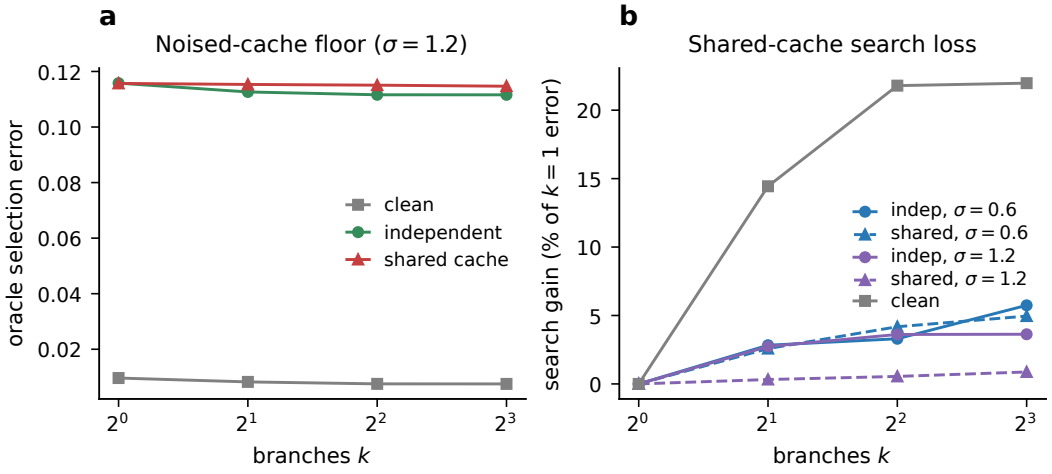

Figure 2: **Real-model best-of-k: two separate injuries.** Left: with a noised cache ($\sigma = 1.2$), selection error acquires a floor $\approx 15\times$ the clean-cache error regardless of protocol—the shared mean offset, most of which is the injected noise itself; no selector, serving protocol, or model-intact training removes it (§5). (Right) On top of the floor, the *shared* (CoW) protocol loses the benefit of k itself: its realized search gain is $\approx 0.24\times$ the independent-branch gain at $\sigma = 1.2$; at matched variance ratio the trained model recovers it to $0.84\times$. Oracle scorer; paired sampling seeds across protocols.

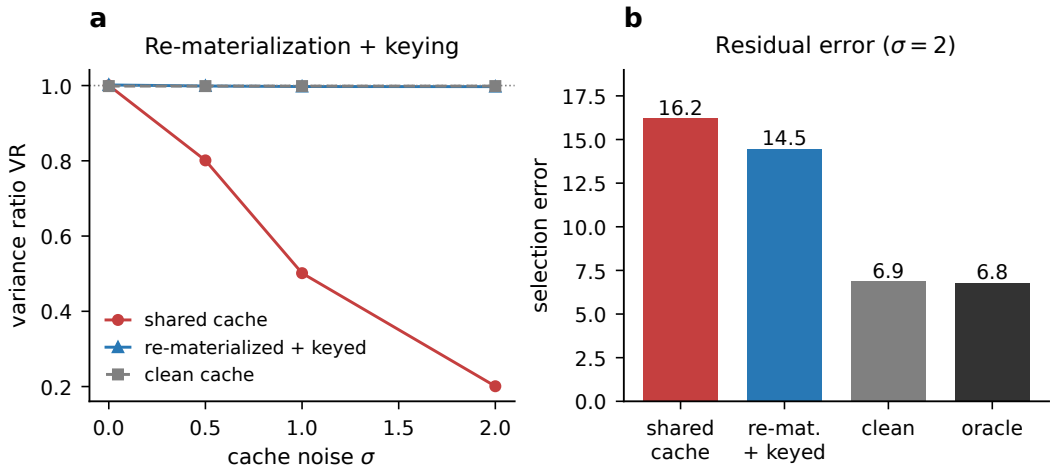

Figure 3: **Keying fixes dispersion, not the offset—and only if the committed prefix noise is keyed too.** Left: keying the noise to the full path restores VR $\to 1$ across cache-noise levels, at zero training cost. Right: yet the selection error at $\sigma = 2$ stays far above clean-cache, because a shared mean offset remains—exchangeable $\neq$ unbiased. On the real model, however, the offset is committed to the shared prefix before any branch-local draw, so serving-side re-keying moves neither quantity (verified in §5); training that keeps the model reading its cache likewise leaves the offset in place, and the one arm in which it moves has stopped reading its cache altogether. Linear-Gaussian analytic model, as in Figure 1; the serving-side null result is on the real 1.4B network.

C Proof of Theorem 1

Fix the true context c. A continuation reads the cached value $c' = c + \varepsilon$, $\varepsilon \sim \mathcal{N}(0, \sigma^2 I)$, and samples $x = Wc' + \eta$ with independent sampling noise $\eta \sim \mathcal{N}(0, \Sigma)$.

Conditional variance. Given both c and a fixed cache realization ε, the only remaining randomness is η, so $\mathrm{Var}[x \mid c, \varepsilon] = \Sigma$.

Shared protocol. All k siblings read one cache, i.e. one realization ε. Conditioned on that ε, sibling i is $x_i = W(c + \varepsilon) + \eta_i$ with i.i.d. η_i; the across-sibling variance is therefore $\mathrm{Var}_{\mathrm{shared}} = \Sigma$, independent of σ.

Fresh protocol. Each sibling draws its own ε_i, giving $x_i = Wc + W\varepsilon_i + \eta_i$ with $W\varepsilon_i \sim \mathcal{N}(0, \sigma^2 WW^\top)$ independent of η_i. Hence $\mathrm{Var}_{\mathrm{fresh}} = \Sigma + \sigma^2 WW^\top$.

Variance ratio. $\mathrm{VR} = \mathrm{Var}_{\mathrm{shared}} (\mathrm{Var}_{\mathrm{fresh}})^{-1} = \Sigma(\Sigma + \sigma^2 WW^\top)^{-1}$ (per-coordinate ratio in the isotropic case). Since $\sigma^2 WW^\top \succeq 0$, $\mathrm{VR} \preceq I$, with equality iff $\sigma = 0$.

Mean offset. Every shared sibling has conditional mean $W(c + \varepsilon)$, offset from the true conditional mean Wc by $W\varepsilon$. Its expected squared magnitude per output coordinate is $\mathbb{E}\|W\varepsilon\|^2/P = \sigma^2 \,\mathrm{tr}(WW^\top)/P$. □

D Proof of Proposition 1

Let the exogenous noise consumed at step t be $\varepsilon_t = \mathrm{PRF}(\mathrm{root}, a_{1:t}, u_{1:t})$, where $u_{1:t}$ is the outcome-index history. For two siblings whose paths differ at any step $\tau \le t$ (in an action a_τ or an outcome index u_τ), the PRF is queried at distinct keys; by PRF security its outputs are computationally indistinguishable from independent uniform draws, hence the induced ε_t are effectively independent. The siblings therefore recover the *fresh*-protocol law: across-sibling variance $\Sigma + \sigma^2 WW^\top$ and per-sibling marginal $\mathcal{N}(W(c + \varepsilon_i), \Sigma)$, which is exactly the marginal an independently re-materialized rollout samples—the definition of exchangeability. A branch replayed along an identical path queries the PRF at identical keys and receives identical ε, giving bit-level reproducibility. Keying on actions alone is insufficient: two branches that diverged at an earlier stochastic node but subsequently choose

the same actions would collide on the PRF key and re-share ε; chaining the full outcome-index history removes the collision. □

E EXPERIMENTAL DETAILS

Model. Self-Forcing (Huang et al., 2025) `generator_ema` (1.419B parameters), Wan2.1-T2V-1.3B DiT backbone (Team Wan et al., 2025), DMD-distilled few-step sampler (Yin et al., 2024) using the shipped `denoising_step_list`, `context_noise`=0.

Hardware / software. Single A100-40GB, PyTorch 2.5.1+cu124, bfloat16, SDPA attention (no FlashAttention). The generation pipeline is bit-deterministic given identical inputs.

Prefix and continuation. The prefix is built with the shipped clean-cache recipe (inference at `context_noise`=0): 2 blocks (6 frames, 9360 tokens across 30 layers). Continuations are 1 block (horizon-1, 3 frames) or 3 blocks (horizon-3, 9 frames). Resolution 480×832; latent dimensions height$/8 \times$ width$/8$; frame-sequence length 1560 tokens per block.

Cache perturbation. σ is a fraction of the per-layer cache RMS. Noise is added in float32 to the used region [0:global_end_index] of the keys and values, then cast back to bfloat16, deterministically as a function of a seed (so one seed reproduces a realization and distinct seeds give independent ones). A *sentinel* check asserts the perturbation reaches the output (mean absolute change $\geq 10^{-3}$; the realized relative change at the tested σ is 0.35).

Estimator. VR is estimated on an $n_s \times n_c = 5 \times 5$ grid: signal $= \mathbb{E}_j \mathrm{Var}_i$ (across sampling seeds at a fixed cache), nuisance $= \mathbb{E}_i \mathrm{Var}_j$ (across independent cache realizations), VR $=$ signal$/$(signal $+$ nuisance). At $\sigma = 0$ all caches coincide, so nuisance $= 0$ and VR $= 1$ by construction.

Multi-prompt sweep. The four prompts of Table 4 (a lakeside pan, a night intersection, a campfire close-up, ocean cliffs) each use their own prefix noise seed, the same 5×5 estimator, and horizon 1.

Table 4: **The law is not a property of one prompt** (four prompts, each with its own prefix; 5×5 grid, horizon 1). Immunity is exact everywhere; the dose response is monotone everywhere; larger sampling spread Σ buys robustness.

prompt	Σ (signal at σ=0)	VR at σ=0	σ=0.3	σ=0.6
lakeside pan	0.013	1.0000	0.881	0.524
night intersection	0.114	1.0000	0.895	0.605
campfire close-up	0.178	1.0000	0.929	0.732
ocean cliffs	0.034	1.0000	0.904	0.639

Native context-noise measurement. Faithful noised-cache serving mirrors the training pipeline's write semantics: after a block is denoised, it is re-noised by the scheduler's forward process at timestep $t =$ `context_noise` (fresh Gaussian noise from a dedicated write stream) and the cache-write pass runs at timestep t. Per cache realization j the *entire* prefix is rebuilt from scratch with identical trajectory randomness (same initial latents and re-noising tape) but write-noise seed W_j, so columns of the ANOVA grid differ only in the exogenous write noise; rows vary the continuation sampling seed as before. At $t = 0$ the write adds no noise and all columns coincide, giving VR $= 1$ exactly. A sentinel verifies that distinct write seeds change the continuation ($|a - b| = 0.089$ at $t = 200$, output scale 0.68).

Best-of-k. $K = 8$ siblings per repeat, $R = 6$ repeats per protocol, all protocols sharing the same per-sibling sampling seeds (paired design). Target $=$ mean of $M = 8$ clean-cache continuations from held-out seeds; scores are per-element MSE to the target. The oracle selector picks the true minimum; the noisy selector adds Gaussian noise of a fixed scale calibrated to the clean-cache score dispersion (SNR 4), identical for every protocol. Shared protocol: one cache-realization seed per repeat; independent: one per sibling; clean: $\sigma = 0$. Reported errors average over repeats; best-of-k for $k < K$ uses the first k siblings. The matched-VR decision comparison uses $R = 10$ repeats and per-repeat paired differences.

Fine-tuning arms. All arms start from the released checkpoint and train at `context_noise` $= 100$ with the shipped DMD recipe (lr 2×10^{-6}, total batch 64, 9-frame rollouts, 7×A100): *ctx100* is

the unmodified recipe; *fork100* physically replicates the prefix KV cache $k{=}2$ ways in-graph and continues siblings on independent draws; the Fork-Invariance arms add the paired objective, running a no-gradient clean-reference arm over the same prefix content and the same pre-drawn sampling tape each step. The gradient share is measured at the rollout latent as the ratio of the two loss gradients' norms (all-reduced across FSDP ranks) and the paired term is capped at 0.2 of the DMD gradient; the measured effective multiplier is stable (0.021–0.033), and a fixed weight of 0.025 reproduces it.

Pixel and prefix-sensitivity checks. Two frozen judges (Qwen2.5-VL-7B and Qwen3-VL-8B) score 9-frame decoded clips on 0–10 prompt-alignment and visual-quality scales, over 4 prompts $\times$ 4 seeds per arm with identical seeds across arms (paired, $n{=}16$). The prefix-sensitivity probe builds two clean prefixes from different trajectory seeds, continues each with the same sampling seed, and reports the continuation difference normalised by the same-prefix, different-seed difference; a cache-reading model scores well above 1 (untrained $3.3\times$), a cache-blind one near or below 1.

F HORIZON ROBUSTNESS

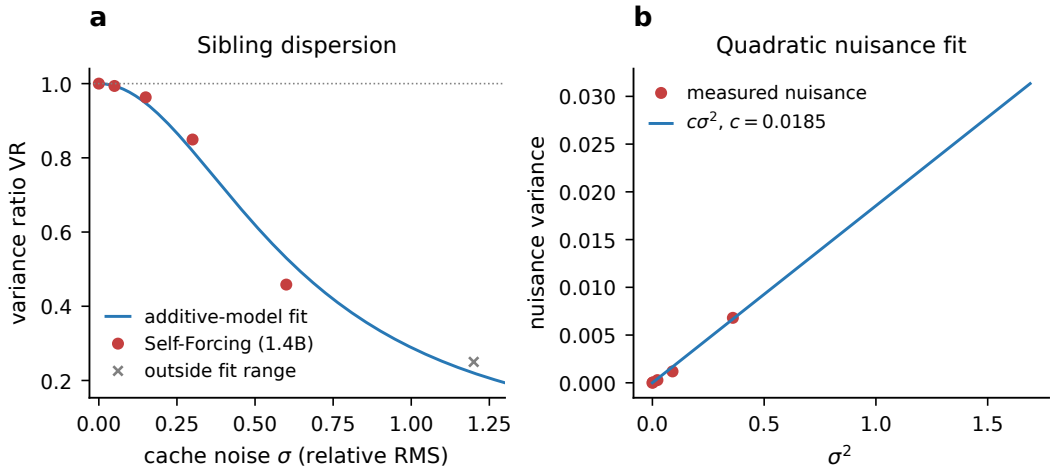

Figure 4: **The mechanism holds on a real** 1.4**B world model.** Left: measured variance ratio (points) falls on the analytic curve $\Sigma/(\Sigma + c\sigma^2)$; the clean-cache point sits exactly at $\mathrm{VR} = 1$. Right: the injected nuisance variance grows as σ^2, matching Theorem 1. Numbers in Table 1.

Table 5 reports the variance ratio at imagination horizons 1 and 3. Immunity is exact at both depths, and the under-dispersion is horizon-stable across the linear regime.

Table 5: Variance ratio versus cache noise at two imagination horizons (Self-Forcing, 1.4B; 5×5 grid). Immunity is horizon-exact; the effect is horizon-stable.

σ	VR (horizon 1, 3 frames)	VR (horizon 3, 9 frames)
0.00	1.0000	1.0000
0.15	0.9631	0.9629
0.30	0.8493	0.8604
0.60	0.4584	0.5481
1.20	0.2501	0.2660