TOWARDS RECURSIVE SELF-IMPROVEMENT: A POST-TRAINING PARADIGM FOR VIDEO GENERATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Post-training has advanced large language models by converting feedback into effective supervision. Extending this process to video generation remains challenging: quality is multidimensional, automated evaluators imperfectly reflect human preferences, and aggregate scores provide limited guidance on what to improve next. We introduce a closed-loop, backbone-agnostic post-training paradigm that connects capability-level evaluation to targeted interventions in data collection and training. The paradigm couples data and training loops through human-calibrated evaluation, combining task-specific metrics, order-controlled pairwise comparisons, and statistically grounded checkpoint promotion criteria. Experimental outcomes guide revisions to training recipes, data mixtures, and collection priorities, while validated improvements are carried forward into subsequent iterations. This feedback-to-intervention mechanism enables adaptation of both the generator and the process used to produce its successors, offering a practical path toward system-level recursive self-improvement. We instantiate the paradigm on several open-weight and licensed video foundation models, including LTX-2.3/2.5, Wan-2.2, and MiniMax-H3. Post-trained LTX checkpoints win 88.2% and 80.6% of decisive human comparisons; Wan-2.2 reduces audio–visual offset from 6.25 to 1.69 frames. H3 achieves an 85.3% advertising reference-fidelity pass rate, the highest among four models and six frontier video-generation APIs, while remaining competitive on general video quality. These results demonstrate the applicability of the paradigm across model families and provide a foundation for studying recursive self-improvement in video generation.

1 INTRODUCTION

Post-training a video generator is increasingly a problem of deciding *what to learn next*. For language models, mathematical answers and executable tests provide cheap feedback on many generated solutions (DeepSeek-AI, 2025). Video has no corresponding general-purpose verifier. A clip can become more realistic while changing a person's identity, deforming a reference product, missing a requested action, or desynchronizing speech. The relevant failure depends on the application, and a single quality score cannot reveal which capability needs new data, which judge needs repair, or which training objective should be tried (Huang et al., 2024; Zheng et al., 2025; Liu et al., 2025d).

This distinction matters because more optimization is useful only when its feedback resolves the failure of interest. A narrow capability such as speech timing admits a repeatable diagnostic and targeted supervised examples. Open-ended visual quality and advertising fidelity instead admit many acceptable outputs and require preferences over generated alternatives. Those preferences can support RL, but a noisy or misaligned video judge can make the advantage between rollouts less reliable than repeated ratings of the same clip (Liu et al., 2025a; Lian et al., 2025). Our recorded campaigns exhibit both regimes: staged supervised training improves speech synchronization, whereas an H3 reward audit finds judge variation comparable to between-rollout variation. The latter calls for changing the *supervision signal* before asking the generator to optimize it.

We ask: *can capability feedback improve the choice of subsequent supervision and interventions?* Our answer is a closed-loop post-training paradigm in which the object that improves is not only the video model, but also the process that produces its successors. At each round, a capability diagnosis identifies a failure and the reliability of its measurement; experiment history determines whether

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.

to collect new examples, revise an evaluator or reward, change the training recipe, or optimize serving; and comparison with the incumbent decides what the next round inherits. This *feedback-to-intervention decision* is the recursive unit. It differs from applying a fixed sequence of SFT, RL, prompting, and inference optimization (Xue et al., 2026): the next stage is chosen using the outcomes of earlier stages.

Figure 1 shows the abstract model layer and its three coupled loops. The data loop updates collection and training mixtures; the evaluation foundation judges outputs. The training loop tests a candidate against the incumbent under capability-specific criteria. The inference loop searches serving changes with parity or quality checks. Their common experiment memory makes a failed hypothesis useful: it changes the next proposal rather than merely terminating a run.

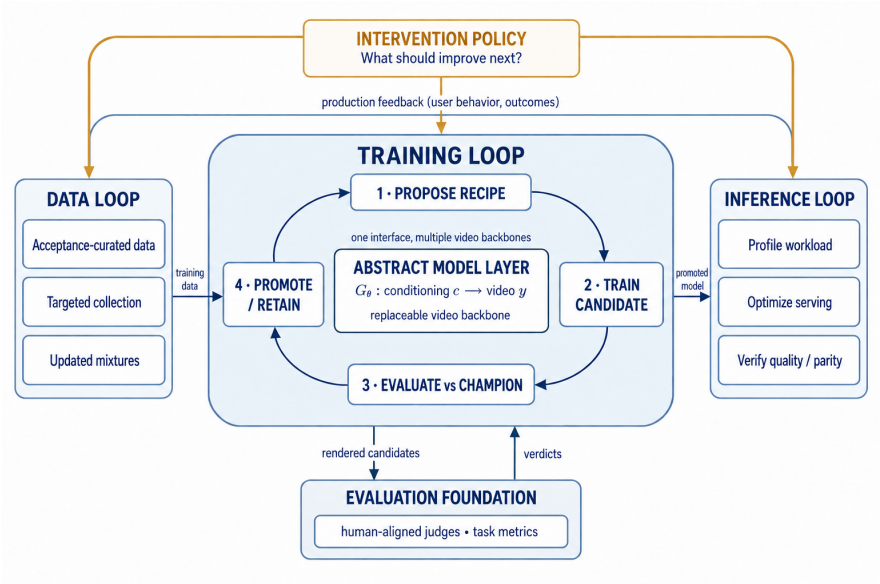


Figure 1: A model-centered three-loop paradigm toward video RSI. Capability feedback and experiment history inform an intervention policy: what should improve next? The data loop supplies supervision, the training loop proposes and evaluates candidates around a replaceable video-model backbone, and the inference loop tests serving changes. Evaluation verdicts determine whether a candidate is promoted or retained.

We study our post-training paradigm in advertising, while the same decision interface is designed to support other domains through domain-specific capability tests and feedback sources. In this setting, we build an ad creative agent and collect user feedback on its outputs, including whether a customer uses a generated creative. This deployment feedback is one of several signals informing the post-training loop, alongside human preference annotations and capability-specific evaluations. Figure 2 makes one product-fidelity decision concrete; Section 5.1 follows recorded changes in evaluator, reward, candidate retention, data supply, and serving. We then test whether the resulting post-trained systems improve their own base models and approach frontier quality.

Our contributions are (i) an effective backbone-agnostic post-training paradigm that makes the *choice of future supervision* a stateful decision across data, training, and inference; (ii) human-calibrated, order-controlled video evaluation and promotion; and (iii) cross-family intervention traces on LTX-2.3/2.5 (HaCohen et al., 2026; Lightricks, 2026b;a), Wan-2.2 (Team Wan, 2025), and MiniMax-H3 (MiniMax, 2026), with human preferences, task-specific gains, and serving profiles.

2 Related Work

Three lines of research make video RSI possible, but leave open the cross-round decision of what to learn from next.

Video post-training. Preference optimization, reward-guided distillation, and flow-based RL improve generated video with learned feedback (Liu et al., 2025c; Wu et al., 2025; Xu et al., 2024; Li

et al., 2025; Liu et al., 2025b). A systematic framework combines SFT, RLHF, prompt enhancement, and inference optimization in one pipeline (Xue et al., 2026). We instead study a policy over these interventions: capability feedback determines which supervision is constructed and which candidate becomes the next incumbent.

Imperfect evaluators and recursive improvement. Language-model post-training benefits from cheap outcome verifiers (Yuan et al., 2024; DeepSeek-AI, 2025). Video rewards face annotation noise, judge disagreement, and reward hacking (Huang et al., 2024; Zheng et al., 2025; Liu et al., 2025d; Lian et al., 2025). Test-time video agents and automated discovery systems also depend on what their verifier resolves (Long et al., 2025; Soni et al., 2024; Novikov et al., 2025; Li et al., 2026). Our loop can revise the evaluator itself, uses human-referenced checks, and conditions further optimization on the diagnosed signal.

Task feedback and transfer. Preferences can come from annotation, arenas, or product choices (Kirstain et al., 2023; Xu et al., 2023; Du et al., 2024). We use acceptance for curation and coverage, while real footage anchors the mixture against synthetic-data feedback effects (Alemohammad et al., 2024; Bertrand et al., 2024). Advertising combines reference identity, requested actions, audio alignment, and runtime in one setting; other domains can replace these capability tests without changing the decision interface.

3 A Shared Model Layer

The abstract model layer in Figure 1 provides a stable decision interface even when architectures, conditioning modalities, and objectives differ. We therefore separate the feedback decision from the generator being adapted. A generator $G_\theta^{(m)}(c)$ belongs to family m and accepts its supported conditioning c (text, images, or audio). Its adapter initializes a base checkpoint, trains a candidate from the incumbent under recipe ρ, and renders a fixed evaluation batch. The recipe specifies the data and learning objective; the adapter retains family-specific objectives, representation, and sampler.

The layer covers joint audio–video LTX-2.3/2.5 (HaCohen et al., 2026; Lightricks, 2026a), speech-driven Wan-2.2 (Team Wan, 2025), and MiniMax-H3 (MiniMax, 2026). Their different modalities test whether the decision structure transfers across families; Appendix G specifies the flow-matching objective.

4 Method: Adaptive Supervision for Video RSI

The shared layer makes the same decision policy available to different generators. The methodological change is to let feedback determine the *next learning problem*, rather than merely score the current model. A diagnosis and experiment history select the evidence to acquire and the intervention to test; comparison with the incumbent determines what is inherited. The three loops implement this decision at the levels of supervision, model weights, and serving, with one shared memory.

The recursive unit is a decision about the next supervision. A fixed recipe chooses its corpus, reward, and stopping rule before seeing the result. Our state additionally contains a data mixture $\mathcal{D}_t$, evaluator protocol e_t, serving strategy s_t, and experiment history H_t alongside generator parameters θ_t. At round t, evaluation produces a diagnosis d_t of capability gaps, evaluator reliability, and serving constraints. The policy $u_t = \Pi(d_t, H_t)$ chooses a concrete intervention and its test: acquire examples, revise an unreliable evaluator, change the training objective or recipe, or modify serving. The incumbent comparison decides whether to retain the candidate; its verdict and reason enter H_{t+1}. Thus a successful update changes the starting model, whereas either success or failure changes the evidence and choices available next time. This second pathway—feedback improving the improvement procedure—is the system-level RSI mechanism.

In our instantiation, Π combines a proposal agent, a coverage-driven collector, recipe validation, and calibrated promotion gates. The ledger records hypotheses, capability-level verdicts, and retention reasons for the next proposal. Researchers can revise evaluator objectives and add action classes; routine data and recipe proposals run through the same recorded decision interface. Figure 2 illustrates one decision, and Section 5.1 follows measured ones.



Figure 2: One feedback-to-intervention decision, illustrated with product fidelity. A label error leads to capability diagnosis, a targeted data request, candidate training, and incumbent comparison. The schematic explains the policy using a fictional NOVA label; the comparison verdict updates experiment memory for the next diagnosis. Section 5.1 evaluates recorded chains.

Deployment target and offline decision. Customer acceptance—the decision to use a delivered creative—is the deployment target, subject to cost and latency. Because it arrives sparsely and after generation, we use it to curate the training mixture, calibrate evaluators, and prioritize coverage rather than as a dense rollout reward. Capability-specific tests then diagnose a failure and an incumbent gate determines what is inherited. The formal objective appears in Appendix B.

4.1 Capability Diagnosis from Imperfect Feedback

The family-specific training loss measures latent prediction, not the deployment objective. In the recorded LTX run its smoothed drift was 0.050 against per-step dispersion of 0.281, while incumbent comparisons changed substantially (Appendix C). A loss threshold cannot identify the capability that improved or regressed. We diagnose three tracks where applicable: general text/image-to-video quality and prompt adherence, audio-driven creator video with lip-sync, and product detail against a reference image. The diagnostic is a vector of failures, not a scalar video reward. This distinction matters when a candidate improves appearance yet damages product identity, or improves synchronization while reducing brief adherence. SyncNet confidence (Chung & Zisserman, 2016) supplies a reproducible learned sync measure; perceptual and product-specific judgments require human review or an MLLM-based evaluator checked against human choices. In one recorded box brief, human reviewers flagged changed shape and action order. A later H3 rendering used a reference-constraining prompt, yet a separate blinded evaluator still found an action-order error. We therefore assess fidelity and adherence separately; the case does not isolate the effect of training (Appendix H).

Where human choices are available, we check evaluator agreement and inspect failures. Reported pairwise analyses show pairs in both orders and retain ties. A decisive win requires both valid orders to select the same video. Where possible, equal-length excerpts and matched rendering seeds limit differences; same-checkpoint, different-seed probes estimate the noise floor. Human review provides calibration and qualitative error analysis. A production multi-axis video QA ensemble, for example, reached 0.764/0.721 balanced accuracy on human-labelled test/holdout splits; the H3 arena also records blind human preferences. Experiments distinguish those choices from model-based comparisons and task-specific scores. Diagnosis proposes an intervention; the incumbent gate decides whether to retain it.

4.2 The Data Loop: Feedback to Supervision

A failure diagnosis becomes useful only when it changes the evidence available for the next update. The data loop converts sparse deployment acceptance and capability-specific evaluation into three forms of supervision: an acceptance-curated training mixture, evaluators aligned to available human choices, and category-level requests for new examples. The same feedback therefore informs *what to learn from*, *how to measure progress*, and *where evidence is missing*. For a product-fidelity gap, the request targets footage with visible product geometry and labels; a creator-realism gap shifts the request toward appropriate human footage. Real footage remains in the mixture to anchor training beyond generated creative.

An agentic collector executes these requests: it reads per-category outcomes, adjusts collection priorities, plans search terms, and performs searching, filtering, deduplication, captioning, and preprocessing. The assembled mixture can enqueue a training hypothesis directly. For example, a measured 55% decisive loss rate on beauty-and-skincare product comparisons raised that category's collection target by one third. Across 26 recorded runs, 40,978 downloaded clips yielded 11,348 accepted clips; a separate seven-wave sequence expanded a targeted class from 3,802 to 48,760 clips (Figure 6). The loop thus turns a measured failure into a new training problem. Appendix D summarizes supervision construction.

4.3 Intervention Search and Incumbent Selection

In the training loop, a supervisor reconciles the training queue and outstanding evaluations on each tick. Proposals come from a language model that reads previous experiments and from new data waves. A validation stage checks whether the recipe tests its hypothesis, duplicates an existing arm, or references an invalid manifest. Accepted proposals branch from an eligible checkpoint and render the applicable benchmark tracks. In-loop scoring supports early termination of clearly losing arms. In the recorded LTX proposer, each automatic proposal changes one recipe component; researchers revise objectives and evaluator protocols (Appendix I).

Algorithm 1 One feedback-to-intervention cycle

1: Diagnose a capability gap and the reliability of its evaluator; read relevant outcomes from the shared ledger.
2: Select a testable intervention: collect supervision, revise a judge/reward, train with recipe ρ, or optimize serving strategy s.
3: Validate the hypothesis, data, resource budget, and incumbent checkpoint; execute the selected intervention.
4: Render task benchmarks and compare the resulting candidate with the incumbent under capability-specific tests.
5: Apply multi-axis guardrails and the versioned promotion screen; confirm promising candidates before incumbent replacement.
6: Store the result and its reason; update the next data request, experiment proposal, or serving search.

The initial action rules map a lagging capability to a testable lever: lip-sync to curriculum order, visual detail to source quality, prompt adherence to caption quality, and coverage to collection. A proposal still has to survive evaluation; the mapping is a hypothesis generator, not a claim that each lever always works. For a narrow capability with paired demonstrations, SFT can be efficient. For open-ended creative preferences, graded model-based comparisons can support preference optimization or RL, as in the H3 GRPO adaptation (Appendix E). The feedback-to-supervision loop determines the evidence and calibration that make either objective meaningful. Each outcome records whether the intervention helped, so the next proposal can change its data source, objective, or recipe instead of repeating the same update on the same evidence. Successful levers become reusable starting points; failed hypotheses are recorded with reasons that the proposer reads before suggesting another arm.

Deployment follows evaluation rather than training loss alone. The selected LTX-2.3 candidate passed pooled two-seed evaluation and was deployed as a base/checkpoint blend; LTX-2.5 likewise compares raw and blended variants. Fixed-checkpoint evaluations identify the selected blend.

4.4 Inference as a Reusable Intervention

After model promotion, the inference loop profiles the deployed workload, proposes runtime changes, and checks latency, computational demand, and output parity or quality before adoption. The intervention changes serving strategy s rather than model weights, but follows the same proposal–test–retain logic. On one fixed LTX streaming workload, graph capture and removal of synchronization reduced time from 1.57 to 0.774 s, with zero mean absolute pixel difference in its parity check. Deployment outcomes can then feed the next capability diagnosis and supervision request. Separate Wan and H3 serving studies appear in Appendix A.

4.5 Reliability of the Incumbent Gate

The recorded screening rule requires win rate at least 0.70, win–loss margin at least 5, and at least 8 decisive outcomes. Under an unbiased independent null, its single-screen false-positive probability is at most 0.1133 for 8–20 decisive pairs; this is not an empirical campaign-wide guarantee because prompts and judges are shared. A screened candidate therefore requires confirmation before promotion. The conditional analysis, evaluator-noise sensitivity, and versioned gate rules appear in Appendix K; Section 5.1 shows a screen that failed on a second seed.

5 Experiments

The experiments follow the decision chain proposed in Section 4: does feedback change the *next decision*, do the resulting systems improve their base models, and do those gains matter at deployment? We lead with evaluator, reward, curriculum, and incumbent-selection decisions; then test four model instantiations and compare their deployed systems with frontier video-generation APIs.

5.1 RSI: Feedback Changes the Next Learning Problem

The central prediction is temporal: once a failure is diagnosed, the next update should use different supervision or a different intervention, and its result should change what survives. Figure 3 shows measured consequences in four parts of the system. The LTX counterfactual below tests a distinct, discrete decision: which branch receives another chance.

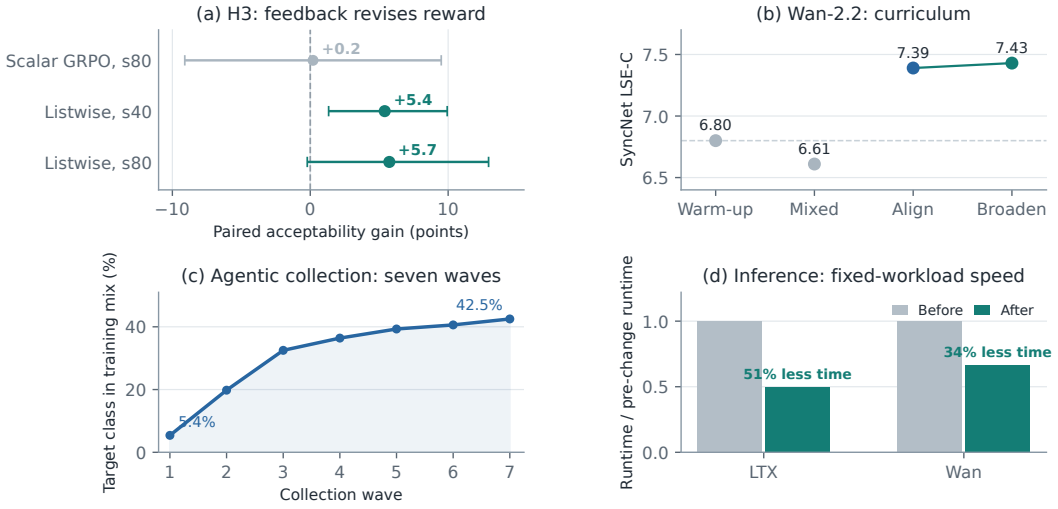


Figure 3: A diagnosed limitation changes the next intervention. (a) H3 reward revision: paired acceptability gain over the same-prompt base (95% case-bootstrap intervals); (b) Wan-2.2 staged alignment curriculum; (c) seven agentic data-collection waves; (d) paired inference optimizations on fixed workloads, normalized within each workload. Each panel measures a different decision and its relevant outcome; protocols appear in Appendices F, I, and A.

The evaluator is part of the object being improved. Human labels exposed errors in the production multi-axis video judge. Revising its visual/language routing increased balanced accuracy from 0.640 to 0.764 on test and from 0.632 to 0.721 on holdout, averaged over three runs per arm (Appendix 3). In a separate H3 arena, the original pairwise judge recognized only five of 21 outputs preferred by blind human reviewers; a defect-aware single-video rubric increased AUC on those development votes from 0.52 to 0.69. Feedback thus changes the instrument that will supervise later updates, rather than simply supplying more labels.

A reward diagnosis changes the next training signal. The variance of repeated ratings of the same H3 clip was 0.0178, at least as large as the 0.014–0.016 variance among rollouts used for the

scalar GRPO advantage. The next recipe replaced a single rating with four order-shuffled listwise judgments of each eight-rollout group and attenuated inconsistent ranks. Under the same prompts and evaluation protocol, its step-40 and step-80 checkpoints gain 5.4 and 5.7 acceptability points over the enhanced base, versus 0.2 for the earlier scalar-reward checkpoint; the step-40 paired 95% interval is [1.33, 9.93]. This is the decisive feedback-to-supervision example: the diagnosed limitation changed the reward, and the changed reward produced better candidates (Figure 3(a); Appendix F).

Different failures call for different objectives. For speech timing, Wan-2.2's alignment-first supervised stage scores 7.39 on SyncNet LSE-C and 7.43 after broadening, versus 6.61 for a recorded broad-mixture arm. For open-ended H3 advertising generation, a fixed SFT recipe lost to its base at all seven evaluated checkpoints; a lower-dose SFT arm increased visual passes from 5 to 9 but raised median audio silence from 0.07 to 0.88. By contrast, blinded human choices favored GRPO v2 by 11 wins to 1 loss and GRPO v5 by 10 to 1 (17 and 18 ties, respectively). The policy therefore selects a supervision type according to what the diagnosed capability can measure, while retaining audiovisual guardrails (Appendix E).

Evaluation also changes which candidate survives. The LTX supervisor launched 84 jobs over 48.09 days. A continuation that screened at 8/10/2 wins/ties/losses failed on a second seed (4/9/7), demonstrating why a first-screen winner is not automatically promoted. Its next branch scored 6/10/4: the former win-rate-and-margin rule would prune it, whereas the revised multi-axis rule retained it because prompt following rose from 0.825 to 0.882 and lip-sync from 0.853 to 0.971. Quality was above the incumbent on both seeds, and the owner preferred the saved candidate. This is an exact replay of the retention decision in the ledger; a separate earlier candidate passed pooled two-seed evaluation and was deployed (Appendix I).

The same policy extends to data and serving. A measured category loss raised its collection target by one third; seven waves expanded clips 12.8-fold. On fixed workloads, serving search cut LTX streaming time from 1.57 to 0.774 s with pixel parity, and Wan cache time from 177.9 to 117.9 s with a perceptual check (Figure 3; Appendix A). Across these traces, diagnosis changes the next evaluator, reward, candidate, data request, or serving strategy. We next test the resulting systems.

5.2 Post-Training Against Its Own Base

We use paired inputs frozen before base/post analysis. LTX-2.3/2.5 use archived advertising briefs, matching seeds, frames, and dimensions. Wan-2.2 uses a speech-driven portrait set scored by one lip-sync stack. H3 uses production briefs whose delivered outputs failed human review. These different target sets assess adaptation within each family; Appendix H gives the protocols.

Both LTX generations improve on the advertising arena. Table 1 shows that a dual-order blinded evaluator favors deployed LTX-2.3 in 83.3% of decisive pairs ($p = 0.0003$). MUSIQ (Ke et al., 2021) rises 5.2% and 3.1% across the two generations; non-rigid warping falls 16.1% and 14.5%, with paired intervals excluding zero. LTX-2.5's cluster-blend 63.0% preference share has an interval spanning 50%, so its stronger evidence is perceptual quality and deformation. In separate four-way human choices among one base and three post-trained variants, the post-trained family receives 88.2% and 80.6% of decisive selections (Table 9); these are not binary win rates.

The other two backbones improve on their own targets. On the speech-driven set, Wan-2.2 post-trained raises lip-sync confidence over the official foundation release by 16.3%, winning three quarters of the scoreable portraits. The clearer effect is timing: the deployed model reduces the absolute audio–visual offset on every scoreable portrait, cutting its mean from 6.25 to 1.69 frames ($p < 10^{-4}$). Both arms take byte-identical inputs and are scored by one SyncNet implementation. On the failed-delivery briefs, H3 post-training raises brief adherence from 27.8% to 50.0%, an 80% relative increase against 16.7% for the delivered commercial outputs, and cuts visual defects per clip by 70%.

Table 1 measures within-family change; the next study uses one task and evaluator for cross-model comparison.

Table 1: Post-training against its own base. Paired inputs. LTX-2.5 preference uses a cluster blend, while its scalar metrics use the deployed service; H3 includes prompt enhancement. Appendix H gives protocols and intervals.

Measure	Instrument	Base	Post-trained	Change	CI / p / scope
LTX-2.3 on the production advertising arena					
Preference vs base	dual-order evaluator	16.7%	**83.3%**	**+66.7 pts**	[66.4, 92.7]%, p=0.0003
MUSIQ ↑	perceptual quality	60.50	**63.63**	**+5.2%**	[+1.01, +5.10]
CLIP-IQA+ ↑	perceptual quality	0.548	**0.567**	**+3.5%**	[+0.006, +0.032]
Warping ↓	non-rigid deformation	1.830	**1.536**	**−16.1%**	[−0.454, −0.146]
LTX-2.5 on the production advertising arena					
Preference vs base	dual-order, cluster blend	37.0%	**63.0%**	**+25.9 pts**	[44.2, 78.5]%, p=0.25
MUSIQ ↑	perceptual quality	61.93	**63.88**	**+3.1%**	[+0.89, +3.14]
Aesthetics ↑	LAION predictor	4.911	**4.983**	**+1.5%**	[+0.007, +0.150]
Warping ↓	non-rigid deformation	1.839	**1.572**	**−14.5%**	[−0.435, −0.106]
Wan-2.2 post-trained on the speech-driven avatar set					
LSE-C ↑	SyncNet lip-sync	6.50	**7.55**	**+16.3%**	p=0.08
|AV offset| ↓	SyncNet timing	6.25	**1.69**	**−73.0%**	every pair, $p < 10^{-4}$
H3 on production briefs whose delivered outputs failed review					
Brief adherence ↑	human review	27.8%	**50.0%**	**+80.0%**	system-level
Defects per clip ↓	human check	1.11	**0.33**	**−70.0%**	system-level

5.3 Comparison with Frontier Video Generators

Advertising quality and serving profiles. The domain endpoint asks whether a generated creative preserves the reference product or creator: the share of blinded fidelity ratings at least 7/10. Table 2 compares four deployed post-trained systems with Seedance 2.0/2.5, MiniMax H3 Max/H3, Gemini Omni 1.1 Flash, and Wan 3.0 on the same briefs, reference images, rubric, and viewing window.

Table 2: Post-trained video models versus frontier APIs. Fidelity is the share of blinded product/creator ratings at least 7/10. Time is generation or API wall time per output-second. Cost index normalizes our estimated GPU expense and API retail prices by the median API retail price; these different accounting bases are not like-for-like inference costs. * denotes H3 serving measured on a different adapter from its scored checkpoint. Best post-trained values are bold; details appear in Appendices H and J.

Model	Ref. fidelity↑	Q-Align	MUSIQ	Motion	Time↓	Cost index↓
Post-trained, deployed (ours)						
H3 post-trained	**85.3**	**4.759**	66.73	0.556	5.20*	0.075*
LTX-2.5 post-trained	70.6	4.615	53.82	**0.623**	1.47	0.021
LTX-2.3 post-trained	67.6	4.589	55.60	0.580	**1.33**	**0.019**
Wan-2.2 post-trained	61.8	4.709	**67.12**	0.521	4.11	0.043
Frontier APIs						
Seedance 2.0	67.6	4.704	66.64	0.544	48.46	1.260
Seedance 2.5	71.9	4.639	63.06	0.573	46.50	1.914
MiniMax H3 Max	79.4	4.675	64.30	0.545	1.82	0.663
MiniMax H3	82.4	4.771	66.61	0.504	32.29	0.663
Gemini Omni 1.1 Flash	70.6	4.736	65.84	0.512	8.01	0.840
Wan 3.0	79.4	4.792	69.06	0.406	9.44	1.160

One of our key findings is that our post-trained H3 model achieves a reference-fidelity pass rate of 85.3%, compared with 82.4% for the strongest of the six frontier APIs. For reference-driven creatives, retaining the supplied product or creator is central to whether an output can be used; the lead therefore concerns a domain-critical capability. H3 also remains competitive on general video quality: its Q-Align of 4.759 lies within 0.033 of the leader, and its MUSIQ of 66.73 exceeds five of six APIs. Together with the within-family gains on open-weight LTX and Wan models in Table 1, these results show that targeted post-training can yield a domain advantage alongside broad video quality. The

development-set case-bootstrap intervals overlap, so the lead is an observed ranking; Appendix H gives continuous and stricter all-axis results.

Table 2 and Figure 4 also report serving. Our cost indices are 0.019/0.021 for LTX-2.3/2.5, 0.043 for Wan-2.2, and 0.075 for H3, versus 0.663–1.914 for the APIs; generation takes 1.33/1.47, 4.11, and 5.20 s per output-second, respectively. Fixed-workload inference changes yield 2.03× LTX and 1.51× Wan speedups with parity or perceptual checks. Cost bases and timer boundaries differ for resident GPUs and APIs (Appendix J). H3's serving adapter differs from its scored checkpoint, so Figure 4 uses an open diamond.

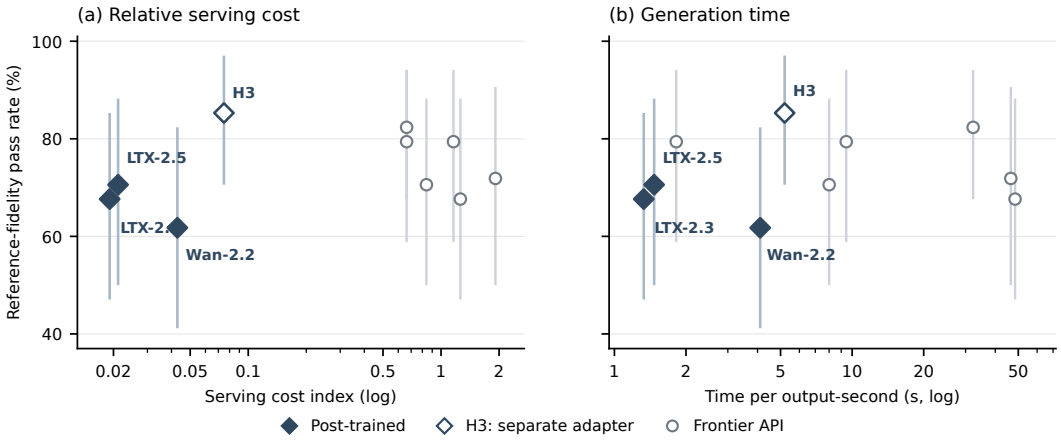


Figure 4: Advertising reference fidelity versus serving cost and time. Upper left is preferable. Cost is indexed to the median frontier-API retail price; self-hosted rows use GPU expense and API rows retail prices. Filled diamonds mark post-trained systems; the open H3 diamond combines separate quality and serving adapters and is not a joint operating point. Bars show 95% case-bootstrap intervals; Appendix J defines accounting and timer boundaries.

Why the domain axis matters. On difficult production briefs, H3 post-trained reaches 50.0% brief adherence, versus 27.8% for its base and 16.7% for delivered commercial outputs, while reducing visual defects per clip by 70%. Appendix M provides brand-free visual comparisons; Appendix 6 reports the dual-order API preferences.

6 Conclusion

Video generation can improve in appearance while regressing in identity, action, or timing. To make such feedback actionable, we presented a backbone-agnostic paradigm that turns capability feedback into targeted data, training, and inference decisions, then tests candidates against an incumbent. Across LTX, Wan, and MiniMax-H3, this process changed subsequent interventions and improved human preference, audio–visual alignment, and advertising performance. Post-trained H3 reached 85.3% reference fidelity, the highest observed against six frontier APIs, while remaining competitive on general video quality. Extensive experiments demonstrate that capability feedback improves not only the models themselves, but also the decisions that produce their successors. This feedback-to-intervention mechanism offers a concrete path toward system-level recursive self-improvement in video generation.

Limitations and next steps. Our loop automates bounded interventions while people define objectives and approve releases. Longer-horizon adaptation of the evaluator and decision policy, and transfer beyond advertising, remain to be tested.

Statement on the use of large language models

Language models participate in the studied workflow as experiment proposers, captioning systems, and MLLM-based evaluators. Their roles and evaluation protocols are described in Sections 4 and 5.

The authors also used generative AI assistants, including Claude and OpenAI Codex, for literature and code inspection, manuscript drafting and revision, methodological critique, mathematical analysis, statistical recomputation, and preparation of LaTeX and figures. These uses extend beyond copy-editing. The authors are responsible for reviewing and approving the final text, analyses, citations, and disclosures.

REPRODUCIBILITY STATEMENT

Algorithm 1 specifies the workflow. Appendix H describes evaluation units, candidate identities, and comparison protocols; Appendix K provides the proofs. The numerical supplement contains anonymous metric tables and a recomputation script. The full training corpus and production checkpoints are not included.

ETHICS STATEMENT

Photorealistic video can support impersonation and misleading advertising. Acceptance feedback expresses platform preferences and does not establish consent or fairness. MLLM-based evaluators can inherit demographic and aesthetic biases. The workflow therefore requires human oversight of data use, evaluator calibration, and deployment decisions.

REFERENCES

Sina Alemohammad, Josue Casco-Rodriguez, Lorenzo Luzi, Ahmed Imtiaz Humayun, Hossein Babaei, Daniel LeJeune, Ali Siahkoohi, and Richard G. Baraniuk. Self-consuming generative models go MAD. In *International Conference on Learning Representations (ICLR)*, 2024.

Quentin Bertrand, Joey Bose, Alexandre Duplessis, Marco Jiralerspong, and Gauthier Gidel. On the stability of iterative retraining of generative models on their own synthetic data. In *International Conference on Learning Representations (ICLR)*, 2024.

Joon Son Chung and Andrew Zisserman. Out of time: automated lip sync in the wild. In *Workshop on Multi-view Lip-reading, ACCV*, 2016.

DeepSeek-AI. DeepSeek-R1: Incentivizing reasoning capability in LLMs via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025.

Zhenbang Du, Wei Feng, Haohan Wang, Yaoyu Li, Jingsen Wang, Jian Li, Zheng Zhang, Jingjing Lv, Xin Zhu, Junsheng Jin, Junjie Shen, Zhangang Lin, and Jingping Shao. Towards reliable advertising image generation using human feedback. In *European Conference on Computer Vision (ECCV)*, 2024.

Yoav HaCohen, Benny Brazowski, Nisan Chiprut, et al. LTX-2: Efficient joint audio-visual foundation model. *arXiv preprint arXiv:2601.03233*, 2026.

Ziqi Huang, Yinan He, Jiashuo Yu, Fan Zhang, Chenyang Si, Yuming Jiang, Yuanhan Zhang, Tianxing Wu, Qingyang Jin, Nattapol Chanpaisit, et al. VBench: Comprehensive benchmark suite for video generative models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024.

Junjie Ke, Qifei Wang, Yilin Wang, Peyman Milanfar, and Feng Yang. Musiq: Multi-scale image quality transformer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021.

Yuval Kirstain, Adam Polyak, Uriel Singer, Shahbuland Matiana, Joe Penna, and Omer Levy. Pick-a-pic: An open dataset of user preferences for text-to-image generation. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

Jiachen Li, Qian Long, Jian Zheng, Xiaofeng Gao, Robinson Piramuthu, Wenhu Chen, and William Yang Wang. T2V-Turbo-v2: Enhancing video generation model post-training through data, reward, and conditional guidance design. In *International Conference on Learning Representations (ICLR)*, 2025.

Yitong Li, Junsong Chen, Haopeng Li, Haozhe Liu, Jincheng Yu, Ligeng Zhu, Ping Luo, Song Han, and Enze Xie. Sol video inference engine: Agent-native full-stack acceleration framework for efficient video generation. *arXiv preprint arXiv:2606.23743*, 2026.

Jiesong Lian, Ruizhe Zhong, Zixiang Zhou, Xiaoyue Mi, Long Hu, Yuan Zhou, Qinglin Lu, Yixue Hao, and Junchi Yan. SoliReward: Mitigating susceptibility to reward hacking and annotation noise in video generation reward models. *arXiv preprint arXiv:2512.22170*, 2025.

Lightricks. LTX-2.5: Open Audio-Video Model. `https://huggingface.co/Lightricks/LTX-2.5`, 2026a.

Lightricks. LTX-2.3 model card. `https://huggingface.co/Lightricks/LTX-2.3`, 2026b.

Yaron Lipman, Ricky T Q Chen, Heli Ben-Hamu, Maximilian Nickel, and Matthew Le. Flow matching for generative modeling. In *International Conference on Learning Representations (ICLR)*, 2023.

Jie Liu, Gongye Liu, Jiajun Liang, Ziyang Yuan, Xiaokun Liu, Mingwu Zheng, Xiele Wu, Qiulin Wang, Menghan Xia, Xintao Wang, Xiaohong Liu, Fei Yang, Pengfei Wan, Di Zhang, Kun Gai, Yujiu Yang, and Wanli Ouyang. Improving video generation with human feedback. *arXiv preprint arXiv:2501.13918*, 2025a.

Jie Liu, Gongye Liu, Jiajun Liang, et al. Flow-GRPO: Training flow matching models via online RL. *arXiv preprint arXiv:2505.05470*, 2025b.

Runtao Liu, Haoyu Wu, Ziqiang Zheng, Chen Wei, Yingqing He, Renjie Pi, and Qifeng Chen. Videodpo: Omni-preference alignment for video diffusion generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025c.

Xingchao Liu, Chengyue Gong, and Qiang Liu. Flow straight and fast: Learning to generate and transfer data with rectified flow. In *International Conference on Learning Representations (ICLR)*, 2023.

Yuanxin Liu, Rui Zhu, Shuhuai Ren, Jiacong Wang, Haoyuan Guo, Xu Sun, and Lu Jiang. UVE: Are MLLMs unified evaluators for AI-generated videos? In *Advances in Neural Information Processing Systems (NeurIPS)*, 2025d.

Do Xuan Long, Xingchen Wan, Hootan Nakhost, Chen-Yu Lee, Tomas Pfister, and Sercan Ö. Arık. VISTA: A test-time self-improving video generation agent. *arXiv preprint arXiv:2510.15831*, 2025.

MiniMax. MiniMax H3: An open model breaking the boundaries between tasks and modalities. `https://www.minimax.io/blog/minimax-h3`, 2026.

Alexander Novikov, Ngân Vũ, Marvin Eisenberger, et al. AlphaEvolve: A coding agent for scientific and algorithmic discovery. *arXiv preprint arXiv:2506.13131*, 2025.

Achint Soni, Sreyas Venkataraman, Abhranil Chandra, Sebastian Fischmeister, Percy Liang, Bo Dai, and Sherry Yang. VideoAgent: Self-improving video generation. *arXiv preprint arXiv:2410.10076*, 2024.

Team Wan. Wan: Open and advanced large-scale video generative models. *arXiv preprint arXiv:2503.20314*, 2025.

Ziyi Wu, Anil Kag, Ivan Skorokhodov, Willi Menapace, Ashkan Mirzaei, Igor Gilitschenski, Sergey Tulyakov, and Aliaksandr Siarohin. DenseDPO: Fine-grained temporal preference optimization for video diffusion models. *arXiv preprint arXiv:2506.03517*, 2025.

Jiazheng Xu, Xiao Liu, Yuchen Wu, Yuxuan Tong, Qinkai Li, Ming Ding, Jie Tang, and Yuxiao Dong. Imagereward: Learning and evaluating human preferences for text-to-image generation. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

Jiazheng Xu, Yu Huang, Jiale Cheng, Yuanming Yang, Jiajun Xu, Yuan Wang, Wenbo Duan, Shen Yang, Qunlin Jin, Shurun Li, Jiayan Teng, Zhuoyi Yang, Wendi Zheng, Xiao Liu, Ming Ding, Xiaohan Zhang, Xiaotao Gu, Shiyu Huang, Minlie Huang, Jie Tang, and Yuxiao Dong. VisionReward: Fine-grained multi-dimensional human preference learning for image and video generation. *arXiv preprint arXiv:2412.21059*, 2024.

Zeyue Xue, Siming Fu, Jie Huang, Shuai Lu, Haoran Li, Yijun Liu, Yuming Li, Xiaoxuan He, Mengzhao Chen, Haoyang Huang, Nan Duan, and Ping Luo. A systematic post-train framework for video generation. *arXiv preprint arXiv:2604.25427*, 2026.

Weizhe Yuan, Richard Yuanzhe Pang, Kyunghyun Cho, Xian Li, Sainbayar Sukhbaatar, Jing Xu, and Jason Weston. Self-rewarding language models. *arXiv preprint arXiv:2401.10020*, 2024.

Dian Zheng, Ziqi Huang, Hongbo Liu, Kai Zou, Yinan He, Fan Zhang, Lulu Gu, Yuanhan Zhang, Jingwen He, Wei-Shi Zheng, Yu Qiao, and Ziwei Liu. VBench-2.0: Advancing video generation benchmark suite for intrinsic faithfulness. *arXiv preprint arXiv:2503.21755*, 2025.

A THE INFERENCE LOOP: FROM INCUMBENT TO PRODUCTION

After a candidate is selected, a serving search profiles the deployed workload, tests an implementation change, and checks whether quality or output parity is preserved. This is a third intervention pathway: it reduces generation cost and latency without requiring another weight update. The search uses isolated implementations and workload-specific measurement, with release review after the quality gate.

On a fixed LTX streaming workload, the retained graph path reduces runtime from 1.57 to 0.774 s with zero mean absolute pixel difference in its parity case. On a separate Wan-2.2 workload, cache reuse reduces 177.9 to 117.9 s and passes the retained perceptual check. H3 has a separately measured one-B200 serving profile used in Table 2. These are engineering interventions on different workloads; the accounting and timer boundaries appear in Appendix J.

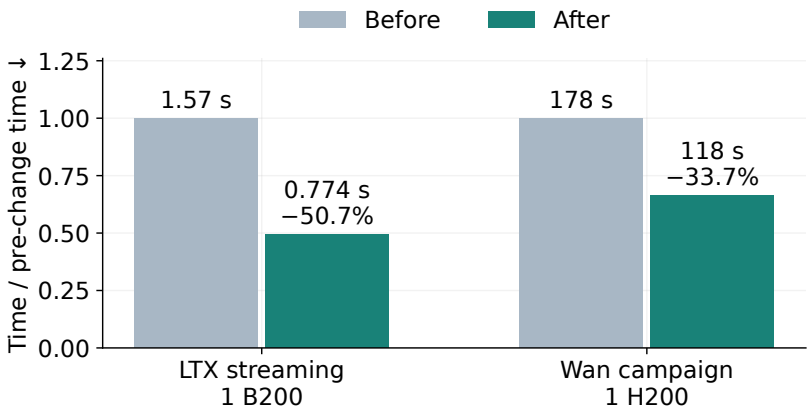


Figure 5: Fixed-workload inference-loop improvements. Runtime is normalized within each unchanged weight configuration. LTX uses a pixel-parity check; Wan uses a perceptual check.

B DEPLOYMENT OBJECTIVE

Let deployment job x provide conditioning $c(x)$, and let $y \sim P_{\theta,s}^{(m)}(\cdot \mid c(x))$ denote an output of backbone m under serving strategy s. If $a(x, y) \in \{0, 1\}$ is the downstream decision to use or discard it, the target is

$$R_m(\theta, s) = \mathbb{E}_{x \sim \text{traffic}} \mathbb{E}_{y \sim P_{\theta,s}^{(m)}(\cdot \mid c(x))}[a(x, y)], \tag{1}$$

under deployment cost and latency constraints. This target is sparse: it is observed only after a customer sees the generation. It can guide curation and evaluator alignment, but does not by itself supply a dense reward for offline video rollouts.

C Capability Feedback and Evaluator Calibration

In the LTX campaign, the smoothed training-loss drift was 0.050, compared with per-step dispersion of 0.281. This diagnostic and the rendered-output comparisons answer different questions: the former measures latent velocity prediction, while the latter measures task preference. The campaign therefore uses capability-level comparisons to decide what to improve next rather than choosing an intervention from loss alone.

Calibration is itself a post-training intervention. The production multi-axis video QA ensemble was compared with human acceptability labels on test and holdout splits. Changing the routing of visual and language calls improved balanced accuracy and Matthews correlation on both splits under three runs per arm (Table 3). The holdout informed configuration choice, so these values describe measured calibration rather than a fresh post-selection generalization estimate. This is a different evaluator from both the H3 training reward and the reference-fidelity rubric used in Table 2.

Table 3: A human-referenced evaluator revision. Means over three runs per split. Balanced accuracy and Matthews correlation coefficient (MCC) use human acceptability labels; false-FAIL is the mean number of human-accepted clips rejected by the judge. The revision routes visual and language dimensions to different judge models.

Split	Evaluator	Balanced acc.	MCC	False-FAIL
Test	Previous	0.640	0.284	11.5
Test	Revised	**0.764**	**0.536**	**8.0**
Holdout	Previous	0.632	0.250	6.0
Holdout	Revised	**0.721**	**0.423**	**5.3**

D Supervision Construction

Acceptance-curated creative and real footage provide complementary sources: the former reflects deployment preferences, while the latter anchors the model to observed video. Capability-level failures change which categories receive further collection and which examples enter the next mixture. Collection, filtering, deduplication, and captioning precede candidate training. The seven-wave targeted-class total includes the assembled class, not only clips accepted in the separate collector record.

E Choosing Supervision from the Diagnosed Failure

The recorded arms support a conditional choice of supervision, not a controlled claim that RL dominates SFT. For Wan-2.2, speech timing has a specific diagnostic: an alignment-first supervised stage followed by broader training gives LSE-C 7.39 then 7.43, versus 6.61 for the recorded broad-mixture arm and 6.80 for its warm-up checkpoint. Those arms differ in initialization, learning rate, and budget. The comparison shows why the controller selected alignment first; it does not isolate curriculum order.

For H3, a fixed SFT arm trained on quality-gated real-ad clips. Across its evaluated checkpoints, it lost to the base on difficult briefs in the recorded visual assessment. At the first assessed checkpoint, mean quality changed by -0.265, prompt following by -0.669, and defects per clip by $+0.958$; the measured comparisons also differed in audio configuration from the preferred H3 arm. A separate low-dose SFT arm on a different 18-brief instrument raised the visual-judge pass count from 5 to 9, but its median audio-silence fraction rose from 0.07 (base) to 0.88. The judge was insensitive to audio, so the visual gain alone is insufficient for promotion. A GRPO-FIX candidate also passed 9 of those 18 briefs under that visual judge, giving no reliable head-to-head advantage over low-dose

SFT on this instrument. These results show the importance of conditioning the intervention on the capability and of evaluating all relevant modalities. They do not support a general RL-versus-SFT ranking. Numerical summaries and qualifications are retained in the supplement.

Blinded human preferences and evaluator calibration. The H3 arena recorded side-by-side human choices at a fixed rendering seed with randomized left/right placement. For one retained GRPO candidate versus base, one reviewer selected the GRPO output 11 times, base once, and a tie 17 times; for a second candidate the record is 10/1/18. Pooling the two checkpoint comparisons gives 21/2/35, but these are not independent votes on one checkpoint. On these human-voted pairs, the original graded pairwise judge had AUC 0.52 for distinguishing GRPO wins from ties. A separately scored single-video judge at 8 fps, with an explicit defect checklist, reached AUC 0.69 for its quality difference (bootstrap 95% interval approximately [0.55,0.82]). The latter rubric was selected using these same human votes; the interval is not an unseen-label validation result. Independently, the production multi-axis video QA ensemble was assessed against human-labelled test and holdout splits: balanced accuracy was 0.764 and 0.721, respectively. The holdout also informed configuration choice. These human-referenced studies concern different evaluators from the Table 2 reference-fidelity threshold; no score or preference from one is substituted for another. The anonymous counts and calibration summaries are in the numerical supplement.

H3 preference-based training. The retained H3 adaptation uses model-based comparative feedback on visual quality and prompt adherence, with audiovisual checks applied at evaluation. The dual-order incumbent comparison in the main method is a separate promotion instrument, not the rollout reward. Development human choices inform evaluator diagnosis; the training reward and deployment-fidelity rubric are distinct instruments.

F A Later H3 Feedback-to-Reward Trace

This research trace is separate from the H3 quality checkpoint in the frontier study. It asks whether a diagnosed evaluator failure changes the next training signal while holding the input treatment fixed. The unified reference-to-video evaluation uses matched development briefs and repeated checklist judgments. All compared training arms receive the same frozen advertising prompt enhancement; the raw-base control alone receives the original brief. Checklist acceptability is not the reference-fidelity pass rate in Table 2; the later reward-revision candidates are not substituted into that table.

The control study first separated prompt enhancement from training: mean acceptability was 53.35 for raw H3 base and 64.93 for identical base weights with the enhancement. The earlier reward arm scored 65.13 under the same enhanced prompts. A later repeated-judgment audit found a same-clip root-mean-square standard deviation of 0.133 on a 0–1 scale (variance 0.0178), comparable to or larger than the 0.014–0.016 reward variance within rollout groups. The diagnosed issue was an unreliable advantage ordering, not simply insufficient optimization steps.

The revised reward uses repeated, order-shuffled listwise judgments and attenuates inconsistent rankings. On the same unified evaluation, two evaluated candidates scored 70.33 and 70.67. Relative to the same-prompt base, paired differences averaged 5.40 and 5.73 points. The 95% case-bootstrap intervals are [1.33, 9.93] and $[-0.22, 12.93]$ after averaging repeated ratings within each brief. The same-pipeline subset for the later candidate averages $+10.70$ points ([1.18, 22.21]). Anonymous per-brief scores and the recomputation script are in the numerical supplement.

G Backbone Training Objective

For the flow-matching instances, clean latents $x_0 \sim \mathcal{D}_\rho$, Gaussian noise x_1, and $x_t = (1-t)x_0 + tx_1$ define the conditional objective (Lipman et al., 2023; Liu et al., 2023)

$$\mathcal{L}_m(\theta; \rho) = \mathbb{E}_{x_0, x_1, t} \| u_\theta^{(m)}(x_t, t, c) - (x_1 - x_0) \|^2. \tag{2}$$

The adapter determines the latent representation and trainable subset; the shared controller chooses and evaluates recipe ρ.

H Evaluation Protocol Details

Shared advertising protocol and input selection. The source benchmark contains 20 product/creator briefs with reference images. The updated comparison uses all 17 briefs whose unchanged prompts fit the strictest requested API's 2,000-character limit; g09, g11, and g19 are excluded by this input rule. Selection was fixed before the new advertising judgments and frontier generations, without using model outcomes. These are development briefs already used during system iteration, not an independent held-out test.

For the domain-specific view in Table 2, we report the advertising reference-fidelity pass rate: the percentage of blinded ratings whose reference-fidelity score is at least 7/10. This axis asks whether the advertised product or creator preserves its appearance, shape, colors, labels, and identity. Seven lies above the rubric's fixed usable anchor of 6 and below its strong anchor of 8. This thresholded summary is an analysis of already frozen ratings; it did not affect prompt, output, or case selection. We also report the originally frozen all-axis pass endpoint, in which all three axes must reach 6, and the arithmetic mean of prompt adherence, reference fidelity, and visual realism on a 0–100 scale. Rates are computed over ratings; uncertainty first averages the two binary ratings for each case and then resamples cases. Because small score changes around a threshold produce discrete rate changes, nearby rates with overlapping intervals are described as observed ordering rather than statistically established superiority. The evaluator is Gemini 3.1 Pro, with model identity hidden, the original prompt and reference image supplied, and video sampled at 8 fps over the first five seconds (or the available excerpt if slightly shorter). The rubric excludes audio and events explicitly scheduled after the excerpt. Both repetitions use the same rubric; temperatures are 0 and 0.3. Nearby rates and overlapping intervals do not establish statistical equivalence. The frozen rubric, numerical outputs, coverage, and paired intervals accompany the supplement. This instrument measures visual advertising execution; it is not a human preference or campaign conversion metric. This new endpoint has not been independently calibrated against a human panel; the controlled human choices and production-task checks provide complementary evidence. Failed generations and evaluator errors are retained in the coverage record, not silently replaced by different cases. Ad-ready rates condition on available outputs, and every successful output is retained. Seedance 2.5 rejected g07 for copyright similarity; this provider rejection is retained and no replacement brief or alternate provider is used for that case.

Q-Align, MUSIQ and Farneback-flow smoothness use the same retained implementations across arms. Generic scores on the selected inputs are reaggregated from the original outputs where available. New outputs use the same metric code. Motion smoothness is the exponential of the negative coefficient of variation of flow magnitude; it measures steadiness, not whether the requested action is correct. Render shapes remain those of the recorded systems, so the table is a system comparison.

LTX-2.3 and LTX-2.5 post-trained denote one fixed recorded production arm each. The H3 base is the matched dense reference implementation. The H3 post-trained arm is the complete deployed adapted system: a preference-trained checkpoint with the pre-existing advertising prompt enhancement. The enhancement sees the original brief and conditioning image; the evaluator receives the original brief. No prompt is selected using a generated video or evaluation score. The Wan-based grouping follows the original third-model campaign. Its foundation arm is reproduced from the archived foundation weights with matched output duration, geometry, reference image, and prompt formatting. The reproduced base runs on the same accelerator class; the post-trained videos are retained production outputs. This controls the recipe without asserting bitwise equivalence across runtime versions.

An illustrative multi-axis failure. Archived brief `f_p03` asks for a box to enter an initially empty frame from the right. Two human annotators marked the delivered output as failing: the box shape changed and the requested sequence was not followed. In a separate H3 research rendering, the system received a frozen advertising prompt enhancement that explicitly described the reference box, its markings, and the intended motion. A blinded evaluator still flagged that output for starting with the box already visible and for unnatural hand motion. These records illustrate why shape fidelity, action sequence, and motion need separate diagnoses. They do not establish that the generator learned a reference-preservation rule, nor do they separate the effect of post-trained weights from prompt enhancement. The case is used as an illustration, not included as an additional quantitative endpoint.

Frontier comparator coverage. The frontier comparator group contains Seedance 2.0, Seedance 2.5, MiniMax H3 Max, MiniMax H3, Gemini Omni 1.1 Flash and Wan 3.0. The public Arena image-to-video snapshot (September 2, 2026) places H3 first; the Artificial Analysis audio-enabled board checked September 10 places H3 Max first and Seedance 2.0 second[1]. Their Elo scales and protocols differ, so their scores are not substituted for our advertising quality. MiniMax H3 Max denotes the post-trained commercial API listed by Artificial Analysis; MiniMax H3 is a separate frontier API, with released weights also available. API requests use five seconds, 720p where supported and 768P for MiniMax, the same input image and unmodified brief, and seed 42 where the API exposes it. The H3 row uses the prompt enhancement described above, so Table 2 is a deployed-system comparison; Table 1 separately tests the adapted system against its own base. The numerical supplement records endpoint/model versions and failures.

Table 4: Advertising evaluation coverage and uncertainty. Reference fidelity applies the ≥ 7 criterion used in Table 2. All-axis pass applies the stricter conjunctive ≥ 6 criterion to all three axes. Intervals resample cases after averaging the two blinded binary ratings within each case. Composite score is the continuous mean of prompt adherence, reference fidelity, and visual realism on a 0–100 scale.

Model	Completed / planned	Ref. fidelity	95% CI	All-axis pass	Composite
LTX-2.3 base	17/17	67.6	[50.0, 85.3]	14.7	51.2
LTX-2.3 post-trained	17/17	67.6	[47.1, 85.3]	5.9	51.5
LTX-2.5 base	17/17	79.4	[58.8, 94.1]	20.6	55.0
LTX-2.5 post-trained	17/17	70.6	[50.0, 88.2]	20.6	53.6
Wan base	17/17	64.7	[44.1, 82.4]	11.8	50.7
Wan-2.2 post-trained	17/17	61.8	[41.2, 82.4]	20.6	51.6
H3 base	17/17	79.4	[61.8, 94.1]	20.6	54.0
H3 post-trained	17/17	85.3	[70.6, 97.1]	35.3	58.0
Seedance 2.0	17/17	67.6	[47.1, 88.2]	32.4	55.8
Seedance 2.5	16/17	71.9	[50.0, 90.6]	31.2	55.2
MiniMax H3 Max	17/17	79.4	[58.8, 94.1]	14.7	55.3
MiniMax H3	17/17	82.4	[67.6, 94.1]	26.5	58.3
Gemini Omni 1.1 Flash	17/17	70.6	[50.0, 88.2]	32.4	57.3
Wan 3.0	17/17	79.4	[58.8, 94.1]	29.4	59.8

Table 5: Paired change in the composite advertising score. Same inputs and evaluator; positive differences favor post-training. A difference is not called statistically significant when its interval includes zero.

Family	Paired cases	Post − base	95% CI
LTX-2.3	17	+0.29	[-3.43, +3.53]
LTX-2.5	17	-1.37	[-7.06, +4.31]
Wan	17	+0.88	[-3.92, +5.59]
H3	17	+4.02	[-1.37, +9.71]

Direct advertising preference. We additionally pair the complete H3 deployed-system output with each of the six retained API outputs on their common assigned briefs. The fixed evaluator sees the original brief, reference image, and equal first-five-second excerpts, with the model names hidden. Each pair is judged in both presentation orders. A win or loss requires agreement across orders; an explicit tie or order disagreement is retained as a tie. This conservative endpoint measures a deployment preference rather than the reference-fidelity pass rate in Table 2. It does not show that H3 is uniformly superior: the net preference is positive against Seedance 2.0 and MiniMax H3 Max, and negative against the other four APIs. It does show that H3 avoids a consistent loss in 68.8–87.5% of comparisons per API at the latency reported in Table 2.

[1] `https://arena.ai/leaderboard/image-to-video`; `https://artificialanalysis.ai/video/leaderboard/image-to-video`

Table 6: Blind advertising deployment preference for H3 post-trained. Each H3/API pair is evaluated in both presentation orders. W or L requires the same winner in both orders; any disagreement or explicit evaluator tie becomes T. Favorable is W+T. The evaluator sees the original brief and reference image, while H3 generation uses the frozen advertising prompt enhancement.

Comparator	H3 W:T:L	Favorable	95% Wilson CI
Seedance 2.0	6:8:2	87.5%	[64.0, 96.5]%
Seedance 2.5	1:10:5	68.8%	[44.4, 85.8]%
MiniMax H3 Max	5:8:3	81.2%	[57.0, 93.4]%
MiniMax H3	2:10:4	75.0%	[50.5, 89.8]%
Gemini Omni 1.1 Flash	3:9:4	75.0%	[50.5, 89.8]%
Wan 3.0	4:7:5	68.8%	[44.4, 85.8]%

Common-output sensitivity. Table 7 restricts all available model arms to exactly the same completed briefs. The planned assignments and failures remain visible in Table 4; this sensitivity view is not presented as an independently sampled validation set.

Table 7: Common-output sensitivity. Every row uses exactly the same 16 briefs with complete advertising and generic scores across the available model arms. This view reports reference fidelity on common outputs and removes differences in output coverage without adding or replacing inputs.

Model	Ref. fidelity	Q-Align	MUSIQ	Motion
LTX-2.3 base	65.6	4.593	54.59	0.480
LTX-2.3 post-trained	65.6	4.582	55.48	0.563
LTX-2.5 base	84.4	4.650	54.12	0.532
LTX-2.5 post-trained	75.0	4.613	53.74	0.617
Wan base	65.6	4.735	68.02	0.495
Wan-2.2 post-trained	65.6	4.744	67.56	0.522
H3 base	84.4	4.754	67.24	0.571
H3 post-trained	84.4	4.750	66.63	0.557
Seedance 2.0	71.9	4.706	66.77	0.535
Seedance 2.5	71.9	4.639	63.06	0.573
MiniMax H3 Max	84.4	4.664	64.15	0.547
MiniMax H3	84.4	4.766	66.65	0.478
Gemini Omni 1.1 Flash	75.0	4.728	65.62	0.512
Wan 3.0	84.4	4.786	69.06	0.425

Controlled LTX arena. The original arena contains 8 product shots, 26 creator scenes, and 6 single-person talking clips. Each controlled base/post pair fixes the image, prompt, random seed, frame count, dimensions, and renderer. Those fixed-renderer comparisons isolate weight changes. Table 1 instead uses the retained deployed-service outputs for LTX-2.5: the deployment validation records all 40 successful renders. Its input manifest exactly matches the original arena, but the service and local base runner differ in implementation. Its gains therefore describe the deployed system. All saved outputs match dimensions within their base/post pair; LTX-2.3 and LTX-2.5 use different spatial-grid rounding across generations. These arena results are not mixed into Table 2. MUSIQ, CLIP-IQA+ and LAION aesthetics average 12 evenly spaced frames; the saved Gemini semantic scorer uses 10 frames for prompt adherence. The main arena columns follow the five-column metric set in the original report. All finite paired measurements are retained: 40 pairs per metric, except LTX-2.3 prompt adherence and VLM aesthetics, which have 39. Paired intervals use case-bootstrap resampling. Table 8 retains additional diagnostics: positive main-table means do not imply uniform improvement in temporal retention or statistical significance on every measure. These are development evaluations with unadjusted intervals. The product-only comparison uses the controlled cluster checkpoint and includes the six product cases with successfully scored outputs for all compared arms; two Seedance files could not be decoded by the scorer.

Human choices. Each decision chooses one of four arms or a tie. LTX-2.3 Post A is the deployed variant (21 selections), with Post B and C receiving 6 and 3. LTX-2.5 Post A, B, and C receive 10, 7, and 8 selections. Bases receive 4 and 6 choices. Thus the family shares are $30/(30+4) = 88.2\%$

Table 8: Production-arena measurements and diagnostics. All available finite pairs are retained. Intervals are unadjusted paired 95% case-bootstrap intervals for post minus base; they do not establish improvement on every metric. Retention ratios describe changes from the first frame, and dynamic degree describes motion amount rather than preference.

Family	Metric	Pairs	Base	Post	Gain 95% interval
LTX-2.3	MUSIQ	40	60.50	63.63	[+1.01, +5.10]
LTX-2.3	CLIP-IQA+	40	0.548	0.567	[+0.006, +0.032]
LTX-2.3	Aesthetics	40	4.923	4.947	[-0.020, +0.068]
LTX-2.3	Prompt	39	0.679	0.724	[-0.064, +0.154]
LTX-2.3	Motion	40	0.498	0.543	[-0.015, +0.108]
LTX-2.3	VLM aesthetics	39	0.891	0.897	[-0.064, +0.077]
LTX-2.3	Color stability	40	0.831	0.796	[-0.073, +0.001]
LTX-2.3	First-frame seam ratio	40	1.036	1.183	[-0.235, +0.466]
LTX-2.3	Dynamic degree	40	0.206	0.227	[+0.002, +0.042]
LTX-2.3	Sharpness retention	40	0.804	1.263	[+0.011, +1.200]
LTX-2.3	HF retention	40	2.328	3.607	[-0.007, +3.547]
LTX-2.3	Fine-detail retention	40	0.769	0.830	[-0.043, +0.190]
LTX-2.3	Nonrigid warping share	40	1.830	1.536	[-0.454, -0.146]
LTX-2.5	MUSIQ	40	61.93	63.88	[+0.89, +3.14]
LTX-2.5	CLIP-IQA+	40	0.574	0.581	[-0.002, +0.017]
LTX-2.5	Aesthetics	40	4.911	4.983	[+0.007, +0.150]
LTX-2.5	Prompt	40	0.713	0.775	[-0.031, +0.150]
LTX-2.5	Motion	40	0.572	0.582	[-0.033, +0.053]
LTX-2.5	VLM aesthetics	40	0.844	0.881	[-0.056, +0.138]
LTX-2.5	Color stability	40	0.832	0.865	[-0.034, +0.101]
LTX-2.5	First-frame seam ratio	40	1.336	0.665	[-2.375, +0.219]
LTX-2.5	Dynamic degree	40	0.190	0.203	[-0.007, +0.033]
LTX-2.5	Sharpness retention	40	2.605	1.074	[-4.200, -0.039]
LTX-2.5	HF retention	40	3.662	2.522	[-2.810, -0.003]
LTX-2.5	Fine-detail retention	40	0.841	0.796	[-0.158, +0.039]
LTX-2.5	Nonrigid warping share	40	1.839	1.572	[-0.435, -0.106]

and $25/(25+6) = 80.6\%$; they are not binary preference win rates. Fixed-checkpoint tables never select the best arm separately for each case.

Fixed-checkpoint pairwise preferences. The controlled LTX arena also retains dual-order Gemini judgments on the same 40 inputs. We use the recorded Gemini 3.1 Pro comparison for both generations. A win requires the same video to win both presentation orders; disagreement counts as a tie. LTX-2.3 records 25 wins, 10 ties and 5 losses, giving 83.3% of decisive comparisons with a Wilson 95% interval of [66.4, 92.7]%. The LTX-2.5 cluster blend records 17/13/10, giving 63.0% with interval [44.2, 78.5]%. With ties assigned half credit, the scores are 75.0% and 58.8%. The earlier LTX-2.3 Flash evaluator gives 19/14/7 (73.1% decisive); its full records remain in the supplement as an evaluator-sensitivity check. Pro was calibrated using development human choices from this arena, so this is not independent held-out calibration. These fixed-checkpoint comparisons are distinct from the four-way human family shares and the new advertising composite. The LTX-2.5 result belongs to the cluster-rendered blend, not the deployed clips in Table 1. The retained deployment/base preference run is incomplete (18 records, including one invalid judgment), so its results do not establish the same preference rate for the deployment.

Retained pairwise judgments and renderer dependence. Earlier pairwise judgments on the full source benchmark use Gemini 3.1 Pro, both presentation orders, equal-length muted pairs, and a tie option. Wins require agreement across orders; inconsistent preferences count as ties. Against human-made references, the retained LTX-2.5 research renderer gives 4 wins, 10 ties, 6 losses (70% favorable), versus its base's 2/8/10 (50%), H3 Max's 5/8/7 (65%), and Seedance's 5/11/4 (80%). The production renderer in Table 2 gives 4/4/12 (40%) under that protocol. The 70% result therefore is not assigned to the production row. These records show why generic quality and prompt adherence are separate axes; the main domain comparison uses the matched difficult-brief study below.

Speech-driven validation set. The Wan-family row of Table 1 extends an earlier four-portrait check to the full curated evaluation set of 17 portraits with driving speech. Both arms were rendered for this paper. The base arm is the official foundation release at its default settings. The post-trained arm is the deployed service through the benchmark's production adapter, so aspect handling, the structured prompt and the refinement stage match what the service serves. Each pair uses the same portrait file and the same audio file, and both sides are trimmed to the same five-second window and scored by one SyncNet implementation. One portrait is excluded because the scorer recovered no face track from the base output; it is reported rather than replaced, and the 16 scoreable pairs carry every number in the table. Lip-sync confidence and the recovered audio-visual offset are separate measures: confidence improves on 12 of 16 portraits, while the offset improves on all 16.

An advertising-fidelity check that does not separate the arms. We also built a second instrument for the base-versus-post question and report it because it is negative. It presents both clips of a pair to a blinded dual-order evaluator together with the conditioning image and the unmodified brief, and asks for a per-clip advertising checklist: whether the advertised subject keeps its reference appearance, whether visible text stays legible, whether the requested action is visible, and how many artefact events appear. Applied to 157 pairs across both pools and all four families, no checklist item separates any family's arms at the 5% level under a McNemar test, and the paired preference it produces is decisive for post-training only on the LTX-2.3 arena block, at 70.0% with an interval that contains the even split. Two readings are consistent with the rest of this appendix. Subject preservation is already high for both arms on these briefs, between 75% and 87% depending on family, so the item has little room to move; and the pooled general briefs, whose reference clips come from external commercial systems, favour neither arm. We keep the checklist as a limitation rather than as support, and the deployment-set measures in Table 1 carry the claim.

H3 production challenge. The 18-case challenge contains 9 product and 9 human briefs whose commercial outputs failed human review. A reference-image/brief screen excludes conflicting inputs. Delivered outputs include Seedance, Veo, and Kling variants. Both H3 arms use matched references and output duration. The base receives the raw brief; the adapted system uses the retained preference-trained checkpoint and advertising prompt-enhancement module. This study measures combined system-level adaptation on a failure-conditioned development set. Checks cover brief adherence, identity, product deformation, label legibility/correctness, hand action, frozen subjects, cuts, and product count. The reported visual-defect count excludes unwanted music and speech/mouth agreement. Qwen PASS uses the fine-tuned Qwen3.6-27B evaluator with majority voting; ties resolve to FAIL. General-quality scores on these same retained outputs are MUSIQ 68.91/69.12/67.44 and motion 0.517/0.553/0.538 for base/post/closed, respectively.

I Training-Loop Operations and Decision Examples

The proposer reads recent experiments and a structured verdict ledger, submits one testable change to a training recipe or data mixture, and receives a recorded result. Validation rejects duplicated or malformed proposals before scheduling. The supervisor then trains eligible candidates, renders the applicable benchmarks, and compares them with the incumbent. Researchers define the capability objectives, revise judges when human review exposes a mismatch, and approve releases; the proposer does not perform those decisions. Across 48.09 days, this system launched 84 training runs. A separate 12.75-day interval with no loop-file edits produced 15 evaluation outcomes and one promotion.

What a failed confirmation changed. One LTX continuation won 8 pairs, tied 10, and lost 2 on its first render seed. The same checkpoint scored 4/9/7 on a second seed, so the first screen did not result in promotion. The following branch scored 6/10/4 and would have been pruned by the prior win-rate-and-margin rule. A revised multi-axis rule retained it for further evaluation: prompt following rose from 0.825 to 0.882 and lip-sync from 0.853 to 0.971 on its first seed, while quality was above the incumbent on two seeds. Owner review preferred the branch, but the recorded ledger did not deploy it. A separate candidate passed pooled two-seed evaluation and was promoted. These outcomes distinguish a proposal, a promising screen, a failed confirmation, retained exploration, and an applied promotion. The old-rule result is a deterministic replay on the same verdict, not a second training campaign.

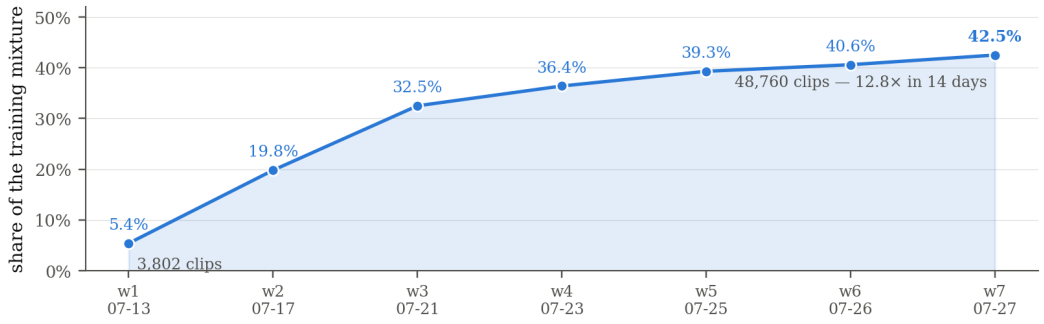


Figure 6: A targeted collection example. Seven waves grew one under-covered class from 3,802 to 48,760 clips and its training-mixture share from 5.4% to 42.5%. This measures data supply, not model quality.

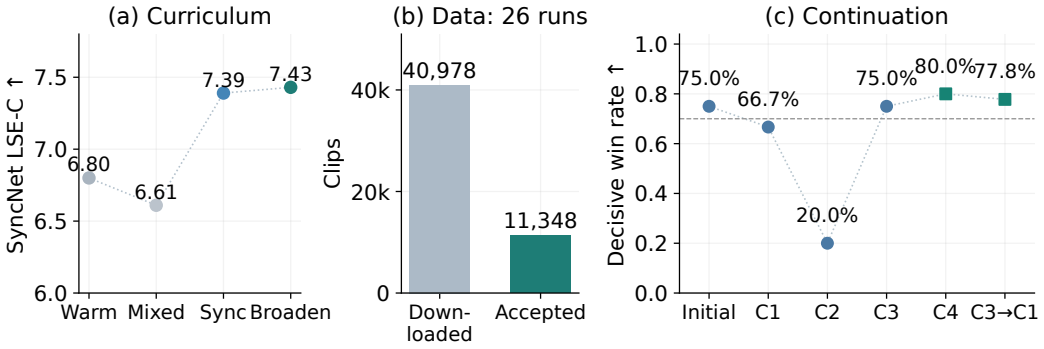


Figure 7: Feedback changes subsequent choices. The retained decision trace shows curriculum revisions and six automatic continuation outcomes, including candidates that failed confirmation. Squares mark promotion-screen passes; they do not imply release.

J Serving Measurement Details

The cost index in Table 2 and Figure 4 divides each arm's cost per output-second by the median API retail cost per output-second among the six frontier APIs in Table 2. For resident self-hosted systems, cost is accelerator occupancy multiplied by the recorded accelerator rate; for closed APIs it is the retail tariff. The index retains the original cost comparison without presenting currency rates as model-quality metrics. These accounting bases differ and are not equivalent estimates of model FLOPs.

The time column of Table 2 normalizes elapsed time by output duration. Self-hosted profiles measure generation on resident accelerators and exclude network transfer. API profiles measure submission to a ready result, including provider queueing and polling; unsuccessful requests are excluded. The two timer boundaries are therefore reported explicitly rather than treated as matched hardware throughput.

The resident LTX-2.3/2.5 profiles take 1.33/1.47 seconds per output-second on one B200. On a fixed streaming workload, the retained inference change reduces LTX runtime from 1.57 to 0.774 seconds (2.03×) with pixel parity. The Wan-2.2 profile takes 4.11 seconds per output-second on one H200; its fixed-workload cache comparison reduces runtime from 177.9 to 117.9 seconds (1.51×) with a perceptual check. The separately measured one-B200 H3 serving profile takes 5.20 seconds per output-second. Its adapter differs from the scored quality checkpoint, as marked in Table 2. These serving studies use different workloads; their normalized times are system profiles, not matched per-prompt accelerator benchmarks.

K Proofs and Computations

Assumption 1 (Null evaluation outcomes). Conditional on a fixed set of D decisive pairs, candidate wins are independent Bernoulli draws with probability p_0 under the null. An unbiased null has $p_0 = 1/2$.

Proposition 1 (Screening probability)**.** *Under Assumption 1, a null candidate clears the screen with probability*

$$\alpha(D, p_0) = \Pr\left[\mathrm{Bin}(D, p_0) \geq \max\left(\lceil 0.7D \rceil, \left\lceil \frac{D+5}{2} \right\rceil\right)\right]. \tag{3}$$

For $p_0 = 1/2$ and $8 \leq D \leq 20$, $\alpha \leq 0.1133$. Two independent tests each satisfying this bound have joint probability at most 0.013.

Seed changes alone do not imply independence when prompts and evaluators are shared; this is a conditional screen analysis, not a campaign-wide error guarantee.

Assumption 2 (Symmetric evaluator noise). The evaluator reverses a latent binary preference with probability $\varepsilon < 1/2$, independently of that preference.

Lemma 1 (Evaluator alignment)**.** *Under Assumption 2, a population preference probability p is observed as $p' = \varepsilon + (1 - 2\varepsilon)p$. Hence an expected threshold $p' \geq \tau$ requires $p \geq (\tau - \varepsilon)/(1 - 2\varepsilon)$.*

For illustration, $\varepsilon = 0.226$ and $\tau = 0.70$ imply $p \gtrsim 0.865$ at the population level. If $\varepsilon = 0.341$, even $p = 1$ gives $p' = 0.659$. Finite samples can still cross a threshold by chance; the identity is not a confidence statement about one candidate.

Proposition 1. Under Assumption 1, $W \sim \mathrm{Bin}(D, p_0)$. The screen requires $W/D \geq 0.7$ and $W - (D - W) \geq 5$, hence $W \geq k(D) = \max(\lceil 0.7D \rceil, \lceil (D + 5)/2 \rceil)$. The binomial upper tail gives Equation 3. For $p_0 = 1/2$ and $D = 8, \ldots, 20$, the probabilities rounded to three decimals are 0.035, 0.090, 0.055, 0.113, 0.073, 0.046, 0.090, 0.059, 0.038, 0.072, 0.048, 0.032, and 0.058. The maximum occurs at $D = 11$. For two independent tests the probabilities multiply, giving the stated bound. Independence is an additional condition on the tests; changing the render seed on shared prompts does not establish it. With $p_0 = 0.6$ the maximum single-test probability becomes approximately 0.296, illustrating the sensitivity to evaluator bias.

Lemma 1. Conditioning on the latent preference gives $p' = p(1 - \varepsilon) + (1 - p)\varepsilon$. The map is increasing for $\varepsilon < 1/2$, has range $[\varepsilon, 1 - \varepsilon]$, and can be inverted to obtain the population requirement in the lemma. Applying that inverse to a sample proportion does not create a finite-sample confidence bound. Heterogeneous evaluator errors and task-dependent disagreements require further calibration.

L Additional Human and Transfer Results

Table 9: Blind human choices favor post-trained variants. Each row comprises 40 four-way decisions: one base, three post-trained variants, and an available tie. Variant identities are detailed in Appendix H; no selections are duplicated.

Base family	Base	Post A	Post B	Post C	Tie
LTX-2.3	4	**21**	6	3	6
LTX-2.5	6	**10**	7	8	9

Table 10: Audio-driven transfer reference. Official Wan2.2-S2V-14B against the deployed post-trained Wan-family model, using identical portrait and audio inputs and the same scored time window; case labels identify the matched comparisons. LSE-C is SyncNet lip-sync confidence (higher is better); offset is the audio-visual misalignment SyncNet recovers, in frames (closer to zero is better).

	LSE-C ↑			|offset| (frames) ↓	
Case	Reference	Post	Gain	Reference	Post
Case 1	7.24	**7.94**	+0.70	6	**2**
Case 2	7.81	**8.69**	+0.88	7	**0**
Case 5	6.03	**8.52**	+2.50	6	**2**
Case 8	3.45	**6.17**	+2.72	10	**1**
Mean	6.13	**7.83**	+1.70 (+27.7%)	7.25	**1.25**

Table 11: H3 base versus post-trained. Top: the loop's 12-case probe, dual-order preference at two seeds; win rates exclude ties and W:T:L is from the post-trained arm's perspective. The first-seed SyncNet LSE-C increases from 7.78 to 7.92. Bottom: 18 production briefs whose delivered commercial outputs (Seedance, Veo, and Kling variants) failed human review; the post-trained arm adds the domain prompt-enhancement module. Brief and identity are explicit visual checks; visual defects exclude audio checks; PASS uses the fine-tuned Qwen evaluator with repeated judgments.

12-case probe (post-trained versus base)			
Seed	W:T:L	Base win rate	Post-trained win rate
A	**5:6:1**	0.167	**0.833**
B	3:8:1	0.250	**0.750**
18 production briefs			
Metric	Delivered commercial	H3 base	H3 post-trained
Follows brief ↑	3/18	5/18	**9/18**
Identity match ↑	**17/18**	15/18	**17/18**
Visual defects per clip ↓	0.56	1.11	**0.33**
Qwen PASS ↑	3/18	5/18	**7/18**

M SELECTED VISUAL COMPARISONS

The following product-only cases were selected from launch reviews for visible differences in reference preservation and requested motion. Each comparison uses the same source image or reference set and creative brief. We show raw frames at matched fractions of each clip, without retouching. These cases illustrate specific failure modes; the aggregate comparisons are reported in Tables 1 and 2. No customer names or brand marks appear in the displayed frames.

Maintaining the requested composition. The source image shows an assortment of prepared meals in several trays. At the matched middle and late frames, Seedance 2.0 switches to close-ups of individual dishes, whereas the post-trained H3 output keeps the multi-tray arrangement visible (Figure 8).

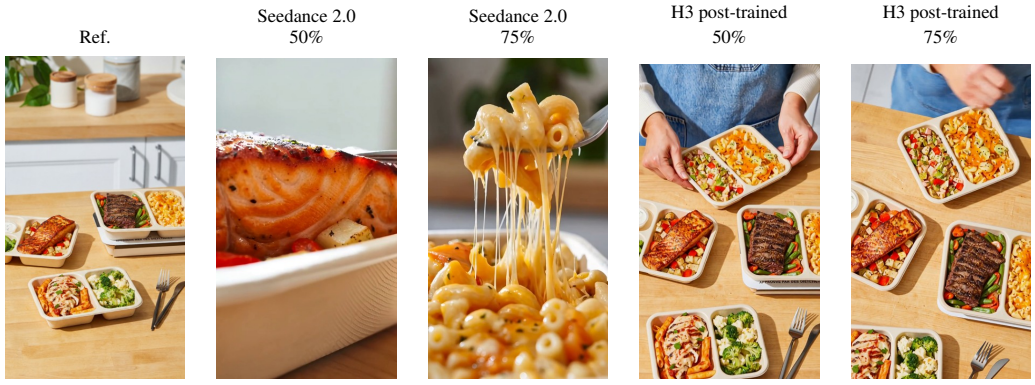


Figure 8: Composition under a product-video brief. The source image and frames at matched fractions of the generated clips. The post-trained output keeps the assortment in view; Seedance 2.0 cuts to individual dishes.

Completing a product turn. The second brief supplies front and back views of a pillow and requests a complete 360° turn. Seedance 2.5 reaches the striped back but ends there; the post-trained H3 output turns through the back and returns to the floral front (Figure 9). The front–back–front sequence makes completion visible in still frames.



Figure 9: Completing the requested 360° turn. Front and back reference images and frames at matched fractions of the two generated clips. The post-trained output shows front, back, and front again; Seedance 2.5 ends on the back rather than completing the turn.