STYLECAST: FUTURE-STYLE GENERATION AND THE ONE-STEP HORIZON

Anonymous authors
Paper under double-blind review

ABSTRACT

Is it possible for generative models to predict the future of artistic style? This question has remained largely unmeasured since modern generators are typically trained on data spanning the past to the present, which already include the historical future relative to any earlier cutoff. In this paper, we study this question by focusing on western painting. We present StyleCast, a cutoff-pure framework in which every image-producing component is trained only on paintings dated up to a cutoff year. To enable directional forecasting, we estimate a stylistic direction from the two most recent pre-cutoff eras and inject it into score space during sampling. The resulting outputs are then evaluated against the authentic, held-out next era. In backtests at two cutoffs (1880 and 1910), the steered ensembles move significantly toward an era the model has never seen, and the pre-specified configuration transfers to the second cutoff without re-tuning. Our central finding can be summarized as a sharp one-step horizon. Within the mechanism family studied, the certified displacement stays within one era step of the cutoff. Stronger steering, iterated generation, and training on generated outputs all fail to extend it, while a small amount of authentic post-cutoff data moves the model where its own generations cannot. Deployed at the present cutoff, StyleCast is able to deliver visually novel compositions. Every displayed image is farther in style space from all training paintings than real paintings are typically from one another.

1 INTRODUCTION

Can a generative model trained only on paintings available before a cutoff year move measurably toward a future artistic era? This question has seldom been addressed by existing research, since it is difficult to evaluate prospectively in the present, where the future ground truth does not yet exist. Also, modern pretrained models have seen artworks created after any historical cutoff, thus apparent foresight may simply be memory. We therefore formulate the problem as historical backtesting. Every generation-side component and design choice is restricted to information available before the cutoff, while post-cutoff artworks are retained exclusively for evaluation. Placing the cutoff in the past turns the held-out future into authentic art history and enables walk-forward scoring against the artworks that actually followed, the discipline forecasting research applies to any predictor (Tashman, 2000; Hyndman & Athanasopoulos, 2021). Prior work on generative novelty rewards deviation from existing data without a directional criterion (Elgammal et al., 2017; Lehman & Stanley, 2011). Rather than asking only whether generated images differ from existing paintings, we evaluate whether the ensemble moves closer to the stylistic distribution of the held-out next era. To prevent future information from entering this comparison, pretrained features, post-cutoff losses, and post-cutoff hyperparameter selection are excluded from the generation pipeline. Modern pretrained encoders are used only on the evaluation side.

Our proposed StyleCast rebuilds the entire generative stack from scratch at each cutoff, without pretrained weights in the image-generation pipeline. It consists of a VQ tokenizer (van den Oord et al., 2017; Esser et al., 2021), a factored style encoder, and a conditional diffusion renderer (Ho et al., 2020; Song et al., 2021b). The steering signal is derived from the renderer itself. Finetuning on the two most recent pre-cutoff eras defines a norm-matched era-delta teacher, whose score difference from the base renderer is injected during sampling following the classifier-free guidance (Ho & Salimans, 2021; Karras et al., 2024). Since the base renderer continues to denoise at every step, every injection is pulled back toward the pre-cutoff distribution. Therefore, a weak temporal

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.

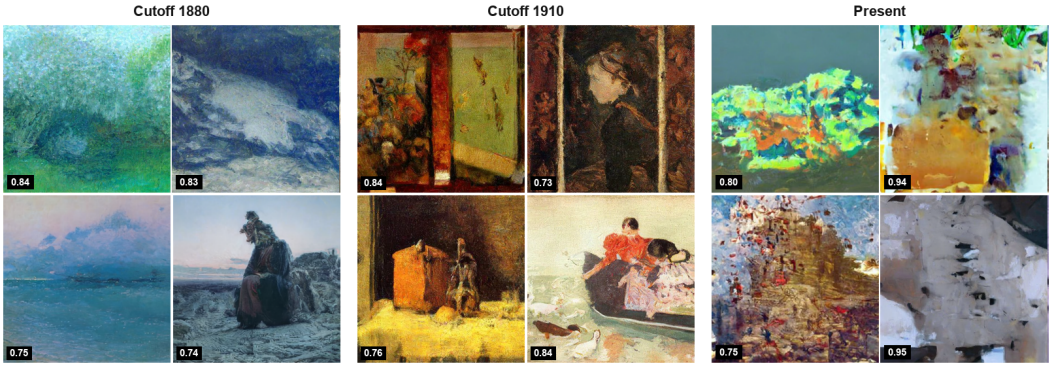


Figure 1: **StyleCast across three eras.** Each stack is trained only on paintings available before its cutoff. The 1910 panel shows source-anchored generations, while the 1880 and present panels show free compositions. Every image satisfies the anti-imitation certificate against both its nearest pre-cutoff neighbor and its conditioning source. The badge reports the smaller distance, and the per-cutoff novelty floors are 0.705, 0.722, and 0.738.

signal can be amplified without producing corrupted images. We use two complementary generation modes. Free generation from noise carries per-image novelty, while source-anchored paired generation uses matched sources, conditioning, and seeds to isolate the effect of steering. We also introduce a method to evaluate whether this steering produces a reliable shift toward the held-out future. Directional certification combines three frozen judges with a validity constraint defined by the displacement observed in the authentic future. A result is certified only when both style judges agree, and each directional claim is compared with five controls (null models) and a sign-asymmetry test.

The backtests provide evidence for one-step foresight. At the 1880 development cutoff, StyleCast moves the paired ensemble significantly toward the held-out next era under both style judges. The pre-specified configuration transfers to the 1910 cutoff without re-tuning, where all three judges certify the shift. The measured displacement corresponds to roughly one fifth to two thirds of a real era step and is detectable at the ensemble level. The main limitation appears beyond this first step. Stronger steering violates the validity constraint, iterative generation restores rather than extends the displacement, and self-training on generated samples fails to move the model further toward the future. In contrast, a small amount of authentic post-cutoff data produces additional directional progress. Within the mechanism family studied, these results are more consistent with an information-limited one-step horizon than with insufficient steering strength. Rolling-origin evaluation further shows that the forecasting model must be updated as history evolves. A model that only assimilates new data fails across the post-1910 stylistic rupture, while rebuilding the full stack restores the correct direction. At the present cutoff, where future ground truth is unavailable, we restrict our claims to per-image novelty for free compositions (Figure 1).

The main contributions of this paper are summarized as follows:

- We introduce StyleCast, a fully cutoff-pure generative pipeline that is rebuilt from scratch at each cutoff and steered by an era delta injected in score space.
- We develop a certification protocol based on held-out art history, combining frozen judges, a validity gate, null controls, and sign asymmetry. Under this protocol, the steered ensembles are certified at two historical cutoffs without re-tuning.
- We identify a one-step prediction horizon. None of the tested extension strategies increases the certified reach, while the contrast between authentic and synthetic fine-tuning supports an information-limited interpretation of this boundary.
- We extend the analysis to rolling-origin backtesting and present-day deployment, where forecasting must be re-anchored as history evolves. Every displayed output at deployment also satisfies an anti-imitation certificate.

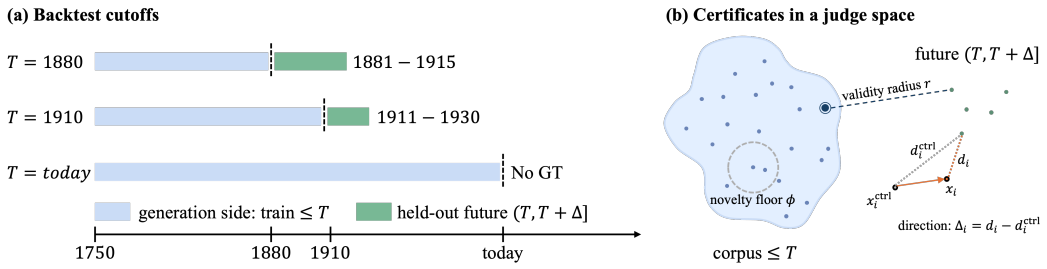


Figure 2: **The backtest formalism.** **(a)** At each cutoff the stack is rebuilt on works dated $\leq T$ (blue) and scored against the held-out future $(T, T+\Delta]$ (green). The present-day cutoff has no ground truth. **(b)** The three certificates in a judge space: paired direction Δ_i, validity radius r, and novelty floor ϕ, defined in the text.

2 Related Work

Steering Diffusion Models. Diffusion models are steered at sampling time by composing scores, from classifier and classifier-free guidance to autoguidance and later compositional, training-free, and theoretical extensions (Dhariwal & Nichol, 2021; Ho & Salimans, 2021; Karras et al., 2024; Liu et al., 2022; Brack et al., 2023; Sadat et al., 2025; Li et al., 2025). A complementary line operates in parameter space through task arithmetic and model merging (Ilharco et al., 2023; Gargiulo et al., 2025; Wortsman et al., 2022; Yang et al., 2026). Era-delta guidance combines the two lines: a delta between era-sliced finetunes, applied as score-space guidance and evaluated by displacement toward a held-out future rather than by sample quality. Image editing, personalization, and style manipulation supply the mechanics of our source-anchored protocol (Meng et al., 2022; Hertz et al., 2023; Brooks et al., 2023; Ruiz et al., 2023; Gal et al., 2023; Ouyang et al., 2025; Gatys et al., 2016; Huang & Belongie, 2017; Karras et al., 2021), in which shared sources, conditioning, and seeds isolate the injected direction. The stack follows the discrete-latent lineage (van den Oord et al., 2017; Esser et al., 2021; Rombach et al., 2022), rebuilt from scratch for cutoff purity.

Evaluating Style, Novelty, and Forecasts. Artistic style is measured by recognition models, dedicated style descriptors, and large contrastive or self-supervised encoders (Karayev et al., 2014; Somepalli et al., 2024; Kumar et al., 2025; Radford et al., 2021; Cherti et al., 2023; Oquab et al., 2024). These families instantiate our frozen judges. Learned representations recover artistic chronology (Elgammal et al., 2018; Van Horn et al., 2026). Creative generation typically rewards deviation from existing data (Elgammal et al., 2017; Lehman & Stanley, 2011; Mouret & Clune, 2015; Pugh et al., 2016; Broad et al., 2021; Jordanous, 2012; Hertzmann, 2018), and distributional metrics score fidelity against a training-time reference (Salimans et al., 2016; Heusel et al., 2017; Sajjadi et al., 2018; Kynkäänniemi et al., 2019; Naeem et al., 2020; Fox et al., 2025). Neither supplies a directional criterion. We instead score motion toward a temporally held-out distribution, adapting rolling-origin forecasting evaluation (Tashman, 2000; Bergmeir & Benítez, 2012; Hyndman & Athanasopoulos, 2021; Makridakis et al., 2020; Yao et al., 2022) to generative models. The anti-imitation certificate connects to memorization and diffusion-novelty findings (Somepalli et al., 2023; Carlini et al., 2023; Jeon et al., 2025; Kamb & Ganguli, 2025).

3 Preliminaries: Historical Backtesting

The Cutoff-Purity Protocol. We fix a cutoff year T and restrict all generation-side components and design choices to information available at or before T. This includes the tokenizer, style encoder, renderer, steering direction, and hyperparameter selection. Post-cutoff information is used only for evaluation. The judges may therefore be modern pretrained encoders, and the targets are authentic works from $(T, T+\Delta]$. In this paper, we define leakage as any post-cutoff information that influences image generation.

Each model configuration is accompanied by a purity ledger recording its training window and information sources. Hyperparameters are selected using internal calibration splits restricted to years $\leq T$. The exceptions are the injection weight, the gate τ, and the delta scale κ (Eq. 1), all

chosen at the development cutoff. Since the judged window overlaps part of the confirmatory target period, we additionally re-evaluated on a trimmed window that removes the overlap. The verdicts remain unchanged.

Cutoffs, Corpus, and Held-Out Futures. Our corpus consists of Western paintings from WikiArt, filtered for reproduction quality, balanced across years, and dated by catalogue date. Historical cutoffs are selected according to a rule fixed in advance: significant stylistic motion before T, movements after T with essentially zero pre-T exposure in the corpus, and dense population on both sides. We evaluate two historical cutoffs and one present-day deployment setting. At $T{=}1880$, the held-out future is 1881–1915. At $T{=}1910$, used for confirmation without re-tuning, the held-out future is 1911–1930, containing Cubism and abstract art, which have essentially no pre-cutoff presence in the corpus. The present-day setting uses paintings through 2015 and has no observable future ground truth. We use approximately 15 years as one *era step*, motivated by the spacing between successive movement cohorts in the corpus. The era slices and rolling increments follow this granularity, and each held-out window spans roughly one to two era steps as listed per cutoff (Figure 2a).

Judges and Certificates. Three frozen, modern, evaluation-side encoders act as judges: a style descriptor specialized to painterly texture (TEXTURE, the CSD descriptor of Somepalli et al., 2024), a large contrastive image encoder capturing semantic style (SEMANTIC, OpenCLIP ViT-bigG; Cherti et al., 2023), and a self-supervised ViT capturing structure (STRUCTURE, DINOv2; Oquab et al., 2024).

Directional claims are carried by a paired certificate (Figure 2b). For generations $\{x_i\}$ and seed-matched controls $\{x_i^{\mathrm{ctrl}}\}$, let d_i be the distance from x_i to its nearest real work of the held-out window in a judge space. The statistic is $\Delta_i = d_i - d_i^{\mathrm{ctrl}}$, bootstrapped over items. An arm is certified under a judge when the 95% confidence interval of $\bar{\Delta}$ lies entirely below zero, indicating that the generated ensemble is closer to the held-out future than its matched control. We report joint certification only when both style judges agree on this direction. STRUCTURE is reported as corroborating evidence. Certification is evaluated only within a validity gate. The median distance of an arm to the pre-cutoff corpus must not exceed the radius r, defined as the median nearest-neighbor distance from authentic future works to the same corpus. This limits admissible displacement to the scale observed in the real future. Each directional result is also compared with five null models: a time-shuffled placebo, two energy-matched random directions, a sign-reversed direction, and a random-teacher variant (the placebo also at an amplified dose in the tables). We additionally require sign asymmetry, meaning that reversed steering must move the ensemble away from the held-out future at matched energy. Displayed samples carry an *anti-imitation certificate* in TEXTURE space when both their distance to the nearest $\leq T$ neighbor and their distance to the conditioning source exceed the novelty floor ϕ, the median nearest-neighbor distance among real works of the corpus.

4 STYLECAST

4.1 OVERVIEW

Figure 3 summarizes the framework: a cutoff-pure stack, an era-delta steering signal injected in score space, and two generation modes that separate per-image novelty from ensemble-level direction. StyleCast rebuilds this stack from scratch at each cutoff without pretrained weights, which excludes perceptual losses backed by pretrained features. The VQ tokenizer supplies a spatial token map ($256^2 \rightarrow 32{\times}32$, codebook 1024). The factored style encoder supplies four 64-dimensional factors (palette, facture, composition, form), defined by transform pairs rather than labels and validated by invariance and artist-holdout tests. The renderer θ is a pixel-space diffusion UNet that attends to the style factors and a genre token and takes the upsampled token map at its input.

The stack is trained on the pre-cutoff corpus without temporal targets and therefore encodes no direction of stylistic change. Section 4.2 constructs this direction as an era delta estimated from the stack's own era-slice finetunes and specifies how it is injected during sampling. Section 4.3 then separates the two claims the framework must support. Direction is an ensemble-level proximity claim and novelty a per-image distance claim, so each is measured in its own generation mode.

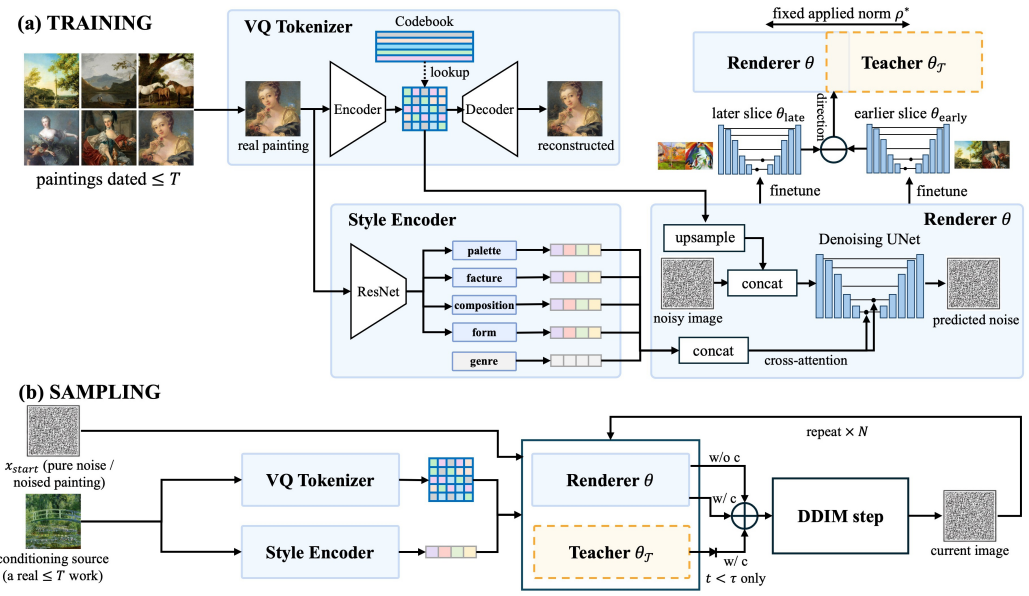


Figure 3: **The StyleCast framework. (a)** Training on $\leq T$ works: VQ tokenizer (1024-entry vocabulary), style encoder (four 64-d factors and a genre label token), renderer (token map upsampled and concatenated at input, condition tokens attended). Era-slice finetunes subtract into the teacher $\theta_{\mathcal{T}}$ at fixed applied norm (Eq. 1), the same network moved one step. **(b)** Sampling: the frozen encoders condition on a real $\leq T$ work, noised into the start state when anchored (sampling then begins at t=640) and replaced by pure noise with a coarsely pooled map when free. Each of N=50 DDIM steps queries the renderer with and without condition and, while the injection is active, the teacher. The gate $t < \tau$ uses τ=640, the start, in anchored passes and τ=400 in from-noise refinement. $\oplus$ combines the three predictions (Eq. 2).

4.2 Era-Delta Score Guidance

The steering direction is derived from the renderer itself. Let θ denote the base renderer, and let $\theta_{\text{early}}, \theta_{\text{late}}$ denote its fine-tuned versions on two adjacent pre-cutoff era slices, whose widths follow the corpus density of the cutoff (1851–1865 and 1866–1880 at T=1880, 1889–1899 and 1900–1910 at T=1910). The era-delta teacher $\theta_{\mathcal{T}}$ is defined as

$$\theta_{\mathcal{T}} \;=\; \theta \;+\; \kappa\,\big(\theta_{\text{late}} - \theta_{\text{early}}\big), \qquad \rho^{*} \;=\; \big\|\theta_{\mathcal{T}} - \theta\big\|, \tag{1}$$

where ρ^* is the recorded applied displacement of the cutoff. The scale κ=0.7 is fixed once and carried unchanged across cutoffs, giving $\rho^* = 97.1$, 70.35, and 127.66, and every competing arm within a comparison is matched to the same ρ^*. The k of the weight-space dose ladder sweeps the scale that κ fixes. The construction is task arithmetic (Ilharco et al., 2023) with era membership as the task variable. At each denoising step t with state x_t, conditioning c, and classifier-free guidance scale g, StyleCast samples with

$$\hat{\epsilon}(x_t) = \epsilon_\theta(x_t, \varnothing) + g\big[\epsilon_\theta(x_t, c) - \epsilon_\theta(x_t, \varnothing)\big] + w\,\mathbb{1}[t < \tau]\,\big[\epsilon_{\theta_{\mathcal{T}}}(x_t, c) - \epsilon_\theta(x_t, c)\big]. \tag{2}$$

The final term injects the teacher-to-base difference, weighted by w and gated by a mode-dependent τ. Source-anchored passes begin at t=640 with composition supplied by the source and inject throughout (τ=640). The from-noise refinement passes of the delivery pipeline take τ=400, after composition is set. Because the base renderer keeps denoising the state, the injection perturbs pre-cutoff sampling rather than replacing it, amplifying a weak temporal signal. At matched displacement and timing, score-space injection outperforms the weight-space addition (Section 5.3).

4.3 Free Composition and Paired Evaluation

Certification employs source-anchored generation, in which an authentic pre-cutoff painting supplies canvas and conditioning. Anchoring isolates the injected direction but keeps the output visually close to the source. For deployment, StyleCast instead uses *free composition*. A 58-word content

Table 1: **Cutoff 1910: paired direction toward the held-out 1911–1930.** Paired displacement is measured toward the held-out 1911–1930 period under three judges. Negative values indicate movement toward the future. ✓ denotes statistical significance and n.s. denotes a non-significant result. Validity requires the median distance to the pre-cutoff corpus to remain within the authentic-future radius. StyleCast with $w{=}1$ is the only configuration certified by all three judges.

Arm	TEXTURE	SEMANTIC	STRUCTURE	Validity
StyleCast ($w{=}1$) ①	$-.0042$ ✓	$-.0046$ ✓	$-.0035$ ✓	pass
StyleCast ($w{=}2$) ②	$-.0052$ ✓	$-.0025$ n.s.	$-.0015$ n.s.	pass
weight-space delta ③	$-.0014$ n.s.	$-.0036$ ✓	$-.0034$ ✓	pass
transition teacher ④	$-.0237$ ✓	+ away	n.s.	pass
placebo (time-shuffled) ⑤	$+.0057$ away	$+.0039$ away	—	pass
placebo, amplified ⑥	$+.0176$ away	$+.0131$ away	—	pass
random / reversed ($\times 4$)	all + away	all + away	—	2×fail

lexicon derived from pre-cutoff titles provides the tag anchor, while a pooled token map provides coarse spatial structure. Style factors, genre token, and pooled map are drawn from a parent pre-cutoff work, and that parent is the conditioning source of the output's anti-imitation certificate. The generation starts from noise. A first pass lays out the composition, a second gated pass re-noises and refines it under the same injection rule, and a 512^2 refiner finetune supplies the delivery resolution. The renderer used here carries a tag-and-pyramid finetune that makes the tag anchor and pooled maps in-distribution inputs. Free composition carries the per-image novelty claim, and source-anchored pairing carries the ensemble direction claim. Direction is a proximity claim and novelty a distance claim, so no single image is selected for both.

5 EXPERIMENTS

5.1 EXPERIMENTAL SETUP

At each backtest cutoff, we select source paintings from the final three decades before the cutoff. Each source is rendered once per arm with shared seeds ($n{=}600$ at the confirmatory cutoff and for the development-cutoff weight-space arms, $n{=}300$ for the development-cutoff score-space arms), which enables paired comparison at the image level. The control arm is the same source reconstructed through the identical sampler with the injection disabled ($w{=}0$). We evaluate each arm using the certificates defined in Section 3. These include paired directional displacement $\bar{\Delta}$, the validity gate, and the plausible-rate, the fraction of items whose distance to the held-out window falls within the radius r. All judges remain frozen across arms. The pre-specified primary configuration is StyleCast with $w{=}1$. All remaining analyses are secondary or descriptive and are reported with per-comparison intervals.

We compare StyleCast at $w{\in}\{1, 2\}$ against the same era delta applied in weight space at matched norm, a movement-transition teacher (Impressionism to Post-Impressionism, at-cutoff knowledge), the five null models, a trend-informed manual baseline that continues measured pre-cutoff attribute trends through calibrated image operations, and two learned temporal-extrapolation baselines, a style-factor trend push and a year-conditioned renderer. The main text reports point estimates. Complete verdict tables with confidence intervals for every arm are collected in Appendix C.

5.2 MAIN RESULTS: FORECASTING THE HELD-OUT NEXT ERA

All reported directional results use the same statistic. For each generated image, we measure its distance to the nearest painting in the held-out future and subtract the corresponding distance of its seed-matched control. Averaging over all pairs gives $\bar{\Delta}$ as defined in Section 3. Negative values indicate movement toward the held-out future. For scale, authentic future works lie roughly 0.8 from the pre-cutoff corpus. One era step of real stylistic change shifts cohort-to-future distances by 0.02 to 0.03 under the style judges. Observed displacements of -0.003 to -0.014 are therefore ensemble-level shifts of a measurable fraction of one era step, not per-image effects.

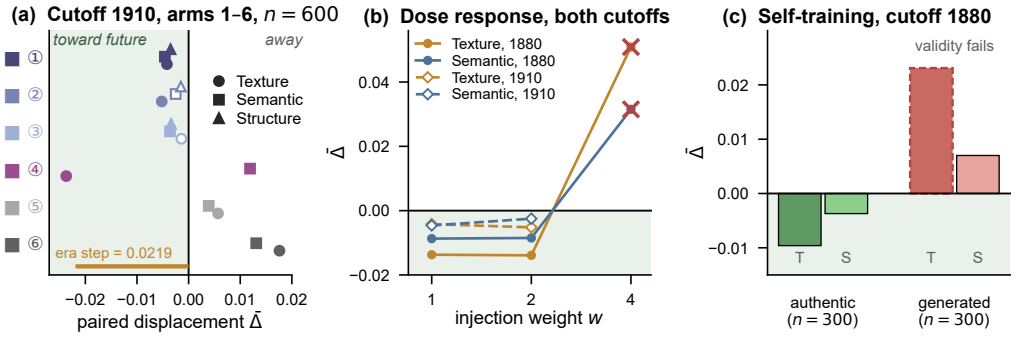


Figure 4: **Certification results.** **(a)** Paired displacement per arm at the 1910 cutoff. Colors identify the arms and key them by circled number to Table 1, marker shapes identify the judges, filled markers are significant, and the orange bar marks one real era step. **(b)** Dose response at both cutoffs, with validity failures crossed in red. **(c)** The self-training contrast under both style judges, with the validity failure dashed.

At the development cutoff (T=1880), StyleCast at w=1 moves the paired ensemble toward the held-out 1881–1915 by -0.0137 under TEXTURE and -0.0087 under SEMANTIC. The corresponding weight-space baseline reaches only -0.0090 and -0.0036 at matched dose. A recency control attributes the displacement to the era delta itself: sampling from the late-slice finetune alone produces no movement under either style judge. However, increasing the injection weight does not improve the result. At w=4, the validity gate fails and the SEMANTIC displacement reverses to $+0.0315$ (Figure 4b).

The same configuration transfers to the confirmatory cutoff T=1910 without re-tuning. At w=1, StyleCast moves significantly toward the held-out 1911–1930 period under all three judges (Table 1, Figure 4a). The magnitudes are smaller than at the development cutoff, as expected when a configuration tuned at one cutoff is applied unchanged at another. The STRUCTURE judge attains significance at the cutoff whose held-out future contains a structural revolution. Increasing the dose to w=2 produces a larger TEXTURE displacement but loses significance under the other two judges. The era step is measured on real paintings as how much closer the 1896–1910 cohort lies to the held-out window than the 1881–1895 cohort does, giving $0.0219/0.0231/0.0073$ under the three judges. Against these steps, the certified displacement at this cutoff is about one fifth under each style judge. The STRUCTURE step is three times smaller than the style-judge steps, so ratios are not comparable across judges. The same construction at the development cutoff, with cohorts 1851–1865 and 1866–1880 against the 1881–1915 window, gives steps of 0.0195 and 0.0323 under the style judges. The development-cutoff displacements are about two thirds of a step under TEXTURE and one quarter under SEMANTIC. The controls support the directional interpretation: all five null variants move away from the future under both style judges, two also fail the validity gate, and the sign-reversed arm departs at matched energy. The conclusion is unchanged under artist-clustered resampling, target-set resampling, cutoff exclusion, and color or luminance normalization. Halving the resolution reverses the TEXTURE verdict, suggesting the signal lies in fine-scale facture. Scored against equal-length windows on either side of the cutoff, the displacement flips sign at the cutoff itself.

Restricted to individual emerging movements under the SEMANTIC judge, the direction approaches Cubism and Expressionism while energy-matched random directions move away from both. None of the pre-specified baselines recovers the shift. The trend-informed manual composite moves away from the held-out future under both style judges while passing the validity gate. Its failure concerns direction rather than plausibility. In its per-operation decomposition, the color operation alone moves toward the future under both style judges, at one third to two thirds of the certified magnitude. A calibrated color trend therefore captures part of the displacement, and the era delta adds the facture-localized remainder. The fragmentation operation moves away under both style judges and dominates the composite. The learned extrapolation baselines leave paired displacement near zero.

5.3 Ablation Studies

We first ablate the steering mechanism. At matched norm and matched timing, score-space injection outperforms the weight-space addition at both cutoffs (Section 5.2 and Table 1). Temporal gating matters at large doses. In a descriptive dose ladder at the present cutoff, ungated double-stage application dissolves structure by $w{=}8$, while the gated single-stage construction remains structurally intact at $w{=}16$ at the cost of reduced texture diversity.

Norm matching is necessary for fair comparison across teachers. The raw era-delta norm inflates as the estimation slice shrinks, and two estimates of the same temporal direction share only $\cos = 0.18$ in weight space. All teacher comparisons therefore fix the applied norm ρ^*. After normalization, the small-data teacher remains competitive. Teacher construction also determines which future distribution is approached. Era-sliced teachers track the overall next-era distribution, while the Impressionism to Post-Impressionism teacher shifts generations toward specific emerging movements.

Deployment-pipeline ablations (source canvas, pooled maps, gated refinement, the 512^2 refiner) change fidelity but not direction, and the conclusions are unchanged under trimmed and single-era-step evaluation windows.

5.4 Prediction Horizon

We test three strategies for extending the certified one-step displacement, each against criteria fixed before execution: amplification against the validity gate and the certified direction, iteration against restoration ratios and a late-window advancement test, and self-training against paired rotation with an authentic-data positive control.

These experiments distinguish two readings. A mechanism-limited horizon should yield to stronger or repeated intervention, while an information-limited one should not.

Amplification increases distance but loses validity. Between $w{=}2$ and $w{=}4$ the ensemble exits the validity envelope and the SEMANTIC judge reverses sign. In a descriptive dose ladder at the present cutoff the corpus distance saturates ($w{=}16$ terminates no farther than $w{=}8$), consistent with departure from the validity region rather than progress in the certified direction.

Iteration does not compound the displacement. When each guided pass's output is fed into the next, an uninjected control shows that the carrier alone removes most of the accumulated displacement. The second injection restores roughly what was removed, and a pre-specified discriminator finds no advantage over a dose-matched single application on later target windows.

Finetuning on authentic future data changes the direction, and synthetic self-training does not. Three hundred authentic post-cutoff paintings rotate the renderer toward the held-out future (-0.0096 and -0.0037). The same protocol on three hundred self-generated samples moves it away ($+0.0231$) and outside the validity region (Figure 4c). A domain-matched control excludes rendering artifacts. The failure mode is consistent with self-consuming degradation (Shumailov et al., 2024; Alemohammad et al., 2024; Dohmatob et al., 2024; 2025; Bertrand et al., 2024; Fu et al., 2025), which only accumulating real data reverses (Gerstgrasser et al., 2024; Kazdan et al., 2025). Within the tested mechanism family the certified reach is one era step. The contrast is most consistent with a bound set by the information content of the corpus rather than by the capacity of the generator. In this setting, sustained forecasting requires periodic authentic data.

5.5 Operational Evaluation

We evaluate StyleCast walk-forward from 1880, revealing authentic data in 15-year increments, rebuilding the teacher at each origin under an identical protocol, and scoring each one-step forecast only after its target period is revealed. A pre-specified validity-only rule selects $w{=}2$ at every origin. Continuation origins retain directional significance under at least one style judge, though none meets the joint two-judge rule at that selected dose. The rupture separates data updating from model rebuilding. A model that incorporates new data while retaining its pre-1880 components predicts in the wrong direction ($+.0059$ under TEXTURE at the selected dose), while its sign-reversed control moves toward the target. The same evaluation window on the earlier anchor reproduces the inversion. Rebuilding the full stack at 1910 restores the joint direction on the same 1911–1925 window

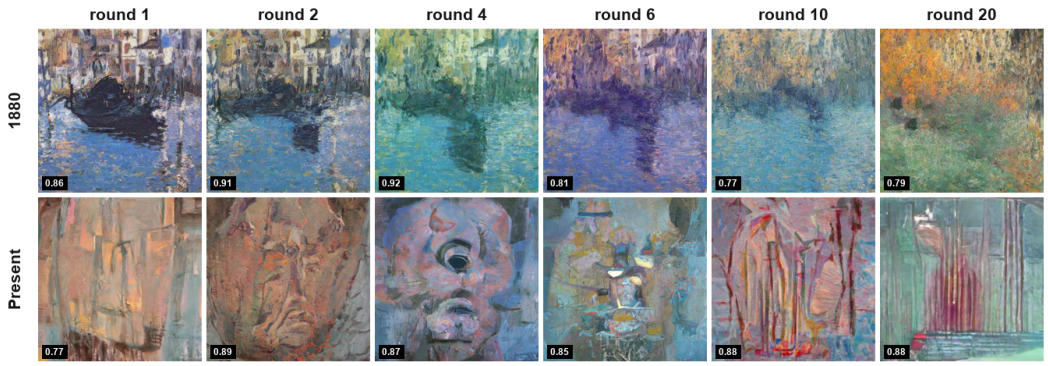


Figure 5: **Free evolution redraws content and composition.** One free-evolution chain per row at the 1880 and present cutoffs, 20 rounds each, every round redrawn from noise. Badges give the TEXTURE distance to the nearest pre-cutoff neighbor, including the conditioning source. All displayed images clear the novelty floor ϕ (0.705 at 1880, 0.738 at the present cutoff).

($-.0043$ and $-.0038$), again at a dose the rule had not picked. Updating the data alone is not enough once the forecasting representation itself is outdated.

At the present cutoff the direction cannot be evaluated, so deployment reports what is verifiable: novelty, validity, and dose stability. Under the same free-composition protocol at the confirmatory cutoff, 49–65% of free generations are simultaneously novel and valid, depending on the judge, and the steered arm matches this rate, so the check certifies the delivery pipeline rather than an added benefit of steering. At the present cutoff itself, the forward direction retains the pre-cutoff distribution better than its reversal or an energy-matched random direction, and the dose response is linear up to the split-half reliability ceiling. Figure 5 shows the delivered free-evolution chains, sampled between the two backtested doses. Each round redraws content and composition within the one-step range supported by the backtests, and every displayed sample passes the anti-imitation certificate.

6 DISCUSSION

Novelty is a per-image property. Directional forecasting is an ensemble-level effect observed through paired comparisons. A larger shift is not better, since stronger steering eventually violates the validity constraint. The experiments also show clear limits. None of the tested strategies extends the forecast beyond one era step, and the generated images do not reproduce the distinctive visual grammar of later artistic ruptures. Rupture-scale novelty also falls outside distance-based gates by construction. Certificates reward proximity to known targets, so ruptures certify only in hindsight. Authentic post-cutoff data improve the forecast, whereas self-generated data do not. This is consistent with an information-limited horizon within the mechanism family studied here. Rolling-origin experiments further show that forecasts must be re-anchored as new history becomes available, especially across major stylistic changes.

These conclusions remain specific to our setting. We study one renderer family, one era-delta steering mechanism, and one corpus of Western painting, and each cutoff's result concerns a single trained realization of the stack. The corpus is also canonized in hindsight, since which pre-cutoff works a modern archive retains reflects post-cutoff judgment, a channel the purity ledger does not close. The norm-matched small-data teacher and the zero-retuning transfer bound the dependence on the canonized corpus from two directions. A decomposition of independently estimated era deltas, whose weight-space agreement is only $\cos = 0.18$, remains an open control. Evaluation relies on frozen embedding-space judges, two of which share a contrastive-pretraining lineage. Transfer to other artistic traditions or generative architectures remains untested. No human evaluation was performed, since the certified effects are expected to lie below per-image perceptibility. An ensemble-level forced-choice study, in which observers match steered and control ensembles to eras, is the natural perceptual test and remains future work. Within this scope, the results support a narrow conclusion: cutoff-pure generation can anticipate the direction of the next stylistic era, but its reach is one era step, and the instrument must be re-anchored as new authentic data arrive.

AI USE STATEMENT

We used generative AI tools to improve the readability and language of the manuscript, assist with code debugging, and support literature search and reference discovery. All AI-assisted text, code modifications, and retrieved references were reviewed and verified by the authors. Generative AI was not used to formulate the research hypotheses, design the methodology or experiments, or interpret the experimental results. We take full responsibility for the final content of this work.

ETHICS STATEMENT

This work analyzes public, canonized artworks and involves no human subjects. Displayed generations carry anti-imitation certificates to avoid uncredited replication of artists' work, and the deployment protocol is scoped to historical-style research rather than the imitation of living artists.

REPRODUCIBILITY STATEMENT

Every confirmatory experiment in this paper was pre-specified: designs, decision criteria, and analyses were fixed in timestamped, version-controlled commits before execution, and verdicts were recorded against those criteria. The adversarial-review process behind this discipline is documented in Appendix D. Robustness re-evaluations are archived under the same manifest discipline. Each experimental arm archives its scripts, random seeds, and per-item manifests. Instrument checkpoints (tokenizer, style encoder, renderer, and era-slice finetunes) and all judge embedding caches are retained, so that every certificate can be recomputed from archived artifacts without regeneration. The purity ledger of each instrument records its training window and every information source (Section 3), and Appendix B summarizes the training and sampling configuration. Anonymized code, the pre-specification ledger, and all manifests will be released upon publication.

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*, 2024.

Christoph Bergmeir and José M. Benítez. On the use of cross-validation for time series predictor evaluation. *Information Sciences*, 191:192–213, 2012.

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

Manuel Brack, Felix Friedrich, Dominik Hintersdorf, Lukas Struppek, Patrick Schramowski, and Kristian Kersting. SEGA: Instructing text-to-image models using semantic guidance. In *Advances in Neural Information Processing Systems*, volume 36, 2023.

Terence Broad, Sebastian Berns, Simon Colton, and Mick Grierson. Active divergence with generative deep learning – a survey and taxonomy. In *Proceedings of the 12th International Conference on Computational Creativity*, 2021.

Tim Brooks, Aleksander Holynski, and Alexei A. Efros. InstructPix2Pix: Learning to follow image editing instructions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023.

Nicholas Carlini, Jamie Hayes, Milad Nasr, Matthew Jagielski, Vikash Sehwag, Florian Tramèr, Borja Balle, Daphne Ippolito, and Eric Wallace. Extracting training data from diffusion models. In *32nd USENIX Security Symposium*, pp. 5253–5270, 2023.

Mehdi Cherti, Romain Beaumont, Ross Wightman, Mitchell Wortsman, Gabriel Ilharco, Cade Gordon, Christoph Schuhmann, Ludwig Schmidt, and Jenia Jitsev. Reproducible scaling laws for contrastive language-image learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2818–2829, 2023.

Prafulla Dhariwal and Alexander Nichol. Diffusion models beat GANs on image synthesis. In *Advances in Neural Information Processing Systems*, volume 34, 2021.

Elvis Dohmatob, Yunzhen Feng, Pu Yang, François Charton, and Julia Kempe. A tale of tails: Model collapse as a change of scaling laws. In *International Conference on Machine Learning*, pp. 11165–11197, 2024.

Elvis Dohmatob, Yunzhen Feng, Arjun Subramonian, and Julia Kempe. Strong model collapse. In *International Conference on Learning Representations*, 2025.

Ahmed Elgammal, Bingchen Liu, Mohamed Elhoseiny, and Marian Mazzone. CAN: Creative adversarial networks, generating "art" by learning about styles and deviating from style norms. In *International Conference on Computational Creativity*, 2017.

Ahmed Elgammal, Bingchen Liu, Diana Kim, Mohamed Elhoseiny, and Marian Mazzone. The shape of art history in the eyes of the machine. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2018.

Patrick Esser, Robin Rombach, and Björn Ommer. Taming transformers for high-resolution image synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 12873–12883, 2021.

Alexis Fox, Samarth Swarup, and Abhijin Adiga. A unifying information-theoretic perspective on evaluating generative models. In *AAAI Conference on Artificial Intelligence*, volume 39, pp. 16630–16638, 2025.

Shi Fu, Yingjie Wang, Yuzhu Chen, Xinmei Tian, and Dacheng Tao. A theoretical perspective: How to prevent model collapse in self-consuming training loops. In *International Conference on Learning Representations*, 2025.

Rinon Gal, Yuval Alaluf, Yuval Atzmon, Or Patashnik, Amit H. Bermano, Gal Chechik, and Daniel Cohen-Or. An image is worth one word: Personalizing text-to-image generation using textual inversion. In *International Conference on Learning Representations*, 2023.

Antonio Andrea Gargiulo, Donato Crisostomi, Maria Sofia Bucarelli, Simone Scardapane, Fabrizio Silvestri, and Emanuele Rodola. Task singular vectors: Reducing task interference in model merging. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2025.

Leon A. Gatys, Alexander S. Ecker, and Matthias Bethge. Image style transfer using convolutional neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2414–2423, 2016.

Matthias Gerstgrasser, Rylan Schaeffer, Apratim Dey, Rafael Rafailov, Henry Sleight, John Hughes, Tomasz Korbak, Rajashree Agrawal, Dhruv Pai, Andrey Gromov, Daniel A. Roberts, Diyi Yang, David L. Donoho, and Sanmi Koyejo. Is model collapse inevitable? Breaking the curse of recursion by accumulating real and synthetic data. In *First Conference on Language Modeling*, 2024.

Amir Hertz, Ron Mokady, Jay Tenenbaum, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. Prompt-to-prompt image editing with cross-attention control. In *International Conference on Learning Representations*, 2023.

Aaron Hertzmann. Can computers create art? *Arts*, 7(2):18, 2018.

Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. GANs trained by a two time-scale update rule converge to a local Nash equilibrium. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. In *NeurIPS 2021 Workshop on Deep Generative Models and Downstream Applications*, 2021.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *Advances in Neural Information Processing Systems*, volume 33, pp. 6840–6851, 2020.

Xun Huang and Serge Belongie. Arbitrary style transfer in real-time with adaptive instance normalization. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 1510–1519, 2017.

Rob J. Hyndman and George Athanasopoulos. *Forecasting: Principles and Practice*. OTexts, Melbourne, Australia, 3rd edition, 2021.

Gabriel Ilharco, Marco Tulio Ribeiro, Mitchell Wortsman, Suchin Gururangan, Ludwig Schmidt, Hannaneh Hajishirzi, and Ali Farhadi. Editing models with task arithmetic. In *International Conference on Learning Representations*, 2023.

Dongjae Jeon, Dueun Kim, and Albert No. Understanding and mitigating memorization in generative models via sharpness of probability landscapes. In *International Conference on Machine Learning*, pp. 27091–27112, 2025.

Anna Jordanous. A standardised procedure for evaluating creative systems: Computational creativity evaluation based on what it is to be creative. *Cognitive Computation*, 4(3):246–279, 2012.

Mason Kamb and Surya Ganguli. An analytic theory of creativity in convolutional diffusion models. In *International Conference on Machine Learning*, pp. 28795–28831, 2025.

Sergey Karayev, Matthew Trentacoste, Helen Han, Aseem Agarwala, Trevor Darrell, Aaron Hertzmann, and Holger Winnemoeller. Recognizing image style. In *British Machine Vision Conference*, 2014.

Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43(12): 4217–4228, 2021.

Tero Karras, Miika Aittala, Tuomas Kynkäänniemi, Jaakko Lehtinen, Timo Aila, and Samuli Laine. Guiding a diffusion model with a bad version of itself. In *Advances in Neural Information Processing Systems*, volume 37, 2024.

Joshua Kazdan, Rylan Schaeffer, Apratim Dey, Matthias Gerstgrasser, Rafael Rafailov, David L. Donoho, and Sanmi Koyejo. Collapse or thrive: Perils and promises of synthetic data in a self-generating world. In *International Conference on Machine Learning*, pp. 29469–29494, 2025.

Anand Kumar, Jiteng Mu, and Nuno Vasconcelos. IntroStyle: Training-free introspective style attribution using diffusion features. In *IEEE/CVF International Conference on Computer Vision*, 2025.

Tuomas Kynkäänniemi, Tero Karras, Samuli Laine, Jaakko Lehtinen, and Timo Aila. Improved precision and recall metric for assessing generative models. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

Joel Lehman and Kenneth O. Stanley. Abandoning objectives: Evolution through the search for novelty alone. *Evolutionary Computation*, 19(2):189–223, 2011.

Xiang Li, Rongrong Wang, and Qing Qu. Towards understanding the mechanisms of classifier-free guidance. In *Advances in Neural Information Processing Systems*, volume 38, 2025.

Nan Liu, Shuang Li, Yilun Du, Antonio Torralba, and Joshua B. Tenenbaum. Compositional visual generation with composable diffusion models. In *European Conference on Computer Vision*, 2022.

Spyros Makridakis, Evangelos Spiliotis, and Vassilios Assimakopoulos. The M4 competition: 100,000 time series and 61 forecasting methods. *International Journal of Forecasting*, 36(1): 54–74, 2020.

Chenlin Meng, Yutong He, Yang Song, Jiaming Song, Jiajun Wu, Jun-Yan Zhu, and Stefano Ermon. SDEdit: Guided image synthesis and editing with stochastic differential equations. In *International Conference on Learning Representations*, 2022.

Jean-Baptiste Mouret and Jeff Clune. Illuminating search spaces by mapping elites. *arXiv preprint arXiv:1504.04909*, 2015.

Muhammad Ferjad Naeem, Seong Joon Oh, Youngjung Uh, Yunjey Choi, and Jaejun Yoo. Reliable fidelity and diversity metrics for generative models. In *International Conference on Machine Learning*, pp. 7176–7185, 2020.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy V. Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. DINOv2: Learning robust visual features without supervision. *Transactions on Machine Learning Research*, 2024.

Ziheng Ouyang, Zhen Li, and Qibin Hou. K-LoRA: Unlocking training-free fusion of any subject and style LoRAs. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2025.

Justin K. Pugh, Lisa B. Soros, and Kenneth O. Stanley. Quality diversity: A new frontier for evolutionary computation. *Frontiers in Robotics and AI*, 3:40, 2016.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, pp. 8748–8763, 2021.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10684–10695, 2022.

Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Yael Pritch, Michael Rubinstein, and Kfir Aberman. DreamBooth: Fine tuning text-to-image diffusion models for subject-driven generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023.

Seyedmorteza Sadat, Manuel Kansy, Otmar Hilliges, and Romann Weber. No training, no problem: Rethinking classifier-free guidance for diffusion models. In *International Conference on Learning Representations*, 2025.

Mehdi S. M. Sajjadi, Olivier Bachem, Mario Lucic, Olivier Bousquet, and Sylvain Gelly. Assessing generative models via precision and recall. In *Advances in Neural Information Processing Systems*, volume 31, 2018.

Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen. Improved techniques for training GANs. In *Advances in Neural Information Processing Systems*, volume 29, 2016.

Ilia Shumailov, Zakhar Shumaylov, Yiren Zhao, Nicolas Papernot, Ross Anderson, and Yarin Gal. AI models collapse when trained on recursively generated data. *Nature*, 631:755–759, 2024.

Gowthami Somepalli, Vasu Singla, Micah Goldblum, Jonas Geiping, and Tom Goldstein. Diffusion art or digital forgery? Investigating data replication in diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6048–6058, 2023.

Gowthami Somepalli, Anubhav Gupta, Kamal Gupta, Shramay Palta, Micah Goldblum, Jonas Geiping, Abhinav Shrivastava, and Tom Goldstein. Investigating style similarity in diffusion models. In *European Conference on Computer Vision*, 2024.

Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. In *International Conference on Learning Representations*, 2021a.

Yang Song, Jascha Sohl-Dickstein, Diederik P. Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. In *International Conference on Learning Representations*, 2021b.

Leonard J. Tashman. Out-of-sample tests of forecasting accuracy: an analysis and review. *International Journal of Forecasting*, 16(4):437–450, 2000.

Aäron van den Oord, Oriol Vinyals, and Koray Kavukcuoglu. Neural discrete representation learning. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

Andrew Van Horn, Lauryn Smith, Mahamad Mahmoud, Michael McMaster, Clara Pinchbeck, Ina Martin, Andrew Lininger, Anthony Ingrisano, Adam Lowe, Carlos Bayod, Elizabeth Bolman, Kenneth Singer, and Michael Hinczewski. PATCH: A deep learning method to assess heterogeneity of artistic practice in historical paintings. *Science Advances*, 12(16):eaea0485, 2026.

Mitchell Wortsman, Gabriel Ilharco, Samir Yitzhak Gadre, Rebecca Roelofs, Raphael Gontijo-Lopes, Ari S. Morcos, Hongseok Namkoong, Ali Farhadi, Yair Carmon, Simon Kornblith, and Ludwig Schmidt. Model soups: Averaging weights of multiple fine-tuned models improves accuracy without increasing inference time. In *International Conference on Machine Learning*, pp. 23965–23998, 2022.

Enneng Yang, Li Shen, Guibing Guo, Xingwei Wang, Xiaochun Cao, Jie Zhang, and Dacheng Tao. Model merging in LLMs, MLLMs, and beyond: Methods, theories, applications, and opportunities. *ACM Computing Surveys*, 58(8), 2026.

Huaxiu Yao, Caroline Choi, Bochuan Cao, Yoonho Lee, Pang Wei Koh, and Chelsea Finn. Wild-time: A benchmark of in-the-wild distribution shift over time. In *Advances in Neural Information Processing Systems*, volume 35, 2022.

A MECHANISM-LOCALIZATION PROBES

This appendix asks which channel of the generative stack carries the steerable temporal signal certified in the main text. Four probe classes cover the channels we could parameterize. Each reparameterizes image production so that the era-delta direction, if present in the exposed channel, could be carried by that channel alone. The recorded outcome then shows whether that channel can carry the signal. The series concludes that, within the tested parameterizations, the signal lies below the discrete token floor, as continuous brushwork micro-texture that smooth low-dimensional operators cannot express and that pixel space reaches only through adversarial solutions. All probes operate at the development cutoff (T=1880), use the certified weight-space era delta as the reference direction, and are scored under the paired protocol of Section 3, with positive paired differences denoting motion away from the held-out future. In the appendices we name the judges concretely: the TEXTURE judge is the CSD contrastive style descriptor (Somepalli et al., 2024), the SEMANTIC judge is a contrastive image encoder of the CLIP-G lineage (Cherti et al., 2023), and the STRUCTURE judge is DINOv2 ViT-L (Oquab et al., 2024). No probe optimizes against any judge: every probe objective is built exclusively from cutoff-pure critics (the factored style encoder, the from-scratch discriminator, and the VQ encoder), so the judges remain fully held out and adversarial solutions that fool only the critics are caught by design.

The signal to be localized is first characterized in the style-factor space of the cutoff-pure encoder. Distilling the certified teacher's edit over 600 paired generations yields a mean edit direction of norm 0.0607 with per-item cosine 0.148 to the mean, a weak but coherent direction. Its projection onto the four style factors is 55.2% facture, 33.5% form, 9.6% palette, and 1.7% composition. The certified direction is thus predominantly a micro-brushwork (facture) signal, and the probe series asks which parameterization of image change can carry it.

The first probe class operates in pixel space: each image is treated as a free parameter and optimized to displace its style-factor embedding along the teacher direction at controlled dose, subject to legality and structure-anchor terms from the cutoff-pure critics. Two designs are tested. In the unconstrained design, all arms exit the validity envelope (median TEXTURE distance to the $\leq$1880 corpus 1.065–1.079 against a radius of 0.838, where the untouched sources sit at 0.210), and both style judges certify significant motion *away* from the held-out future (TEXTURE $+0.245$ to $+0.259$, SEMANTIC $+0.143$ to $+0.155$), with the plausible-rate falling by 0.46 and 0.52 under the two style judges. The outcome is both direction-blind and dose-blind: the teacher-direction arms are indistinguishable from matched-norm random directions, and doses $k \in \{1, 4, 16\}$ share one signature. The constrained design adds a Laplacian-pyramid parameterization restricting fine scales to luminance, strong augmentation, tripled discriminator and smoothness weights, and an explicit orthogonal-drift penalty. The verdict is unchanged: validity 1.076–1.082 (fail), TEXTURE $+0.218$ to $+0.229$, SEMANTIC $+0.162$ to $+0.169$, both away. Direct optimization of pixels against cutoff-pure critics is an adversarial-example mechanism: the displacement it achieves in the critic space does not transfer to any held-out judge. The pixel-probe outputs appear strongly "stylized" to the eye while all arms, including random directions, share a single signature in judge space. This illustrates why the protocol admits no per-image visual evidence.

The second probe class replaces free pixels with a smooth, low-dimensional operator family (approximately 270 parameters per image: a 3×4 color affine, an 8×8 control-grid warp field, and two-band 8×8 frequency-gain maps), chosen so that the adversarial texture channel is inexpressible by construction. Validity is fully restored (TEXTURE 0.364–0.443, pass), but the direction is not reached: the residual displacement is an order of magnitude smaller than in pixel space and still significantly misdirected (TEXTURE $+0.021$ to $+0.029$). Teacher-direction and random-direction arms are again indistinguishable, and the achieved style-space alignment is at most $+0.18$ (cosine at the highest dose, -0.09 at dose 4). The reachable set of legal operators in this family does not contain the direction.

The third probe class gives the same operator family the densest possible supervision: for each certified pair, operator parameters are fit per-image to reproduce the teacher's own edit, and the fitted edit is then dose-extrapolated. The fit captures roughly 24% of the teacher edit, leaving approximately 76% as residual. Applying the captured component at λ=1 fails to reproduce the teacher's certificate (TEXTURE $+0.0127$ away, where the teacher itself scores -0.0090 under the identical control, SEMANTIC $+0.0152$ versus -0.0036). Extrapolation worsens monotonically, and a cross-

Table 2: **The mechanism-localization series** (development cutoff, $T{=}1880$). Each probe attempts to carry the certified era-delta direction through an alternative channel. Direction entries are paired differences against the probe's control, negative toward the held-out future and positive away. Validity is the median TEXTURE distance to the $\leq$1880 corpus (radius 0.838, untouched sources 0.210). Dashes denote entries not recorded at this granularity in the archived verdicts.

Probe	Channel	Validity	TEXTURE	SEMANTIC
pixel, unconstrained	free pixels under cutoff-pure critics	1.065–1.079 fail	$+0.245$ to $+0.259$	$+0.143$ to $+0.155$
pixel, constrained	pyramid parameterization + penalties	1.076–1.082 fail	$+0.218$ to $+0.229$	$+0.162$ to $+0.169$
operator	$\sim$270-d smooth operator parameters	0.364–0.443 pass	$+0.021$ to $+0.029$	—
distillation	operators fit to teacher edits ($\lambda{=}1$)	0.66–0.83 pass	$+0.0127$	$+0.0152$
weight-space delta (teacher)	reference under identical control	pass	-0.0090	-0.0036
token, carrier	teacher token grid re-rendered	—	-0.0016 n.s.	$+0.0001$ n.s.
token, resampling	transition semigroup P^{λ}	0.90–0.93 fail	—	$+0.1252$ vs. $+0.1264$ shuffled

image parameter shuffle at dose 4 behaves like the fitted edit itself ($+0.0495$ versus $+0.0424$). The fitted component carries generic operator energy with no teacher specificity. The direction therefore resides in the 76% residual that the operator family cannot express: fine-grained reorganization of brushwork texture, consistent with the 55.2% facture share of the distilled direction.

The fourth probe class tests the discrete token vocabulary as a carrier. Re-rendering the teacher's own token grids erases the certified signal entirely (TEXTURE -0.0016, SEMANTIC $+0.0001$, both n.s.), even though 55% of token positions flip under the teacher. The systematic component of the flip statistics (a 3.7% code-frequency drift) lies below the rendering noise floor. Editing tokens directly, via a transition-matrix resampling with semigroup dose P^{λ}, is destructive and grammar-blind: validity fails at all doses (0.90–0.93), and the arm is indistinguishable from a row-shuffled transition control (SEMANTIC $+0.1252$ versus $+0.1264$). At the $32{\times}32$, 1024-code granularity of the vocabulary, the token space cannot carry the temporal direction.

Table 2 summarizes the series. No probe channel both passes validity and moves toward the held-out future, while the weight-space teacher reference under the identical control does both. Within the tested channels, the probes locate the steerable temporal signal in the *sub-token continuous domain*: continuous micro-texture below the token floor, concentrated in the facture factor, inexpressible by smooth low-dimensional operators, and reachable in pixel space only through adversarial solutions. The success of the score-space mechanism of Eq. 2 is consistent with this localization: score-space injection operates in the sub-token continuous domain, and its per-step signal $\|\Delta\epsilon(t)\|$ concentrates in the detail-formation phase (rising from 0.019 at $t{=}640$ to 0.058 at $t{=}0$), motivating the gate $t < \tau$. The series also motivates the design rule used throughout the framework: every shortcut channel is closed by parameterization, not by assumption or by regularization. The two pixel designs show that penalties, even at tripled strength with explicit orthogonality constraints, do not close the adversarial channel, whereas the operator and token parameterizations close it constructively. Each subsequent instrument in the framework follows the same rule.

Every probe objective above is built from the cutoff-pure instrument family. Table 3 reproduces its purity ledger (Section 3). The training pools contain 6,708 works at $T{=}1880$, 27,187 at $T{=}1910$, and 44,556 at the present cutoff. The same construction repeats at each cutoff with only the training pool changed. Tokenizer health statistics show that purity does not come at the cost of degeneracy: holdout ℓ_1 reconstruction error is 0.0501 at the 1880 cutoff and 0.0704 at the 1910 cutoff, quantization error 0.004 and 0.0018 respectively, with the 1024-entry codebook fully utilized at both cutoffs. Every hyperparameter that selects among instruments is chosen on internal calibration splits strictly inside $\leq T$, and no information flows from any judge or from the held-out window into any instrument.

B TRAINING AND SAMPLING CONFIGURATION

This appendix records the configuration under which the instrument family was trained and all reported generations were sampled. The first paragraph fixes the architecture, optimizer, and training schedule of each trained component (tokenizer, style encoder, renderer). The second fixes the sampling, guidance, and measurement settings shared across arms. Together with the per-arm manifests, these settings allow every instrument and every certificate to be rebuilt from the released artifacts.

Table 3: **Purity ledger.** Training window and information sources per instrument. Windows repeat across the three cutoffs with only the training pool changed (6,708 / 27,187 / 44,556 works at T=1880 / 1910 / present). The era-slice row lists the cutoff-specific slice definitions. All discriminators are trained from scratch, and no component uses pretrained weights of any kind.

Instrument	Training window	Information sources
VQ tokenizer	works $\leq T$ (all three cutoffs)	corpus pixels; ℓ_1 + from-scratch patch discriminator where enabled (disabled at the 1880 cutoff, where it degraded holdout reconstruction); pretrained perceptual features excluded
factored style encoder	works $\leq T$ (all three cutoffs)	transform-pair supervision; attribute anchors re-standardized on each cutoff's $\leq T$ pool
conditional renderer	works $\leq T$ (all three cutoffs)	VQ tokens, style factors, and genre tokens of the $\leq T$ pool; from-scratch adversarial finetune
era-slice finetunes	1851–1865 / 1866–1880 (1,037 works per slice); 1889–1899 / 1900–1910 (3,903 per side); a pair centered on 1987 and 2005 (595 per side), the present-cutoff analogue of the adjacent-slice pair	era-sliced subsets of the $\leq T$ pool; applied norm ρ^* fixed by Eq. 1 and recorded per manifest
512^2 refiner	works $\leq T$ (cutoffs 1910 and present)	renderer weights + $\leq T$ images at 512^2 (+40k steps)

The visual tokenizer is a convolutional VQ autoencoder. Its codebook is maintained by spherical EMA vector quantization rather than by gradient, with cosine lookup in a factorized 32-d code space. It trains with Adam for 70k–100k steps per cutoff. The factored style encoder is a ResNet-18 trunk trained from scratch with four linear factor heads. It is optimized with AdamW for 25k–30k steps per cutoff under an InfoNCE objective (temperature 0.2) with artist-masked in-batch negatives, a gradient-reversal genre adversary, and the attribute-anchor regressions of Table 3. The conditional renderer is a 224M-parameter pixel-space diffusion UNet (ϵ prediction, cosine schedule, 1,000 training timesteps). It is trained in three stages totaling 220k steps: base conditioning, adversarial finetune, and token-map conditioning. The from-scratch patch discriminator is applied to the x_0 prediction at low noise levels only ($t < 400$). The 512^2 refiner adds 40k steps.

All experimental arms share the following sampling and measurement settings. Sampling uses DDIM (Song et al., 2021a) with 50 steps and classifier-free guidance g=1.5 throughout. Source-anchored arms noise the source canvas at strength 0.65, so sampling begins at t=640. The injection is active over the whole anchored trajectory. Free composition starts from pure noise, and its refinement passes gate the injection at t<400. Free-evolution chains run at w=1.5, between the two backtested doses (w=1 and w=2). Weight-space arms load the merged checkpoint's full EMA triple and sample under the standard classifier-free rule, with both branches from the merged renderer. Judge distances are Euclidean nearest-neighbor distances in the L2-normalized embedding spaces of the three judges (768-d, 1280-d, and 1024-d respectively). All intervals are percentile bootstrap intervals over 2,000 resamples. Per-arm manifests record every remaining configuration value. The released code reproduces each instrument from the corpus and each certificate from the archived generations.

C Full Verdict Tables

This appendix prints the complete verdict tables behind the summary results of the main text. The tables report, for each experimental stage, the full arm spectrum with paired displacements, bootstrap intervals, validity outcomes, and pre-specification identifiers. They support the same conclusions as the summaries. The certified one-step displacements stand at both backtest cutoffs, and none of the pre-specified extension strategies extends the certified reach.

The section expands Table 1 of the main text and the horizon-strategy outcomes into complete verdict tables: the development-cutoff arms (Table 4), the full confirmatory-cutoff arm spectrum including the trend-informed manual baseline and the null spectrum (Table 5), the horizon-extension arms (Table 6), the rolling-origin backtest with its attribution block (Table 7), and the deployment-cutoff certificates (Table 8).

Conventions are shared across the tables. The three judges of Section 3 are instantiated as follows. The Texture judge is the CSD contrastive style descriptor (the released ViT-L checkpoint of Somepalli et al., 2024). The Semantic judge is the OpenCLIP ViT-bigG-14 image encoder

Table 4: **Development cutoff (T=1880): full verdict table.** Paired displacement toward the held-out 1881–1915 (negative = toward the future). Validity reports pass/fail against the real-future radii 0.838 (TEXTURE) / 0.781 (SEMANTIC), with the arm's median distances to the ≤1880 corpus in parentheses (TEXTURE/SEMANTIC) where recorded. The bottom panel is evaluated at the confirmatory cutoff under the identical protocol at matched applied norm 70.3. Identifiers in row order are `deps1`, `deps2`, and `deps4` for the first panel, the `a0` ladder (k=0.5–2.0) with `b30u070` and `slicelate` for the second, and `time_k070` with `ctrl_n070` for the third.

Arm	TEXTURE	SEMANTIC	Validity
Score-space injection (Eq. 2); n=300			
StyleCast (w=1)	$-.0137$ [$-.0184, -.0089$] ✓	$-.0087$ [$-.0123, -.0050$] ✓	pass (.698/.493)
StyleCast (w=2)	$-.0139$ [$-.0201, -.0075$] ✓	$-.0085$ [$-.0134, -.0035$] ✓	pass (.788/.586)
StyleCast (w=4)	$+.0509$ away	$+.0315$ away	fail (.945)
Weight-space era delta (Eq. 1); n=600			
era delta, k=0.5	$-.0016$ [$-.0040, +.0008$]	$-.0026$ [$-.0047, -.0006$]	pass (.628)
era delta, k=1.0	$-.0073$ ✓	$-.0035$ ✓	pass (.676)
era delta, k=1.5	$-.0183$ [$-.0225, -.0141$]	$+.0004$ [$-.0031, +.0040$]	pass (.761)
era delta, k=2.0	$-.0276$ [$-.0333, -.0220$]	$+.0122$ [$+.0076, +.0164$]	pass (.830)
weight-space delta (k=0.7, 30k)	$-.0090$ ✓	$-.0036$ ✓	pass (.658)
late-slice renderer (recency control)	$+.0015$ n.s.	$-.0005$ n.s.	pass (.618)
Norm-matched data scaling (confirmatory cutoff; applied norm 70.3)			
full-data delta	$-.0014$ n.s.	$-.0036$ ✓	pass
small-data delta (3.7× less)	$-.0086$ ✓	$-.0023$ ✓	pass

(Cherti et al., 2023) as packaged in the Kandinsky 2.2 prior. The STRUCTURE judge is DINOv2 ViT-L/14 (Oquab et al., 2024). All three are frozen, evaluation-side instruments. No information flows from any judge into image production (Section 3). Unless a caption states otherwise, each cell reports the paired displacement $\bar{\Delta}$ toward the stated held-out window (negative = toward the future), bootstrapped over items with 2,000 resamples. A ✓ marks a 95% interval that excludes zero toward the future, "away" marks a 95% interval that excludes zero in the opposite direction, and n.s. marks an interval containing zero. In cells that print an interval, the interval itself carries the verdict. Values and intervals are transcribed verbatim from the execution log and the archived verdict files, or recomputed from the archived embedding caches under the identical pipeline, which reproduces every printed value exactly. Cells marked *arch.* have their value or interval only in the arm's archived manifest. Arms carry their pre-specification identifiers, listed in row order at the end of each caption. The identifier is the key under which the arm's scripts, seeds, per-item manifest, and judge embedding caches are archived, so that every row can be recomputed from archived artifacts (Reproducibility Statement).

Table 4 reports the development cutoff. The score-space arms apply Eq. 2 at doses $w \in \{1, 2, 4\}$ over n=300 paired sources against the reconstruction control. The w=4 row prints the archived trimmed-window re-evaluation, identical in sign and significance to the original-window verdict of the robustness appendix. The plausible-rate gains accompanying the w=1 arm are $+0.0600$ $[+0.0200, +0.1000]$ (TEXTURE) and $+0.0533$ $[+0.0067, +0.1033]$ (SEMANTIC). The weight-space ladder applies the era delta of Eq. 1 directly to the renderer weights at $k \in \{0.5, 1.0, 1.5, 2.0\}$ (the freely swept counterpart of Eq. 1's κ) over n=600 paired sources. At k=2.0 the recorded contrast is the plausible-rate, $+8.3$pp and still rising under TEXTURE against a significant -8.2pp reversal under SEMANTIC. The two finetunes are genre-balanced era slices 1851–1865 and 1866–1880 (1,037 works per side, 15k steps). The weight-space delta of Section 5.2 is the norm-matched champion of this family: the 30k-step delta applied at k=0.7, matching the applied displacement of the 15k-step delta at k=1 (raw delta norms 138.7 and 95.4, respectively). The STRUCTURE judge served at this cutoff as a legality instrument and is direction-neutral for texture-level deltas (champion arm: paired displacement n.s., plausible-rate $+2.8$pp). Its directional verdicts first appear at the confirmatory cutoff. The bottom panel archives the norm-matched data-scaling contrast of Section 5.3, evaluated under the identical confirmatory-cutoff protocol: the full-data delta and a delta estimated from 3.7× fewer works, both applied at displacement norm 70.3. The small-data delta certifies outright, while the full-data delta retains the SEMANTIC verdict with TEXTURE not significant. Scaling the estimation set thus does not extend the certified reach.

Two further baselines are reported in prose only. Both are learned temporal-extrapolation baselines, run at the development cutoff during instrument development on the same cutoff-pure stack family. Pushing a trend direction fit in the style-factor space at vector doses $0.15/0.30/0.60$ is inert: validity

Table 5: **Confirmatory cutoff (T=1910): full arm spectrum.** All arms use n=600, and negative values point toward the held-out 1911–1930. A dash marks a judge not run for that arm. Identifiers in row order are `deps_time_w1`, `deps_time_w2`, `time_k070`, `trans_n140`, and `deps_trans_w1` for the first panel, the dose-ladder arms indexed by applied norm with `ctrl_k210` for the second, `manual_trend`, `manual_color`, `manual_edge`, and `manual_frag` for the manual panel, and `placebo_nT`, `deps_placebo_w2`, `random_nT`, `random_n140`, `reverse_k070`, and `deps_random_w2` for the null panel.

Arm	TEXTURE	SEMANTIC	STRUCTURE	Validity
StyleCast and teacher variants				
StyleCast (w=1)	$-.0042$ [$-.0071, -.0014$]	$-.0046$ [$-.0066, -.0026$]	$-.0035$ [$-.0062, -.0007$]	pass
StyleCast (w=2)	$-.0052$ ✓	$-.0025$ [$-.0054, +.0005$] n.s.	$-.0015$ n.s.	pass
weight-space delta (k=0.7)	$-.0014$ n.s.	$-.0036$ ✓	$-.0034$ ✓	pass
transition teacher (weight)	$-.0237$ ✓	$+.0119$ [$+.0094, +.0145$]	n.s.	pass
transition teacher (score)	$-.0268$ ✓	$+.0126$ [$+.0100, +.0153$]	n.s.	pass
Weight-space dose ladder (by applied norm)				
norm 113 (small-data, k=0.7)	$-.0161$ ✓	n.s.	—	arch.
norm 141	$-.0230$ ✓	+ away (arch.)	—	arch.
norm 211	$-.0561$ ✓	+ away (arch.)	—	arch.
norm 226	$-.0471$ ✓	+ away (arch.)	—	arch.
small-data, k=2.1	arch.	arch.	—	fail (.870)
Trend-informed manual baseline				
composite	$+.0036$ [$+.0020, +.0053$]	$+.0029$ [$+.0010, +.0049$]	$-.0021$ [$-.0047, +.0006$]	pass
color operation	$-.0029$ [$-.0036, -.0023$] ✓	$-.0014$ [$-.0018, -.0010$] ✓	$+.0001$ [$-.0002, +.0003$]	pass
edge operation	$-.0001$ [$-.0004, +.0003$]	$-.0010$ [$-.0013, -.0008$] ✓	$+.0001$ [$-.0001, +.0003$]	pass
fragmentation operation	$+.0071$ [$+.0045, +.0097$]	$+.0092$ [$+.0068, +.0116$]	$-.0007$ [$-.0047, +.0030$]	pass
Null spectrum				
placebo (time-shuffled)	$+.0057$ away	$+.0039$ away	—	pass
placebo, amplified	$+.0176$ away	$+.0131$ away	—	pass
random direction	$+.0110$ [$+.0072, +.0148$]	$+.0105$ [$+.0083, +.0125$]	—	pass
random direction (norm 140)	$+.0337$ [$+.0288, +.0388$]	$+.0321$ [$+.0282, +.0359$]	—	fail
reversed	$+.0041$ [$+.0018, +.0065$]	$+.0064$ [$+.0047, +.0083$]	—	pass
random teacher (score)	$+.0407$ [$+.0355, +.0458$]	$+.0342$ [$+.0307, +.0380$]	—	fail

barely moves ($0.605 \to 0.613$), the TEXTURE displacement peaks at -0.003, and the SEMANTIC judge records zero at every dose. Amplifying the conditioning channel (guidance $1.5 \to 3.0$) improves this only to -0.0048 with SEMANTIC still at zero. A year-conditioned renderer (a trained year-token modulation, finetuned 30k steps) is also inert: a five-arm year sweep (n=600 per arm), including a falsification year predating the corpus slice, leaves paired displacement at approximately zero under both style judges with validity static (0.602–0.610). None of these conditioning channels (style-factor pushes, amplified guidance, and trained year tokens) moved facture-level statistics in this stack, which motivated the function-space program of Eqs. 1–2.

Table 5 expands Table 1. All arms share n=600 paired sources from 1881–1910, the reconstruction control, and the held-out 1911–1930 target. The per-judge radii recorded for this protocol are 0.8623 (TEXTURE), 0.7971 (SEMANTIC), and 0.8749 (STRUCTURE). The plausible-rate gain accompanying StyleCast (w=1) under the SEMANTIC judge is +4.8pp (✓). The weight-space dose ladder is indexed by applied displacement norm. Its norm-70 point is the weight-space delta row of the first block. The manual block reports the trend-informed manual baseline in full: all four calibrated operations pass every validity gate (TEXTURE medians 0.805–0.830 against the radius 0.8623). The fragmentation operation additionally shows significant plausible-rate drops -0.0317 $[-0.0583, -0.0067]$ (TEXTURE) and -0.0517 $[-0.0817, -0.0217]$ (SEMANTIC), complementing the displacement verdicts of Section 5.2. Restricting the target set to individual movements gives two further verdicts, both under the SEMANTIC judge with the energy-matched random direction as control: Cubism (n=72) -0.0092 ✓ against random $+0.0030$ away, and Expressionism -0.0124 ✓ against random $+0.0131$ away.

Table 6 collects the horizon-extension arms of Section 5.4. Amplification at the two backtest cutoffs is carried by the dose rows of Tables 4 and 5. At the deployment cutoff, the extreme dose ladder (the delivery teacher at applied norm 127.7, $w \in \{4, 8, 16\}$ at n=60, descriptive, no paired certificate) yields median SEMANTIC distances to the $\leq$2015 corpus of 0.813 at w=4 (30% of items within the era-step radius 0.785), 0.875 at w=8 (0%), and 0.856 at w=16 (0%, terminating no farther than w=8). The delivered dose w=1, measured on its own n=300 arm, sits at 0.725, inside the radius. This is the saturation cited in the main text. In the iteration panels, the uninjected re-rendering control removes approximately 69% of the single-step displacement under the re-encoded carrier

Table 6: **Horizon-extension arms** under the two style judges. Negative values point toward the future. Both iteration panels use the development-cutoff protocol, with single-step reference arms from the top panel of Table 4. Identifiers in row order are `deps1`, `comp2`, `drift2`, and `iddrift` for the first panel, `comp_w1`, `drift0`, `iddrift0`, and `comp_w1_zadv` for the second, and `reallate`, `synD2`, `reconD`, and `early` for the third.

Arm / contrast	TEXTURE	SEMANTIC	Outcome
Iteration, re-encoded carrier (n=126)			
single step	$-.0130$ ✓	$-.0075$ ✓	one-step reference
composed (two steps)	$-.0134$ ✓	$-.0047$ n.s.	$\approx$ one step
carrier drift (uninjected)	$-.0040$ n.s.	arch.	carrier removes step
identity drift	n.s.	arch.	protocol zero
Iteration, score carrier (n=300)			
composed (two steps)	$-.0156$ $[-.0202, -.0113]$ ✓	$-.0053$ $[-.0091, -.0015]$ ✓	refill, not advance
carrier drift (uninjected)	$-.0020$ $[-.0054, +.0011]$	$+.0005$ $[-.0020, +.0031]$	step removed
identity drift	$+.0003$ $[-.0028, +.0034]$	$+.0016$ $[-.0009, +.0041]$	protocol zero
composed, advanced cond.	$-.0172$ $[-.0218, -.0126]$ ✓	$-.0063$ $[-.0102, -.0023]$ ✓	refill, not advance
restoration ratio	0.997 $[0.763, 1.333]$	0.664 $[0.329, 1.086]$	—
restoration ratio, adv. cond.	1.111 $[0.867, 1.481]$	0.782 $[0.442, 1.258]$	—
composed vs. dose-matched (late)	$+.0104$ $[+.0058, +.0151]$	$+.0085$ $[+.0046, +.0125]$	no advancement
Self-training (n=300; 300 works per arm; 10^4 steps)			
real futures vs. early control	$-.0096$ $[-.0133, -.0059]$ ✓	$-.0037$ $[-.0068, -.0004]$ ✓	toward, certified
synthetic futures vs. domain control	$+.0231$ $[+.0164, +.0298]$	$+.0070$ $[+.0016, +.0129]$	away; validity fail (.860)
domain control vs. early control	$+.0451$ $[+.0405, +.0501]$	$+.0341$ $[+.0299, +.0385]$	fingerprint bound
real futures, late window	$-.0099$ $[-.0136, -.0061]$ ✓	$-.0034$ $[-.0064, -.0004]$ ✓	descriptive

and 85–100% under the score carrier, and the second injection restores approximately the amount removed. The late-window row is the pre-specified advancement discriminator: positive values mean the composed arm approaches later targets (1901–1915) *less* closely than the dose-matched single application. The restoration-ratio rows report the second-step gain over the first-step gain against a pre-specified accumulation line of 70%. All score-carrier arms pass the validity gate in both judge spaces. The self-training panel reports the paired finetuning contrast (300 works per arm, 10^4 steps, probe ensembles of n=300, target window 1881–1895). The positive control on authentic post-cutoff works rotates the renderer toward the future. The identical protocol on generated futures rotates it away and exits the validity envelope (0.860 against radius 0.809, TEXTURE). The domain contrast bounds the rendering-fingerprint effect that any synthetic finetune inherits. The data-scaling side of this attribution is archived in Table 4: the raw estimated delta inflates in norm as the slice shrinks (161.6 from 3.7$\times$ less data against 100.5, at the confirmatory cutoff), two same-protocol estimates of the same direction share only $\cos = 0.178$ in weight space, and the norm-matched small-data teacher certifies at least as strongly (-0.0086/-0.0023).

Table 7 reports the rolling-origin backtest in full. At each origin the teacher is rebuilt from adjacent pre-origin windows under an identical protocol (applied norms 60.8/60.8/61.3 across the three origins, construction-symmetry gates 1.01/1.25/1.16, all pass), the dose is selected by the pre-specified validity-only rule using $\leq T_k$ information (the rule selects w=2 at every step), and the instrument gate (reconstruction-lane validity) passes at every step in both judge spaces. At step 1 the w=1 arm is certified by both style judges, and all four null lanes move away from the truth window under the SEMANTIC judge. At step 2 the TEXTURE judge certifies both doses. At step 3, the rupture step, the true-direction arms move away from the revealed future and the sign-reversed lanes move toward it. The attribution block rebuilds the certified window pair of the confirmatory cutoff (1889–1899/1900–1910) on the rolled anchor (teacher applied norm 63.5, symmetry gate 1.16). The inversion reproduces and strengthens, which argues against window timescale as the cause. In contrast, the natively rebuilt 1910 stack certifies the same held-out era (Table 5).

Table 8 archives the free-composition distribution statistics together with the present-cutoff dose-robustness certificates of Section 5.5. The distribution statistics were established on the confirmatory-cutoff stack, whose teachers supply the direction arms, before the protocol was transplanted to the present cutoff. The distribution panel reports, over n=300 shared free drafts per arm, the fraction of compositions simultaneously novel (beyond the novelty floor) and valid (inside the real-future envelope). The gated StyleCast arm at w=2 is statistically indistinguishable from the free pipeline in all three spaces (rates archived). The exploration-regime transition arm at w=4 is novel on 98–100% of items but collapses the validity rate. The placebo and random lanes collapse comparably (rates archived), and the sign-reversed arm collapses to 0–6% validity. The dose panel

Table 7: **Rolling-origin backtest: per-step verdicts.** Negative values point toward the subsequently revealed truth window, and ✓ marks a 95% interval that excludes zero toward it. Validity failures print the failing median and judge, and all other arms pass both gates. Identifiers per step in row order are `real/deps1`, `real/deps2`, `bridge/deps1` at step 1 only, `rand/deps1`, `rand/deps2`, `rev/deps1`, and `rev/deps2`, with the attribution block `b_real_w1`, `b_real_w2`, `b_rev_w1`, and `b_rev_w2`. The native-rebuild rows score the natively rebuilt 1910 stack on the step-3 window, recomputed from the archived embedding caches.

Arm	TEXTURE	SEMANTIC	Validity
Step 1: origin 1880 → truth 1881–1895 (n=300; radii .8094/.7450)			
forecast (w=1)	$-.0025$ $[-.0047, -.0001]$ ✓	$-.0021$ $[-.0038, -.0005]$ ✓	pass
forecast (w=2, selected)	$+.0007$ $[-.0033, +.0047]$	$-.0038$ $[-.0065, -.0012]$ ✓	pass
bridge (development-cutoff teacher)	$-.0017$ $[-.0047, +.0013]$	$-.0004$ $[-.0027, +.0019]$	pass
random (w=1)	$+.0036$ $[-.0012, +.0082]$	$+.0049$ $[+.0024, +.0076]$	pass
random (w=2)	$+.0275$ $[+.0207, +.0344]$	$+.0280$ $[+.0231, +.0330]$	pass
reversed (w=1)	$-.0014$ $[-.0041, +.0012]$	$+.0023$ $[+.0002, +.0046]$	pass
reversed (w=2)	$+.0043$ $[-.0002, +.0088]$	$+.0086$ $[+.0047, +.0126]$	pass
Step 2: origin 1895 → truth 1896–1910 (n=300; radii .8168/.7575)			
forecast (w=1)	$-.0044$ $[-.0063, -.0022]$ ✓	$-.0006$ $[-.0023, +.0012]$	pass
forecast (w=2, selected)	$-.0046$ $[-.0077, -.0015]$ ✓	$-.0000$ $[-.0031, +.0032]$	pass
random (w=1)	$+.0102$ $[+.0052, +.0150]$	$+.0078$ $[+.0049, +.0106]$	pass
random (w=2)	$+.0227$ $[+.0150, +.0300]$	$+.0319$ $[+.0267, +.0372]$	fail (.845, TEXTURE)
reversed (w=1)	$-.0022$ $[-.0045, +.0001]$	$+.0015$ $[-.0010, +.0037]$	pass
reversed (w=2)	$+.0046$ $[+.0003, +.0092]$	$+.0070$ $[+.0031, +.0111]$	pass
Step 3 (rupture): origin 1910 → truth 1911–1925 (n=600; radii .8502/.7866)			
forecast (w=1)	$+.0023$ $[+.0007, +.0038]$	$+.0027$ $[+.0015, +.0039]$	pass
forecast (w=2, selected)	$+.0059$ $[+.0033, +.0087]$	$+.0045$ $[+.0024, +.0065]$	pass
random (w=1)	$+.0112$ $[+.0077, +.0147]$	$+.0054$ $[+.0034, +.0075]$	pass
random (w=2)	$+.0116$ $[+.0064, +.0169]$	$+.0213$ $[+.0181, +.0247]$	fail (.853, TEXTURE)
reversed (w=1)	$-.0081$ $[-.0102, -.0059]$ ✓	$-.0024$ $[-.0038, -.0009]$ ✓	pass
reversed (w=2)	$-.0105$ $[-.0146, -.0066]$ ✓	$-.0020$ $[-.0044, +.0004]$	pass
native rebuild (w=1)	$-.0043$ $[-.0071, -.0016]$ ✓	$-.0038$ $[-.0057, -.0017]$ ✓	pass
native rebuild (w=2, selected)	$-.0037$ $[-.0080, +.0004]$	$-.0013$ $[-.0043, +.0018]$	pass
Attribution: certified window pair (1889–1899/1900–1910) on the rolled anchor			
rebuilt pair (w=1)	$+.0055$ $[+.0036, +.0073]$	$+.0020$ $[+.0004, +.0037]$	pass
rebuilt pair (w=2)	$+.0124$ $[+.0093, +.0157]$	$+.0052$ $[+.0024, +.0079]$	pass
rebuilt pair, reversed (w=1)	$-.0137$ $[-.0162, -.0112]$ ✓	$-.0023$ $[-.0040, -.0006]$ ✓	pass
rebuilt pair, reversed (w=2)	$-.0183$ $[-.0229, -.0138]$ ✓	$+.0001$ $[-.0027, +.0030]$	pass

reports the paired validity-shift of each arm against the free base under the primary SEMANTIC judge (novelty floor 0.6913, era-step radius 0.7853, negative = toward the $\leq$2015 corpus). The radius is the median nearest-neighbor distance of the most recent fifteen-year cohort of real works to the corpus preceding it, the present-cutoff transplant of the backtest radius construction. Direction at this cutoff is carried by sign asymmetry rather than by a target window. The asymmetry rows compare the reversed and random arms against the matched-norm forward arm (positive = farther off-manifold than forward). The dose response is linear up to the split-half estimation ceiling: $\cos = 0.914$ between the matched-norm and delivered-norm responses, against ceiling 0.967 and random floor 0.443. At the 512^2 delivery layer (n=100) the same signature holds: -0.0200 $[-0.0286, -0.0116]$ ✓ and -0.0133 $[-0.0194, -0.0070]$ ✓ for the deployed and matched doses respectively, with linearity 0.880 against its ceiling 0.896.

D INTERNAL ADVERSARIAL REVIEW

This appendix explains how the certificate chain was protected against errors that leave no trace in final statistics. It first states the two-stage adversarial-review protocol under which every confirmatory experiment was pre-specified, attacked, executed, and re-audited. It then reconstructs four cases in which the review caught a flaw that would otherwise have produced a wrong result, among them a withdrawn certification. The conclusion is that the review records are themselves part of the released scientific product.

Every confirmatory experiment behind the certificate chain of StyleCast was executed under a two-stage internal adversarial-review protocol. Before execution, each experiment was pre-specified: a design document fixing the arms, sample sizes, seed pairing, decision criteria (negative and inconclusive branches included), and the analysis to be run was frozen in a timestamped, version-controlled

Table 8: **Free-composition distribution statistics (confirmatory cutoff) and dose robustness at the present cutoff.** The top panel reports the novel-and-valid rate over n=300 shared drafts, followed by validity rates of the exploration and reversed arms. The bottom panel reports the paired validity shift against the free base under the SEMANTIC judge, negative toward the $\leq$2015 corpus. Identifiers are `ctrl`, `time_w2g`, `trans_w4g`, and `revtrans_w4g` for the top panel and `recon`, `w1_n182`, `w1_n070`, `w2_n070`, `rand_w1_n070`, and `rev_w1_n070` for the bottom.

Arm	TEXTURE	SEMANTIC	STRUCTURE
Novel-and-valid rate			
free pipeline	48.7%	65.3%	63.0%
StyleCast (w=2, gated)	no significant difference from the free pipeline (arch.)		
Validity rate (novelty 98–100%)			
transition (w=4, exploration)	26.7%	4%	n.s. shift
placebo / random lanes	comparable collapse (arch.)		
reversed (w=4)	0–6%		

Arm	$\bar{\Delta}$ vs. free base (SEMANTIC)	novel $\times$ valid
free base	—	60.3%
deployed dose (w=1, norm 127.7)	$-.0146$ $[-.0196, -.0095]$ ✓	64.0%
matched dose (w=1, norm 70.34)	$-.0100$ $[-.0138, -.0063]$ ✓	61.0%
matched dose (w=2)	$-.0067$ $[-.0112, -.0022]$ ✓	63.3%
random (matched norm)	$+.0008$ $[-.0059, +.0080]$	62.7%
reversed (matched norm)	$-.0031$ $[-.0065, +.0004]$	61.7%
reversed vs. forward	$+.0069$ $[+.0030, +.0111]$ ✓	—
random vs. forward	$+.0109$ $[+.0048, +.0174]$ ✓	—

Table 9: **Four representative interventions of the internal adversarial review.** Design review precedes execution, delta review precedes entry of a verdict into the certificate chain, and the in-flight audit is continuous. Details in the text.

Entry	Stage caught	Flaw	Disposition
N16	in-flight audit	silent conditioning-library mismatch	runs re-executed; hard assertion added
N32	delta review	single-judge certification of an artifact	certification withdrawn; cross-judge rule
N68	design review	dose matched to the wrong norm	matching target corrected; deployment erratum
N73	design review	three fatal findings (gate, baseline, exclusion)	redesigned, re-frozen, then executed

commit. No confirmatory run was launched from an unfrozen design. Between drafting and freezing, the design was attacked in an internal adversarial review against a standing threat catalogue: information leakage across the cutoff, baselines transcribed from the wrong cutoff or judge, uncalibrated gates and instruments, dose and energy confounds, selection and reporting effects, seed and pairing structure, and statistical power. Findings were graded fatal, major, or minor. A design proceeded only once every fatal finding was resolved and the resolution folded back into the frozen pre-specification, a disposition the ledger records as *go-after-fixes*.

After execution, verdicts were recorded strictly against the pre-specified criteria. Each completed run received a delta review, including cross-judge replication and visual inspection of unselected outputs, before its result entered the certificate chain. Certifications could be withdrawn, and one was. The review findings and their resolutions are archived alongside the designs they modified. The remainder of this appendix documents four representative interventions in which review caught a flaw that would otherwise have produced a statistically significant but wrong result. Table 9 summarizes the four cases, and experiments are referenced by their pre-specification identifiers throughout.

The first case is an instrument incident caught by the in-flight audit during the development-cutoff instrument phase (N16). The conditional renderer of that instrument generation had been trained against one archived extraction of the factored style codes, while two evaluation scripts defaulted to a second extraction. The two versions share an identical shape of four 64-dimensional factors, so the mismatch raised no error and silently scrambled the conditioning semantics. The audit reconstructed a contamination matrix run by run: one formal backtest round and the first six cells of a twelve-cell operating-point sweep had consumed mismatched codes, two finetuning arms had trained against them for approximately one thousand steps, and the remaining runs were verified clean. All contaminated runs were killed and re-executed from a clean state, and the contaminated outputs were quarantined under a mismatch-labeled archive rather than deleted. The clean re-run

reversed a conclusion that had already been drafted: under mismatched codes the sweep had failed the validity gate in every cell (0.855–0.896 against the radius 0.838), a pattern that had been read as a tension between validity and conditioning strength. Under matched codes all twelve cells passed (0.662–0.768), with the heaviest conditioning configuration attaining the best validity in the grid (0.662). The apparent tension was an artifact of scrambled conditioning, which acts as noise rather than as a usable control signal. The standing fix is a hard assertion, wired into all three sampling entry points, that compares the code-library identity recorded in the training configuration against the library supplied at sampling time and refuses to run on mismatch unless explicitly overridden.

The second case is the withdrawn certification (N31/N32). An early structural-operator study at the development cutoff was about to record its first fully certified structural operator: a token-grid fragmentation operator at its lightest dose had passed the validity gate and moved the paired ensemble toward the held-out future under the TEXTURE judge, the CSD contrastive style descriptor (Somepalli et al., 2024), by -0.0111 $[-0.0144, -0.0078]$ with a plausible-rate gain of $+3.17$ percentage points $[+0.50, +5.83]$ at n=600, and its style-recombination variant appeared stronger still (-0.0124 $[-0.0171, -0.0081]$; $+5.17$ $[+2.00, +8.17]$). The delta review caught it before the claim hardened: the mandated inspection of unselected outputs found patch-displacement artifacts with textures copied from the source, and a paired recheck under the SEMANTIC judge, a CLIP-G-lineage contrastive image encoder (Cherti et al., 2023), reversed the direction: $+0.0077$ $[+0.0042, +0.0113]$ significantly *away* from the future for the base operator, and $+0.0071$ with a plausible-rate drop of -5.5 percentage points $[-9.2, -1.7]$ for the variant. The TEXTURE judge had scored block-displacement artifacts as unlike the past. The SEMANTIC judge, consistent with the inspection, placed the outputs farther from the authentic future. In contrast, a genuine effect certified in the same period, the adversarial-refinement gain, moved both style judges in the same direction ($+16.8$ and $+13.3$ percentage points of plausible-rate). The certification was withdrawn. The rule that no directional claim may rest on a single judge (both style judges significant, in the same direction) was adopted as a precondition for certification and is the origin of the protocol of Section 3. Fragmentation-style artifacts have since reappeared on independent material: the fragmentation component of the trend-informed manual baseline moves away from the future under both style judges ($+0.0071$/$+0.0092$, Section 5.2), and under the rule it cannot certify.

The third case is a design-review fatal in the deployment-cutoff dose-robustness arm (N68). The delivered present-cutoff ensemble had been generated with an era-delta teacher whose recorded norm was 182.4, with no norm argument connecting the deployed dose to the certified backtest dose. The robustness arm was designed to close this gap with norm-matched lanes. As first drafted, the lanes were matched to 100.5, the norm of the cutoff-1910 era delta itself. Review flagged this target as fatal: the quantity the cutoff-1910 backtest certified is the displacement actually applied at the certified operating point, $\|\theta_{\mathcal{T}} - \theta\| \approx 70.35$ (Eq. 1; the delta norm 100.5 scaled by the certified injection factor 0.7), and 100.5 itself lies inside the untested 70–113 gray zone bracketing the weight-space validity cliff. A lane matched there would have probed neither the certified point nor the deployed dose, and any failure there would have been uninterpretable. The matching target was corrected to the applied norm before freezing. The same review pass surfaced an erratum in the deployment record itself: the archived figure of 182.4 was likewise a delta norm before scaling. The norm actually applied at delivery, measured at build time, is 127.66. The manifest was corrected. Under the corrected design, the pre-specified dose-robust branch held: the deployed dose remains within the manifold under the deployment cutoff's primary judge (SEMANTIC; -0.0146, certified), sign asymmetry holds against the reversal ($+0.0069$) and the energy-matched random direction ($+0.0109$), and the dose response is linear up to the split-half estimation ceiling ($\cos 0.914$ vs. 0.967). These numbers, reported in Appendix E and Table 8, exist in this form only because the matching target was corrected before execution.

The fourth case is the review of the deployment-cutoff operator league (N73), an exploration-side design probing whether inference-time recombination operators can populate the region beyond the real-future radius while remaining legible. Review returned three fatal findings. First, the verdict gate was uncalibrated: operator lanes were to be scored under the same distance-based validity gate as the forecasting arms, without any measurement of how authentic revolutionary works fare under that gate. The required calibration became a result in its own right: authentic held-out avant-garde works at the confirmatory cutoff (Cubism, abstract art, and Surrealism; n=187) pass the validity gates at rates of only 19.8%, 14.4%, and 9.6% under the TEXTURE, SEMANTIC, and STRUCTURE judges (the last being DINOv2 ViT-L; Oquab et al., 2024), while continuation movements pass at

54–55%. A distance gate thus rejects, by construction, the works that constitute a rupture. This calibration underlies the statement in Section 6 that rupture-level novelty sits structurally outside distance-based gates. The design was rebuilt around a two-band scoring in which the band beyond the real-future radius is scored by legibility rather than by distance, and the pre-specified negative branch was rewritten so that failure would constitute calibrated evidence rather than the artifact of an unpassable gate. Second, a decision baseline had been transcribed from the wrong cutoff: the design carried 63.0% as the free-composition control lane's continuation-band rate under the STRUCTURE judge, whereas recomputation during review gave 31.7% at the deployment cutoff. Every lane would have been held to a threshold roughly twice the true baseline, guaranteeing a spurious global failure. Third, the hybrid lane composes two parent works, but the novelty accounting as drafted excluded only the primary parent from the nearest-neighbor reference set, so a lane could have registered as novel while copying its second parent. Dual-parent and same-artist near-neighbor exclusion was added. With all three findings resolved and the design re-frozen, the league executed and the negative branch triggered exactly as pre-specified: no operator lane exceeded the free-composition control, the strongest lane being the control itself at readable rupture-band rates of 20.0%/19.0% under SEMANTIC/STRUCTURE. The verdict is interpretable only because the gate had been calibrated against authentic revolutions and the baseline corrected before execution.

These four interventions share a structure. In each, the uncorrected design would have run to completion, produced clean-looking statistics, and supported a wrong conclusion: a spurious validity–conditioning tension (N16), a certified operator whose gains were artifacts of the certifying judge (N32), a robustness lane aimed at an uninterpretable dose (N68), and a league verdict measured against an uncalibrated gate and a doubled baseline (N73). None of these failure modes is visible in final statistics. They are visible only in the records of the process. We therefore treat the review records as part of the scientific product. The release includes, alongside code and manifests, the pre-specification ledger with its timestamped freezes and go-after-fixes revisions, the review findings with their gradings and resolutions, the contamination matrix and quarantined outputs of the N16 incident, and the withdrawn certification of N32, retained rather than deleted. The reliability of the certificate chain rests on the process that produced it as much as on the statistics it reports. Only a process whose failures are preserved can be audited end to end.

E ROBUSTNESS RE-EVALUATIONS

This appendix re-evaluates the certified conclusions of the main text under stricter windows, transferred doses, and altered scoring conditions. It archives the re-evaluations referenced in Sections 5.3 and 5.5: two re-analyses of the judged window, a dose-transfer check at the deployment cutoff, the supporting analyses for the two deployment components (map-pyramid augmentation and the 512^2 refiner), and two re-analyses of inferential threats to the primary confirmatory certificate. Each re-evaluation is archived under the same manifest discipline as the confirmatory experiments, and each paragraph ends with a plain statement of its outcome. Every certified conclusion is unchanged, and the single correction on record is an erratum in one reported norm.

The first re-analysis addresses a pre-specified caveat: the doses $w\in\{1, 2\}$ carried to the confirmatory cutoff were selected at the development cutoff, whose judged window (1881–1915) overlaps the confirmatory target window (1911–1930) in the years 1911–1915, a scalar, selection-level channel from evaluation to generation. Re-scoring the development-cutoff verdicts with the judged window trimmed to $\leq$1910 leaves them essentially unchanged under both style judges. SEMANTIC gives -0.0087 at w=1 and -0.0087 at w=2 (full window: -0.0087 / -0.0085). TEXTURE gives -0.0137 $[-0.0185, -0.0094]$ at w=1 and -0.0127 $[-0.0190, -0.0065]$ at w=2 (full window: -0.0137 / -0.0139), both significant. The w=4 arm remains rejected under the trimmed window ($+0.0315$ SEMANTIC, $+0.0509$ TEXTURE). Outcome: the dose selection has no dependence on the overlap years, and all certified conclusions are unchanged.

The second re-analysis restricts the target window itself. The development-cutoff window 1881–1915 spans roughly two era steps. Re-scoring the certified arms against the single era step 1881–1895 under the TEXTURE judge leaves the operating point certified at both doses (paired displacement toward the future -0.0093 at w=1 and -0.0085 at w=2, both significant). Effect magnitudes are not commensurable across teacher constructions. The rolling origin-1880 forecast toward the same window (-0.0025, Table 7) uses a teacher rebuilt from the adjacent pre-origin windows at

Table 10: **Dose-transfer check at the present cutoff** (SEMANTIC judge, n=300 per arm, paired against the free-base control). ✓ marks a 95% interval excluding zero. Novel×valid is the fraction of items simultaneously beyond the novelty floor and inside the real-future envelope (Section 5.5).

Arm	Paired Δ to ≤-cutoff corpus	Novel×valid
StyleCast (w=1, delivered, norm 127.66)	−.0146 ✓	64.0%
StyleCast (w=1, matched, norm 70.34)	−.0100 ✓	61.0%
StyleCast (w=2, matched, norm 70.34)	−.0067 ✓	63.3%
random direction (norm 70.34)	+.0008 n.s.	62.7%
reversed (norm 70.34)	−.0031 n.s.	61.7%
free base (control)	—	60.3%

applied norm 60.8. The development-cutoff arms inject the 30k-step slice-pair teacher at applied norm $0.7 \times 138.7 = 97.1$. The pre-specification accordingly bars cross-construction magnitude comparisons, so only within-protocol contrasts are interpreted. Outcome: the certified displacement is not an artifact of the wider window's later years, and the same window semantics carry over to the rolling-origin protocol of Section 5.5, which scores one era step per origin.

The third re-evaluation is the dose-transfer check at the present cutoff. Two corrections and one question motivated it. First, an erratum: the deployment record initially reported the raw delta norm (182.4) where the applied norm belonged. The applied norm measured at build time is 127.66. Second, the applied dose at deployment had no norm-level justification: the confirmatory cutoff certified an applied norm of approximately 70.35, and the weight-space dose ladder at that cutoff had located a SEMANTIC-judge cliff between applied norms 70 and 113, leaving open whether the delivered dose sat beyond a cliff. The check therefore adds arms norm-matched to the confirmatory certified dose (a parameter-wise rescaling of the delivered direction by 0.5509, applied norm 70.3), together with matched-norm random and reversed directions, all generated through a self-contained controlled chain sharing sources and seeds. With no future window at this cutoff, the distance statistic is manifold retention, negative when the arm sits closer to the ≤-cutoff corpus than its paired control. Table 10 reports the pre-specified primary judge (SEMANTIC). The delivered dose is *closer* to the ≤-cutoff corpus than the paired free base (-0.0146 $[-0.0196, -0.0095]$, plausible-rate +9.7pp). The norm-matched tier is concordant (-0.0100 $[-0.0138, -0.0063]$, +7.3pp). Sign asymmetry holds: the forward direction exceeds its reversal by $+0.0069$ $[+0.0030, +0.0111]$ and an energy-matched random direction by $+0.0109$ $[+0.0048, +0.0174]$. Dose response is linear up to the estimation ceiling ($\cos(70.34, 127.66) = 0.914$ against a split-half ceiling of 0.967 and a random floor of 0.443). The TEXTURE judge records no significant motion for any true arm and flags only the random arm as departing ($+0.0117$ $[+0.0026, +0.0209]$), corroborating the asymmetry. The STRUCTURE judge is not direction-specific at this cutoff and shows a monotone decline of structural novelty with dose ($54.0\% \rightarrow 45.3\% \rightarrow 31.7\%$ from free base to matched to delivered). This decline motivates retaining the norm-matched tier as a documented conservative option rather than as evidence against the delivered dose. Outcome: the weight-space cliff between applied norms 70 and 113 does not transfer to score-space injection gated to $t < \tau$. The delivered configuration stands, with one corrected value (127.66) and a conservative tier on record.

The next two analyses turn from the certificates to the deployment pipeline itself. The fourth analysis documents why map-pyramid augmentation is necessary for free composition. The token-map plane steers structure: the renderer's token-map agreement is 0.137 against a random baseline of 0.0014, and a deliberately mismatched map is followed instead (0.126). However, an 8×8-pooled coarse map is out-of-distribution for a renderer finetuned on full-resolution maps, and free drafts under pooled maps degrade into structureless haze. This failure is recorded twice independently: at the confirmatory cutoff before the pyramid finetune, and again at the development cutoff (whose renderer has no pyramid finetune), where the pooled map had to be replaced by a structural-variation operator to obtain usable drafts. Training the renderer with map-pyramid augmentation (the 250k-step tag-and-pyramid finetune) makes pooled maps an in-distribution input, enabling the free-composition protocol and the distribution-level check of Section 5.5. Outcome: the augmentation is a necessity for the deployment pipeline, not a tuning choice.

The fifth analysis concerns the 512^2 refiner, a +40k-step finetune of the pyramid-augmented renderer at 512^2 (the VQ tokenizer natively yields a 64×64 token grid at this resolution, and all conditioning channels are inherited). Qualitatively, the refiner resolves faces into features, hair

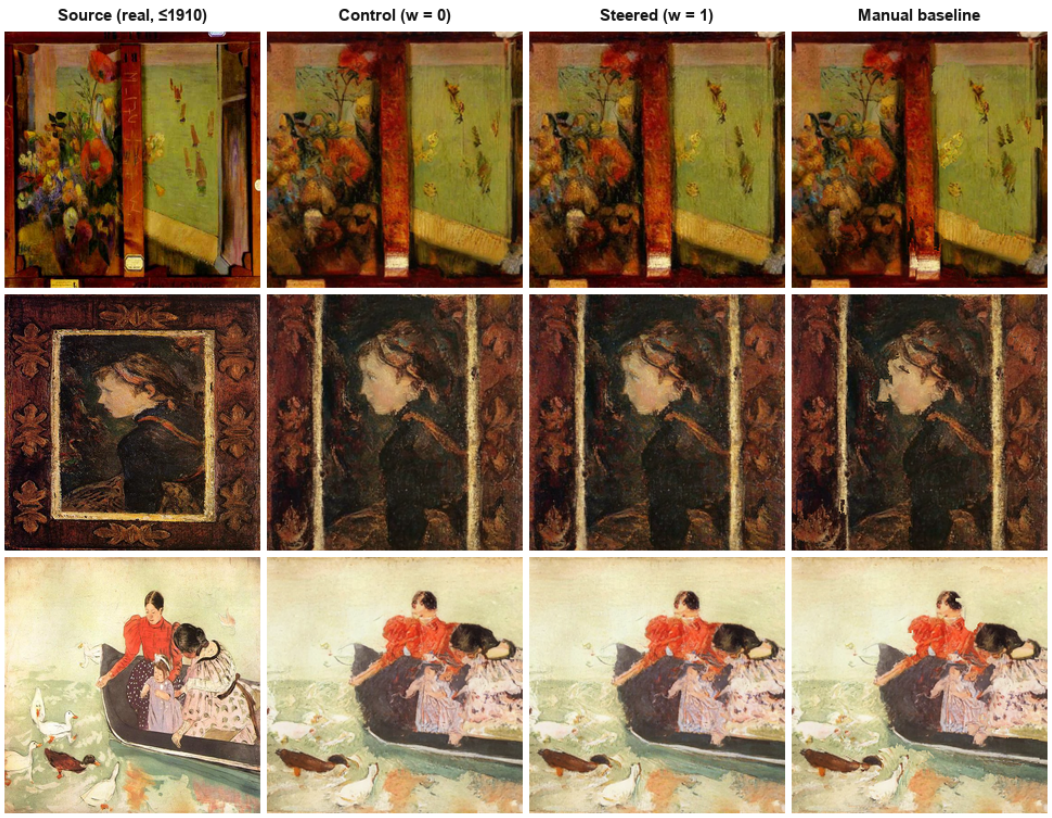


Figure 6: **Paired qualitative results at the confirmatory cutoff (certified samples of Figure 1).** Each row anchors on one real $\leq$1910 source (left): the seed-matched control (w=0), the StyleCast arm (w=1), and the trend-informed manual composite baseline all render the same conditioning. The three generated columns are nearly indistinguishable per image by design.

into strands, and fabric into brushwork, without altering the certified configuration upstream. Quantitatively, the dose-transfer check was repeated at the 512 delivery layer: the delivered and matched tiers reproduce the judging-layer signature (-0.0200 $[-0.0286, -0.0116]$ and -0.0133 $[-0.0194, -0.0070]$ under the SEMANTIC judge), with dose linearity 0.880 against a split-half ceiling of 0.896. Outcome: the gap between the judging layer and the delivery layer is closed. The refiner changes resolution, not conclusions.

The final two re-analyses address two inferential threats to the primary confirmatory certificate. First, because canonized paintings cluster by artist, the paired bootstrap was re-run resampling artist clusters of the source works (570 clusters over 600 sources) in place of items: the intervals are essentially unchanged and all three judges remain significant (TEXTURE -0.0042 $[-0.0070, -0.0015]$; SEMANTIC -0.0046 $[-0.0066, -0.0027]$; STRUCTURE -0.0035 $[-0.0061, -0.0007]$). Second, because works of different eras enter digital corpora through different reproduction pipelines, the certificate was re-scored under the TEXTURE judge after converting generations, controls, and both real pools to grayscale, and again after additionally equalizing each image's luminance histogram: the paired displacement strengthens to -0.0147 $[-0.0173, -0.0122]$ and -0.0146 $[-0.0171, -0.0121]$ respectively, roughly $3.5\times$ the full-color verdict. The same grayscale re-scoring under the SEMANTIC judge strengthens its verdict comparably (-0.0154 $[-0.0173, -0.0134]$ vs. -0.0046). Band-limiting completes the profile: under a Gaussian blur of σ=1, which suppresses reproduction-sharpness differences while preserving coarse facture, the TEXTURE certificate again survives and strengthens (-0.0058 $[-0.0080, -0.0036]$). Halving the resolution ($256 \rightarrow 128 \rightarrow 256$), which removes the fine-texture band outright, eliminates and reverses it ($+0.0071$ $[+0.0045, +0.0096]$). This profile is consistent with the facture-dominated localization of Appendix A: the certified direction requires the fine-texture band to exist, yet strengthens whenever the color, luminance-distribution, and sharpness statistics through which reproduction pipelines dif-

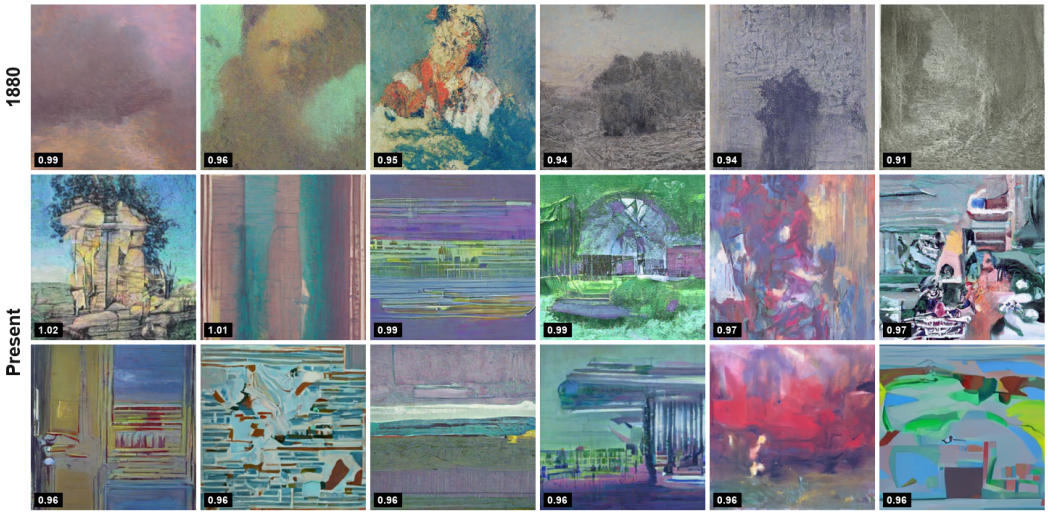


Figure 7: **Further free-evolution chains after six rounds**, extending Figure 5. The top row shows 1880-cutoff chains and the lower two rows present-cutoff chains. Badges follow that figure, and every displayed image clears its cutoff's novelty floor.

fer are suppressed with the band preserved. Acquisition drift in those channels therefore does not drive the certified direction and, if anything, dilutes it.

A final group of re-evaluations probes the certificate's dependence on the held-out target set. Excluding all works dated within two years of the cutoff slightly strengthens the confirmatory verdicts (-0.0057, -0.0054, and -0.0036 under the three judges), so boundary dating noise does not explain them. Resampling the 1911–1930 target corpus itself, rather than the generated pairs, leaves both style-judge certificates intact (-0.0042 and -0.0050). Scored against equal-length fifteen-year windows on either side of the cutoff, the displacement flips sign at the cutoff. It moves significantly away from 1876–1890 and 1891–1905 and significantly toward 1911–1925, 1926–1940, and 1941–1955 under both style judges, and STRUCTURE agrees in four of its five windows. The direction is temporally oriented rather than a generic departure from the source period. Movement toward the remote windows reflects this orientation rather than reach. Proximity is certified only for the next era, and the extension experiments bound the certified reach at one step.

In summary, the certified conclusions of the main text are unchanged under both window re-analyses. The dose-transfer check corrects one recorded value (applied norm 127.66 in place of the raw delta norm 182.4) and adds a documented conservative dose tier. The deployment analyses corroborate the necessity of map-pyramid augmentation and the neutrality of the 512^2 refiner with respect to every certificate.

F ADDITIONAL QUALITATIVE BOARDS

This appendix shows what the certified deployment statistics look like at the level of individual images. It collects three kinds of qualitative boards in support of the deployment claims: paired generation–control boards at the certified doses (Figure 6), further present-cutoff free-composition boards (Figure 7), and nearest-neighbor proof panels for the displayed generations (Figure 8). Every displayed sample passes the anti-imitation certificate of Section 3. The boards illustrate the certified statistics and carry no directional claim of their own.

Figure 6 pairs seed-matched arms on the same conditioning. Its near-identical columns are consistent with a displacement resolvable only at the ensemble level: a certified one-step displacement is 0.5–1.6% of the validity radius depending on the cutoff. Over the paired ensemble, the middle pair differs *toward* the held-out future under all three judges, and the right pair differs *away* under both style judges. The *visible* novelty of the framework appears in the free-composition delivery protocol (Figures 1 and 5).

Figure 7 adds further free-evolution chains at the 1880 and present cutoffs beyond the boards of Figures 1 and 5, illustrating the range of compositions the free-composition protocol produces.

Figure 8 then documents the anti-imitation certificate for the displayed samples, placing each generation beside its conditioning source and its nearest pre-cutoff neighbor so the certified distances can be read directly. The floor is the median nearest-neighbor distance between real pre-cutoff paintings, so no pre-cutoff painting, the source included, lies within the typical distance separating real paintings from one another. Compositional inheritance from the conditioning source is by design.

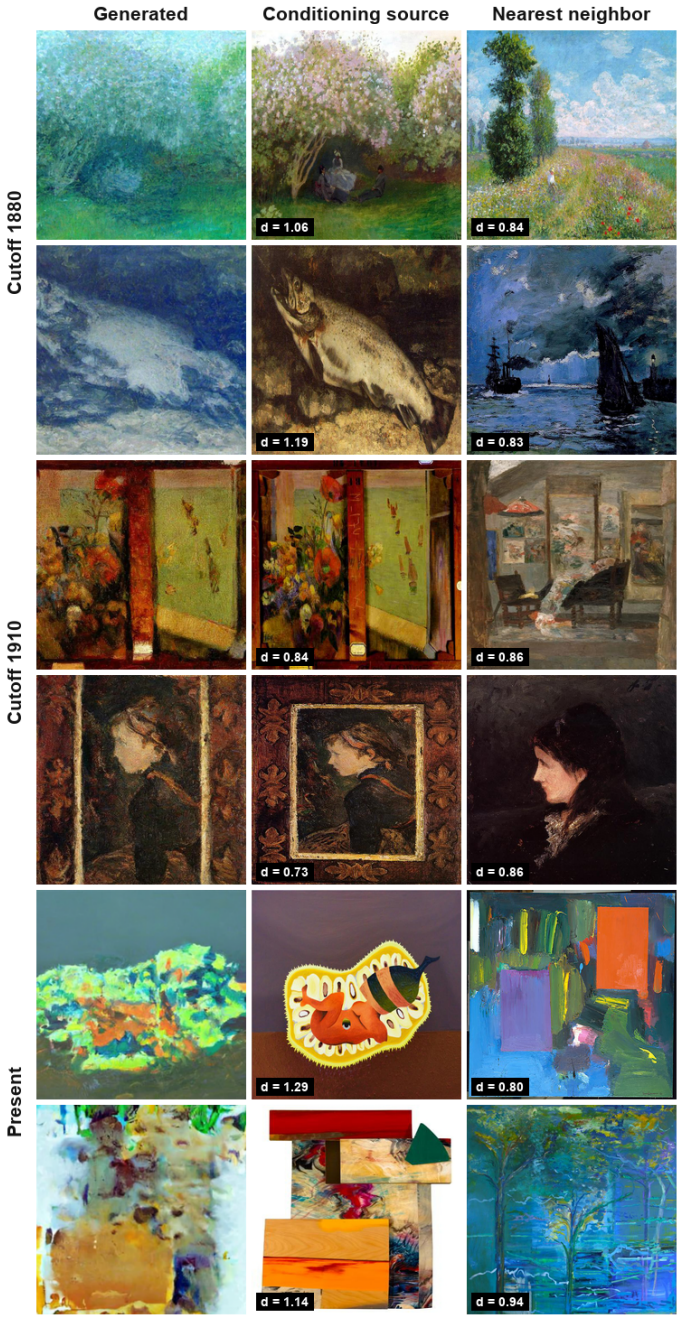


Figure 8: **Nearest-neighbor proof for the displayed generations.** For two samples per cutoff from Figure 1: the generated image (left), the real work supplying its semantic conditioning (middle), and its nearest neighbor in the pre-cutoff corpus under the TEXTURE judge (right), with distances from the generation to each. Every distance exceeds the cutoff's novelty floor (0.705/0.722/0.738).