Controlling Photometric Drift in Diffusion-Generated Video via Neuron-Level Bias

Ajitesh Bankula[1]

Rensselaer Polytechnic Institute, 110 8th St, Troy, NY 12180, USA
bankua@rpi.edu

Abstract. Color continuity is central to cinematography: small shifts in saturation, contrast, or lightness can break immersion and inflate grading time. Diffusion-based video generators accelerate ideation but often exhibit, *photometric drift*[1], undesired frame-to-frame changes in color statistics even when content is fixed. We quantify drift using eight appearance descriptors across multiple seeds and shots, observing up to **3%** cumulative deviation and **1.3%** run-to-run *coefficient of variation (CV)*[2] within 16 frames. We considered hard constraints such as *freeze*[3] and *pin*[4], but argue they are fragile in practice (ghosting, over-constrained dynamics, VRAM sensitivity) and not photometric-specific. We instead propose a tiny inference-only intervention: a *four-channel neuron-level bias*[5] derived from *activation traces*[6] and injected late in the *decoder*[7]. The bias reduces saturation drift by **74%** and contrast by **14%**, while increasing colourfulness and luminosity drift by **19%** and **709%**. The design should work with any model; we just tested it on a standard image-conditioned diffusion pipeline to show it works. We discuss trade-offs versus manual fixing, hard constraints, and retraining.

Resource website: `https://ajiteshbankulaa.github.io/BiasedSDVideoGeneration/`

Keywords: Video diffusion · Color continuity · Photometric drift · Cinematography

[1] Systematic frame-to-frame change in colour/tonal statistics (e.g., saturation, contrast, lightness) even when scene content is fixed.

[2] Standard deviation divided by the mean; here computed per frame across seeds to quantify run-to-run spread.

[3] Lock a layer's output to its frame-0 value for all subsequent frames.

[4] Blend features or pixels toward frame 0 with weight $\alpha \in [0, 1]$.

[5] A constant vector of four additive offsets applied to selected decoder channels at inference to nudge photometrics.

[6] Per-frame statistics (here, spatial means) of selected neural channels across time.

[7] The final stage that projects latent features to image space (e.g., UNet conv_out + upscaler).

1 Introduction

1.1 Why continuity of appearance matters

Color is a crucial narrative instrument: it primes expectations, steers attention, and shows intent in a film, a look that persists across a scene. Production teams therefore target *continuity of appearance* stable statistics such as saturation, contrast, and lightness—so that editorial cuts feel invisible and emotional beats land with precision [8, 2]. When continuity slips, audience experience and immersion are broken; as such, in post production, colorists typically spend many hours matching shots and building the intended look [7].

1.2 Diffusion video and the continuity gap

Text-to-video diffusion systems now produce striking imagery and are entering previsualization and idea creation workflows. Yet they are inherently stochastic: successive frames sampled from the same conditioning signal often wander. This means that saturation can creep upward, lightness can trend darker, and contrast can transiently flatten before recovering. This *appearance drift*[8] complicates downstream color grading and undermines repeatability across runs.

1.3 Problem statement and desiderata

We address two questions: **(Q1)** How large is photometric drift in typical diffusion pipelines when the scene content is fixed? **(Q2)** Can we reduce drift without retraining and with negligible runtime overhead? For practical adoption, the solution should be: (i) *not dependent* on the model; (ii) *controllable* (expose a knob like "keep lightness steady"); (iii) *composable* with standard grading; and (iv) *transparent*, showing what is being changed.

1.4 Terminology and definitions (quick reference)

We use the following terms consistently:

- **Photometrics / photometric statistics**: frame-level numeric descriptors of color and tone (e.g., saturation, lightness, contrast, colorfulness, hue entropy).
- **Photometric (appearance) drift**: systematic change of photometrics across frames with fixed content; we quantify with $\Delta m_t = (m_t - m_0)/m_0$.
- **Baseline run**: the short reference generation used to compute activation traces and fit a bias.
- **Decoder**: the last feature-to-image projection stage (e.g., UNet conv_out and upsampling).
- **Activation trace**: per-frame summary statistic (spatial mean) of a channel over time.

[8] We use "appearance drift" synonymously with photometric drift; the focus is color/tonal statistics, not geometry.

- **Neuron-level bias**: constant additive offsets applied to selected decoder channels at inference.
- **Freeze / Pin**: hard constraints; freeze copies frame-0 activations; pin blends features or pixels toward frame 0 with weight α.
- **CV (coefficient of variation)**: CV $= \sigma/\mu$, used per frame across seeds to quantify run-to-run spread.
- **Exposure**: average lightness; we use HLS lightness as a proxy.

1.5 Approach and contributions

We propose a minimal intervention: where we use a four-scalar neuron-level bias from one short baseline run and inject it late in the decoder at inference time. The bias nudges selected channels whose temporal activity correlates with drift in target metrics. Our contributions are:

1. **Protocol & reference suite** for measuring photometric drift using eight descriptors (Sec. 3).
2. **Quantification** of frame-wise drift and run-to-run variability with concrete plots and statistics (Sec. 4).
3. **Neuron-level bias** that reduces targeted drift with near-zero compute and no training (Secs. 3–5).
4. **Comparisons and trade-offs** with manual grading, freeze/pin, and retraining (Sec. 5).

The design is independent of any particular model; implementation details only appear in Sec. 3 to demonstrate feasibility.

2 Literature Review

2.1 Colour control in film practice

Color has evolved from photochemical timing to digital intermediate and node-based grading. Modern workflows mix scene-referred transforms, LUTs, and targeted corrections. Across decades, the main constant though is maintaining a *consistent appearance* across shots to preserve continuity and emotional coherence [7]. Even small deviations in saturation or exposure can mess up story beats and increase revision cycles [2]. Further research shows clearly that the control of color statistics can modulate affect and attention [1].

2.2 Video diffusion: fidelity vs. continuity

There has been rapid progression in video diffusion quality, conditioning methods, and controllability [9, 3]. Coherence is an active area: synchronized sampling and better context packing improve content stability across frames [5, 10]. However, *photometric continuity* is typically treated qualitatively or delegated to post-production. Our work focuses precisely on quantifying and controlling these photometric statistics.

2.3 Photometric descriptors and measurement

Classical image processing offers interpretable descriptors: *RMS contrast*[9], mean saturation in HSV/HLS, perceptual *colorfulness (Hasler–Süsstrunk)*[10], sharpness via Laplacian variance, hue-entropy for palette spread, and percentile spans for dynamic range. These descriptors are cheap, continuous, and map to colorist vocabulary, making them useful for diagnosing drift.

2.4 Activation-level interventions

Intervening in neural activations e.g., channel-wise nudges or steering along concept directions—offers direct control without full fine-tuning [11, 6]. For diffusion models, inference time hooks are quite attractive: they avoid weight storage, minimize latency, and preserve most of the model. We adopt this spirit, targeting late-decoder channels as a practical point for photometric control.

2.5 Positioning relative to alternatives

Manual grading is precise but labor-intensive and must be repeated for each new sample. Freezing and pinning constraints are simple but can suppress legitimate dynamics or introduce ghosting(faint visual artifacts from prior frames). Retraining/fine-tuning may internalize continuity but is compute- and data-heavy and can be brittle across scenes. A tiny inference-time bias-based approach, though, offers a complementary point in this design space. That could help with the issues of the other methods.

3 Methods and Data

3.1 Reference data

We curate ten color-rich frames from *The Grand Budapest Hotel*. Each still is center-cropped to 500×377 and resized to 1024×576 (16:9) to match a standard working resolution. For each still we render five 16-frame clips with different seeds, yielding 35 clips (560 frames) plus the seven originals.

3.2 Photometric descriptors

We compute eight descriptors per frame; all are real-valued, continuous, and cheap to evaluate:

- **Saturation** (S): mean of the HSV/HLS saturation channel in $[0, 1]$.
- **Luminosity** (L): mean of HLS lightness in $[0, 1]$ (proxy for exposure).

[9] Root-mean-square (standard deviation) of grayscale luminance.

[10] A perceptual measure combining the standard deviations and means of $rg = R - G$ and $yb = \frac{1}{2}(R + G) - B$; see [4].

- **Contrast** (C): *RMS contrast* on grayscale luminance Y normalized to $[0, 1]$.
- **Colourfulness** ($\mathcal{C}$): Hasler–Süsstrunk [4].
- **Hue entropy** (H_e): entropy of the hue histogram (36 bins), normalized by $\log_2 36$.
- **Warmth ratio** (W_r): warm-band counts (15°–35°) over teal-band (90°–110°).
- **Sharpness** (S_h): variance of the Laplacian of Y.
- **Dynamic range** (D): percentile span $P_{95}(Y) - P_5(Y)$.

We analyze two types of variability: *Drift* relative to frame 0, $\Delta m_t = (m_t - m_0)/m_0$, and *run-to-run variability*, measured as CV per frame across seeds.

3.3 Design: neuron-level bias

Let $a_c(t)$ be the spatial mean of decoder channel c at frame t.

(1) Channel selection. Pick K=4 channels with largest temporal variance $\mathrm{Var}_t[a_c(t)]$ (moving channels are better levers).

(2) Sensitivity fit. Build $X \in \mathbb{R}^{T\times K}$ from their traces and fit, for each target metric $m \in \{\text{saturation, contrast, colourfulness}\}$,

$$y_m = X\beta_m, \quad y_m(t) = \Delta m_t.$$

(3) Bias construction. To counter combined drift use

$$b = -(\beta_{\text{sat}} + \beta_{\text{ctr}} + \beta_{\text{col}}) \odot \sigma_X,$$

where σ_X is the column-wise stdev of X (unit-matching).

(4) Injection. At inference, add b to the selected channels after the final convolution and before RGB upscaling for all diffusion steps $t > 0$.

Complexity. One short baseline pass to log $a_c(t)$ and metrics; at generation time add only four scalars per step. No training or stored weights.

3.4 Other Low-level constraints considered: freeze and pin (why we did not adopt them)

Two straightforward, low-level constraints can be applied to diffusion pipelines: **Freeze** copies a target layer's activations from frame 0 to all subsequent frames. **Pin** blends either *in feature space* (decoder output) or *in pixel space* toward frame 0:

$$\text{out}_t \leftarrow (1-\alpha)\cdot \text{out}_t + \alpha \cdot \text{frame0} \quad (\alpha \in [0, 1]).$$

Why these are fragile in practice.

1. **Not photometric-specific.** Neither freeze nor pin targets saturation, contrast, or lightness directly; both indiscriminately damp *all* variation in the chosen tensor or pixels towards the starting frame. As a result, they can stabilize exposure while simultaneously suppressing desirable micro-contrast or color nuance.
2. **Suppress legitimate dynamics.** Freezing late-decoder features forces a stationary appearance even when subtle temporal changes are part of the intended motion (e.g., specular flicker), producing a "stuck" look.
3. **Ghosting and smear.** Pixel-space pin blends toward frame 0 regardless of motion, leading to ghost trails when objects move or when the camera reframes.
4. **VRAM sensitivity.** Using the pin method at the decoder often increases activation lifetimes and copied amounts; on lower end commodity GPUs this can trigger OOMs or have the downside of a slow offload, which heavily harms its usability.
5. **Global coupling side-effects.** Because the intervention is broad, adjusting α to fix lightness frequently modifies saturation/contrast as collateral changes; there is no per-metric knob.

Rationale for our choice. These caveats motivated a *photometric-aware* alternative: a tiny neuron-level bias fit from activation traces to target specific drift metrics. It preserves legitimate dynamics, exposes a controllable knob (Sec. 4–5), and keeps runtime costs negligible.

4 Results

4.1 How much drift occurs?

Figure 1 reports Δm_t averaged over seeds and stills. **Colourfulness** rises steadily (+2.8% by frame 15), indicating increasing chroma spread. **Saturation** shows a similar but smaller monotonic increase (+2.1%). **Contrast** tends to dip in the first few frames (likely denoising transients) and then recover. **Luminosity** trends slightly negative (mild darkening). Hue entropy (not shown) rises $\sim 1.4\%$, implying a broader hue distribution.

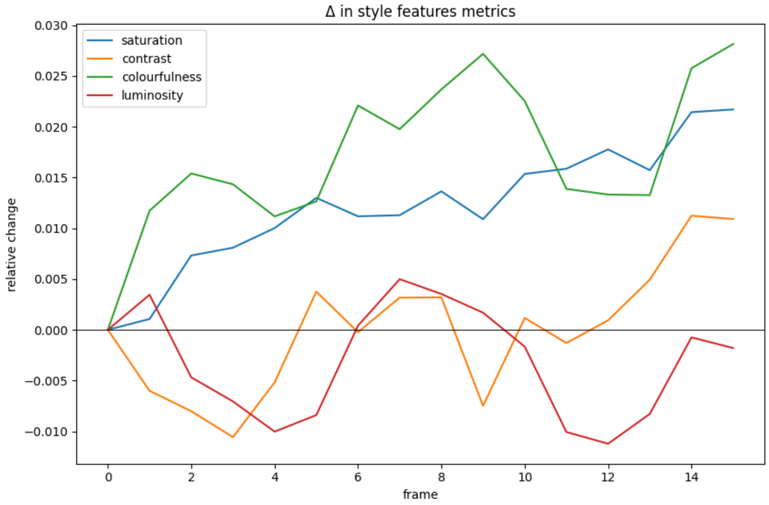

Fig. 1. Frame-wise photometric drift (mean of 5 seeds, 10 shots). Positive values indicate an increase relative to frame 0.

4.2 How variable are different runs?

Run-to-run variability (Fig. 2) is highest in early frames for most metrics (CV up to 1.3% in colourfulness), then decays as the process stabilizes toward the conditioned scene. Luminosity generally exhibits lower CV, suggesting exposure is less seed-sensitive than chroma attributes.

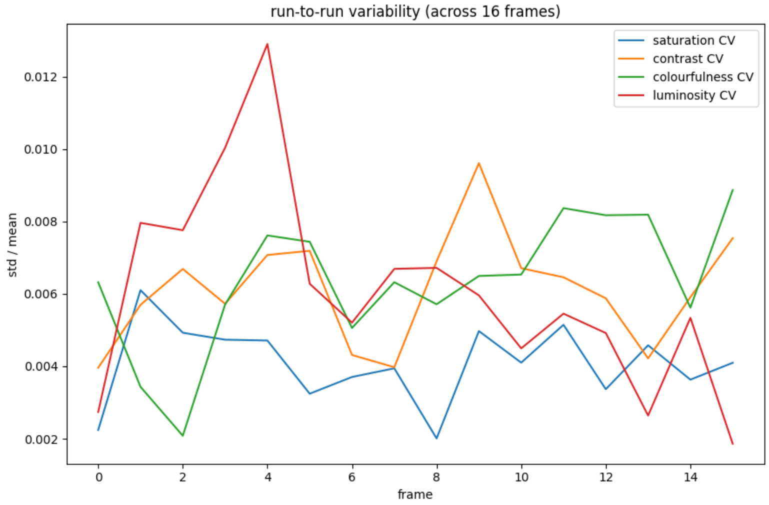

Fig. 2. Run-to-run variability. Coefficient of variation (CV) per frame across five seeds (shaded: $\pm 1\sigma$).

4.3 What does the neuron bias change?

Table 1 summarizes absolute drift between frames 0 and 15 for four primary metrics. The bias *reduces* saturation by 74% and contrast by 14% relative to baseline, while *increasing* colourfulness and the magnitude of luminosity drift. Qualitatively, clips retain their intent yet look slightly more vivid and darker.

Table 1. Absolute drift at frame 15 (15 − 0). $\Delta = (\text{Bias} - \text{Baseline})/|\text{Baseline}|$.

Metric	Baseline	+Bias	Δ (%)
Saturation	0.01559	0.00400	−**74**
Contrast	0.00230	0.00198	−**14**
Colourfulness	2.14702	2.54642	+19
Luminosity	−0.00055	−0.00445	+709

4.4 Is there a usable control knob?

We sweep the bias scale $s \in [-1, 1]$ and measure mean lightness. Figure 3 shows a smooth, monotonic response, indicating the bias can serve as a practical *exposure knob* without retraining.

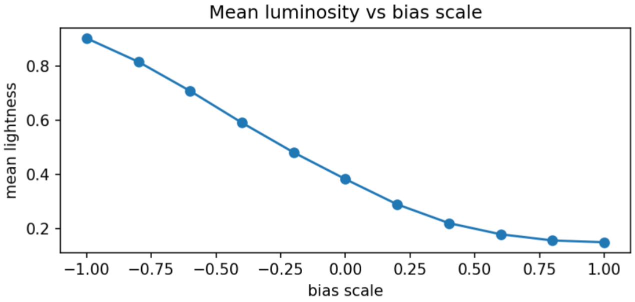

Fig. 3. Bias strength vs. mean lightness. A monotonic relationship suggests a usable control knob.

5 Evaluation

5.1 Comparing against alternatives

Table 2 positions our method against common alternatives. Manual grading is precise but costly; freeze/pin offers strong continuity at the risk of over-constraining dynamics or adding ghosting; retraining promises broader improvements but at significant data/compute cost. Our bias introduces a minimal, composable knob with small side-effects (e.g., lightness).

Table 2. Qualitative comparison of continuity strategies.

Method	Training	Runtime Overhead	Continuity Strength	Typical Side-Effects
Manual grading	No	Per shot/clip	High (post)	Time cost, redo per run
Freeze	No	Low	Very high	Suppressed dynamics
Pin (pixel/feature)	No	Low–Med	High	Ghosting/VRAM usage
Retraining/fine-tune	Yes	High (train)	Potentially high	Data/compute cost
Neuron bias (ours)	No	**Very low**	Medium → High	Color/brightness coupling

5.2 Deployment and protocol

Two practical checks for deployment: **(i) Seeds/early frames:** recompute drift for new seeds; report distribution of Δm_t at $t \in \{1, 2, 3, 15\}$. **(ii) Channel count** K**:** test $K \in \{2, 4, 8\}$; we observed best stability and least coupling at K=4. **(iii) Selection criterion:** variance vs. correlation-to-target; variance was more robust to short probes. **(iv) Bias scale:** use Fig. 3 to pick an exposure target, then refit per scene.

5.3 Practical integration

Compute the bias on a *look shot* that defines the desired appearance; then reuse it across takes with different seeds.

6 Discussion

6.1 Why is this important for filmmakers?

The bias-based approach provides a *low-cost continuity lever*. It is fast, requires no weight edits, and coexists with grading. This allows directors to stabilize saturation and contrast early in the pipeline, while preserving flexibility for creative changes later. Further, the method exposes interpretable diagnostics (metrics over time) for look matching.

6.2 Failure and mitigation

Because the bias is linear and shared across frames, it may over correct in scenes with deliberate photometric evolution. Metric coupling (e.g., colourfulness vs. lightness) can yield darker yet more vivid frames. Mitigations include per-metric biases, a soft pixel pin, or dialing the scale via the sweep.

6.3 Limitations

Scope and data. Findings are based on ten color-rich frames from a single film and short 16-frame clips with image-conditioned generation. This emphasizes appearance statistics under relatively static content. Generalization to text-only prompts, multi-shot sequences, long takes, or strongly dynamic scenes remains to be tested.

Measurement validity. All descriptors are computed in display-referred sRGB with 8-bit conversions (OpenCV HSV/HLS/Lab), not in scene-referred linear space. *Lightness*[11] and *colourfulness*[12] are convenient but imperfect stand-ins for perceptual judgments; small gamut/clipping differences or gamma can bias values. Global frame statistics ignore local regions (e.g., skin tones) and do not capture temporal frequency artifacts (flicker). Reported percentages can inflate when the baseline magnitude is near zero (e.g., luminosity's +709% corresponds to an absolute change of ≈ 0.0039).

External validity and compute. Results were obtained under specific sampling schedules (frames, steps, FPS) and commodity-GPU memory settings. Different schedulers, guidance scales, resolutions, or decode chunking can change drift profiles and the bias's effect size.

[11] HLS L channel used as an exposure proxy

[12] Hasler–Süsstrunk measure

Human perception and workflow fit. We did not conduct user studies with colourists. Metric improvements do not guarantee perceived continuity improvements in editorial context; integration with standard color-management (e.g., ACES, scene-referred grading) and look-development pipelines is future work.

7 Conclusions and Future Work

We quantified photometric drift in diffusion-generated video and introduced a minimal, neuron-level bias that suppresses targeted drift with negligible overhead. The approach is model-agnostic, composable with colour grading, and offers a practical control knob. Future directions: (i) per-metric biases fitted jointly to decouple lightness from chroma, (ii) lightweight nonlinear controllers, (iii) automatic target-setting from a colour reference, and (iv) perceptual validation with professional colourists.

Acknowledgements

I thank my peers(Momir Petrovic and Zihao Zhan) and the staff (Prof. Neha Keshan and TA Nipun Pathirage) for feedback on continuity, metrics, and writing. Thank you.

References

1. Cao, Z., et al.: Exploring the combined impact of color and editing on emotional perception in authentic films. Humanities and Social Sciences Communications **11**, 1349 (2024)
2. C&I Studios: Why color consistency is crucial for cinematic storytelling (2025), accessed 2025
3. Ehtesham, A., et al.: Moviegen: Swot analysis of meta's generative AI foundation model. In: IEEE CCWC (2025)
4. Hasler, D., Süsstrunk, S.: Measuring colorfulness in natural images. In: Proc. SPIE Human Vision and Electronic Imaging. vol. 5007, pp. 87–95 (2003)
5. Kim, S., et al.: Tuning-free multi-event long video generation via synchronized coupled sampling (2025)
6. Liu, Z., et al.: Cones: Concept neurons in diffusion models for customized generation (2023)
7. Pixflow: From black-and-white to technicolor dreams: A history of color grading in cinema (2025), accessed 2025
8. Videomaker: Continuity editing – the importance of consistency in cinema (2023), accessed 2025
9. Xing, Z., et al.: A survey on video diffusion models. ACM Computing Surveys (2025), to appear
10. Zhang, L., Agrawala, M.: Packing input frame context in next-frame prediction models (2025)
11. Zhou, B., et al.: Interpreting deep visual representations via network dissection. IEEE TPAMI **41**(9), 2131–2145 (2019)

Appendix A: Metric Definitions and Implementation Notes

Saturation & Luminosity. OpenCV conversions to HSV/HLS; average S and L scaled to $[0, 1]$.

Contrast. RMS contrast on luminance: $\mathrm{std}(Y)$.

Colourfulness. Hasler–Süsstrunk [4]; see Sec. 3.

Hue entropy. 36-bin histogram, Laplace-smoothed, entropy normalized by $\log_2 36$.

Warmth ratio. Ratio of counts in $15° - 35°$ to $90° - 110°$.

Sharpness. Variance of the Laplacian.

Dynamic range. $P_{95}(Y) - P_5(Y)$.

Drift and CV. Drift Δm_t and CV as defined in Sec. 3.

Appendix B: Bias Computation and Injection (Pseudo-Algorithm)

Input: baseline clip (frames $0..T-1$), per-frame metrics m_t, decoder channel means $a_c(t)$.
Select channels: top K by $\mathrm{Var}_t[a_c(t)]$.
Fit sensitivities: build $X \in \mathbb{R}^{T \times K}$; fit $y_m = X\beta_m$ for $m \in \{\text{sat, ctr, col}\}$.
Bias: $b = -(\beta_{\text{sat}} + \beta_{\text{ctr}} + \beta_{\text{col}}) \odot \sigma_X$.
Inject at inference: for steps $t > 0$, add b to those channels after the final convolution.
Render: generate the clip with identical settings and a held-out seed.