BROADCAST, REINFORCE, REPEAT: ATTENTION DYNAMICS BEHIND REPETITION COLLAPSE IN DIFFUSION LANGUAGE MODEL FORECASTING

Anonymous authors
Paper under double-blind review

ABSTRACT

Repetition is a persistent failure mode of language generation, yet its mechanisms in diffusion language models (DLMs) remain poorly understood because bidirectional attention and iterative unmasking introduce dynamics absent from causal decoding. In this work, we study repetition collapse in diffusion-based time-series forecasting, a setting in which predictions can progressively lose temporal variation and converge to pathological numerical repetitions. Through systematic analysis across diffusion steps and Transformer depth, we identify a recurring attention geometry consisting of shared vertical stripes (anchor attention), shifted-diagonal predecessor feedback, and residual structure. We formalize these observations with an idealized reduced dynamical model. Rather than predicting monotonic collapse at every layer, the model shows that positive cumulative anchor–background gain across depth drives anchor concentration and, under appropriate alignment conditions, promotes cross-position representation homogenization. To test this mechanism interventionally, we construct two complementary training-free perturbations that act through distinct computational pathways: attention-level spectral SVD truncation and history-sparsified classifier-free guidance. Both interventions reduce the measured cumulative gain in the predicted direction and substantially suppress repetition while preserving forecasting accuracy. Together, these results connect observed attention geometry, cumulative collapse dynamics, and intervention-based validation into a coherent mechanistic account of repetition collapse in diffusion-based time-series forecasting.

1 INTRODUCTION

While repetitive degeneration is well-documented in autoregressive (AR) models (Holtzman et al., 2020; Welleck et al., 2020), diffusion language models (DLMs) (Lou et al., 2024; Nie et al., 2025) reconstruct sequences via bidirectional attention across unresolved positions, creating fundamentally different dynamics. Predictions made at one denoising step alter the shared context, causing dominant candidates to amplify over steps and layers—a phenomenon we study as *repetition collapse* in digit-wise time-series forecasting (Gruver et al., 2023; Pei et al., 2025). Unlike AR repetition driven by causal prefixes, DLM repetition evolves along two interacting axes: the *diffusion step* (revealing newly unmasked context) and the *Transformer layer* (redistributing representations within a forward pass). Understanding this failure requires modeling how these two axes couple unresolved tokens before final commitment.

The collapse develops progressively: candidate predictions homogenize while forecasting error has largely plateaued, temporally aligned with the scheduler's slot-wise reveal phases. Internally, repetitive trajectories consistently exhibit a structured attention geometry combining shared vertical stripes (queries attending to high-influence *anchor* keys) and shifted-predecessor feedback (queries attending to their serialized predecessor). We model this geometry as $A^{(\ell)} \approx \mathbf{1}c^{(\ell)\top} + \gamma_\ell P + E^{(\ell)}$, where the first term represents shared anchor broadcasting, the second captures predecessor feedback, and $E^{(\ell)}$ collects remaining interactions.

Together, these components produce a Broadcast–Reinforce–Repeat process. Anchor attention first broadcasts a common digit preference across positions. The resulting hidden-state updates generate

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.

an anchor–background gain that reinforces this preference over Transformer depth. Gain accumulation determines the overall amplification. Positive cumulative gain increases anchor concentration and promotes cross-position representation alignment. Empirically, the measured gain closely tracks changes in anchor-mass log-odds ($r \approx 0.838$ in LLaDA), and repeating trajectories accumulate more gain than non-repeating trajectories. Once high-confidence digits are committed, they become part of the context for the next diffusion step, extending the within-pass amplification into an across-step feedback process. We further examine this mechanism through two training-free interventions that act at different stages of generation. Spectral SVD truncation modifies attention within a forward pass, while history-sparsified classifier-free guidance (HSCFG) changes the token history carried into subsequent diffusion steps. Across these distinct pathways, both interventions reduce cumulative gain together with repetition while preserving forecasting accuracy. This shared pattern provides further support for the proposed feedback mechanism.

Our contributions are three-fold:

- **Empirical Characterization & Geometry.** We characterize progressive repetition collapse in digit-wise diffusion forecasting and identify a recurring shared-anchor and predecessor-attention geometry ($\mathbf{1}c^\top + \gamma P$).
- **Cumulative Dynamics & Empirical Evidence.** We derive a finite-depth model connecting cumulative gain to anchor amplification across depth and show that the measured gain closely tracks changes in anchor-mass log-odds.
- **Intervention-Based Evidence.** We further examine the mechanism using pass-level SVD truncation and step-level history-sparsified guidance. Both reduce cumulative gain together with repetition while preserving forecasting accuracy, providing complementary support for the proposed feedback process.

2 Related Work

Repetition in Autoregressive and Diffusion Language Models. Repetitive degeneration has been extensively studied in autoregressive generation, with explanations involving decoding behavior, representation anisotropy, prefix-level self-reinforcement, and repetition-related internal features or attention circuits (Holtzman et al., 2020; Welleck et al., 2020; Fu et al., 2021; Xu et al., 2022; Su et al., 2022; Hiraoka & Inui, 2025; Yao et al., 2025; McDougall et al., 2024). Related failures also arise in non-autoregressive generation (Ran et al., 2020; Cheng & Zhang, 2022). More recently, repetition in diffusion models has been attributed to disrupted context flow under KV caching, mask-prior drift, and self-conditioning attractors (Zhao et al., 2026; Hong et al., 2026; Zhang et al., 2026). In contrast, we study uncached masked-DLM forecasting and focus on how shared high-attention columns interact with predecessor-aligned attention during iterative unmasking.

Attention Geometry, Representation Collapse, and Spectral Intervention. Self-attention can induce representation homogenization through rank collapse and clustering dynamics, motivating spectral analyses and singular-value interventions (Dong et al., 2021; Geshkovski et al., 2023; Saada et al., 2024; Yan et al., 2022). Recent work also identifies dynamically shifting attention sinks in diffusion language models (Rulli et al., 2026). Our analysis differs by modeling the joint effect of shared query-independent attention and shifted-predecessor feedback through the structured form $A \approx \mathbf{1}c^\top + \gamma P + E$, and by applying spectral truncation directly to the attention operator to perturb the feedback structure implicated by this decomposition.

Masked Diffusion Forecasting and Guidance. Masked diffusion models generate through iterative denoising, with performance depending strongly on token ordering and unmasking policies (Austin et al., 2021; Lou et al., 2024; Sahoo et al., 2024; Nie et al., 2025; Kim et al., 2025; Jazbec et al., 2026). For time-series forecasting, numerical trajectories can be serialized into token sequences (Gruver et al., 2023), and LEAF (Pei et al., 2025) adapts pretrained diffusion language models to zero-shot forecasting through iterative unmasking and classifier-free guidance. Separately, classifier-free guidance and related contrastive decoding methods steer generation by comparing predictions under different conditioning states (Ho & Salimans, 2022; Schiff et al., 2025; He et al., 2026; Rojas et al., 2026; Li et al., 2025; Shi et al., 2024). We build on this forecasting setting but use guidance as a mechanism-targeted intervention: the sparse-history branch perturbs

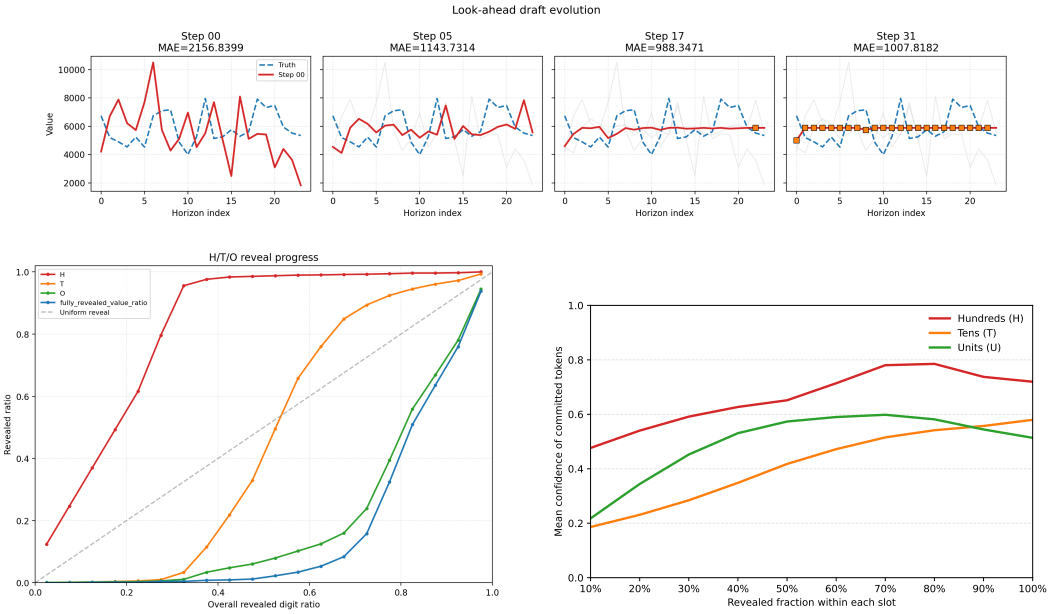

Figure 1: **Progressive emergence of repetition during iterative decoding.** (a) Representative intermediate forecasts lose temporal variation before decoding completes. (b) Unmasking proceeds asynchronously across digit slots, with hundreds resolved first, followed by tens and units. (c) Confidence rises rapidly within the successive hundreds, tens, and units reveal phases, producing stage-wise confidence explosion. Formal definitions, scheduler controls, unresolved-position collision, and additional trajectories are provided in Appendix A.

the revealed-token trajectory while retaining the current decoding state, allowing us to test whether weakening the predicted feedback dynamics also reduces repetition.

3 Problem Setup and Empirical Characterization

3.1 Problem Setup and Generation Dynamics

Masked Diffusion Forecasting. Given history $\mathbf{y}^{\text{hist}} = (y_{-P+1}, \ldots, y_0)$, we forecast F future values. Observations and targets are serialized into a token sequence $x_t = (x^{\text{hist}}, x_t^{\text{fut}}) \in (\mathcal{V} \cup \{[\texttt{MASK}]\})^N$, where x^{hist} is observed and x_t^{fut} is unmasked across iterations $t \in \{0, \ldots, T\}$. At step t, let $\mathcal{M}_t$ denote masked forecast positions. The model produces logits $z_{t,i}$ and predictions $\pi_{t,i} = \text{softmax}(z_{t,i})$ for $i \in \mathcal{M}_t$. A scheduler selects a subset $\mathcal{U}_t \subseteq \mathcal{M}_t$ to reveal, updating $\mathcal{M}_{t+1} = \mathcal{M}_t \setminus \mathcal{U}_t$. We group digit positions by numerical slot $s \in \mathcal{S} = \{\text{hundreds}, \text{tens}, \text{units}\}$ via $\mathcal{I}_s = \{i \in \mathcal{I}_{\text{dig}} : \text{slot}(i) = s\}$. Throughout the paper, t indexes diffusion iterations, whereas ℓ and r index Transformer layers and attention heads, respectively.

Attention Dynamics. At iteration t, the Transformer computes layer-ℓ hidden states $\mathbf{H}_t^{(\ell)} \in \mathbb{R}^{N \times d}$. For head r, $A_{t,r}^{(\ell)} \in \mathbb{R}^{N \times N}$ denotes the standard row-softmax attention matrix. The iteration index t captures token commitment dynamics across diffusion steps, while layer ℓ captures spatial redistribution within a forward pass. Indices t, ℓ, r are omitted when fixed in context.

Experimental setup. We evaluate LLaDA-8B (Nie et al., 2025) and Dream-7B (Ye et al., 2025) on the Monash Time Series Forecasting Archive (Godahewa et al., 2021).Section 3.2 aggregates decoding behavior across both models; subsequent mechanistic results use LLaDA, with Dream replications in Appendix F. Following prior language-model forecasting protocols(Gruver et al., 2023), and LEAF (Pei et al., 2025), we map the bulk of each series into a common numerical range before serialization. The resulting hundreds, tens, and units positions therefore arise naturally from

the numerical representation. Full model, decoding, data, and analysis settings are provided in Appendix G.

3.2 Progressive Repetition During Iterative Unmasking

Repetition emerges progressively. Repetition emerges well before decoding completes. During the first few unmasking steps, the look-ahead forecast remains heterogeneous across the prediction horizon. As decoding proceeds, however, repeated candidates accumulate progressively: the look-ahead repetition rate increases from 0.045 near the beginning of decoding to 0.326 when approximately 27.5% of forecast values have been fully revealed, and reaches 0.497 near completion. In contrast, Draft NMAE improves primarily during the early stage, decreasing from 1.038 to approximately 1.010, and subsequently remains nearly unchanged. Thus, the later growth of repetition is not accompanied by a corresponding improvement in predictive accuracy; instead, the remaining decoding steps increasingly consolidate an already homogenized forecast. Metric definitions and more detailed data are provided in Appendix A.2.

Collapse follows the slotwise unmasking order. Across decoding trajectories, consecutive unmasking steps predominantly target tokens within the same numerical-order group, with 78.7% of consecutive reveal pairs remaining localized within that group. Decoding starts almost exclusively from the highest-order positions, which are selected first in 98.8% of rollouts. Unmasking then proceeds through ordered phases, moving from higher-order positions to intermediate- and lower-order positions over the course of decoding. Repetition follows the same progression, cascading sequentially across these groups and taking hold one position at a time. Detailed reveal statistics and phase-transition tables are provided in Appendix A.3.

Slotwise convergence is accompanied by confidence explosion. Figure 1c shows the confidence of tokens committed during the hundreds, tens, and units phases. Within each phase, confidence rises rapidly as additional positions from the same slot family are revealed. Consequently, later same-slot decisions are made substantially more decisively than the initial commitments. We refer to this rapid stage-wise amplification as *confidence explosion*. This upward trend emerges directly alongside candidate homogenization, active long before the repeated forecast stabilizes.

These results show that repetition develops gradually during decoding and follows the slot-wise reveal order. As more positions from the same slot are committed, their predictions become increasingly similar and more confident.

3.3 Structured Attention Geometry and Empirical Adequacy

The progressive output-space homogenization in Section 3.2 motivates us to examine how information is redistributed within each Transformer forward pass. Across repetition-prone trajectories, attention maps exhibit two recurring structures: shared vertical stripes, where multiple queries attend to common high-influence key columns, and a shifted diagonal following the preceding-token relation induced by the serialized forecast.

We summarize this geometry with the constrained decomposition

$$A^{(\ell)} = \mathbf{1}{c^{(\ell)}}^{\top} + \gamma_\ell P + E^{(\ell)}, \qquad c^{(\ell)} \in \mathbb{R}_+^N, \quad \gamma_\ell \geq 0, \quad \mathbf{1}^\top c^{(\ell)} + \gamma_\ell \leq 1, \tag{1}$$

where $A^{(\ell)} \in \mathbb{R}^{N\times N}$ is row-stochastic, $A^{(\ell)}\mathbf{1} = \mathbf{1}$, $c^{(\ell)}$ captures query-independent shared broadcasting, $P_{i,p(i)} = 1$ with $p(i) = i-1$ represents attention to the preceding serialized position, and $E^{(\ell)}$ collects the remaining query-specific interactions. For an effective anchor set $\mathcal{A}$, we denote the shared-stripe mass assigned to it by $C_{\mathcal{A}}^{(\ell)} = \sum_{j\in\mathcal{A}} c_j^{(\ell)}$; the normalization consequences of this decomposition are given in Appendix A.4.

We measure approximation quality by the relative residual $\rho_E = \|E\|_F/\|A\|_F$. The mean residual is lowest in shallow layers (0.303), rises substantially through the middle of the network, and remains elevated at greater depth. Thus, the two modeled components are most faithful in the shallow network, while interactions outside the shared-stripe and predecessor components become more prominent at intermediate depth. The middle layers provide a natural out-of-regime control. Where the stripe–predecessor approximation deteriorates, anchor amplification weakens and hidden-state

distances expand. The predicted signatures reappear only when structural fidelity partially recovers at later depth. A more detailed analysis of this behavior is presented in Section 4. At greater depth, additional periodic shifted diagonals partly explain the remaining residual; including stride-4 components, details are given in Appendix A.4.

We therefore treat equation 1 as a depth-dependent reduced description of the structured attention regime and next analyze its cumulative dynamical consequences.

4 Anchor Concentration and Representation Convergence

We analyze how the structured attention geometry in Section 3.3 amplifies anchor influence across Transformer depth. We first derive a reduced query-level anchor-mass dynamics linking the shared stripe and shifted-predecessor feedback to an effective anchor–background gain, and then characterize when the resulting anchor concentration leads to representation alignment.

4.1 From Structured Attention to Anchor-Mass Dynamics

The derivation follows three steps: structured attention determines the hidden-state update, the resulting update induces an anchor–background gain, and this gain controls the query-level total anchor mass.

Lemma 1 (Structured attention update)**.** *For $C_A^{(\ell)} > 0$, the analyzed hidden-state update of query i can be written as*

$$\boxed{\Delta h_i^{(\ell)} = C_A^{(\ell)} W_O^{(\ell)} v_A^{(\ell)} + \gamma_\ell W_O^{(\ell)} v_{p(i)}^{(\ell)} + u_{R,i}^{(\ell)}.} \tag{2}$$

The first two terms isolate the shared-broadcast and shifted-predecessor pathways. Precise definitions and the proof are given in Appendix B.1.

We next ask how these pathways change the relative score of anchor and background keys. Under a two-group reduction, let $G_i^{(\ell)}$ denote the resulting layerwise anchor–background gain; its precise definition is given in Appendix B.3.

Lemma 2 (Layerwise gain decomposition)**.** *Under this reduction,*

$$\boxed{G_i^{(\ell)} \approx C_A^{(\ell)} S_i^{(\ell)} + \gamma_\ell D_i^{(\ell)} + R_i^{(\ell)}.} \tag{3}$$

Here $S_i^{(\ell)}$, $D_i^{(\ell)}$, and $R_i^{(\ell)}$ respectively capture the broadcast, predecessor, and remaining contributions to the gain. Their exact forms and the proof are given in Appendix B.3.

Lemma 3 (Two-group replicator reduction)**.** *Let $C_i^{(\ell)}$ denote the true query-level attention mass assigned to the anchor set. Under a two-group approximation in which score increments are taken to be homogeneous within the anchor and background groups,*

$$\boxed{C_i^{(\ell+1)} \approx \frac{C_i^{(\ell)} e^{\beta G_i^{(\ell)}}}{C_i^{(\ell)} e^{\beta G_i^{(\ell)}} + 1 - C_i^{(\ell)}}.} \tag{4}$$

Thus, the gain determines whether query-level anchor attention is amplified or suppressed. The exact token-level identity, group reduction, and proof are given in Appendix B.2.

For the subsequent querywise analysis, we fix an arbitrary query i and write $C_\ell := C_i^{(\ell)}$, $G_\ell := G_i^{(\ell)}$, with the query index on S_ℓ, D_ℓ, and R_ℓ suppressed similarly. Under this notation, the two-group recurrence in equation 4 becomes

$$C_{\ell+1} = \frac{C_\ell e^{\beta G_\ell}}{C_\ell e^{\beta G_\ell} + 1 - C_\ell}. \tag{5}$$

This defines the reduced querywise dynamics used below. This is just a notation convention.

4.2 Cumulative Anchor Amplification

Having established how G_ℓ controls the one-step evolution of the query-level anchor mass, we now study how these gains accumulate across Transformer depth. We now move from the local effect of a single layer to the net effect accumulated over a finite depth interval.

Corollary 1 (Layerwise monotonicity). *For the reduced interior dynamics,* $\operatorname{sign}(C_{\ell+1} - C_\ell) = \operatorname{sign}(G_\ell)$. *Thus, positive gain increases the anchor mass, whereas negative gain decreases it.*

The proof is given in Appendix B.4.

Lemma 4 (Cumulative odds identity). *For every* $L > \ell_0$,

$$\boxed{\frac{1 - C_L}{C_L} = \frac{1 - C_{\ell_0}}{C_{\ell_0}} \exp\left(-\beta \sum_{\ell=\ell_0}^{L-1} G_\ell\right).} \tag{6}$$

The proof is given in Appendix B.5. Hence the finite-depth anchor mass is controlled by the cumulative gain $\sum_{\ell=\ell_0}^{L-1} G_\ell$, instead of just by layerwise monotonicity.

Theorem 1 (Finite-Depth Anchor Concentration). *Under the reduced querywise dynamics equation 6 initialized at layer* ℓ_0 *with anchor mass* $C_{\ell_0} \in (0, 1)$, *the remaining non-anchor mass at depth* $L > \ell_0$ *admits the cumulative-gain bound:*

$$1 - C_L \le \frac{1 - C_{\ell_0}}{C_{\ell_0}} \exp\left(-\beta \sum_{\ell=\ell_0}^{L-1} G_\ell\right). \tag{7}$$

Consequently, whenever the cumulative gain grows unbounded ($\sum_{\ell=\ell_0}^{L-1} G_\ell \to +\infty$*), there exists a finite threshold depth* L_ε *for any target error* $\varepsilon \in (0, 1)$ *such that* $C_L \ge 1 - \varepsilon$ *for all* $L \ge L_\varepsilon$, *with* $C_L \to 1$ *in the ideal asymptotic limit.*

The proof is given in Appendix B.6. The cumulative condition allows intermediate layers with $G_\ell < 0$ and therefore accommodates the empirically observed mid-layer decrease in anchor mass. For a finite Transformer, the finite-depth bound is the relevant result.

4.3 From Anchor Concentration to Representation Convergence

The previous two sections characterize when query-level anchor mass is amplified. We now ask what this concentration implies for the corresponding hidden state convergence and define the attention readout $H_i^{(\ell+1)} := \sum_{j=1}^{N} A_{ij}^{(\ell)} V_j^{(\ell)}$.

Assumption 1 (Uniformly bounded projected values). *Let* $V_j^{(\ell)} := W_O^{(\ell)} v_j^{(\ell)}$. *There exists* $B < \infty$, *independent of* i, j, *and* ℓ, *such that* $\|V_j^{(\ell)}\| \le B$ *for all* j, ℓ.

A sufficient condition for Assumption 1 in a pre-RMSNorm Transformer is derived in Appendix C.1.

Lemma 5 (Anchor-conditional readout bound). *Let* $V_{A,i}^{(\ell)}$ *and* $V_{O,i}^{(\ell)}$ *denote the anchor- and background-conditioned readout centers. Under Assumption 1,*

$$\boxed{H_i^{(\ell+1)} = C_i^{(\ell)} V_{A,i}^{(\ell)} + (1 - C_i^{(\ell)}) V_{O,i}^{(\ell)}, \qquad \left\| H_i^{(\ell+1)} - V_{A,i}^{(\ell)} \right\| \le 2B(1 - C_i^{(\ell)})} \tag{8}$$

The proof and definition are given in Appendix C.2. Thus anchor concentration forces each query readout toward its own anchor-conditioned center; the corresponding positionwise convergence and rate are given in Appendix C.3.

Beyond individual positionwise alignment, the structured attention matrix further drives stronger cross-position representation alignment and even convergence.

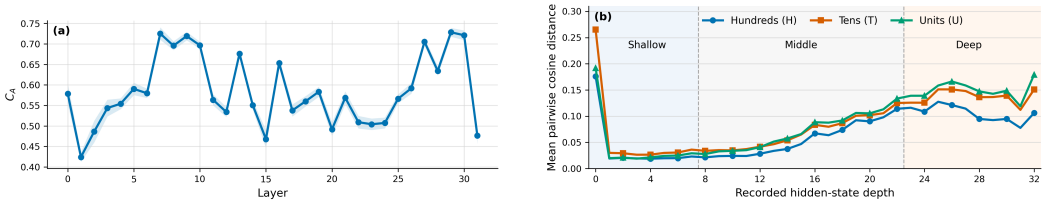

Figure 2: **Depth-dependent anchor concentration and representation alignment.** **(a)** Pooled fitted anchor mass C_A and **(b)** mean within-slot pairwise cosine distance (for hundreds, tens, and units). Both panels exhibit synchronized, non-monotone dynamics across depth: anchor concentration peaks in shallow and deep layers where representations contract, while relaxing in middle layers—supporting the finite-depth dynamics characterized in Theorem 1 and aligning with Section 3.3.

Corollary 2 (Cross-position convergence)**.** *Let $\mathcal{M}$ be a set of query positions. Under Assumption 1 and the cumulative-gain regime of Theorem 1, when the query-specific corrections to the shared stripe vanish with depth,*

$$\max_{i,k\in\mathcal{M}} \left\| H_i^{(\ell+1)} - H_k^{(\ell+1)} \right\| \longrightarrow 0. \tag{9}$$

The correction terms and the corresponding finite-depth bound are given in Appendix C.4.

4.4 Empirical Tests of the Reduced Dynamics

Gain and anchor amplification. We first examine the gain dynamics underlying Section 4.2. The depth-accumulated gain is positive at all diffusion steps in the tested dataset. Layerwise gain also closely tracks changes in fitted anchor-mass log-odds (Pearson $r = 0.838$, Spearman $\rho = 0.864$). Moreover, rollouts exhibiting repetition have a larger total gain than non-repeating rollouts (42.298 vs. 40.45); detailed statistics are reported in Appendix E. This difference is consistent with the cumulative-gain mechanism, in which stronger gain accumulation produces stronger anchor amplification. The fitted anchor mass in Fig. 2(a) further shows a distinctly stage-wise trajectory across depth: concentration strengthens over the shallow layers, weakens and fluctuates through the middle regime, and increases again in deep layers. We return to this depth structure below.

Depth-wise representation dynamics. Figure 2(b) reveals a closely related stage-wise pattern in the hidden states. Within-slot pairwise cosine distance contracts sharply at the network entrance, with H/T/U distances decreasing from $0.176/0.265/0.192$ at depth 0 to $0.019/0.030/0.020$ at depth 1. The distances then expand through the middle layers, contract again in late depth, and partially rebound at the final state. Strikingly, these representation dynamics align qualitatively with both the anchor-mass trajectory in Fig. 2(a) and the depth-dependent adequacy of the structured attention model reported in Section 3.3: shallow layers exhibit stronger anchor concentration and rapid representation contraction where the stripe–predecessor decomposition is most accurate; the middle regime shows weaker and more irregular anchor concentration together with representation expansion as the residual grows; and the deep regime exhibits a partial recovery of the structured fit alongside renewed anchor amplification and local contraction. The two measurements do not track each other layer by layer, but both exhibit a similar shallow–middle–deep trend. The reduced dynamics therefore appear to describe the network better at depths where the structured attention pattern is clearer.

5 Mechanistic Interventions and Validation

Using the fixed-query notation introduced above, our theory isolates the layerwise anchor–background gain

$$G_\ell = C_A^{(\ell)} S_\ell + \gamma_\ell D_\ell + R_\ell$$

as the driver of the query-level anchor dynamics. We use this relation to turn the proposed mechanism into an interventional test. We first quantify how reducing G_ℓ weakens anchor concentration,

and then use this comparison as a common theoretical interface for two training-free interventions: SVD truncation and history-sparsified guidance. The two interventions act at different stages of inference, but both can be evaluated through their effect on cumulative gain.

Lemma 6 (Gain–contraction interface)**.** *For the reduced anchor dynamics, the exact decay factor of the background mass is*

$$\eta_\ell(C_\ell, G_\ell) := \frac{1 - C_{\ell+1}}{1 - C_\ell} = \frac{1}{C_\ell e^{\beta G_\ell} + 1 - C_\ell}, \qquad \frac{\partial \eta_\ell}{\partial G_\ell} < 0. \tag{10}$$

Hence, at the same current C_ℓ, $G_\ell^{\mathrm{int}} < G_\ell$ implies $\eta_\ell^{\mathrm{int}} > \eta_\ell$ and therefore slower anchor concentration. Near $C_\ell = 1$, $\eta_\ell \approx e^{-\beta G_\ell}$. For two trajectories starting from the same C_{ℓ_0},

$$\operatorname{logit} C_L^{\mathrm{int}} - \operatorname{logit} C_L = \beta \sum_{\ell=\ell_0}^{L-1} \left(G_\ell^{\mathrm{int}} - G_\ell \right). \tag{11}$$

The proof is given in Appendix E.1. The one-step comparison fixes C_ℓ; after the two trajectories diverge, the cumulative identity in equation 11 is the relevant finite-depth criterion.

5.1 SVD Truncation

For $A^{(\ell)} = U^{(\ell)} \Sigma^{(\ell)} V^{(\ell)\top}$, we replace the attention readout operator by

$$\widetilde{A}_k^{(\ell)} = U^{(\ell)}_{[:,1:k]} \Sigma^{(\ell)}_{1:k,1:k} V^{(\ell)\top}_{[:,1:k]}. \tag{12}$$

Numerical full rank does not imply that every singular direction carries an independent contextual interaction. For $A_0 = \mathbf{1}c^\top + \gamma P$, at least $N - 3$ singular values equal $|\gamma|$ (Appendix E.3). Thus, the shifted-predecessor component generates a high-dimensional spectral plateau, whereas the shared stripe changes at most two singular directions. Empirically, our measured spectra and singular-vector reconstructions confirm that these predecessor modes sit below the leading stripe modes.

Let $k = \lceil N/4 \rceil$ or even $k = 8$, the structure preserves the observed leading stripe while removing many plateau directions. Denoting the surviving diagonal channel by an aggregate effective strength $\gamma_{k,\ell} < \gamma_\ell$, the focused attribution is

$$G_\ell - G_\ell^{\mathrm{SVD}} \approx (\gamma_\ell - \gamma_{k,\ell}) D_\ell > 0 \quad (D_\ell > 0). \tag{13}$$

Here $\gamma_{k,\ell}$ does not imply uniform shrinkage of every diagonal entry. The general diagnostic remains the measured total or cumulative gain. The results indicate that SVD truncation effectively weakens the predecessor-supported feedback channel, thereby lowering the gain Emprical analysis will be discussed in section 5.3.

5.2 History-Sparsified Guidance

At denoising step t, HSCFG contrasts a full-history branch with a history-sparsified reference branch while keeping the prompt, mask pattern, and current forecast canvas identical:

$$z_t^{\mathrm{HSCFG}} = z_t^{\mathrm{sp}} + s \left(z_t^{\mathrm{full}} - z_t^{\mathrm{sp}} \right), \qquad s > 1.$$

Let a denote the repeated digit and b a competing digit, and define $m_t = z_t(a) - z_t(b)$. Writing $\lambda = s - 1$, the guided margin satisfies

$$m_t^{\mathrm{HSCFG}} = m_t^{\mathrm{full}} - \lambda \left(m_t^{\mathrm{sp}} - m_t^{\mathrm{full}} \right).$$

After repeated tokens accumulate, we observe $m_t^{\mathrm{sp}} > m_t^{\mathrm{full}}$; under this condition, HSCFG strictly reduces the repeated digit's margin. Thus, the sparse-history contrast provides a direct mechanism for interrupting repetition at token commitment.

Table 1: **Interventional diagnostics of the gain mechanism.** Both training-free interventions reduce matched cumulative gain while preserving its association with layerwise anchor-mass amplification. Full paired statistics and temporal diagnostics are reported in Appendix E.

Condition	Δ Cum. Gain	Repetition rate	Pearson r	Spearman ρ
Vanilla	–	1.000	0.838	0.864
Quarter-rank SVD	-3.22%	0.701	0.823	0.862
HSCFG	-3.03%	0.132	0.820	0.868

Unlike SVD, HSCFG does not directly modify the attention operator. We therefore test its connection to the proposed gain dynamics empirically: if changing the committed trajectory disrupts the same feedback process, the subsequent cumulative gain should decrease. Section 5.3 shows that it does; detailed channel-wise diagnostics are deferred to Appendix E.6.

5.3 Interventional Validation

For the paired intervention analysis, vanilla, SVD, and HSCFG are evaluated on matched inputs. We first identify repetition-prone vanilla rollouts using $R_H > 0.99$, $R_T > 0.95$, and $R_U > 0.95$, and then retrieve the corresponding SVD and HSCFG trajectories for the same series.

We finally test the common prediction of the two interventions: if the proposed gain mechanism is relevant to repetition collapse, successful interventions should reduce the cumulative gain governing anchor concentration. On matched paired rollouts, quarter-rank SVD reduces cumulative gain by 3.22%, while HSCFG reduces it by 3.03% relative to their corresponding vanilla trajectories (Table 1). While modest in absolute terms, these reductions bridge a major portion of the 4.569% gain gap between repeating and non-repeating sequences. Moreover, these changes are in the direction predicted by Lemma 6.

Importantly, the gain–anchor coupling remains stable after intervention. The association between measured gain and the layerwise change in fitted anchor-mass log-odds remains strong under both SVD (Pearson $r = 0.823$, Spearman $\rho = 0.862$) and HSCFG (Pearson $r = 0.820$, Spearman $\rho = 0.868$), close to the vanilla relationship (Pearson $r = 0.838$, Spearman $\rho = 0.864$). Thus, both interventions reduce cumulative gain while preserving the gain–concentration relationship identified in Section 4.4.

These internal changes are accompanied by substantial reductions in repetition. value-level repetition from 1.000 to 0.132, while the corresponding SVD evaluation likewise shows reduced repetition under spectral truncation; complete repetition and forecasting metrics are reported in Appendix E. Together, these results provide interventional support for the proposed mechanism: two interventions acting at different stages of inference both reduce cumulative gain, preserve its coupling to anchor amplification, and suppress the repetition failure mode. Detailed paired protocols, temporal gain profiles, correlation diagnostics, and complete outcome tables are provided in Appendix E.

6 Conclusion and Future Work

Conclusion. Starting from the observed repetition-collapse phenomenon, we provide, to our knowledge, the first mathematical abstraction of its underlying attention dynamics. Our finite-depth model links the shared-anchor stripes and shifted-predecessor feedback geometry to cumulative gain, anchor concentration, and cross-position alignment. Repeating trajectories accumulate more gain, whereas SVD truncation and history-sparsified guidance reduce both gain and repetition through distinct pathways. Together, these findings provide a mechanistic account of repetition collapse in diffusion-based time-series forecasting.

Future work. The decomposition is depth dependent, with fidelity varying across heads and layers. MLP pathways and other unmodeled operations enter the residual term $R_i^{(\ell)}$. More explicit pathway models and component-level interventions remain for future work.

AI USE STATEMENT

Generative AI tools were used in this work to assist with drafting manuscript sections, language polishing, and writing code scripts for data cleaning and statistical analysis. All AI-assisted content, code, and statistical results were thoroughly verified and finalized by the authors, who take full responsibility for the paper's final content.

ETHICS STATEMENT

This work studies repetition dynamics in diffusion language models for time-series forecasting using publicly available benchmark datasets. The study does not involve human subjects, personally identifiable information, or experiments in high-stakes decision-making settings.

REPRODUCIBILITY STATEMENT

We provide sufficient methodological detail to reproduce the analyses and experiments reported in this work. The paper and appendices contain the theoretical definitions, complete proofs, evaluation metrics, and core experimental procedures. The supplementary material provides the full model and decoding configurations, data preprocessing details, intervention hyperparameters, paired-evaluation protocol, and analysis settings used in our experiments. We intend to release the complete codebase and analysis pipeline to support future research and facilitate reproducibility.

REFERENCES

Jacob Austin, Daniel D. Johnson, Jonathan Ho, Daniel Tarlow, and Rianne van den Berg. Structured denoising diffusion models in discrete state-spaces. In *Advances in Neural Information Processing Systems*, volume 34, pp. 17981–17993, 2021. URL https://proceedings.neurips.cc/paper/2021/hash/958c530554f78bcd8e97125b70e6973d-Abstract.html.

Hao Cheng and Zhihua Zhang. MR-P: A parallel decoding algorithm for iterative refinement non-autoregressive translation. In *Findings of the Association for Computational Linguistics: ACL 2022*, pp. 285–296. Association for Computational Linguistics, 2022. doi: 10.18653/v1/2022.findings-acl.25. URL https://aclanthology.org/2022.findings-acl.25/.

Yihe Dong, Jean-Baptiste Cordonnier, and Andreas Loukas. Attention is not all you need: Pure attention loses rank doubly exponentially with depth. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 2793–2803. PMLR, 2021. URL https://proceedings.mlr.press/v139/dong21a.html.

Zihao Fu, Wai Lam, Anthony Man-Cho So, and Bei Shi. A theoretical analysis of the repetition problem in text generation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pp. 12848–12856, 2021.

Borjan Geshkovski, Cyril Letrouit, Yury Polyanskiy, and Philippe Rigollet. The emergence of clusters in self-attention dynamics. In *Advances in Neural Information Processing Systems*, volume 36, pp. 57026–57037, 2023. URL https://openreview.net/forum?id=aMjaEkkXJx.

Rakshitha Godahewa, Christoph Bergmeir, Geoffrey I. Webb, Rob J. Hyndman, and Pablo Montero-Manso. Monash time series forecasting archive. In *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*, volume 1, 2021. URL https://papers.nips.cc/paper/2021,https://nips.cc/Conferences/2021.

Nate Gruver, Marc Finzi, Shikai Qiu, and Andrew Gordon Wilson. Large language models are zero-shot time series forecasters. In *Advances in Neural Information Processing Systems*, volume 36, pp. 19622–19635, 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/3eb7ca52e8207697361b2c0fb3926511-Abstract-Conference.html.

Ye He, Kevin Rojas, and Molei Tao. What exactly does guidance do in masked discrete diffusion models. In *International Conference on Learning Representations*, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/a14193e9d9fb0b03af0b717de1cac8ac-Abstract-Conference.html.

Tatsuya Hiraoka and Kentaro Inui. Repetition neurons: How do language models produce repetitions? In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 2: Short Papers)*, pp. 483–495. Association for Computational Linguistics, 2025. doi: 10.18653/v1/2025.naacl-short.41. URL https://aclanthology.org/2025.naacl-short.41/.

Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance, 2022. URL https://arxiv.org/abs/2207.12598.

Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. The curious case of neural text degeneration. In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=rygGQyrFvH.

Sujung Hong, Chanyong Yoon, and Seong Jae Hwang. Mitigating mask prior drift and positional attention collapse in large diffusion vision-language models, 2026. URL https://arxiv.org/abs/2605.14530.

Metod Jazbec, Theo X. Olausson, Louis Béthune, Pierre Ablin, Michael Kirchhof, João Monteiro, Victor Turrisi, Jason Ramapuram, and Marco Cuturi. Learning unmasking policies for diffusion language models, 2026. URL https://arxiv.org/abs/2512.09106. Accepted as an oral spotlight at ICML 2026.

Jaeyeon Kim, Kulin Shah, Vasilis Kontonis, Sham M. Kakade, and Sitan Chen. Train for the worst, plan for the best: Understanding token ordering in masked diffusions. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 30749–30768. PMLR, 2025. URL https://openreview.net/forum?id=DjJmre5IkP.

Pengxiang Li, Shilin Yan, Jiayin Cai, Renrui Zhang, Ruichuan An, Ziyu Guo, and Xiaowei Gao. Adaptive classifier-free guidance via dynamic low-confidence masking. In *Advances in Neural Information Processing Systems*, volume 38, pp. 52874–52893, 2025. URL https://papers.neurips.cc/paper_files/paper/2025/hash/43ba0466af2b1ac76aa85d8fbec714e3-Abstract-Conference.html.

Aaron Lou, Chenlin Meng, and Stefano Ermon. Discrete diffusion modeling by estimating the ratios of the data distribution. In Ruslan Salakhutdinov, Zico Kolter, Katherine Heller, Adrian Weller, Nuria Oliver, Jonathan Scarlett, and Felix Berkenkamp (eds.), *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 32819–32848, 21–27 Jul 2024.

Callum Stuart McDougall, Arthur Conmy, Cody Rushing, Thomas McGrath, and Neel Nanda. Copy suppression: Comprehensively understanding a motif in language model attention heads. In *Proceedings of the 7th BlackboxNLP Workshop: Analyzing and Interpreting Neural Networks for NLP*, pp. 337–363. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.blackboxnlp-1.22. URL https://aclanthology.org/2024.blackboxnlp-1.22/.

Shen Nie, Fengqi Zhu, Zebin You, Xiaolu Zhang, Jingyang Ou, Jun Hu, Jun Zhou, Yankai Lin, Ji-Rong Wen, and Chongxuan LI. Large language diffusion models. In D. Belgrave, C. Zhang, H. Lin, R. Pascanu, P. Koniusz, M. Ghassemi, and N. Chen (eds.), *Advances in Neural Information Processing Systems*, volume 38, Main Conference, pp. 50608–50646. Curran Associates, Inc., 2025. doi: 10.52202/085713-1689. URL https://proceedings.neurips.cc/paper_files/paper/2025/file/48b383b24230e0e6e649d9c98dae4d8c-Paper-Conference.pdf.

Yuhang Pei, Tao Ren, Yifan Wang, Zhipeng Sun, Wei Ju, Chong Chen, Xian-Sheng Hua, and Xiao Luo. LEAF: Large language diffusion model for time series forecasting. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pp. 1076–1091, 2025.

Qiu Ran, Yankai Lin, Peng Li, and Jie Zhou. Learning to recover from multi-modality errors for non-autoregressive neural machine translation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 3059–3069. Association for Computational Linguistics, 2020. doi: 10.18653/v1/2020.acl-main.277. URL https://aclanthology.org/2020.acl-main.277/.

Kevin Rojas, Ye He, Chieh-Hsin Lai, Yuhta Takida, Yuki Mitsufuji, and Molei Tao. Improving classifier-free guidance in masked diffusion: Low-dim theoretical insights with high-dim impact. In *International Conference on Learning Representations*, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/391f1a5319c929173bddce4cd9367056-Abstract-Conference.html.

Maximo Eduardo Rulli, Simone Petruzzi, Edoardo Michielon, Fabrizio Silvestri, Simone Scardapane, and Alessio Devoto. Attention sinks in diffusion language models. In *Findings of the Association for Computational Linguistics: ACL 2026*, pp. 43244–43256. Association for Computational Linguistics, 2026. doi: 10.18653/v1/2026.findings-acl.2145. URL https://aclanthology.org/2026.findings-acl.2145/.

Thiziri Nait Saada, Alireza Naderi, and Jared Tanner. Mind the gap: a spectral analysis of rank collapse and signal propagation in attention layers. *arXiv preprint arXiv:2410.07799*, 2024.

Subham Sekhar Sahoo, Marianne Arriola, Yair Schiff, Aaron Gokaslan, Edgar Marroquin, Justin T. Chiu, Alexander Rush, and Volodymyr Kuleshov. Simple and effective masked diffusion language models. In *Advances in Neural Information Processing Systems*, volume 37, pp. 130136–130184, 2024. doi: 10.52202/079017-4135. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/eb0b13cc515724ab8015bc978fdde0ad-Abstract-Conference.html.

Yair Schiff, Subham Sekhar Sahoo, Hao Phung, Guanghan Wang, Sam Boshar, Hugo Dalla-Torre, Bernardo P. de Almeida, Alexander Rush, Thomas Pierrot, and Volodymyr Kuleshov. Simple guidance mechanisms for discrete diffusion models. In *International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=i5MrJ6g5G1.

Weijia Shi, Xiaochuang Han, Mike Lewis, Yulia Tsvetkov, Luke Zettlemoyer, and Wen-tau Yih. Trusting your evidence: Hallucinate less with context-aware decoding. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 2: Short Papers)*, pp. 783–791. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.naacl-short.69. URL https://aclanthology.org/2024.naacl-short.69/.

Yixuan Su, Tian Lan, Yan Wang, Dani Yogatama, Lingpeng Kong, and Nigel Collier. A contrastive framework for neural text generation. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh (eds.), *Advances in Neural Information Processing Systems*, volume 35, pp. 21548–21561. Curran Associates, Inc., 2022. doi: 10.52202/068431-1566. URL https://proceedings.neurips.cc/paper_files/paper/2022/file/871cae8f599cb8bbfcb0f58fe1af95ad-Paper-Conference.pdf.

Sean Welleck, Ilia Kulikov, Stephen Roller, Emily Dinan, Kyunghyun Cho, and Jason Weston. Neural text generation with unlikelihood training. In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=SJeYe0NtvH.

Jin Xu, Xiaojiang Liu, Jianhao Yan, Deng Cai, Huayang Li, and Jian Li. Learning to break the loop: Analyzing and mitigating repetitions for neural text generation. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh (eds.), *Advances in Neural Information Processing Systems*, volume 35, pp. 3082–3095. Curran Associates, Inc., 2022. doi: 10.52202/068431-0223. URL https://proceedings.neurips.cc/paper_files/paper/2022/file/148c0aeea1c5da82f4fa86a09d4190da-Paper-Conference.pdf.

Hanqi Yan, Lin Gui, Wenjie Li, and Yulan He. Addressing token uniformity in transformers via singular value transformation. In *Proceedings of the Thirty-Eighth Conference on Uncertainty in Artificial Intelligence*, volume 180 of *Proceedings of Machine Learning Research*, pp. 2181–2191. PMLR, 2022. URL https://proceedings.mlr.press/v180/yan22b.html.

Junchi Yao, Shu Yang, Jianhua Xu, Lijie Hu, Mengdi Li, and Di Wang. Understanding the repeat curse in large language models from a feature perspective. In *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 7787–7815. Association for Computational Linguistics, 2025. doi: 10.18653/v1/2025.findings-acl.406. URL https://aclanthology.org/2025.findings-acl.406/.

Jiacheng Ye, Zhihui Xie, Lin Zheng, Jiahui Gao, Zirui Wu, Xin Jiang, Zhenguo Li, and Lingpeng Kong. Dream 7b: Diffusion large language models, 2025. URL https://arxiv.org/abs/2508.15487.

Shuai Zhang, Zijie Chen, Hongliang He, Lun Du, and Zhenzhong Lan. Low perplexity is repetition: A one-dimensional self-conditioning attractor in continuous diffusion LMs, 2026. URL https://arxiv.org/abs/2607.00588.

Qiyan Zhao, Xiaofeng Zhang, Shuochen Chang, Qianyu Chen, Xiaosong Yuan, Xuhang Chen, Luoqi Liu, Jiajun Zhang, Xu-Yao Zhang, and Da-Han Wang. Context tokens are anchors: Understanding the repetition curse in dMLLMs from an information flow perspective, 2026. URL https://arxiv.org/abs/2601.20520.

A Additional Characterization of Progressive Repetition

A.1 Common Measurement Framework

We first define the measurements used throughout the paper. These definitions apply to arbitrary complete forecasts and therefore cover the final outputs of vanilla decoding, the intervention results, and the intermediate forecasts constructed by the look-ahead analysis in Appendix A.2.

Let F denote the forecasting horizon and $\mathbf{y} = (y_1, \ldots, y_F)$ an arbitrary complete numerical forecast, with ground-truth counterpart $\mathbf{y}^\star$. Each value is represented by digit slots $s \in \{H, T, U\}$, corresponding to hundreds, tens, and units. We write $d_{q,s}$ for the digit in slot s of forecast position q, and $\mathbf{d}_s = (d_{1,s}, \ldots, d_{F,s})$ for the corresponding slot sequence.

Value-level repetition. We measure numerical repetition using the fraction of ordered pairs of distinct forecast positions that contain the same complete value:

$$\mathcal{R}_{\text{value}}(\mathbf{y}) = \frac{1}{F(F-1)} \sum_{\substack{q,r=1 \\ q \neq r}}^{F} \mathbb{1}\{y_q = y_r\}. \tag{14}$$

For the model prediction $\widehat{\mathbf{y}}$ and ground-truth future $\mathbf{y}^\star$, respectively, we use

$$R^{\text{pred}}_{\text{value}} = \mathcal{R}_{\text{value}}(\widehat{\mathbf{y}}), \qquad R^{\text{gt}}_{\text{value}} = \mathcal{R}_{\text{value}}(\mathbf{y}^\star). \tag{15}$$

The ground-truth statistic captures legitimate repetition already present in the data, whereas the predicted statistic may additionally contain excess repetition introduced during decoding.

Slot-wise repetition. Complete values may differ even after one of their digit slots has largely collapsed. We therefore compute pairwise repetition separately for each slot:

$$\mathcal{R}_{\text{slot},s}(\mathbf{d}_s) = \frac{1}{F(F-1)} \sum_{\substack{q,r=1 \\ q \neq r}}^{F} \mathbb{1}\{d_{q,s} = d_{r,s}\}, \qquad s \in \{H, T, U\}. \tag{16}$$

The final predicted and ground-truth slot-wise repetition rates are

$$R^{s,\text{pred}}_{\text{slot}} = \mathcal{R}_{\text{slot},s}(\widehat{\mathbf{d}}_s), \qquad R^{s,\text{gt}}_{\text{slot}} = \mathcal{R}_{\text{slot},s}(\mathbf{d}^\star_s). \tag{17}$$

These measurements identify whether repetition is concentrated in a particular numerical slot. For example, a high $R^{H,\text{pred}}_{\text{slot}}$ with lower tens- and units-slot rates indicates that the hundreds digits have homogenized even when the complete predicted values are not yet identical.

Normalized forecasting error. To make forecasting errors comparable across time series with different scales, we normalize each prediction against a last-value persistence baseline. Let y_0 be the final observed value before the forecasting horizon. The persistence forecast is

$$y_q^{\text{pers}} = y_0, \qquad q = 1, \dots, F.$$

For an arbitrary forecast $y = (y_1, \dots, y_F)$ with ground-truth future $y^\star = (y_1^\star, \dots, y_F^\star)$, we define

$$\text{NMAE}(y, y^\star) = \frac{\sum_{q=1}^{F} |y_q - y_q^\star|}{\sum_{q=1}^{F} |y_0 - y_q^\star|}.$$

Reveal progress. The fraction of slot s that has been revealed is

$$\text{Prog}_s(t) = 1 - \frac{|\mathcal{M}_t \cap \mathcal{I}_s|}{|\mathcal{I}_s|}. \tag{18}$$

Writing $\mathcal{I}_{\text{dig}} = \bigcup_{s \in \{H,T,U\}} \mathcal{I}_s$, the overall digit progress is

$$\text{Prog}_{\text{dig}}(t) = 1 - \frac{|\mathcal{M}_t \cap \mathcal{I}_{\text{dig}}|}{|\mathcal{I}_{\text{dig}}|}. \tag{19}$$

Confidence at commitment. Consider a reveal event e occurring at diffusion step t_e. Let i_e be the selected position, s_e its slot, and d_e the digit committed to the canvas. Its commitment confidence is

$$c_e = \pi_{t_e, i_e}(d_e). \tag{20}$$

Under vanilla Top-1 decoding, $c_e = \max_{a \in \mathcal{V}_{\text{dig}}} \pi_{t_e, i_e}(a)$. For an intervention that commits a non-Top-1 token, c_e remains the probability assigned to the digit actually written. We associate event e with its post-commitment within-slot progress,

$$u_e = \text{Prog}_{s_e}(t_e + 1). \tag{21}$$

The slot-wise confidence curves therefore compare committed-token confidence at matched fractions of each slot's own reveal process rather than at matched global diffusion steps.

Scheduler Diagnostics The reveal progress of slot family s is

$$\text{Prog}_s(t) = 1 - \frac{|\mathcal{M}_t \cap \mathcal{I}_s|}{|\mathcal{I}_s|}. \tag{22}$$

This statistic equals zero before any position in slot s is committed and one after the slot is completely revealed.

To measure persistence across reveal events, let s_e denote the slot family selected at event e. Let $a_e = 1$ when slot s_e still contains unresolved positions after that event, and $a_e = 0$ otherwise. We define

$$P_{\text{same}} = \frac{\sum_e a_e \, \mathbb{1}\{s_{e+1} = s_e\}}{\sum_e a_e}. \tag{23}$$

The availability condition removes events for which same-slot continuation is impossible because the preceding slot has been exhausted. We compare the observed statistic with an availability-matched randomized scheduler and with non-repetitive control trajectories. Accordingly, P_{same} measures scheduler persistence but does not by itself establish that the scheduler causes repetition.

A.2 Candidate Predictions and Repetition Metrics

Unresolved candidate predictions. At time step t, we denote the unresolved forecast positions by$\mathcal{M}_t$ t, the positions associated with numerical slot family $s \in \{H, T, U\}$ by $\mathcal{I}_s$, and let $\mathcal{V}_{\text{dig}}$ denote the digit vocabulary. For every unresolved numerical position $i \in \mathcal{M}_t \cap \mathcal{I}_s$, its Top-1 candidate and confidence are

$$\hat{d}_{t,i} = \arg\max_{a \in \mathcal{V}_{\text{dig}}} \pi_{t,i}(a), \qquad p_{t,i} = \pi_{t,i}(\hat{d}_{t,i}), \tag{24}$$

where $\pi_{t,i}$ is the predictive distribution at position i.

To track the visible state of the entire forecast canvas, we combine committed digits with the current candidates at unresolved positions:

$$\bar{d}_{t,i} = \begin{cases} x_{t,i}, & i \notin \mathcal{M}_t, \\ \hat{d}_{t,i}, & i \in \mathcal{M}_t. \end{cases} \tag{25}$$

Instantiation of the common measurements. Rather than introducing separate repetition and error definitions for the look-ahead analysis, we apply the functionals from Appendix A.1 to the candidate-completed forecast:

$$R_{\mathrm{draft}}(t) = \mathcal{R}_{\mathrm{value}}(\overline{\mathbf{y}}_t), \qquad \mathrm{NMAE}_{\mathrm{draft}}(t) = \mathrm{NMAE}(y_t, y^\star). \tag{26}$$

Thus, final and look-ahead repetition differ only in the forecast to which the same functional is applied:

$$R^{\mathrm{pred}}_{\mathrm{value}} = \mathcal{R}_{\mathrm{value}}(\widehat{\mathbf{y}}), \qquad R_{\mathrm{draft}}(t) = \mathcal{R}_{\mathrm{value}}(\overline{\mathbf{y}}_t). \tag{27}$$

At the terminal step, $\overline{\mathbf{y}}_T = \widehat{\mathbf{y}}$, so

$$R_{\mathrm{draft}}(T) = R^{\mathrm{pred}}_{\mathrm{value}}, \qquad \mathrm{NMAE}_{\mathrm{draft}}(T) = \mathrm{NMAE}_{\mathrm{pred}}. \tag{28}$$

A progress-bin average near 100% need not equal the exact terminal statistic because the bin may contain both terminal and near-terminal records.

Detailed repetition data. Tables 2, 3, and 4 provide quantitative evidence for the progressive nature of repetition collapse. Specifically, Table 4 confirms the decoupling between repetition and accuracy: while draft repetition rises monotonically from 0.0412 to 0.4670, draft NMAE plateaus early at approximately 0.997. Table 3 details the underlying scheduling mechanism, showing that digit positions are revealed in a strict $H \to T \to U$ sequence (e.g., 95.55% of hundreds digits are committed at 30% progress, compared to only 3.35% for tens and 1.10% for units). Consequently, Table 2 demonstrates that slot-wise repetition follows this exact cascade, with the hundreds slot homogenizing early (reaching over 89% repeat rate by 12.5% progress) while tens and units align progressively during later decoding stages.

Table 2: Slot-wise repetition rates across different progress ratios.

Progress Ratio	Slot H Repeat Rate	Slot T Repeat Rate	Slot U Repeat Rate
0.025	0.7384	0.2151	0.1344
0.125	0.8960	0.5136	0.2504
0.275	0.8967	0.5549	0.3674
0.425	0.8978	0.5660	0.4284
0.575	0.8977	0.5679	0.4649
0.725	0.8977	0.5748	0.4834
0.875	0.9002	0.5835	0.5092
0.975	0.9016	0.5875	0.5311

A.3 Detailed Slotwise Unmasking Statistics

To quantify the sequential nature of the unmasking schedule, we partition the total decoding process into three equal stages ($1/3$, $2/3$, and $3/3$ progress) and analyze the distribution of consecutive token reveal pairs.

As detailed in Table 5, unmasking decisions exhibit strong temporal localization within individual digit families across all decoding phases. Stage 1 is overwhelmingly dominated by hundreds digits (93.6% H–H pairs), Stage 2 shifts decisively to tens digits (63.6% T–T pairs), and Stage 3 resolves the remaining units digits (67.7% U–U pairs). Across all stages, the vast majority of consecutive unmasking steps remain within the same slot family (up to 93.8%), quantitatively confirming that model itself tends to cascade through a strict, asynchronous $H \to T \to U$ sequence.

Table 3: Slot-wise reveal state at matched overall digit progress. "Target" is the requested overall progress or above that target. "Full values" is the fraction of forecast values for which all three digit slots have been committed.

Target	H	T	U	Full values
10%	36.99%	0.26%	0.25%	0.12%
20%	61.57%	0.58%	0.35%	0.15%
30%	95.55%	3.35%	1.10%	0.45%
40%	98.35%	21.84%	4.81%	0.93%
50%	98.68%	44.00%	7.32%	1.94%
60%	99.03%	75.96%	12.51%	5.32%
70%	99.25%	89.36%	23.90%	15.79%
80%	99.61%	94.52%	55.87%	50.88%
90%	99.72%	97.22%	78.06%	75.94%
100%	100.00%	100.00%	100.00%	100.00%

Table 4: Look-ahead measurements over fully revealed value progress. Column headers denote the progress of the revealing process

Metric	0%	25%	50%	75%	100%
Draft repetition	0.0412	0.2988	0.3934	0.4137	0.4670
Draft NMAE	1.038	0.997	0.998	1.000	1.010

A.4 Empirical Validation of the Structured Attention Model

Section 3.3 introduces the constrained stripe–predecessor decomposition. Here we record its normalization consequences, report the full reconstruction statistics across depth, and examine additional geometric components that account for part of the residual structure in deeper layers.

Normalization consequences. Because $A^{(\ell)}$ is row-stochastic and P is the non-cyclic predecessor-shift operator,

$$A^{(\ell)}\mathbf{1} = \mathbf{1}, \qquad P\mathbf{1} = \mathbf{1} - e_1,$$

where e_1 is the first standard basis vector. Substituting equation 1 gives

$$E^{(\ell)}\mathbf{1} = \left(1 - \mathbf{1}^\top c^{(\ell)} - \gamma_\ell\right)\mathbf{1} + \gamma_\ell e_1 \tag{29}$$

Hence, for every interior row $i \geq 2$,

$$\sum_j E^{(\ell)}_{ij} = 1 - \mathbf{1}^\top c^{(\ell)} - \gamma_\ell,$$

while the first row contains the boundary correction $\sum_j E^{(\ell)}_{1j} = 1 - \mathbf{1}^\top c^{(\ell)}$. Thus the original normalization relation holds exactly away from the single sequence-boundary row.

For an effective anchor set $\mathcal{A}$,

$$C^{(\ell)}_{\mathcal{A}} := \sum_{j \in \mathcal{A}} c^{(\ell)}_j.$$

Table 5: Distribution of consecutive token reveal pairs across decoding stages. All entries are percentages of the reveal pairs within each stage.

Stage	**H–H**	**T–T**	**U–U**	**Same-Slot**
1/3	93.6%	0.1%	0.1%	93.8%
2/3	0.4%	63.6%	4.9%	69.0%
3/3	0.0%	7.3%	67.7%	75.0%

Table 6: Relative reconstruction residual of the stripe–permutation model across depth groups. Lower values indicate a closer fit. We divide the 32 Transformer layers into shallow ($0 \le \ell < 8$), middle ($8 \le \ell < 23$), and deep ($23 \le \ell < 32$) groups.

Depth group	Mean ρ_E	Median ρ_E
Shallow	0.303	0.286
Middle	0.702	0.707
Deep	0.638	0.627

Under $c^{(\ell)} \ge 0$, $\gamma_\ell \ge 0$, and $\mathbf{1}^\top c^{(\ell)} + \gamma_\ell \le 1$, the constrained decomposition gives

$$0 \le C_{\mathcal{A}}^{(\ell)} \le \mathbf{1}^\top c^{(\ell)} \le 1 - \gamma_\ell \le 1. \tag{30}$$

Hence $C_{\mathcal{A}}^{(\ell)}$ is a bounded shared-stripe mass assigned to the anchor set, as used in the reduced dynamics of Section 4.

Reconstruction fidelity across depth. We evaluate the structured attention model introduced in Section 3.3 using the relative reconstruction residual

$$\rho_E = \frac{\|E\|_F}{\|A\|_F}, \qquad E = A - \mathbf{1}c^\top - \gamma P. \tag{31}$$

This quantity measures how closely the shared stripe and shifted-diagonal components approximate the complete attention operator. We compute it separately for each recorded attention matrix and report the mean and median within each depth group. As shown in Table 6, the approximation is most accurate in the shallow layers, where the mean and median relative residuals are 0.303 and 0.286, respectively. These layers are also the region targeted by our SVD intervention. In addition, during the experiment we find that in the shallow layers 70.15% of the observation have $\rho_E \le 0.5$ and 29.27% of the observation have $\rho_E \le 0.2$. The residual increases in the middle layers and decreases somewhat in the deep layers, while remaining substantially above the shallow-layer level. Thus, the measured geometry supports a shallow-layer structural approximation, with more substantial interactions outside the two modeled components at greater depth.

Additional periodic structure at depth. However, this higher residual does not signal a breakdown of our baseline geometry. Rather, the primary anchor and diagonal patterns persist in deeper layers, augmented by auxiliary shifted diagonals P_r with stride $\tau = 4$ (corresponding to matching identical digit slots across prior steps):

$$A = \underbrace{\mathbf{1}c^\top + \gamma P}_{\text{baseline model}} + \underbrace{\sum_{r \in \mathcal{R}} \eta_r P_r}_{\text{auxiliary diagonals}} + E, \tag{32}$$

where $\mathcal{R}$ denotes the set of periodic stride offsets. Incorporating these stride-4 diagonals lowers the deep-layer median relative residual substantially to 0.354.

While multi-stride diagonals offer a tighter numerical fit at depth, we restrict our formal dynamic model (Section 3.3) to the primary diagonal P. We acknowledge this as a theoretical simplification. Nevertheless, because higher-order diagonals induce qualitatively identical positive-feedback mechanisms as P, omitting them maintains analytical tractability without altering the fundamental physics of the representation collapse.

B Detailed Proofs for Anchor Concentration

B.1 Proof of Lemma 1

The i-th row of equation 1 gives

$$A_{ij}^{(\ell)} = c_j^{(\ell)} + \gamma_\ell P_{ij} + E_{ij}^{(\ell)}, \qquad P_{ij} = \mathbf{1}\{j = p(i)\}. \tag{33}$$

The analyzed layer update can therefore be written as

$$\begin{aligned}\Delta h_i^{(\ell)} &= W_O^{(\ell)} \sum_{j=1}^{N} A_{ij}^{(\ell)} v_j^{(\ell)} \\ &= W_O^{(\ell)} \sum_{j=1}^{N} c_j^{(\ell)} v_j^{(\ell)} + \gamma_\ell W_O^{(\ell)} v_{p(i)}^{(\ell)} + W_O^{(\ell)} \sum_{j=1}^{N} E_{ij}^{(\ell)} v_j^{(\ell)}. \end{aligned} \tag{34}$$

For $C_A^{(\ell)} > 0$, define the stripe-weighted anchor value

$$v_A^{(\ell)} := \frac{1}{C_A^{(\ell)}} \sum_{j \in \mathcal{A}} c_j^{(\ell)} v_j^{(\ell)}. \tag{35}$$

Split the shared stripe into anchor and background parts:

$$\begin{aligned}\sum_{j=1}^{N} c_j^{(\ell)} v_j^{(\ell)} &= \sum_{j \in \mathcal{A}} c_j^{(\ell)} v_j^{(\ell)} + \sum_{j \in \mathcal{O}} c_j^{(\ell)} v_j^{(\ell)} \\ &= C_A^{(\ell)} v_A^{(\ell)} + \sum_{j \in \mathcal{O}} c_j^{(\ell)} v_j^{(\ell)}. \end{aligned} \tag{36}$$

Define

$$u_{R,i}^{(\ell)} := W_O^{(\ell)} \sum_{j \in \mathcal{O}} c_j^{(\ell)} v_j^{(\ell)} + W_O^{(\ell)} \sum_{j=1}^{N} E_{ij}^{(\ell)} v_j^{(\ell)} + r_{\mathrm{blk},i}^{(\ell)}, \tag{37}$$

where $r_{\mathrm{blk},i}^{(\ell)}$ contains any additional block contribution included in the transition. Substituting equation 36 into equation 34 proves equation 2. If the analyzed $\Delta h_i^{(\ell)}$ contains only the attention output, set $r_{\mathrm{blk},i}^{(\ell)} = 0$.

B.2 Proof of Lemma 3

Recall the score increment $\Delta s_{ij}^{(\ell)}$ and the anchor–background gain $G_i^{(\ell)}$ defined in Appendix B.3. Define the corresponding softmax attention weights and query-level anchor mass by

$$p_{ij}^{(\ell)} := \frac{e^{\beta s_{ij}^{(\ell)}}}{\sum_m e^{\beta s_{im}^{(\ell)}}}, \qquad C_i^{(\ell)} := \sum_{j \in \mathcal{A}} p_{ij}^{(\ell)}.$$

We take $\mathcal{A}$ and $\mathcal{O}$ to be nonempty. For finite softmax scores, $0 < C_i^{(\ell)} < 1$. Under the same two-group reduction,

$$\Delta s_{ij}^{(\ell)} \approx \Delta s_{i,A}^{(\ell)} \quad (j \in \mathcal{A}), \qquad \Delta s_{ij}^{(\ell)} \approx \Delta s_{i,O}^{(\ell)} \quad (j \in \mathcal{O}),$$

with $G_i^{(\ell)} = \Delta s_{i,A}^{(\ell)} - \Delta s_{i,O}^{(\ell)}$.

By definition,

$$s_{ij}^{(\ell+1)} = s_{ij}^{(\ell)} + \Delta s_{ij}^{(\ell)}. \tag{38}$$

Thus

$$p_{ij}^{(\ell+1)} = \frac{\exp\left(\beta s_{ij}^{(\ell)} + \beta \Delta s_{ij}^{(\ell)}\right)}{\sum_m \exp\left(\beta s_{im}^{(\ell)} + \beta \Delta s_{im}^{(\ell)}\right)} = \frac{e^{\beta s_{ij}^{(\ell)}} e^{\beta \Delta s_{ij}^{(\ell)}}}{\sum_m e^{\beta s_{im}^{(\ell)}} e^{\beta \Delta s_{im}^{(\ell)}}}. \tag{39}$$

Divide the numerator and denominator by $\sum_n e^{\beta s_{in}^{(\ell)}}$. Since

$$p_{ij}^{(\ell)} = \frac{e^{\beta s_{ij}^{(\ell)}}}{\sum_n e^{\beta s_{in}^{(\ell)}}}, \tag{40}$$

we obtain the exact identity

$$p_{ij}^{(\ell+1)} = \frac{p_{ij}^{(\ell)} e^{\beta \Delta s_{ij}^{(\ell)}}}{\sum_m p_{im}^{(\ell)} e^{\beta \Delta s_{im}^{(\ell)}}}. \tag{41}$$

Sum over $j \in \mathcal{A}$. Under the group-homogeneous approximation stated in Lemma 3, this gives

$$C_i^{(\ell+1)} \approx \frac{e^{\beta \Delta s_{i,A}^{(\ell)}} \sum_{j \in \mathcal{A}} p_{ij}^{(\ell)}}{e^{\beta \Delta s_{i,A}^{(\ell)}} \sum_{m \in \mathcal{A}} p_{im}^{(\ell)} + e^{\beta \Delta s_{i,O}^{(\ell)}} \sum_{m \in \mathcal{O}} p_{im}^{(\ell)}} = \frac{C_i^{(\ell)} e^{\beta \Delta s_{i,A}^{(\ell)}}}{C_i^{(\ell)} e^{\beta \Delta s_{i,A}^{(\ell)}} + \left(1 - C_i^{(\ell)}\right) e^{\beta \Delta s_{i,O}^{(\ell)}}}. \tag{42}$$

Dividing by $e^{\beta \Delta s_{i,O}^{(\ell)}}$ and using $G_i^{(\ell)} = \Delta s_{i,A}^{(\ell)} - \Delta s_{i,O}^{(\ell)}$ proves equation 4.

B.3 Proof of Lemma 2

For query i, define the attention score

$$s_{ij}^{(\ell)} := k_j^{(\ell)\top} W_Q^{(\ell)} h_i^{(\ell)},$$

and its layerwise increment

$$\Delta s_{ij}^{(\ell)} := s_{ij}^{(\ell+1)} - s_{ij}^{(\ell)}.$$

Under the two-group reduction, we approximate the increments within the anchor and background sets by representative values,

$$\Delta s_{ij}^{(\ell)} \approx \Delta s_{i,A}^{(\ell)} \quad (j \in \mathcal{A}), \qquad \Delta s_{ij}^{(\ell)} \approx \Delta s_{i,O}^{(\ell)} \quad (j \in \mathcal{O}),$$

and define the anchor–background gain

$$G_i^{(\ell)} := \Delta s_{i,A}^{(\ell)} - \Delta s_{i,O}^{(\ell)}.$$

Let $k_A^{(\ell)}$ and $k_O^{(\ell)}$ denote the representative key directions under the group-homogeneous reduction. Then

$$s_{i,A}^{(\ell)} \approx k_A^{(\ell)\top} W_Q^{(\ell)} h_i^{(\ell)}, \qquad s_{i,O}^{(\ell)} \approx k_O^{(\ell)\top} W_Q^{(\ell)} h_i^{(\ell)}. \tag{43}$$

Using $G_i^{(\ell)} = \Delta s_{i,A}^{(\ell)} - \Delta s_{i,O}^{(\ell)}$ gives

$$G_i^{(\ell)} \approx \left(k_A^{(\ell+1)} - k_O^{(\ell+1)}\right)^\top W_Q^{(\ell+1)} h_i^{(\ell+1)} - \left(k_A^{(\ell)} - k_O^{(\ell)}\right)^\top W_Q^{(\ell)} h_i^{(\ell)}. \tag{44}$$

Substitute $h_i^{(\ell+1)} = h_i^{(\ell)} + \Delta h_i^{(\ell)}$:

$$G_i^{(\ell)} \approx \left(k_A^{(\ell+1)} - k_O^{(\ell+1)}\right)^\top W_Q^{(\ell+1)} \Delta h_i^{(\ell)} + \Big[\left(k_A^{(\ell+1)} - k_O^{(\ell+1)}\right)^\top W_Q^{(\ell+1)} - \left(k_A^{(\ell)} - k_O^{(\ell)}\right)^\top W_Q^{(\ell)}\Big] h_i^{(\ell)}. \tag{45}$$

Define

$$S_i^{(\ell)} := \left(k_A^{(\ell+1)} - k_O^{(\ell+1)}\right)^\top W_Q^{(\ell+1)} W_O^{(\ell)} v_A^{(\ell)},$$

and

$$D_i^{(\ell)} := \left(k_A^{(\ell+1)} - k_O^{(\ell+1)}\right)^\top W_Q^{(\ell+1)} W_O^{(\ell)} v_{p(i)}^{(\ell)}.$$

Now substitute equation 2. The stripe and diagonal terms become

$$C_A^{(\ell)} \left(k_A^{(\ell+1)} - k_O^{(\ell+1)}\right)^\top W_Q^{(\ell+1)} W_O^{(\ell)} v_A^{(\ell)} = C_A^{(\ell)} S_i^{(\ell)}, \tag{46}$$

$$\gamma_\ell \left(k_A^{(\ell+1)} - k_O^{(\ell+1)}\right)^\top W_Q^{(\ell+1)} W_O^{(\ell)} v_{p(i)}^{(\ell)} = \gamma_\ell D_i^{(\ell)}. \tag{47}$$

Define the remaining contribution as

$$R_i^{(\ell)} := \left(k_A^{(\ell+1)} - k_O^{(\ell+1)}\right)^\top W_Q^{(\ell+1)} u_{R,i}^{(\ell)} + \Big[\left(k_A^{(\ell+1)} - k_O^{(\ell+1)}\right)^\top W_Q^{(\ell+1)} - \left(k_A^{(\ell)} - k_O^{(\ell)}\right)^\top W_Q^{(\ell)}\Big] h_i^{(\ell)}. \tag{48}$$

Collecting equation 46– equation 48 proves equation 3.

B.4 Proof of Corollary 1

Starting from equation 4,

$$\begin{aligned} C_{\ell+1} - C_\ell &= \frac{C_\ell e^{\beta G_\ell}}{C_\ell e^{\beta G_\ell} + 1 - C_\ell} - C_\ell \\ &= \frac{C_\ell e^{\beta G_\ell} - C_\ell \left(C_\ell e^{\beta G_\ell} + 1 - C_\ell\right)}{C_\ell e^{\beta G_\ell} + 1 - C_\ell} \\ &= \frac{C_\ell (1 - C_\ell)\left(e^{\beta G_\ell} - 1\right)}{C_\ell e^{\beta G_\ell} + 1 - C_\ell}. \end{aligned} \tag{49}$$

Because the reduced state is defined on $(0, 1)$, the denominator and $C_\ell(1 - C_\ell)$ are positive. Hence the increment has the sign of G_ℓ, proving Corollary 1.

B.5 Proof of Lemma 4

Next,

$$1 - C_{\ell+1} = \frac{1 - C_\ell}{C_\ell e^{\beta G_\ell} + 1 - C_\ell}. \tag{50}$$

Dividing this expression by $C_{\ell+1}$ gives

$$\frac{1 - C_{\ell+1}}{C_{\ell+1}} = \frac{1 - C_\ell}{C_\ell} e^{-\beta G_\ell}. \tag{51}$$

Iterating the complement-to-anchor ratio in equation 51 from ℓ_0 to $L-1$ proves equation 6.

B.6 Proof of Theorem 1

equation 6 gives the exact finite-depth identity

$$\frac{1 - C_L}{C_L} = \frac{1 - C_{\ell_0}}{C_{\ell_0}} \exp\left(-\beta \sum_{\ell=\ell_0}^{L-1} G_\ell\right). \tag{52}$$

Solving for C_L yields

$$\boxed{C_L = \left[1 + \frac{1 - C_{\ell_0}}{C_{\ell_0}} \exp\left(-\beta \sum_{\ell=\ell_0}^{L-1} G_\ell\right)\right]^{-1}}. \tag{53}$$

Consequently, for any finite depth L,

$$\sum_{\ell=\ell_0}^{L-1} G_\ell > 0 \quad \Longrightarrow \quad C_L > C_{\ell_0}. \tag{54}$$

This conclusion only requires positive cumulative gain; individual layers may have negative gain. If instead $G_\ell > 0$ at every analyzed layer, Corollary 1 additionally gives strict layerwise monotonicity.

Because $0 < C_L \leq 1$, equation 52 also gives

$$1 - C_L \leq \frac{1 - C_{\ell_0}}{C_{\ell_0}} \exp\left(-\beta \sum_{\ell=\ell_0}^{L-1} G_\ell\right), \tag{55}$$

which is the relevant concentration bound for a finite-depth Transformer.

For completeness, if

$$\sum_{\ell=\ell_0}^{L-1} G_\ell \longrightarrow +\infty \tag{56}$$

as $L \to \infty$, the exponential term in equation 53 converges to zero and therefore $C_L \to 1$. A sufficient condition is the existence of $\delta > 0$ such that $G_\ell \geq \delta$ for every $\ell \geq \ell_0$, because then

$$\sum_{\ell=\ell_0}^{L-1} G_\ell \geq \delta(L - \ell_0). \tag{57}$$

Thus, positive cumulative gain controls finite-depth anchor amplification, while divergence of cumulative gain gives the corresponding idealized infinite-depth concentration result.

Empirical connection. The theorem remains conditional because its gain is defined at the query-level reduced dynamics. However, the measured depth-cumulative gain is positive throughout the analyzed repetition-prone rollout, placing the observed model dynamics in the finite-depth positive-gain regime of equation 54. The measurement protocol, statistical granularity, and correspondence with fitted anchor-mass changes are reported in Appendix D.

C Detailed Proofs for Representation Convergence

C.1 Sufficient Value Bound under RMSNorm

Suppose the model uses pre-RMSNorm,

$$v_j^{(\ell)} = W_V^{(\ell)} \,\mathrm{RMSNorm}^{(\ell)}\left(h_j^{(\ell)}\right), \tag{58}$$

with

$$\mathrm{RMSNorm}^{(\ell)}(h) = g^{(\ell)} \odot \frac{h}{\sqrt{\|h\|_2^2/d + \varepsilon}}. \tag{59}$$

Then

$$\begin{aligned}\left\|\mathrm{RMSNorm}^{(\ell)}(h)\right\|_2 &\le \|g^{(\ell)}\|_\infty \frac{\|h\|_2}{\sqrt{\|h\|_2^2/d + \varepsilon}} \\ &\le \sqrt{d}\,\|g^{(\ell)}\|_\infty.\end{aligned} \tag{60}$$

Since $V_j^{(\ell)} = W_O^{(\ell)} v_j^{(\ell)}$,

$$\begin{aligned}\|V_j^{(\ell)}\|_2 &\le \|W_O^{(\ell)}\|_{\mathrm{op}} \|W_V^{(\ell)}\|_{\mathrm{op}} \left\|\mathrm{RMSNorm}^{(\ell)}\left(h_j^{(\ell)}\right)\right\|_2 \\ &\le \sqrt{d}\,\|W_O^{(\ell)}\|_{\mathrm{op}} \|W_V^{(\ell)}\|_{\mathrm{op}} \|g^{(\ell)}\|_\infty.\end{aligned} \tag{61}$$

Therefore, Assumption 1 holds whenever

$$B := \sup_\ell \left[\sqrt{d}\,\|W_O^{(\ell)}\|_{\mathrm{op}} \|W_V^{(\ell)}\|_{\mathrm{op}} \|g^{(\ell)}\|_\infty\right] < \infty. \tag{62}$$

C.2 Proof of Lemma 5

Recall the projected values

$$V_j^{(\ell)} := W_O^{(\ell)} v_j^{(\ell)}.$$

For query i, define the attention readout

$$H_i^{(\ell+1)} := \sum_{j=1}^{N} A_{ij}^{(\ell)} V_j^{(\ell)},$$

and the query-level anchor mass

$$C_i^{(\ell)} := \sum_{j \in \mathcal{A}} A_{ij}^{(\ell)}.$$

Since $\mathcal{A}$ and $\mathcal{O}$ are nonempty and the attention weights are strictly positive, $0 < C_i^{(\ell)} < 1$. We therefore define the anchor- and background-conditioned readout centers by

$$V_{A,i}^{(\ell)} := \frac{\sum_{j\in\mathcal{A}} A_{ij}^{(\ell)} V_j^{(\ell)}}{C_i^{(\ell)}}, \qquad V_{O,i}^{(\ell)} := \frac{\sum_{j\in\mathcal{O}} A_{ij}^{(\ell)} V_j^{(\ell)}}{1 - C_i^{(\ell)}}. \tag{63}$$

These are the normalized attention-weighted means of the projected values within the anchor and background sets, respectively.

Splitting the attention readout over $\mathcal{A}$ and $\mathcal{O}$ gives

$$\begin{aligned} H_i^{(\ell+1)} &= \sum_{j\in\mathcal{A}} A_{ij}^{(\ell)} V_j^{(\ell)} + \sum_{j\in\mathcal{O}} A_{ij}^{(\ell)} V_j^{(\ell)} \\ &= C_i^{(\ell)} V_{A,i}^{(\ell)} + \left(1 - C_i^{(\ell)}\right) V_{O,i}^{(\ell)}. \end{aligned} \tag{64}$$

The normalized weights

$$\alpha_{ij}^{(\ell)} := \frac{A_{ij}^{(\ell)}}{C_i^{(\ell)}}, \qquad j \in \mathcal{A}, \tag{65}$$

are nonnegative and sum to one. Thus $V_{A,i}^{(\ell)}$ is a convex combination of the anchor values and

$$\|V_{A,i}^{(\ell)}\| \le B. \tag{66}$$

The same argument gives

$$\|V_{O,i}^{(\ell)}\| \le B. \tag{67}$$

Subtracting $V_{A,i}^{(\ell)}$ from equation 64,

$$H_i^{(\ell+1)} - V_{A,i}^{(\ell)} = \left(1 - C_i^{(\ell)}\right)\left(V_{O,i}^{(\ell)} - V_{A,i}^{(\ell)}\right). \tag{68}$$

Taking norms and applying equation 66 and equation 67 yields

$$\begin{aligned} \left\| H_i^{(\ell+1)} - V_{A,i}^{(\ell)} \right\| &\le \left(1 - C_i^{(\ell)}\right)\left(\|V_{O,i}^{(\ell)}\| + \|V_{A,i}^{(\ell)}\|\right) \\ &\le 2B\left(1 - C_i^{(\ell)}\right), \end{aligned} \tag{69}$$

which proves equation 8.

C.3 Positionwise Alignment from Cumulative Gain

Lemma 5 directly yields a positionwise consequence of anchor concentration. For every query position i,

$$\left\| H_i^{(\ell+1)} - V_{A,i}^{(\ell)} \right\| \le 2B\left(1 - C_i^{(\ell)}\right).$$

as mentioned in theorem 1, when $C_i^{(\ell)} \to 1$,

$$\boxed{\left\| H_i^{(\ell+1)} - V_{A,i}^{(\ell)} \right\| \longrightarrow 0.}$$

Thus, increasing anchor concentration forces the attention readout at each query toward its own anchor-conditioned center.

A finite-depth rate follows by combining Lemma 5 with the cumulative-gain result of Theorem 1. Suppose that, from some layer ℓ_0 onward,

$$G_i^{(\ell)} \geq \delta > 0.$$

Then the query-level version of the cumulative odds identity gives

$$\frac{1 - C_i^{(L)}}{C_i^{(L)}} = \frac{1 - C_i^{(\ell_0)}}{C_i^{(\ell_0)}} \exp\left(-\beta \sum_{\ell=\ell_0}^{L-1} G_i^{(\ell)} \right).$$

Since $G_i^{(\ell)} \geq \delta$,

$$\frac{1 - C_i^{(L)}}{C_i^{(L)}} \leq \frac{1 - C_i^{(\ell_0)}}{C_i^{(\ell_0)}} e^{-\beta\delta(L-\ell_0)}.$$

Using $C_i^{(L)} \leq 1$ therefore yields

$$1 - C_i^{(L)} \leq \frac{1 - C_i^{(\ell_0)}}{C_i^{(\ell_0)}} e^{-\beta\delta(L-\ell_0)}.$$

Substituting this bound into Lemma 5 gives

$$\boxed{\left\| H_i^{(L+1)} - V_{A,i}^{(L)} \right\| \leq 2B \frac{1 - C_i^{(\ell_0)}}{C_i^{(\ell_0)}} e^{-\beta\delta(L-\ell_0)}.}$$

This result establishes not only local positionwise alignment, but also an explicit exponential convergence rate with respect to network depth L.

C.4 Proof of Corollary 2

We make precise the two uniform conditions stated in Corollary 2. For $C_A^{(\ell)} > 0$, define the stripe-induced anchor center

$$\bar{V}_A^{(\ell)} := \frac{\sum_{j \in \mathcal{A}} c_j^{(\ell)} V_j^{(\ell)}}{C_A^{(\ell)}}. \tag{70}$$

Because the stripe profile $c^{(\ell)}$ is shared across queries, $\bar{V}_A^{(\ell)}$ is independent of i. Moreover, since $c_j^{(\ell)} \geq 0$ and $\sum_{j \in \mathcal{A}} c_j^{(\ell)} = C_A^{(\ell)}$, it is a convex combination of the projected anchor values. Assumption 1 therefore gives

$$\left\| \bar{V}_A^{(\ell)} \right\| \leq B. \tag{71}$$

For readability, temporarily suppress the layer index. Using

$$A_{ij} = c_j + \gamma P_{ij} + E_{ij},$$

the query-specific anchor center defined in Appendix C.2 satisfies

$$C_i V_{A,i} = C_A \bar{V}_A + \gamma \sum_{j \in \mathcal{A}} P_{ij} V_j + \sum_{j \in \mathcal{A}} E_{ij} V_j. \tag{72}$$

Similarly,

$$C_i - C_A = \gamma \sum_{j \in \mathcal{A}} P_{ij} + \sum_{j \in \mathcal{A}} E_{ij}. \tag{73}$$

Combining equation 72 and equation 73 yields the exact identity

$$\boxed{C_i \left(V_{A,i} - \bar{V}_A \right) = \gamma \sum_{j \in \mathcal{A}} P_{ij} \left(V_j - \bar{V}_A \right) + \sum_{j \in \mathcal{A}} E_{ij} \left(V_j - \bar{V}_A \right).} \tag{74}$$

equation 74 shows that deviations of the query-specific anchor centers from the common stripe center arise only from the shifted-predecessor and residual components. Define their combined anchor-restricted correction by

$$\xi_i^{(\ell)} := \gamma_\ell \sum_{j\in\mathcal{A}} P_{ij} \left(V_j^{(\ell)} - \bar{V}_A^{(\ell)}\right) + \sum_{j\in\mathcal{A}} E_{ij}^{(\ell)} \left(V_j^{(\ell)} - \bar{V}_A^{(\ell)}\right). \tag{75}$$

Then

$$V_{A,i}^{(\ell)} - \bar{V}_A^{(\ell)} = \frac{\xi_i^{(\ell)}}{C_i^{(\ell)}}. \tag{76}$$

Now let

$$\underline{C}_\ell := \min_{i\in\mathcal{M}} C_i^{(\ell)}, \qquad \xi_{\mathcal{M},\ell} := \max_{i\in\mathcal{M}} \left\|\xi_i^{(\ell)}\right\|.$$

Uniform anchor concentration means

$$\underline{C}_\ell \longrightarrow 1,$$

while vanishing query-specific deviations from the shared stripe means

$$\xi_{\mathcal{M},\ell} \longrightarrow 0.$$

equation 76 therefore gives

$$\max_{i\in\mathcal{M}} \left\|V_{A,i}^{(\ell)} - \bar{V}_A^{(\ell)}\right\| \le \frac{\xi_{\mathcal{M},\ell}}{\underline{C}_\ell} \longrightarrow 0. \tag{77}$$

Lemma 5 further gives

$$\max_{i\in\mathcal{M}} \left\|H_i^{(\ell+1)} - V_{A,i}^{(\ell)}\right\| \le 2B\,(1-\underline{C}_\ell) \longrightarrow 0. \tag{78}$$

For any $i, k \in \mathcal{M}$, the triangle inequality gives

$$\begin{aligned}\left\|H_i^{(\ell+1)} - H_k^{(\ell+1)}\right\| \le\ & \left\|H_i^{(\ell+1)} - V_{A,i}^{(\ell)}\right\| \\ &+ \left\|V_{A,i}^{(\ell)} - \bar{V}_A^{(\ell)}\right\| \\ &+ \left\|V_{A,k}^{(\ell)} - \bar{V}_A^{(\ell)}\right\| \\ &+ \left\|V_{A,k}^{(\ell)} - H_k^{(\ell+1)}\right\|.\end{aligned} \tag{79}$$

Taking the maximum over $i, k \in \mathcal{M}$ and applying equation 77 and equation 78 yields the finite-depth bound

$$\boxed{\max_{i,k\in\mathcal{M}} \left\|H_i^{(\ell+1)} - H_k^{(\ell+1)}\right\| \le 4B\,(1-\underline{C}_\ell) + \frac{2\xi_{\mathcal{M},\ell}}{\underline{C}_\ell}.} \tag{80}$$

Hence, if $\underline{C}_\ell \to 1$ and $\xi_{\mathcal{M},\ell} \to 0$, then

$$\max_{i,k\in\mathcal{M}} \left\|H_i^{(\ell+1)} - H_k^{(\ell+1)}\right\| \longrightarrow 0,$$

which proves Corollary 2.

C.5 Empirical hidden-state geometry across depth

To examine the empirical relevance of the convergence analysis, we measure cross-position cosine distance within each recorded hidden-state depth. For rollout r, diffusion step t, and recorded depth j, let $h_{r,t,i}^{(j)}$ denote the hidden state at forecast position i. For a position group $\mathcal{I}_g$ of size $n_g \ge 2$, we compute

$$D_g^{(j)}(r,t) = \frac{2}{n_g(n_g-1)} \sum_{\substack{i<k \\ i,k\in\mathcal{I}_g}} \left(1 - \frac{\langle h_{r,t,i}^{(j)}, h_{r,t,k}^{(j)}\rangle}{\|h_{r,t,i}^{(j)}\|_2 \|h_{r,t,k}^{(j)}\|_2}\right). \tag{81}$$

Here, $g \in \{H, T, U\}$ denotes hundreds, tens, or units positions. We aggregate the distances across rollouts at each diffusion step and then average the step-level results at each depth. This statistic measures angular alignment rather than the maximum Euclidean distance appearing in the theoretical convergence result. While our formal proof employs the L_2 norm to maintain analytical tractability, evaluating directional alignment via cosine distance is empirically more faithful in high-dimensional representation spaces, where raw L_2 metrics are easily distorted by layer-wise norm scaling and magnitude drift.

Table 7: Mean pairwise cosine distance at selected recorded depths. Depth indices follow `hidden_index` in the measurement records. Lower values indicate stronger directional alignment.

Recorded depth j	**Hundreds** H	**Tens** T	**Units** U
0	0.1762	0.2652	0.1922
1	0.0192	0.0299	0.0198
8	0.0214	0.0338	0.0278
16	0.0673	0.0834	0.0885
23	0.1161	0.1258	0.1390
25	0.1275	0.1511	0.1585
28	0.0947	0.1363	0.1473
31	0.0777	0.1120	0.1189
32	0.0961	0.1209	0.1693

Depth-dependent alignment. The trajectories are non-monotonic. A pronounced initial contraction is followed by increasing distances through much of the network. Local contraction reappears at greater depth: between recorded depths 25 and 31, within-slot distances decrease by 39.1%, 25.9%, and 25.0% for hundreds, tens, and units respectively. All groups subsequently increase at the final recorded depth. Thus, the measurements reveal localized angular homogenization, particularly within slot families, rather than uniform contraction throughout the network.

Connection to the structured attention model. Table 6 shows that the stripe–permutation approximation also has depth-dependent fidelity: the mean relative residual is 0.303 in shallow layers, increases to 0.702 in middle layers, and decreases to 0.638 in deep layers. The stronger shallow-layer fit and partial deep-layer recovery, considered alongside the initial and local late-stage contractions, are qualitatively consistent with a mechanism whose empirical relevance varies across depth.

The theoretical results characterize a sufficient route to homogenization under the stated dynamical and center-alignment conditions; they do not require every observed distance to decrease at every layer. Departures from the structured regime therefore limit the direct applicability of the reduced dynamics without constituting a counterexample to the conditional results. Together with the measured gain reductions and repetition mitigation under SVD and history-sparsified guidance, these observations provide qualitative and intervention-based support for the proposed mechanism, while not establishing uniform Euclidean convergence of the full network.

D Empirical Verification of the Positive-Gain Regime

The anchor-concentration theorem is stated in terms of the query-level anchor–background gain $G_i^{(\ell)}$. We therefore evaluate whether the analyzed model trajectories empirically occupy a positive-gain regime and whether the measured gain tracks the corresponding change in anchor mass.

D.1 Gain Estimation

At diffusion step t, layer ℓ, head h, and query position i, we compute the empirical anchor–background score gap

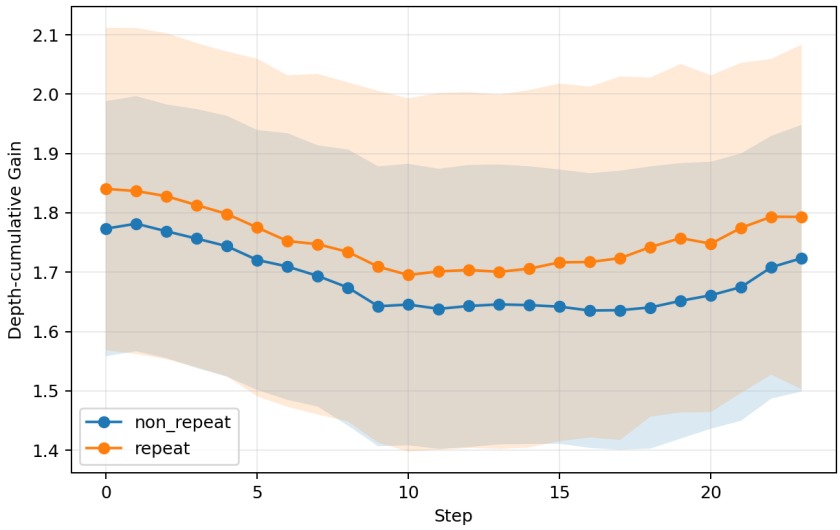

Figure 3: **Depth-cumulative gain across diffusion steps for repeating and non-repeating rollouts.** At each diffusion step, gain is accumulated over Transformer depth. Repeating rollouts exhibit consistently larger depth-cumulative gain than non-repeating rollouts across the decoding trajectory. Shaded regions show the corresponding variability across evaluated rollouts.

Table 8: **Cumulative-gain comparison between repeating and non-repeating rollouts.** Rollouts are grouped according to whether the vanilla prediction exhibits repetition. Repeating rollouts show a larger cumulative gain than non-repeating rollouts, consistent with the cumulative-gain mechanism in Section 4.2.

Rollout group	Cumulative gain	Gain difference	Total repeat rate	R^{H}_{slot}	R^{T}_{slot}	R^{U}_{slot}
Non-repeating	40.450	–	0.239	0.856	0.413	0.327
Repeating	42.298	+4.569%	1.000	1.000	0.997	0.994

$$\widehat{g}_{t,\ell,h,i} = \frac{q_{t,\ell,h,i}^{\top}\left(k_{A,t,\ell,h} - k_{O,t,\ell,h}\right)}{\sqrt{d_h}}, \tag{82}$$

where k_A and k_O are the detected stripe-weighted anchor and background key prototypes.The measured layerwise gain is

$$\widehat{G}_{t,\ell,h,i} = \widehat{g}_{t,\ell+1,h,i} - \widehat{g}_{t,\ell,h,i}. \tag{83}$$

A head is considered valid when at least two anchor positions are detected and both the anchor and background stripe masses exceed 10^{-8}. Let $\mathcal{H}_{t,\ell}$ denote the valid-head set and let Ω denote the reported query scope. We define the depth-cumulative gain at diffusion step t as

$$\widehat{\mathcal{G}}_t = \sum_{\ell=0}^{L-2} \operatorname*{Mean}_{h\in\mathcal{H}_{t,\ell},\, i\in\Omega} \widehat{G}_{t,\ell,h,i}. \tag{84}$$

For intervention comparisons, $\mathcal{H}_{t,\ell}$ is restricted to the intersection of valid layer–head transitions shared by the paired conditions. This prevents apparent gain differences from being produced by different numbers of valid heads.

Table 9: Association between measured gain and the layerwise change in fitted anchor-mass log-odds, pooled across valid transitions in the tested rollout.

Condition	Pearson r	Spearman ρ
Vanilla	0.838	0.864
Quarter-rank SVD	0.823	0.862
History-sparsified CFG	0.820	0.868

D.2 Positive Gain in Repetition-Prone Decoding

The gain summed over the ten analyzed diffusion steps is positive in the representative vanilla rollout.

$$\widehat{\mathcal{G}}t > 0, \qquad \forall t \in \mathcal{T}\text{obs}. \tag{85}$$

where $\mathcal{T}_{\text{obs}}$ denotes the set of diffusion steps actually executed and analyzed along the rollout. Thus, every analyzed forward pass lies in the positive cumulative-gain regime of equation 54. Under the reduced dynamics, this predicts an increase in anchor log-odds through Transformer depth at each diffusion step.

This statement concerns the aggregate statistic in equation 84. It does not claim that $\widehat{G}_{t,\ell,h,i} > 0$ for every individual layer, head, and query. Accordingly, the experiment directly supports finite-depth anchor amplification, while the infinite-depth limit remains an idealized corollary of the theory.

D.3 Association with Anchor-Mass Amplification

For every valid layer transition, we compare the measured gain with

$$\Delta \operatorname{logit} \widehat{C}_A^{(\ell)} = \operatorname{logit} \widehat{C}_A^{(\ell+1)} - \operatorname{logit} \widehat{C}_A^{(\ell)}, \tag{86}$$

where $\widehat{C}_A$ is the fitted stripe-anchor mass.

The consistently strong association supports the monotone gain–concentration relationship predicted by the reduced dynamics (Table 9).

E Additional Analysis of Gain-Reducing Interventions

E.1 Proof of Lemma 6

The scalar anchor dynamics are

$$C_{\ell+1} = \frac{C_\ell e^{\beta G_\ell}}{C_\ell e^{\beta G_\ell} + 1 - C_\ell}. \tag{87}$$

Subtracting this expression from one gives

$$1 - C_{\ell+1} = \frac{1 - C_\ell}{C_\ell e^{\beta G_\ell} + 1 - C_\ell}, \tag{88}$$

which proves equation 10. At fixed C_ℓ,

$$\frac{\partial \eta_\ell}{\partial G_\ell} = -\frac{\beta C_\ell e^{\beta G_\ell}}{(C_\ell e^{\beta G_\ell} + 1 - C_\ell)^2} < 0. \tag{89}$$

If $G(C)$ is continuous at $C = 1$, then

Table 10: **Cumulative-gain reduction on repeating rollouts.** Results are computed on paired repeating rollouts. Both SVD and HSCFG reduce cumulative gain and repetition relative to their matched vanilla counterparts.

	Cumulative gain	Gain reduction	NMAE	Total repeat rate	R^H_{slot}	R^T_{slot}	R^U_{slot}
Vanilla	42.298	–	1.010	1.000	1.000	0.997	0.994
Quarter-rank SVD	40.936	-3.22%	1.007	0.701	0.913	0.770	0.747
History-sparsified CFG	41.016	-3.03%	1.010	0.132	0.519	0.228	0.197

$$\lim_{C_\ell \to 1} \frac{1 - C_{\ell+1}}{1 - C_\ell} = e^{-\beta G(1)}. \tag{90}$$

For a non-autonomous layer sequence, this gives the local approximation $\eta_\ell \approx e^{-\beta G_\ell}$ when $C_\ell \approx 1$. equation 87 also yields the exact log-odds update

$$\operatorname{logit} C_{\ell+1} = \operatorname{logit} C_\ell + \beta G_\ell. \tag{91}$$

Iterating and subtracting two trajectories with the same initial mass proves equation 11. If their initial masses differ, the initial-logit difference must also be retained.

E.2 Paired Gain-Reduction Analysis

We evaluate SVD and HSCFG using paired rollouts with the same prompt, initial forecast canvas, and random seed. For each comparison, gain is computed only over the intersection of valid layer–head transitions shared by vanilla and the corresponding intervention. This prevents changes in anchor-detection coverage from being interpreted as changes in gain.

Table 10 summarizes the matched rollout-level results. Quarter-rank SVD reduces cumulative gain by 3.22%, while HSCFG reduces it by 3.03%. These results provide the shared empirical connection to the gain–contraction interface: although the interventions act on different objects, both decrease the effective anchor-amplification gain.

The temporal profiles reflect the different intervention points. SVD acts within every Transformer forward but produces its largest gain reduction during later repetition-prone decoding steps. HSCFG acts through token commitments: its effect is negligible before the generated canvases diverge and becomes persistent after guided tokens enter subsequent forwards.

For visualization, we report the paired relative gain change

$$\Delta^{\text{int}}_{\text{gain}}(t) = \frac{\widehat{\mathcal{G}}^{\text{int}}_t - \widehat{\mathcal{G}}^{\text{van}}_t}{\left|\widehat{\mathcal{G}}^{\text{van}}_t\right|}. \tag{92}$$

Connection to anchor concentration. The reduction in measured gain is accompanied by a strong association between gain and fitted anchor-mass amplification. Under quarter-rank SVD, the correlation between $\widehat{G}$ and $\Delta \operatorname{logit} \widehat{C}_A$ remains high (Pearson $r = 0.823$, Spearman $\rho = 0.862$; Table 9). Thus, spectral truncation lowers the cumulative gain without breaking the empirical gain–concentration relationship established for vanilla decoding.

E.3 Permutation-Induced Spectral Plateau

For the non-cyclic predecessor-shift operator P,

$$P_{i,i-1} = 1, \qquad i = 2, \ldots, N,$$

with all other entries zero. Hence,

$$PP^\top = I - e_1 e_1^\top,$$

where e_1 is the first standard basis vector. For $A_0 = \mathbf{1}c^\top + \gamma P$,

$$A_0 A_0^\top = \gamma^2 I - \gamma^2 e_1 e_1^\top + \gamma \left[\mathbf{1}(Pc)^\top + (Pc)\mathbf{1}^\top\right] + \|c\|_2^2 \mathbf{1}\mathbf{1}^\top. \tag{93}$$

For every $x \perp \mathrm{span}\{e_1, \mathbf{1}, Pc\}$, all correction terms vanish and $A_0 A_0^\top x = \gamma^2 x$. This subspace has dimension at least $N - 3$; hence at least $N - 3$ singular values of A_0 equal $|\gamma|$. Therefore, the predecessor-shift component induces a high-dimensional spectral plateau.

For the isolated predecessor channel $\mathcal{D} = \gamma P$, the singular values consist of $|\gamma|$ with multiplicity $N - 1$ and one zero singular value. For $1 \le k \le N - 1$, an optimal rank-k approximation $\mathcal{D}_k$ therefore retains k of these nonzero singular directions, so

$$\frac{\|\mathcal{D}_k\|_F^2}{\|\mathcal{D}\|_F^2} = \frac{k\gamma^2}{(N-1)\gamma^2} = \frac{k}{N-1}. \tag{94}$$

The corresponding aggregate RMS strength is $|\gamma|\sqrt{k/N}$. This is a position-averaged summary: SVD compresses the independent shifted-predecessor feedback into a k-dimensional subspace rather than multiplying every diagonal entry by a common scalar. The plateau proof does not show that $|\gamma|$ is small or that stripe-associated singular values must be larger; that ordering is empirical.

E.4 Empirical Spectral and Attention Diagnostics after SVD

Matched attention reconstructions. The effect of spectral truncation varies with the pre-intervention structure of each head. Figure 4 illustrates two recurring regimes using matched attention operators. In a stripe-dominant head, both quarter-rank and rank-8 truncation preserve the leading vertical broadcast structure. In a diagonal-dominant head, quarter-rank truncation retains part of the shifted-diagonal pattern, whereas rank-8 removes it much more aggressively. Thus, SVD does not transform every head into the same low-rank pattern; it selectively preserves the dominant modes already present in each head while discarding increasing numbers of weaker spectral directions.

These heatmaps are intended as qualitative illustrations rather than estimates of how frequently each regime occurs. The aggregate singular-value statistics and gain measurements provide the corresponding quantitative evidence across the analyzed heads.

The spectral plateau result explains why the identity-like channel occupies a large number of singular directions, but it does not establish that these directions lie below the stripe-associated modes. We therefore examine the measured singular-value spectra and matched low-rank reconstructions.

Figure 4 shows that the leading singular modes are separated from a broader lower-magnitude plateau. The rank-8 and quarter-rank cutoffs both retain these leading modes while removing different fractions of the plateau.

E.5 Detailed SVD Gain Attribution

Let $\gamma_{k,\ell}$ denote an aggregate effective strength of the surviving shifted-predecessor feedback. Under the focused approximation that the stripe response and residual change less than this channel,

$$G_\ell^{\mathrm{SVD}} \approx C_A^{(\ell)} S_\ell + \gamma_{k,\ell} D_\ell + R_\ell, \qquad \gamma_{k,\ell} < \gamma_\ell. \tag{95}$$

For $D_\ell > 0$,

$$G_\ell - G_\ell^{\mathrm{SVD}} \approx (\gamma_\ell - \gamma_{k,\ell}) D_\ell > 0, \qquad \frac{\eta_\ell^{\mathrm{SVD}}}{\eta_\ell} \approx e^{\beta(\gamma_\ell - \gamma_{k,\ell}) D_\ell} > 1. \tag{96}$$

Because truncation can also alter C_ℓ, S_ℓ, D_ℓ, and R_ℓ, the total and cumulative gain comparisons remain the primary diagnostics.

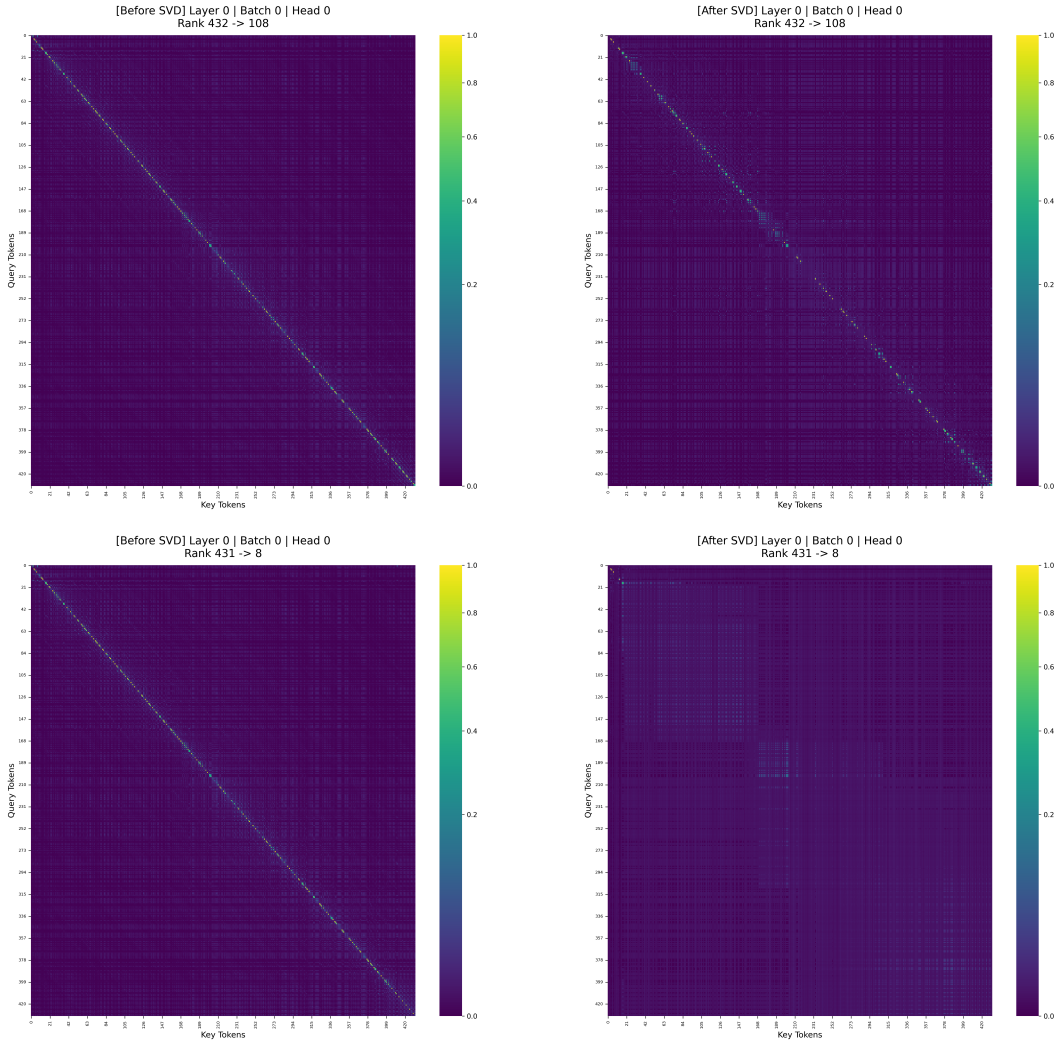

Figure 4: **Matched attention reconstructions under spectral truncation.** All four panels within each row use the same input, diffusion step, layer, head, token ordering, and color range. The dominant vertical stripe is preserved where it is present, whereas the shifted-diagonal structure is progressively attenuated as the retained rank decreases. The heterogeneous responses across heads indicate that SVD preserves head-specific leading modes rather than imposing a universal attention pattern.

Connection to anchor concentration. The reduction in measured gain is accompanied by a strong association between gain and fitted anchor-mass amplification. Under quarter-rank SVD, the correlation between $\widehat{G}$ and $\Delta \operatorname{logit} \widehat{C}_A$ remains high (Pearson $r = 0.823$, Spearman $\rho = 0.862$; Table 9). Thus, spectral truncation lowers the cumulative gain without breaking the empirical gain–concentration relationship established for vanilla decoding.

E.6 Channel-Wise Diagnostics after CFG

Suppose vanilla decoding writes the repeated digit a, whereas HSCFG writes $b \neq a$. The current forwards are unchanged; the following quantities are recomputed only after b enters Y_{t+1}. Suppressing the diffusion-step index, the diagonal alignments before and after replacement are

$$D_\ell = (k_A^{(\ell+1)} - k_O^{(\ell+1)})^\top W_Q^{(\ell+1)} W_O^{(\ell)} v_{p(i)}^{(\ell)},$$
$$D_\ell^{\mathrm{CFG}} = (k_{A,\mathrm{CFG}}^{(\ell+1)} - k_{O,\mathrm{CFG}}^{(\ell+1)})^\top W_Q^{(\ell+1)} W_O^{(\ell)} v_{p(i),\mathrm{CFG}}^{(\ell)}. \tag{97}$$

If the dominant measured change is in k_A, while k_O and v_i remain comparatively stable, then

$$D_\ell^{\mathrm{CFG}} - D_\ell \approx (k_{A,\mathrm{CFG}}^{(\ell+1)} - k_A^{(\ell+1)})^\top W_Q^{(\ell+1)} W_O^{(\ell)} v_{p(i)}^{(\ell)}. \tag{98}$$

A negative measured alignment supports $D_\ell^{\mathrm{CFG}} < D_\ell$; token replacement alone does not guarantee it. More generally,

$$\begin{aligned} G_\ell^{\mathrm{CFG}} - G_\ell = (C_{A,\mathrm{CFG}}^{(\ell)} S_\ell^{\mathrm{CFG}} - C_A^{(\ell)} S_\ell) \\ + (\gamma_\ell^{\mathrm{CFG}} D_\ell^{\mathrm{CFG}} - \gamma_\ell D_\ell) + (R_\ell^{\mathrm{CFG}} - R_\ell). \end{aligned} \tag{99}$$

We therefore measure the left-hand side first and use the three bracketed terms only to attribute its source. The finite-depth statistic is

$$\Delta\mathcal{G}_{\ell_0:L} := \sum_{\ell=\ell_0}^{L-1} (G_\ell^{\mathrm{CFG}} - G_\ell). \tag{100}$$

For matched initial mass, $\Delta\mathcal{G}_{\ell_0:L} < 0$ implies smaller final anchor log-odds. This diagnostic does not infer an output digit from anchor concentration and does not require a token-content homogeneity theorem.

Connection to anchor concentration. HSCFG exhibits the same gain–concentration correspondence despite changing the generated token trajectory. Its measured gain remains strongly associated with the layerwise change in fitted anchor-mass log-odds (Pearson $r = 0.820$, Spearman $\rho = 0.868$; Table 9). This supports interpreting the observed gain reduction as weaker anchor amplification rather than as an unrelated change in score geometry.

F Cross-Model Replication on Dream

The main-text mechanistic analyses use LLaDA-8B-Instruct for a consistent presentation across attention, hidden-state, gain, and intervention measurements. Here, we report the corresponding results for Dream-v0-Instruct-7B. Dream is evaluated using the same digit-wise serialization, forecasting tasks, decoding protocol, repetition criteria, and analysis pipeline described in Appendix G. The results reproduce the three main empirical relationships observed with LLaDA: repeating rollouts accumulate greater gain than non-repeating rollouts, gain remains coupled to changes in fitted anchor mass, and both interventions reduce repetition together with cumulative gain.

F.1 Cumulative Gain in Repeating and Non-Repeating Rollouts

We first divide vanilla Dream rollouts according to whether their final predictions satisfy the repetition criterion defined in Appendix G. As shown in Table 11, repeating rollouts accumulate greater gain than non-repeating rollouts, consistent with the cumulative-gain mechanism in Section 4.

F.2 Matched Intervention Results

We next evaluate SVD truncation and history-sparsified CFG on matched Dream forecasting instances whose vanilla rollouts satisfy the repetition criterion. Table 12 reports the paired results. Both interventions reduce cumulative gain and repetition, while their effects on forecasting accuracy are measured using NMAE.

Table 11: Cumulative gain and repetition statistics for repeating and non-repeating Dream rollouts.

Rollout group	Total repeat	R^{H}_{slot}	R^{T}_{slot}	R^{U}_{slot}	Cumulative gain
Repeating	1.000	1.000	1.000	0.997	111.970
Non-repeating	0.228	0.846	0.410	0.305	107.827

Table 12: intervention results on repetition-prone Dream rollouts. Gain reduction is computed relative to the matched vanilla condition.

Condition	Cumulative gain	Gain reduction	NMAE	Total repeat	R^{H}_{slot}	R^{T}_{slot}	R^{U}_{slot}
Vanilla	111.970	–	0.976	1.000	1.000	1.000	1.000
Quarter-rank SVD	108.644	-2.97%	1.000	0.599	0.879	0.744	0.591
History-sparsified CFG	108.498	-3.10%	1.008	0.045	0.374	0.130	0.126

F.3 Gain–Anchor Coupling

We examine whether the measured layerwise gain tracks the corresponding change in fitted anchor-mass log-odds. As shown in Table 13, all conditions exhibit moderate-to-strong positive correlations. Although weaker than those observed on LLaDA, the consistent association across vanilla, SVD, and HSCFG shows that the gain–anchor relationship also persists on Dream.

G Experimental Details

Models and dataset. We conduct experiments with two pretrained diffusion language models: `GSAI-ML/LLaDA-8B-Instruct` and `Dream-org/Dream-v0-Instruct-7B`. Both models are evaluated on time-series forecasting tasks from the Monash benchmark using the digit-wise serialization .

Forecasting and decoding. Following the standard numerical serialization pipeline for LLM-based time-series forecasting, each series is first scale-normalized and then converted into a fixed-precision base-10 representation. Under this scaling, the bulk of normalized values falls within a common numerical range, so the resulting serialized values naturally occupy three digit positions. For convenience, we denote these positions by H, T, and U, corresponding to the most significant, intermediate, and least significant digits, respectively. Consecutive values are separated by commas. We use three samples per series with batch size one and temperature 1.0. The reverse process is given a budget of 32 denoising steps. Masked positions are selected using low-confidence remasking, and the number of transferred tokens is computed deterministically rather than sampled stochastically. Zero-transfer iterations may be removed by the scheduler, so the number of effective commitment steps can be smaller than the nominal denoising budget. Unless explicitly stated otherwise, the forecast horizon is determined by the complete target sequence of each series.

Matched repetition protocol. To evaluate whether the interventions mitigate an already established repetition failure, we run vanilla, SVD, and HSCFG decoding on matched forecasting instances. Selection of the evaluation subset depends only on the vanilla prediction. Specifically, a vanilla rollout is classified as repetition-prone when

$$R_H > 0.99, \qquad R_T > 0.95, \qquad R_U > 0.95, \tag{101}$$

where R_H, R_T, and R_U are the final pairwise repetition rates of the hundreds, tens, and units digit sequences, respectively. Once a vanilla rollout satisfies this criterion, we retrieve the corresponding SVD and HSCFG trajectories for the same series. Intervention outcomes are never used to determine membership in the repetition-prone subset. Although the repetition-prone criterion uses explicit thresholds, it does not isolate a vanishingly small set of pathological examples: 29.79% of vanilla rollouts satisfy the criterion. We use the thresholds only to obtain a high-purity subset in which the repetition failure is already clearly established before intervention; intervention outcomes are never used for subset selection.

Table 13: Mean correlation between measured layerwise gain $\widehat{G}$ and the change in fitted anchor-mass log-odds $\Delta \operatorname{logit}(\widehat{C}_A)$ on Dream.

Correlation	Vanilla	Quarter-rank SVD	HSCFG
Pearson r	0.698	0.657	0.611
Spearman ρ	0.692	0.637	0.650

SVD intervention. For the primary spectral intervention, we apply quarter-rank truncated SVD to the attention readout operator. For an attention matrix with numerical rank r, the retained rank is

$$k = \max\{1, \lfloor N/4 \rfloor\}. \tag{102}$$

Numerical rank is determined using a threshold of 10^{-5}. The implementation selects the top 8 intervention heads according to the predefined SVD selection rule and applies the truncation to the full-history branch. All remaining decoding settings are kept identical to vanilla decoding. Rank-8 truncation is additionally used as an ablation, while the main intervention results reported in the paper use quarter-rank SVD.

History-sparsified guidance. HSCFG constructs a full-history branch and a history-sparsified reference branch from the same diffusion state. The prompt, current forecast canvas, and already revealed forecast tokens are identical between the two branches; sparsification is restricted to the historical conditioning tokens. We use the `prompt_token_overlap` keep policy, which preserves the designated prompt-overlapping history tokens while masking the remaining eligible historical content.

Let z_t^{full} and z_t^{sp} denote the logits from the full- and sparse-history branches. The implementation computes

$$z_t^{\text{HSCFG}} = z_t^{\text{sp}} + (1+\lambda)\left(z_t^{\text{full}} - z_t^{\text{sp}}\right), \tag{103}$$

with guidance strength $\lambda = 3.0$ in the HSCFG experiments.

Attention decomposition and anchor estimation. For each analyzed layer and attention head, we fit the empirical attention matrix using

$$A \approx \mathbf{1}c^\top + \gamma P + E, \tag{104}$$

where P is the one-step predecessor shift, $P_{i,i-1} = 1$ for $i = 2, \ldots, M$. The shared-stripe coefficient of key position j is estimated after excluding the shifted-diagonal entry,

$$c_j = \frac{\sum_{i:P_{ij}=0} A_{ij}}{\sum_i (1 - P_{ij})}, \tag{105}$$

and the predecessor coefficient is obtained by the Frobenius projection

$$\gamma = \frac{\langle P, A - \mathbf{1}c^\top \rangle_F}{\|P\|_F^2}. \tag{106}$$

The residual is then $E = A - \mathbf{1}c^\top - \gamma P$.

To detect shared anchor columns, we first mask the support of P and operate on log attention. Each key column is compared against a local median background computed over a window whose radius scales as $\lceil \sqrt{M} \rceil$. The column statistic is the 0.10 quantile of the resulting log-residual across valid queries. We use a lower-tail fraction of 0.25, a tail margin of 0.20, exclude mask-token keys, and require at least two detected anchors for a valid head. The same detector and thresholds are used for all experimental conditions. Given the detected anchor set $\mathcal{A}$, the fitted shared-anchor mass is

$$C_A = \sum_{j \in \mathcal{A}} c_j. \tag{107}$$