MEAN MODE SCREAMING: MEAN–VARIANCE SPLIT RESIDUALS FOR 1000-LAYER DIFFUSION TRANSFORMERS

Anonymous authors
Paper under double-blind review



Figure 1: **Text-to-image generation samples from our 1000-layer MV-Split DiT.** More samples are provided in Appendix M.

ABSTRACT

Scaling Diffusion Transformers (DiTs) to hundreds of layers introduces a structural vulnerability: networks can enter a silent, mean-dominated collapse state that homogenizes token representations and suppresses centered variation. Through mechanistic auditing, we isolate the trigger event of this collapse as **Mean Mode Screaming (MMS)**. MMS can occur even when training appears stable, with a mean-coherent backward shock on residual writers that opens deep residual branches and drives the network into a mean-dominated state. We show this behavior is driven by an exact decomposition of these gradients into mean-coherent and centered components, compounded by the structural suppression of attention-logit gradients through the null space of the Softmax Jacobian once values homogenize. To address this, we propose **Mean–Variance Split (MV-Split) Residuals**, which combine a separately gained centered residual update with a leaky trunk-mean replacement. On a 400-layer single-stream DiT, MV-Split prevents the divergent collapse that crashes the un-stabilized baseline; it tracks close to the baseline's pre-crash trajectory while remaining substantially better than token-isotropic gating methods such as LayerScale across the full schedule. Finally, we present a 1000-layer DiT as a scale-validation run at boundary scales, establishing that the architecture remains stably trainable at extreme depth.

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.

1 INTRODUCTION

Scaling laws for generative modeling (Liang et al., 2024) indicate that depth is an important dimension of capacity and model performance. Training ultra-deep Diffusion Transformers (DiTs) (Ho et al., 2020; Song et al., 2021; Peebles & Xie, 2023), however, introduces structural reliability issues that are not well described by standard exploding or vanishing gradient heuristics. In some runs, optimization remains stable for thousands of steps and then diverges within a few updates, with the loss returning near its initialization level and not recovering. These events can occur without NaNs or obvious forward saturation.

In this work, we study a **mean-dominated collapse state** in ultra-deep DiTs, in which token representations homogenize and centered token variation is suppressed. We reserve the term **Mean Mode Screaming (MMS)** for the abrupt entry event into this state: a spike in the mean-coherent gradient component, rapid residual branch opening, and subsequent Q/K gradient suppression.

Mechanistically, this failure exploits a geometric asymmetry between the token-mean and centered subspaces. Row-stochastic attention strictly preserves pure-mean states, while the centered component is propagated by a separate mixing operator and can become contractive in deep layers. On the backward pass, gradients admit an exact decomposition into mean-coherent and centered components; as token alignment increases, the mean-coherent component accumulates coherently with sequence length and can dominate the residual branch update. Once values homogenize, attention-logit gradients are suppressed through the null space of the Softmax Jacobian, suppressing Q/K learning and locking the network into the collapsed state.

Existing depth stabilizers suppress the entire residual branch isotropically in token space: ReZero (Bachlechner et al., 2020) and LayerScale (Touvron et al., 2021) apply scalar and per-channel learnable gates respectively, shrinking the mean and centered components together. This stabilizes training but slows convergence by also damping the centered signal responsible for spatially varying feature learning.

These observations motivate **MV-Split Residuals**, which combine a separately gained centered residual update with a leaky trunk-mean replacement. By damping the mean path without shrinking the centered path by the same factor, MV-Split stabilizes training without the convergence cost of isotropic residual gating.

Our contributions are:

1. **Characterization.** We characterize a mean-dominated collapse state and distinguish it from MMS, the abrupt entry event into this state. A standard-initialization control reaches the same collapse state more progressively across depth.
2. **Mechanism.** We show that row-stochastic attention preserves pure-mean states, that gradients split exactly into mean-coherent and centered components, with the mean-coherent component entering an $\mathcal{O}(T)$ coherent regime when tokens align, and that value homogenization suppresses attention-logit gradients through the null space of the Softmax Jacobian.
3. **Method and result.** We propose **MV-Split Residuals**, which combine a separately gained centered residual update with a leaky trunk-mean replacement. In matched 400-layer quantitative evaluation, MV-Split removes collapse events and converges faster than LayerScale; in a separate 1000-layer run, the same design remains stably trainable and serves as a scale-validation run at boundary scales.

2 PRELIMINARIES

We first describe the backbone, initialization, and training objective used in the main training runs.

2.1 MINIMAL SINGLE-STREAM MULTI-MODAL DIFFUSION TRANSFORMER

We use a deliberately stripped-down single-stream DiT (Peebles & Xie, 2023) so that deep residual propagation, rather than external modulation or skip pathways, remains the dominant carrier of both signal and gradients. Concretely, we employ a **Post-Norm** residual chain (Vaswani et al., 2017; Xiong et al., 2020) ($X_{l+1} = \mathrm{RMSNorm}(X_l + f_l(X_l))$ (Zhang & Sennrich, 2019)) without AdaLN (Peebles

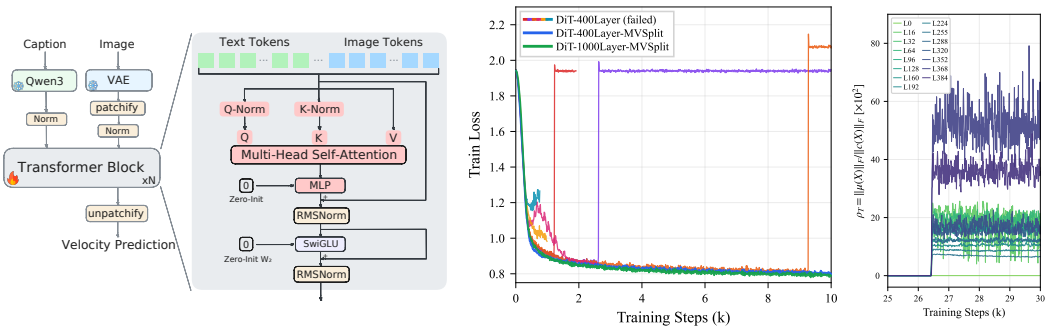

Figure 2: **Baseline DiT and representative training diagnostics. (Left)** Single-stream DiT backbone. **(Middle)** Training loss over the first 10k steps for the un-stabilized 400-layer baseline and the MV-Split 400-/1000-layer runs. **(Right)** Per-layer energy ratio $\rho_T = \|\mu(X)\|_F / \|c(X)\|_F$ (Appendix A) across L0–L384 in a baseline run.

& Xie, 2023) or other per-layer modulation mechanisms, to avoid introducing alternative depthwise control channels that would complicate attribution of the collapse dynamics. Instead of cross-attention, we concatenate VAE-encoded (Kingma & Welling, 2013; Rombach et al., 2022) image tokens X_{img} and text embedding tokens X_{txt} into a unified sequence $X_{in} = [X_{img}; X_{txt}]$ (Bao et al., 2022; Esser et al., 2024; Qin et al., 2025; Z-Image Team et al., 2025), forcing self-attention (Vaswani et al., 2017) to handle all multimodal interaction. For positional encoding, we apply a 2D extension of RoPE (Su et al., 2024) to image tokens following recent vision/diffusion Transformer practice (Chu et al., 2024; Lu et al., 2024), and leave text tokens without rotary positional encoding. The left panel of Figure 2 gives the corresponding backbone schematic.

2.2 Residual Writer Zero Initialization

For the main training runs used in the main text, except the LayerScale control, we zero-initialize the **residual writers** (W_O and W_2), following the broader practice of identity-initialized residual branches and zero-initialized output pathways in residual and diffusion architectures (Goyal et al., 2017; Zhang et al., 2019; Peebles & Xie, 2023; Zhang et al., 2023; Zhu et al., 2025). Here W_O is the attention output projection. For the FFN branch, we write the SwiGLU (Shazeer, 2020) feed-forward transformation as

$$[g_l, v_l] = W_{13} x_l, \qquad \mathrm{FFN}(x_l) = W_2\big(\mathrm{SiLU}(g_l) \odot v_l\big), \tag{1}$$

so W_2 is the residual writer of the FFN block. In these zero-writer training runs, the internal branch parameters (e.g., Q, K, V and W_{13}) remain at their standard initialization. Appendix B shows that standard initialization does not avoid the mean-dominated regime; the same collapse appears from the start as a depth-progressive front, rather than through the delayed writer-opening spike that defines MMS in the zero-writer training runs.

2.3 Rectified Flow Matching

We train the model using a **Rectified Flow** (Liu et al., 2022; Lipman et al., 2022) objective. Given a data distribution x_0 (VAE latents) and a Gaussian noise distribution $x_1 \sim \mathcal{N}(0, I)$, we define a linear interpolation path $z_t = (1-t)x_0 + t x_1$ for $t \in [0, 1]$. The model v_θ is trained to predict the vector field pointing from noise toward data:

$$\mathcal{L} = \mathbb{E}_{t, x_0, x_1}\left[\|v_\theta(z_t, X_{txt}) - (x_0 - x_1)\|^2\right] \tag{2}$$

3 Failure Dynamics: Mean-Dominated Collapse

To understand the failure mode limiting depth scaling, we analyze a representative abrupt-failure run from the main diagnostic regime. We first introduce a token-space decomposition that separates sequence-mean and centered variation. We then use this decomposition to trace the observed

divergence sequence: a mean-coherent gradient shock, residual branch opening, mean-dominated forward collapse, and Q/K gradient suppression. Section 4 explains why this sequence occurs.

3.1 Geometric Preliminaries: Token-Space Asymmetry

The failure dynamics are fundamentally tied to how information is distributed across tokens. Let $\mathbf{1} \in \mathbb{R}^T$ denote the all-ones vector, and define $J \triangleq \frac{1}{T}\mathbf{1}\mathbf{1}^\top$ and $P \triangleq I - J$. For any token sequence $X \in \mathbb{R}^{T\times D}$, we write

$$X = JX + PX \;\equiv\; \mu(X) + c(X), \tag{3}$$

where $\mu(X) \triangleq JX$ is the sequence-mean component and $c(X) \triangleq PX$ is the centered variation component. Row-stochastic attention acts asymmetrically on these two subspaces.

Proposition 1 (Pure-mean component is preserved)**.** *For any row-stochastic attention matrix A satisfying $A\mathbf{1} = \mathbf{1}$, $A\,\mu(X) = \mu(X)$.*

Note that Proposition 1 governs only the pure-mean component of the input. For a general input $X = JX + PX$, the output mean satisfies $\mu(AX) = JAX = JX + JAPX$; centered variation can therefore contribute to the output mean through the leakage term $JAPX$.

Proposition 2 (Centered component is governed by PAP)**.** *For any row-stochastic attention matrix A satisfying $A\mathbf{1} = \mathbf{1}$,*

$$c(AX) = PAX = PAPX, \quad \textit{and therefore} \quad \|c(AX)\|_F \le \|PAP\|_2\,\|c(X)\|_F. \tag{4}$$

We denote $\mu_{\text{eff}}(A) \triangleq \|PAP\|_2$. When $\mu_{\text{eff}}(A) < 1$, the layer is strictly contractive on the centered subspace. This geometric asymmetry imposes a structural vulnerability: row-stochastic attention leaves pure-mean states invariant, while its action on token-specific variation is governed by μ_{eff} and can become contractive. Consequently, the network must rely on residual branches to continuously replenish the centered subspace. If the residual updates become dominated by the mean component, the representation is driven toward a pure-mean state.

3.2 Tracing the Divergence Event: From Trigger to Lock-in

Figure 3 traces the divergence in a 400-layer baseline through a tight chronological sequence. The backward pass exhibits a mode-selective shock: the gradient spike is concentrated primarily in the mean-coherent component while Q/K gradients collapse in lockstep, leaving residual writers as the dominant active learning channel. This shock then locks in across the forward pass — branches open into a mean-dominated regime (ρ_T explodes), and with attention contractive on the centered subspace and no branch-side variance replenishment, tokens homogenize across depth into a trivial mean-prediction baseline.

This empirical sequence isolates two questions for the mechanistic analysis in Section 4: (1) why the gradient amplifies specifically in the mean-coherent direction, and (2) why token homogenization structurally suppresses Q/K gradients.

4 Mechanism

4.1 Gradient Decomposition and Backward Alignment Amplification Law

Consider a token-wise linear map W (e.g., residual writers W_O, W_2) whose gradient takes the form $\nabla_W \mathcal{L} = \sum_{t=1}^T \delta_t y_t^\top$. Decomposing the forward inputs y_t and backward gradients δ_t into their sequence means $(\bar{y}, \bar{\delta})$ and centered residuals $(\tilde{y}_t, \tilde{\delta}_t)$, the cross-terms vanish identically under summation (proof in Appendix C.1), yielding an exact additive decomposition:

$$\nabla_W \mathcal{L} \;=\; \underbrace{T\,\bar{\delta}\,\bar{y}^\top}_{\Delta W_\mu \text{ (mean-coherent, } \mathcal{O}(T) \text{ when aligned)}} \;+\; \underbrace{\textstyle\sum_{t=1}^T \tilde{\delta}_t \tilde{y}_t^\top}_{\Delta W_c \text{ (centered, diffusive)}} . \tag{5}$$

We denote $G_{\text{mean}} \triangleq \|\Delta W_\mu\|_F$ and $G_{\text{ctr}} \triangleq \|\Delta W_c\|_F$. This decomposition exposes a scaling transition. The mean component has norm $\|\Delta W_\mu\|_F = T\,\|\bar{\delta}\|\,\|\bar{y}\|$, so it remains small when sequence means

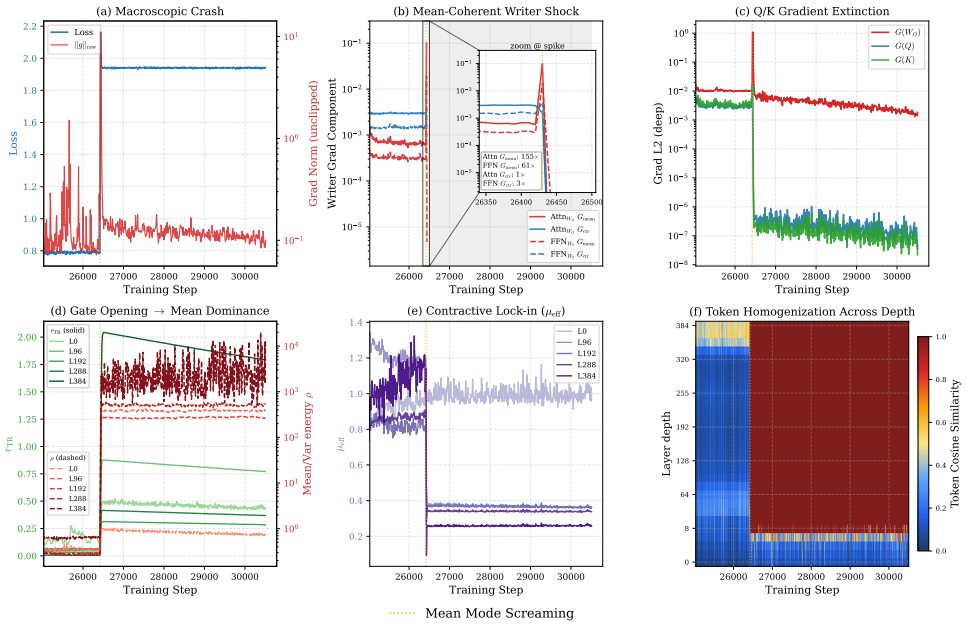

Figure 3: **Empirical trajectory of a representative divergence event (400-layer).** The vertical dashed line marks the divergence step. **(a–c) Backward trigger:** The global gradient norm spikes (a). The spike is concentrated in the mean-coherent gradient component G_{mean}, while the centered component G_{ctr} shows no comparable amplification (b). After the spike, Q/K gradients drop by roughly four orders of magnitude while W_O gradients remain nonzero (c). **(d–f) Forward lock-in:** The residual branch opens and the mean/centered energy ratio ρ_T rises sharply (d). Deep attention remains contractive on the centered subspace, with limited variance replenishment (e). Token representations homogenize across depth, with cosine similarity approaching one in deep layers (f).

cancel; under weak centered alignment, ΔW_c sums diffusively. As representations and adjoints homogenize, however, the sequence means stop canceling, $\|\bar{y}\|$ and $\|\bar{\delta}\|$ become order-one, and the rank-1 mean mode enters its coherent $\mathcal{O}(T)$ regime. Operationally, Mean Mode Screaming acts as a sharp transition from diffusive cancellation to coherent accumulation.

To quantify this transition, we define the dimensionless *alignment amplification* $\mathcal{A}$ as the ratio of the true gradient energy to the independent-token baseline. As derived in Appendix C.2, expanding this ratio yields an identity linking the cross-token coherent amplification of gradients to microscopic token alignment. Under an equal-magnitude proxy, it takes the compact form:

$$\underbrace{\frac{\|\nabla_W \mathcal{L}\|_F^2}{\sum_t \|\delta_t\|^2 \|y_t\|^2}}_{\text{Amplification } \mathcal{A}} - 1 \;=\; \frac{\sum_{s\neq t} (\delta_s^\top \delta_t)(y_s^\top y_t)}{\sum_t \|\delta_t\|^2 \|y_t\|^2} \;\approx\; (T-1)\, \underbrace{\mathbb{E}_{s\neq t}\big[\cos(y_s, y_t)\cos(\delta_s, \delta_t)\big]}_{\text{Pairwise alignment } \kappa}. \tag{6}$$

Equation 6 identifies when token-wise gradients stop canceling and enter a coherent accumulation regime. When tokens are heterogeneous, signed off-diagonal terms cancel ($\kappa \approx 0$) and $\mathcal{A} \approx 1$. As both representations and adjoints become aligned in deep layers, the signed off-diagonal terms stop canceling; in the limiting case $\cos(y_s, y_t) \to 1$ and $\cos(\delta_s, \delta_t) \to 1$, giving $\kappa \to 1$ and the gradient enters its $\mathcal{O}(T)$ coherent-amplification regime. We empirically audit this transition in Section 6.1 using the absolute-coherence upper-envelope proxy $\hat{\kappa} \triangleq \mathbb{E}_{s\neq t}[|\cos(y_s, y_t)|\,|\cos(\delta_s, \delta_t)|]$.

4.2 Q/K Gradient Extinction via the Softmax Null Space

A gradient spike alone would not lock in the failure if the attention path could restore token variation. However, once the residual stream becomes mean-dominated, the value vectors homogenize. Consequently, the Softmax Jacobian zeroes out the constant component of the attention-weight gradient.

Lemma 1 (Softmax null space under value collapse)**.** *For one attention row i, if $V_j = \bar{v}$ for all j, then $\partial\mathcal{L}/\partial S_i = \mathbf{0}$, where S_i is the vector of pre-softmax logits.*

By the chain rule, $\partial\mathcal{L}/\partial a_{ij} = \langle \partial\mathcal{L}/\partial Y_i, V_j \rangle$ is independent of j when $V_j = \bar{v}$, yielding $\partial\mathcal{L}/\partial a_i \propto \mathbf{1}$. Because $J_{\text{sm}}(a_i)\mathbf{1} = \mathbf{0}$, the logit gradient strictly vanishes. Under approximate homogeneity, this null space still removes the constant component, strongly suppressing Q/K learning while the residual-writer gradient (Eq. 5) is not zeroed by this Softmax null space (proof in Appendix C.3).

5 Method: MV-Split Residuals

Section 4 isolates a single unstable mode: the rank-one mean-coherent gradient update ΔW_μ. We therefore decouple its residual gain from the centered update. Let $X_l \in \mathbb{R}^{T\times D}$ be the trunk and $F_l \triangleq f_l(X_l)$ the branch output. Using the orthogonal projectors J and $P = I - J$ from Section 3.1, we replace the standard Post-Norm merge $X_{l+1} = \text{RMSNorm}(X_l + F_l)$ with a subspace-routed merge:

$$Z_l \triangleq X_l \; + \; \underbrace{\beta \odot (PF_l)}_{\text{centered path}} \; + \; \underbrace{\alpha \odot J(F_l - X_l)}_{\text{mean path}}, \tag{7}$$

$$X_{l+1} = \text{RMSNorm}(Z_l), \tag{8}$$

where $\alpha, \beta \in \mathbb{R}^D$ are per-block learnable vectors broadcast across tokens. Our multimodal transformer implementation applies the residual projectors segment-wise ($J_{\text{seg}}, P_{\text{seg}}$) to avoid directly mixing image and text means in the residual control path (Appendix E).

Forward dynamics. Prior to token-dependent RMS normalization, projecting Eq. 7 exactly decouples the pre-normalization merge:

$$PZ_l = PX_l + \beta \odot (PF_l), \qquad JZ_l = (1-\alpha) \odot (JX_l) + \alpha \odot (JF_l). \tag{9}$$

The centered subspace follows a standard residual update with gain β, while the mean subspace becomes a per-feature leaky integrator (when $0 < \alpha_d \leq 1$): each layer contracts the trunk mean by $1 - \alpha_d$ before adding a fresh correction.

Backward dynamics. Let $G_l \triangleq \partial\mathcal{L}/\partial Z_l$. Because J, P are self-adjoint and orthogonal, the gradient flowing back into the branch factors along the same split:

$$\frac{\partial\mathcal{L}}{\partial F_l} \; = \; \beta \odot (PG_l) + \alpha \odot (JG_l). \tag{10}$$

Centered and mean-coherent gradients receive independent gains. Together with (9), a small α both damps mean-coherent forward accumulation and shrinks the ΔW_μ component of the gradient (Eq. 5) by the same factor, without tying the local centered branch-gradient to the small mean gain α.

Comparison to other residual-gain methods. LayerScale (Touvron et al., 2021) and ReZero (Bachlechner et al., 2020) apply a single residual gain (per-channel and scalar, respectively) that does not distinguish the mean and centered subspaces, so ΔW_μ and ΔW_c are suppressed jointly. We elaborate on the structural distinctions between MV-Split and these residual-gain methods in Appendix D.

6 Experiments

All 400-layer runs share backbone, optimizer, data, batch size, and non-residual primitives: ImageNet-2012 (Russakovsky et al., 2015) latents from a frozen FLUX.2 VAE (Rombach et al., 2022; Labs, 2025), conditioned on a frozen Qwen3-0.6B text encoder (Yang et al., 2025); each stabilizer uses its standard residual initialization (Appendix G). The 1000-layer scale-validation run uses the same residual design and is trained through post-training on a curated ∼50k-image set (Figure 1, Appendix M). Appendix F rules out alternative explanations for the loss spike and localizes the failure to MMS.

6.1 Testing the Alignment-Amplification Law

Figure 4 tests Eq. 6 in a representative unstable 400-layer run whose writer-gradient norm spikes at step $t^\star = 3400$. Before the spike, both writers lie well below the saturation envelope. Absolute

cross-token coherence is present, but the signed off-diagonal terms in Eq. 6 still cancel. The small pre-spike slopes therefore measure how loose the envelope is in this run, not new constants.

At step $t^\star$, the active layers lie close to the saturation envelope for both `Attn_WO` and `FFN_W2`. The main observation is that the spike occurs when signed cancellation at the residual writer largely disappears. The same near-saturation appears in the attention and FFN writers, supporting a writer-interface explanation rather than an attention-specific one.

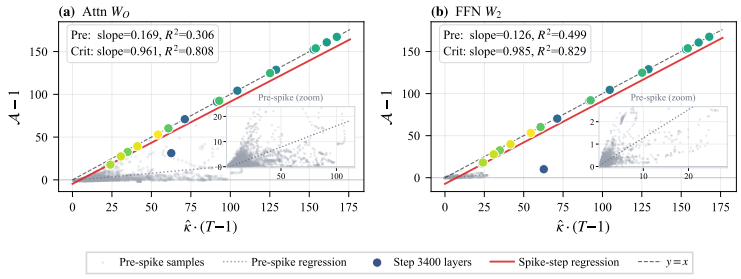

Figure 4: **Writer amplification at the gradient spike (400-layer Base η run, $t^\star$=3400, measured on the T=256 image-token segment).** Each point plots $\mathcal{A}-1$ against the equal-magnitude absolute-coherence upper-envelope proxy $(T-1)\hat{\kappa}$ for **(a)** `Attn_WO` and **(b)** `FFN_W2`. Gray points are pre-spike layer-step samples; colored points are active layers at $t^\star$ ($\mathcal{A}-1>0.5$). The dashed line is the absolute-coherence saturation envelope; fitted slopes and R^2 values are shown in each panel.

The largest active-layer values reach $\mathcal{A}-1 \approx 167$, corresponding to a $\sim 13\times$ writer-gradient norm amplification relative to the independent-token baseline. The shallowest active layer remains below the saturation envelope, consistent with a boundary region where absolute coherence is already high but sign cancellation has not fully disappeared. These measurements support the mechanism in Section 4.1: MMS occurs when residual writers lose signed cancellation across tokens, allowing the mean-coherent update ΔW_μ to approach its coherent $\mathcal{O}(T)$ scaling regime.

6.2 MV-Split Shifts the Stability-Constrained Quality Frontier

We next evaluate whether MV-Split changes the usable quality frontier under an explicit stability constraint: a run is treated as usable only if it remains non-divergent over the measured training horizon.

Figure 5 and Table 1 show the resulting stability-constrained quality frontier. The un-stabilized baselines are useful references for early learning speed, but they do not define stable frontier points: both enter the mean-dominated failure state. Reducing the learning rate delays this failure rather than removing it. LayerScale remains stable over the measured horizon, but its token-isotropic per-channel gain also reduces the centered residual updates needed for token-varying feature learning.

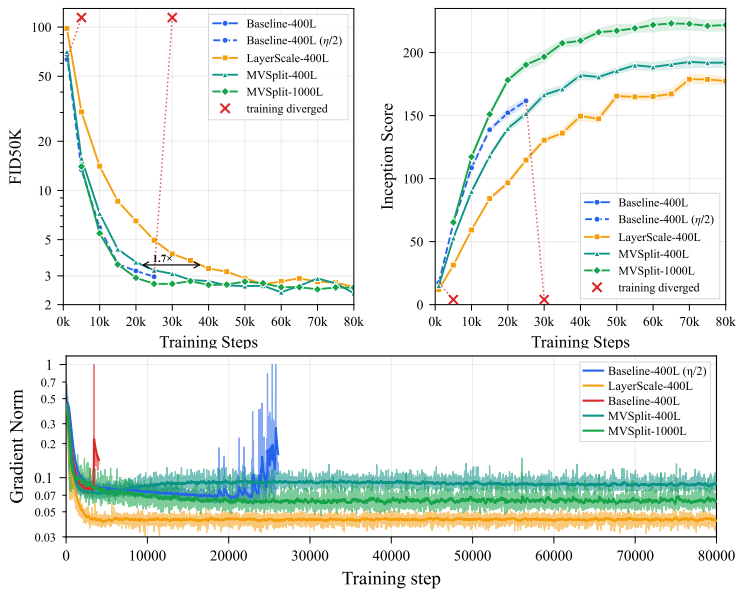

Figure 5: **Quality and optimizer stability over 80k steps (ImageNet 256×256). (Top)** FID-50K and Inception Score. **(Bottom)** Post-clipping global gradient norm. The 400-layer curves define the controlled comparison: among the non-divergent 400-layer runs, MV-Split preserves a higher bounded gradient band than LayerScale while avoiding the spikes of the un-stabilized baselines. The 1000-layer MV-Split trace is included as scale validation.

Under this stability constraint, MV-Split shifts the controlled 400-layer frontier. It does not uniformly dominate the unstable baselines at early checkpoints, but those trajectories leave the stable set; MV-Split preserves much of their early convergence speed while avoiding their collapse. Among the non-divergent 400-layer runs, MV-Split is already substantially ahead of LayerScale by 20k–30k steps, and the added 40k/50k checkpoints show that this advantage persists rather than reflecting a short early transient. The gradient-norm trace also separates MV-Split from simple global shrinkage: it operates in a higher bounded gradient band than LayerScale, while avoiding the spikes seen in the un-stabilized runs.

Table 1: **Stability and convergence across 400-/1000-layer DiT runs.** The 400-layer rows define the matched stability-constrained comparison. The 1000-layer row is a separate scale-validation point and is not part of the matched 400-layer frontier comparison. FID-50K and IS are computed with Euler sampling, 25 NFE, and CFG scale $w = 2.0$ for all rows. "—" denotes divergence before the checkpoint or failure to produce a valid evaluation. Bold highlights the best non-divergent result within the matched 400-layer comparison. The default-LR baseline diverges before the first checkpoint. † The lower-LR baseline diverges later in training and is shown only as a pre-crash speed reference; it is not counted as a stable frontier point. 400L LayerScale reports the best stable member of the λ_{init} sweep.

	FID↓ / IS↑				
Method	@10k	@20k	@30k	@40k	@50k
400L Base (η)	—	—	—	—	—
400L Base ($\eta/2$)†	5.92 / 108.6	3.22 / 152.2	—	—	—
400L LayerScale	14.08 / 59.2	6.50 / 96.6	4.09 / 130.5	3.33 / 149.6	2.90 / 165.5
400L MV-Split	**7.23 / 89.8**	**3.64 / 139.9**	**3.09 / 166.5**	**2.79 / 182.0**	**2.60 / 185.5**
1000L MV-Split	5.47 / 117.3	2.92 / 178.2	2.68 / 196.6	2.64 / 209.4	2.77 / 217.3

The 1000-layer run extends this observation to boundary depth. The same residual design remains stable over the measured training horizon and reaches strong fixed-checkpoint FID/IS values at the reported boundary depth. Because this run uses a separate training and post-training pipeline, we do not use it as a matched frontier point against the 400-layer controls. Instead, it serves as scale validation: the residual mechanism that shifts the controlled 400-layer frontier remains usable at 1000 layers. Additional GenEval and DPG-Bench measurements for the post-trained checkpoint are reported in Appendix K.1 as calibration rather than as state-of-the-art comparison.

6.3 Writer-Gradient Mode Decomposition

The convergence curves alone do not distinguish mode-selective control from a smaller effective learning rate. We therefore measure the two writer-gradient components from Eq. 5: the mean-coherent component G_{mean} and the centered component G_{ctr}.

Figure 6 shows that LayerScale bounds the mean-coherent writer component, but does so by shrinking the centered component as well. This is expected from a token-isotropic residual gain: the same per-channel multiplier is applied before any token-space split, so the method provides no explicit mechanism to preserve centered variation while damping the token-mean component. The resulting low centered-gradient band is consistent with the slower convergence observed in Figure 5.

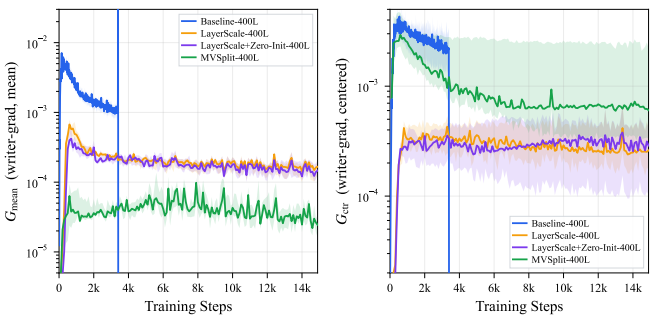

Figure 6: **Residual-writer gradient mode decomposition.** Per-step median across depth of the mean-coherent (G_{mean}, left) and centered (G_{ctr}, right) writer-gradient magnitudes; shaded regions denote the interquartile range (IQR; 25–75% across depth). Token-isotropic per-channel gating compresses both modes; MV-Split bounds the mean-coherent component while preserving a higher, stable centered band.

MV-Split changes this pattern. The mean-coherent component remains bounded, while the centered component stays in a higher stable band. This supports the intended mechanism of Eq. 10: the mean and centered components receive separate gains at the residual merge. Thus the improved stability in Section 6.2 is not explained by uniformly smaller gradients, but by damping the writer-gradient mode associated with the collapse.

Deferred analyses. Beyond stability, a linear probe confirms **the token mean acts as an implicit global timestep carrier** (near-perfect R^2 predicting t across depth), justifying our design to gain-limit rather than strictly project out the mean subspace (Appendix H). Infrastructure-level optimizations for ultra-deep training are deferred to Appendix I.

7 Related Work

Diffusion Transformers replace U-Net backbones (Dhariwal & Nichol, 2021) with Transformer blocks over latent or image patches. DiT (Peebles & Xie, 2023) showed that increasing Transformer compute through depth, width, or token count improves generative quality, while U-ViT (Bao et al., 2022) and MMDiT/Stable Diffusion 3 (Esser et al., 2024) demonstrate that token-based diffusion backbones can support long skips, multimodal token mixing, and rectified-flow text-to-image generation. Unlike standard DiT conditioning stacks that inject the noise or timestep level through AdaLN or related modulation paths, recent work suggests that explicit noise/timestep conditioning is not always required for denoising generative models (Sun et al., 2025; Sahraee-Ardakan et al., 2026). Our focus is complementary to this objective-level question: we use a noise-agnostic backbone to study a depthwise residual-stream failure mode in ultra-deep DiTs and a residual merge that stabilizes this signal path. Appendix H further shows that our trained network implicitly carries the continuous timestep in the token-mean subspace.

Training instability in deep Transformers is often addressed by changing normalization placement, residual scaling, or residual connectivity. Post-LN Transformers can require warmup because large gradients appear near the output layers at initialization, whereas Pre-LN changes this gradient geometry (Xiong et al., 2020). Admin (Liu et al., 2020) attributes instability to residual-branch dependence that amplifies update perturbations. ReZero (Bachlechner et al., 2020), LayerScale (Touvron et al., 2021), DeepNorm (Wang et al., 2022), and Keel (Chen & Wei, 2026) stabilize depth by gating, rescaling, or altering the residual/carry path, with DeepNorm and Keel reporting 1,000-layer-scale Transformer training. These methods control the residual branch or carry path as a whole. MV-Split targets a different axis: it combines a separately gained centered residual update with a leaky trunk-mean replacement, damping the mean-coherent channel without applying the same shrinkage to centered token variation.

MMS is also related to work on training spikes and attention/representation collapse, but its diagnostic object is different. Loss-spike and proxy studies connect large-scale instabilities to sub-layer Jacobian bounds, attention-logit growth, output-logit divergence, and predictive activation/gradient-norm trends (Takase et al., 2023; Wortsman et al., 2023). Other work shows that self-attention can drive token representations toward rank-one uniformity with depth (Dong et al., 2021), that rank collapse can vanish Q/K gradients via signal-propagation arguments (Noci et al., 2022), and that low attention entropy is associated with unstable or divergent Transformer training (Zhai et al., 2023). MMS, in contrast, is diagnosed through an exact gradient decomposition into mean-coherent and centered components, followed by a mean-dominated forward state in which centered token variation is suppressed and Q/K learning is reduced through the Softmax Jacobian null space. We therefore do not claim that MMS explains all deep-Transformer spikes; it identifies a specific residual-subspace failure pathway in ultra-deep Post-Norm DiTs, and MV-Split acts at the residual interface rather than as a global residual shrinkage or attention-operator correction. Appendix J reports negative controls that test several superficially related alternatives.

8 Conclusion

We study a **mean-dominated collapse state** that limits depth scaling in very deep Diffusion Transformers, and we use **Mean Mode Screaming (MMS)** for the abrupt writer-gradient event that accompanies entry into this state in zero-writer training runs. The main mechanism is an imbalance between a mean-coherent writer update that can grow as $\mathcal{O}(T)$ and a centered path that is not sufficiently replenished once deep layers become contractive. MV-Split addresses this by combining a separately gained centered residual update with a leaky trunk-mean replacement. Under matched-backbone stabilizer protocols at 400 layers, MV-Split removes collapse events and gives the best stable frontier among the methods we evaluate; a separate 1000-layer run shows that the same design remains trainable at that depth.

AI USE STATEMENT

We used generative AI tools in the following ways. **Writing:** LLM assistants were used to polish the wording of author-written drafts and to draft BibTeX entries. Every reference was then checked against its authoritative source (arXiv, DOI, or publisher page) by an automated field-by-field comparison, followed by manual verification of titles and author lists. **Research ideation and execution:** LLM assistants were used to discuss candidate explanations of the observed collapse, to check derivations, and to help write and debug training, diagnostic, and plotting code. All code behind the reported results was reviewed and tested by the authors, all derivations in the paper were verified by the authors, and every reported number comes from our own training runs. **Model components:** separately from the above, a frozen Qwen3-0.6B text encoder (Yang et al., 2025) provides text conditioning (Section 2.1), and the ImageNet-2012 captions used for text-conditioned training were generated offline by an LLM (Appendix M). We take full responsibility for the content of this paper, including text, claims, and artifacts produced with the aid of generative AI.

ETHICS STATEMENT

This work involves no human subjects and no personal data collection. Training uses the ImageNet-2012 dataset (Russakovsky et al., 2015) under its research license and a curated $\sim$50k-image set used only for post-training; the FLUX.2 VAE (Labs, 2025) and the Qwen3-0.6B text encoder (Yang et al., 2025) are used under the Apache 2.0 license. As with other text-to-image models, improved generation quality could be misused to produce disinformation or non-consensual content. Our training data contain no web-scraped face data, which limits high-fidelity person generation, and any released checkpoint will be accompanied by a model card and an acceptable-use policy that prohibits such uses.

REPRODUCIBILITY STATEMENT

The backbone is specified in Section 2.1, the residual-writer initialization in Section 2.2, and the MV-Split update in closed form in Section 5 (Eqs. 7–10). The segment-wise multimodal variant and its backward pass are derived in Appendices E and I.1. Appendix G lists the architecture and all training hyperparameters for every reported run, including the μP-derived learning rate, optimizer settings, batch size, warmup, weight decay, gradient clipping, and hardware. The fused kernels are described in Appendix I, the failure-attribution procedure in Appendix F, and the evaluation protocol in Section 6 and Appendix K.1.

REFERENCES

Thomas Bachlechner, Bodhisattwa Prasad Majumder, Huanru Henry Mao, Garrison W. Cottrell, and Julian McAuley. Rezero is all you need: Fast convergence at large depth, 2020. URL https://arxiv.org/abs/2003.04887.

Fan Bao, Shen Nie, Kaiwen Xue, Yue Cao, Chongxuan Li, Hang Su, and Jun Zhu. All are worth words: A vit backbone for diffusion models, 2022. URL https://arxiv.org/abs/2209.12152.

Chen Chen and Lai Wei. Post-layernorm is back: Stable, expressive, and deep, 2026. URL https://arxiv.org/abs/2601.19895.

Tianqi Chen, Bing Xu, Chiyuan Zhang, and Carlos Guestrin. Training deep nets with sublinear memory cost, 2016. URL https://arxiv.org/abs/1604.06174.

Xiangxiang Chu, Jianlin Su, Bo Zhang, and Chunhua Shen. Visionllama: A unified llama backbone for vision tasks, 2024. URL https://arxiv.org/abs/2403.00522.

Tri Dao, Daniel Y. Fu, Stefano Ermon, Atri Rudra, and Christopher Ré. Flashattention: Fast and memory-efficient exact attention with io-awareness, 2022. URL https://arxiv.org/abs/2205.14135.

Mostafa Dehghani, Josip Djolonga, Basil Mustafa, Piotr Padlewski, Jonathan Heek, Justin Gilmer, Andreas Steiner, Mathilde Caron, Robert Geirhos, Ibrahim Alabdulmohsin, Rodolphe Jenatton, Lucas Beyer, Michael Tschannen, Anurag Arnab, Xiao Wang, Carlos Riquelme, Matthias Minderer, Joan Puigcerver, Utku Evci,

Manoj Kumar, Sjoerd van Steenkiste, Gamaleldin F. Elsayed, Aravindh Mahendran, Fisher Yu, Avital Oliver, Fantine Huot, Jasmijn Bastings, Mark Patrick Collier, Alexey Gritsenko, Vighnesh Birodkar, Cristina Vasconcelos, Yi Tay, Thomas Mensink, Alexander Kolesnikov, Filip Pavetić, Dustin Tran, Thomas Kipf, Mario Lučić, Xiaohua Zhai, Daniel Keysers, Jeremiah Harmsen, and Neil Houlsby. Scaling vision transformers to 22 billion parameters, 2023. URL https://arxiv.org/abs/2302.05442.

Prafulla Dhariwal and Alex Nichol. Diffusion models beat gans on image synthesis, 2021. URL https://arxiv.org/abs/2105.05233.

Yihe Dong, Jean-Baptiste Cordonnier, and Andreas Loukas. Attention is not all you need: Pure attention loses rank doubly exponentially with depth, 2021. URL https://arxiv.org/abs/2103.03404.

Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, Dustin Podell, Tim Dockhorn, Zion English, Kyle Lacey, Alex Goodwin, Yannik Marek, and Robin Rombach. Scaling rectified flow transformers for high-resolution image synthesis, 2024. URL https://arxiv.org/abs/2403.03206.

Priya Goyal, Piotr Dollár, Ross Girshick, Pieter Noordhuis, Lukasz Wesolowski, Aapo Kyrola, Andrew Tulloch, Yangqing Jia, and Kaiming He. Accurate, large minibatch sgd: Training imagenet in 1 hour, 2017. URL https://arxiv.org/abs/1706.02677.

Albert Gu and Tri Dao. Mamba: Linear-time sequence modeling with selective state spaces, 2023. URL https://arxiv.org/abs/2312.00752.

Alex Henry, Prudhvi Raj Dachapally, Shubham Pawar, and Yuxuan Chen. Query-key normalization for transformers, 2020. URL https://arxiv.org/abs/2010.04245.

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

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models, 2020. URL https://arxiv.org/abs/2006.11239.

Keller Jordan, Yuchen Jin, Vlado Boza, Jiacheng You, Franz Cesista, Laker Newhouse, and Jeremy Bernstein. Muon: An optimizer for hidden layers in neural networks. https://kellerjordan.github.io/posts/muon/, 2024. URL https://kellerjordan.github.io/posts/muon/.

Tero Karras, Miika Aittala, Timo Aila, and Samuli Laine. Elucidating the design space of diffusion-based generative models, 2022. URL https://arxiv.org/abs/2206.00364.

Diederik P. Kingma and Max Welling. Auto-encoding variational bayes, 2013. URL https://arxiv.org/abs/1312.6114.

Black Forest Labs. FLUX.2: Frontier Visual Intelligence. https://bfl.ai/blog/flux-2, 2025. URL https://github.com/black-forest-labs/flux2.

Zhengyang Liang, Hao He, Ceyuan Yang, and Bo Dai. Scaling laws for diffusion transformers, 2024. URL https://arxiv.org/abs/2410.08184.

Yaron Lipman, Ricky T. Q. Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow matching for generative modeling, 2022. URL https://arxiv.org/abs/2210.02747.

Jingyuan Liu, Jianlin Su, Xingcheng Yao, Zhejun Jiang, Guokun Lai, Yulun Du, Yidao Qin, Weixin Xu, Enzhe Lu, Junjie Yan, Yanru Chen, Huabin Zheng, Yibo Liu, Shaowei Liu, Bohong Yin, Weiran He, Han Zhu, Yuzhi Wang, Jianzhou Wang, Mengnan Dong, Zheng Zhang, Yongsheng Kang, Hao Zhang, Xinran Xu, Yutao Zhang, Yuxin Wu, Xinyu Zhou, and Zhilin Yang. Muon is scalable for llm training, 2025. URL https://arxiv.org/abs/2502.16982.

Liyuan Liu, Xiaodong Liu, Jianfeng Gao, Weizhu Chen, and Jiawei Han. Understanding the difficulty of training transformers, 2020. URL https://arxiv.org/abs/2004.08249.

Xingchao Liu, Chengyue Gong, and Qiang Liu. Flow straight and fast: Learning to generate and transfer data with rectified flow, 2022. URL https://arxiv.org/abs/2209.03003.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017. URL https://arxiv.org/abs/1711.05101.

Zeyu Lu, Zidong Wang, Di Huang, Chengyue Wu, Xihui Liu, Wanli Ouyang, and Lei Bai. Fit: Flexible vision transformer for diffusion model, 2024. URL https://arxiv.org/abs/2402.12376.

Lorenzo Noci, Sotiris Anagnostidis, Luca Biggio, Antonio Orvieto, Sidak Pal Singh, and Aurelien Lucchi. Signal propagation in transformers: Theoretical perspectives and the role of rank collapse, 2022. URL https://arxiv.org/abs/2206.03126.

William Peebles and Saining Xie. Scalable diffusion models with transformers, 2023. URL https://arxiv.org/abs/2212.09748.

Qi Qin, Le Zhuo, Yi Xin, Ruoyi Du, Zhen Li, Bin Fu, Yiting Lu, Jiakang Yuan, Xinyue Li, Dongyang Liu, Xiangyang Zhu, Manyuan Zhang, Will Beddow, Erwann Millon, Victor Perez, Wenhai Wang, Conghui He, Bo Zhang, Xiaohong Liu, Hongsheng Li, Yu Qiao, Chang Xu, and Peng Gao. Lumina-image 2.0: A unified and efficient image generative framework, 2025. URL https://arxiv.org/abs/2503.21758.

Zihan Qiu, Zekun Wang, Bo Zheng, Zeyu Huang, Kaiyue Wen, Songlin Yang, Rui Men, Le Yu, Fei Huang, Suozhi Huang, Dayiheng Liu, Jingren Zhou, and Junyang Lin. Gated attention for large language models: Non-linearity, sparsity, and attention-sink-free, 2025. URL https://arxiv.org/abs/2505.06708.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Stefano Ermon, Christopher D. Manning, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model, 2023. URL https://arxiv.org/abs/2305.18290.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models, 2022. URL https://arxiv.org/abs/2112.10752.

Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. Imagenet large scale visual recognition challenge, 2015. URL https://arxiv.org/abs/1409.0575.

Mojtaba Sahraee-Ardakan, Mauricio Delbracio, and Peyman Milanfar. The geometry of noise: Why diffusion models don't need noise conditioning, 2026. URL https://arxiv.org/abs/2602.18428.

Noam Shazeer. Glu variants improve transformer, 2020. URL https://arxiv.org/abs/2002.05202.

Mohammad Shoeybi, Mostofa Patwary, Raul Puri, Patrick LeGresley, Jared Casper, and Bryan Catanzaro. Megatron-lm: Training multi-billion parameter language models using model parallelism, 2019. URL https://arxiv.org/abs/1909.08053.

Yang Song, Jascha Sohl-Dickstein, Diederik P. Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations, 2021. URL https://arxiv.org/abs/2011.13456.

Jianlin Su, Murtadha Ahmed, Yu Lu, Shengfeng Pan, Wen Bo, and Yunfeng Liu. Roformer: Enhanced transformer with rotary position embedding. *Neurocomputing*, 568:127063, 2024. URL https://www.sciencedirect.com/science/article/abs/pii/S0925231223011864.

Qiao Sun, Zhicheng Jiang, Hanhong Zhao, and Kaiming He. Is noise conditioning necessary for denoising generative models? In *Proceedings of the 42nd International Conference on Machine Learning*, 2025. URL https://arxiv.org/abs/2502.13129.

Sho Takase, Shun Kiyono, Sosuke Kobayashi, and Jun Suzuki. Spike no more: Stabilizing the pre-training of large language models, 2023. URL https://arxiv.org/abs/2312.16903. Published at COLM 2025.

Philippe Tillet, H. T. Kung, and David Cox. Triton: an intermediate language and compiler for tiled neural network computations. In *Proceedings of the 3rd ACM SIGPLAN International Workshop on Machine Learning and Programming Languages*, MAPL 2019, pp. 10–19, New York, NY, USA, 2019. Association for Computing Machinery. ISBN 9781450367196. doi:10.1145/3315508.3329973. URL https://doi.org/10.1145/3315508.3329973.

Hugo Touvron, Matthieu Cord, Alexandre Sablayrolles, Gabriel Synnaeve, and Hervé Jégou. Going deeper with image transformers, 2021. URL https://arxiv.org/abs/2103.17239.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need, 2017. URL https://arxiv.org/abs/1706.03762.

Bram Wallace, Meihua Dang, Rafael Rafailov, Linqi Zhou, Aaron Lou, Senthil Purushwalkam, Stefano Ermon, Caiming Xiong, Shafiq Joty, and Nikhil Naik. Diffusion model alignment using direct preference optimization, 2023. URL https://arxiv.org/abs/2311.12908.

Hongyu Wang, Shuming Ma, Li Dong, Shaohan Huang, Dongdong Zhang, and Furu Wei. Deepnet: Scaling transformers to 1,000 layers, 2022. URL https://arxiv.org/abs/2203.00555.

Mitchell Wortsman, Peter J. Liu, Lechao Xiao, Katie Everett, Alex Alemi, Ben Adlam, John D. Co-Reyes, Izzeddin Gur, Abhishek Kumar, Roman Novak, Jeffrey Pennington, Jascha Sohl-dickstein, Kelvin Xu, Jaehoon Lee, Justin Gilmer, and Simon Kornblith. Small-scale proxies for large-scale transformer training instabilities, 2023. URL https://arxiv.org/abs/2309.14322.

Guangxuan Xiao, Yuandong Tian, Beidi Chen, Song Han, and Mike Lewis. Efficient streaming language models with attention sinks, 2023. URL https://arxiv.org/abs/2309.17453.

Ruibin Xiong, Yunchang Yang, Di He, Kai Zheng, Shuxin Zheng, Chen Xing, Huishuai Zhang, Yanyan Lan, Liwei Wang, and Tie-Yan Liu. On layer normalization in the transformer architecture, 2020. URL https://arxiv.org/abs/2002.04745.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jing Zhou, Jingren Zhou, Junyang Lin, Kai Dang, Keqin Bao, Kexin Yang, Le Yu, Lianghao Deng, Mei Li, Mingfeng Xue, Mingze Li, Pei Zhang, Peng Wang, Qin Zhu, Rui Men, Ruize Gao, Shixuan Liu, Shuang Luo, Tianhao Li, Tianyi Tang, Wenbiao Yin, Xingzhang Ren, Xinyu Wang, Xinyu Zhang, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yinger Zhang, Yu Wan, Yuqiong Liu, Zekun Wang, Zeyu Cui, Zhenru Zhang, Zhipeng Zhou, and Zihan Qiu. Qwen3 technical report, 2025. URL https://arxiv.org/abs/2505.09388.

Greg Yang, Edward J. Hu, Igor Babuschkin, Szymon Sidor, Xiaodong Liu, David Farhi, Nick Ryder, Jakub Pachocki, Weizhu Chen, and Jianfeng Gao. Tensor programs v: Tuning large neural networks via zero-shot hyperparameter transfer, 2022. URL https://arxiv.org/abs/2203.03466.

Z-Image Team, Huanqia Cai, Sihan Cao, Ruoyi Du, Peng Gao, Steven Hoi, Zhaohui Hou, Shijie Huang, Dengyang Jiang, Xin Jin, Liangchen Li, Zhen Li, Zhong-Yu Li, David Liu, Dongyang Liu, Junhan Shi, Qilong Wu, Feng Yu, Chi Zhang, Shifeng Zhang, and Shilin Zhou. Z-image: An efficient image generation foundation model with single-stream diffusion transformer, 2025. URL https://arxiv.org/abs/2511.22699.

Shuangfei Zhai, Tatiana Likhomanenko, Etai Littwin, Dan Busbridge, Jason Ramapuram, Yizhe Zhang, Jiatao Gu, and Josh Susskind. Stabilizing transformer training by preventing attention entropy collapse, 2023. URL https://arxiv.org/abs/2303.06296.

Biao Zhang and Rico Sennrich. Root mean square layer normalization. *Advances in Neural Information Processing Systems*, 32, 2019. URL https://neurips.cc/virtual/2019/poster/14195.

Hongyi Zhang, Yann N. Dauphin, and Tengyu Ma. Fixup initialization: Residual learning without normalization, 2019. URL https://arxiv.org/abs/1901.09321.

Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models, 2023. URL https://arxiv.org/abs/2302.05543.

Jie Zhu, Mingyu Ding, Boqiang Duan, Leye Wang, and Jingdong Wang. Unveiling the secret of adaln-zero in diffusion transformer. https://openreview.net/forum?id=E4roJSM9RM, 2025. URL https://openreview.net/forum?id=E4roJSM9RM. ICLR 2025.

A Diagnostic Metrics and Definitions

Table 2 provides the mathematical definitions for all diagnostic metrics referenced in our analysis. Spatially coherent metrics are estimated robustly on sampled token subsets during the live training pass.

Table 2: **Diagnostic Metrics Glossary.** Definitions for all diagnostic metrics referenced in our analysis. X denotes forward representations; A denotes attention matrices; Δ denotes backward gradients.

Metric	Formal Definition	Description
Writer GMD ($G_{\text{mean}}, G_{\text{ctr}}$)	$\|\Delta W_\mu\|_F$, $\|\Delta W_c\|_F$	Frobenius norms of the matrix components $\Delta W_\mu = T\bar{\delta}\bar{y}^\top$ and $\Delta W_c = \sum_t \tilde{\delta}_t \tilde{y}_t^\top$, decoupling the mean-coherent and centered components of the writer-weight update.
Q/K Grad Norm ($G(Q), G(K)$)	$\text{RMS}(\nabla_{W_{Q,K}} \mathcal{L})$	Root-mean-square of the gradients with respect to the query and key projection weights.
Energy Ratio (ρ_T)	$\frac{\|\mu(X)\|_F}{\|c(X)\|_F + \epsilon}$	Ratio between the energy of the mean component $\mu(X)$ and the centered component $c(X)$ of the token representation.
TR Ratio (r_{TR})	$\frac{\|U(X)\|_F}{\|X\|_F + \epsilon}$	Ratio between the residual-branch update $U(X)$ and the residual-stream state X.
Variance Gain (VarGain)	$\frac{\|c(U(X))\|_F}{\|c(X)\|_F + \epsilon}$	Ratio between the centered energy of the branch update $c(U(X))$ and the centered energy of the input $c(X)$.
Attn Contraction (μ_{eff})	$\|PAP\|_2$ (centered power iter.)	Spectral norm of the attention operator restricted to the centered subspace, where $P = I - J$ projects out the token mean.
Row Diversity (RowDiv)	$\frac{\|A - JA\|_F}{\|A\|_F}$	Relative deviation of the attention rows from their column-mean profile.
Centered Retention (Ret($c \leftarrow c$))	$\frac{\|c(A \cdot c(X))\|_F}{\|c(X)\|_F + \epsilon}$	Fraction of centered input energy that remains in the centered subspace after one attention operation.
Mean Leakage (Leakage($\mu \leftarrow c$))	$\frac{\|\mu(A \cdot c(X))\|_F}{\|c(X)\|_F + \epsilon}$	Fraction of centered input energy mapped into the mean subspace by the attention operator via JAP.
Token Cosine Similarity (TCS)	$\mathbb{E}_{i \neq j}[\cos(X_i, X_j)]$	Average pairwise cosine similarity between token representations, estimated on sampled token pairs when needed. High values indicate token homogenization.

B STANDARD INITIALIZATION ENTERS THE SAME MEAN-DOMINATED STATE

The main diagnostic runs use zero-initialized residual writers, a standard identity-start choice for deep residual networks. This choice keeps the early trajectory well behaved and makes the delayed MMS event easy to isolate. We now ask how the same backbone behaves when residual writers are initialized open from the first step. To test this, we train a 128-layer single-stream DiT with no residual gating and Gaussian initialization, $\mathcal{W} \sim \mathcal{N}(0, 0.02^2)$ (Shoeybi et al., 2019), for the residual writers and final projection. This control is not a matched convergence comparison; it tests whether the mean-dominated state is specific to the identity-start writer schedule.

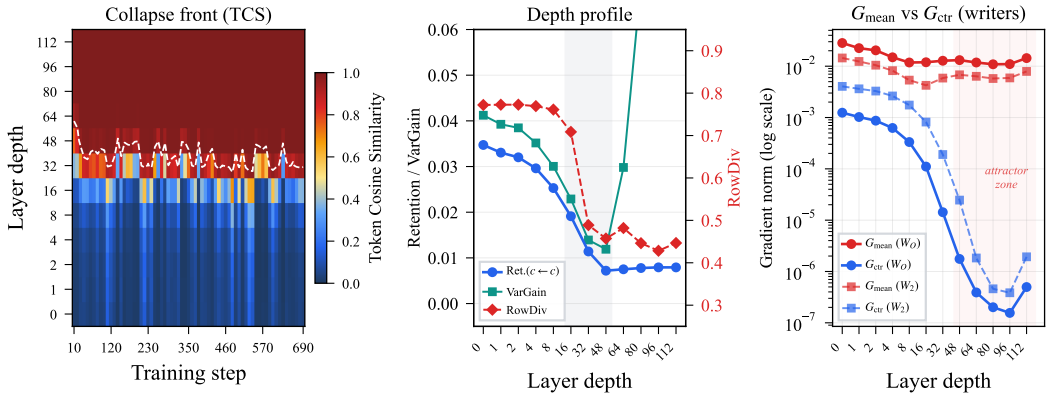

Figure 7: **Standard-initialization control for a 128-layer DiT. (a)** Token cosine similarity (TCS) over training steps and layer depth. The dashed white contour marks $\mathrm{TCS} = 0.9$. **(b)** Depth profiles of centered retention $\mathrm{Ret}(c \leftarrow c)$, centered branch replenishment VarGain, and attention row diversity RowDiv; curves report the median over diagnostic checkpoints from steps 10–690. **(c)** Median writer-gradient decomposition for the attention output projection W_O and FFN output projection W_2, reporting G_{mean} and G_{ctr} across depth.

Figure 7 shows that standard initialization enters the same mean-dominated regime. The temporal pattern differs from the zero-writer runs: instead of a delayed writer-gradient spike after an initially stable period, deep layers have high token similarity from the beginning of training, and the high-similarity region forms a depth-wise collapse front. In this run, the loss quickly reaches a high plateau, consistent with a residual stream that has lost most of its token-varying information in deep layers.

The forward diagnostics connect this behavior to the same subspace mechanism studied in the main text. In deep layers, centered retention and branch-side centered replenishment are both small, so centered variation is not maintained through depth. RowDiv remains nonzero, ruling out the simpler explanation that attention has literally collapsed to identical rows. The failure is instead a subspace imbalance: row-stochastic attention preserves pure-mean states, while the centered component is weakly retained and weakly replenished.

The writer-gradient decomposition shows the same imbalance on the backward path. In the deep collapsed layers, the mean-coherent writer component G_{mean} dominates the centered component G_{ctr} by several orders of magnitude for both W_O and W_2. Thus the standard-init run reaches the same endpoint as the zero-writer MMS runs: token variation is suppressed and residual writer updates become mean-dominated.

Writer initialization therefore changes the temporal presentation of the failure, not the subspace-level failure mode. Identity-start runs expose the failure as a delayed MMS event, while standard initialization exposes it as an early depth-wise collapse front. In both cases, the common failure channel is the imbalance between the invariant pure-mean direction and insufficient centered-subspace maintenance. MV-Split therefore targets the residual writer geometry rather than a peculiarity of the zero-writer schedule.

C Derivations of Writer Gradient Scaling

C.1 Proof of the Gradient Mode Decomposition

In Section 4.1, we state that the gradient admits an exact additive decomposition into a rank-1 mean-coherent component ΔW_μ and a centered variation component ΔW_c, with cross-terms identically vanishing. We provide the brief algebraic proof here.

Let the forward input $y_t \in \mathbb{R}^n$ and the backward gradient $\delta_t \in \mathbb{R}^m$ for token t be decomposed into their sequence means and zero-mean centered components:

$$y_t = \bar{y} + \tilde{y}_t, \quad \text{where} \quad \bar{y} = \frac{1}{T}\sum_{t=1}^{T} y_t, \quad \sum_{t=1}^{T} \tilde{y}_t = \mathbf{0}, \tag{11}$$

$$\delta_t = \bar{\delta} + \tilde{\delta}_t, \quad \text{where} \quad \bar{\delta} = \frac{1}{T}\sum_{t=1}^{T} \delta_t, \quad \sum_{t=1}^{T} \tilde{\delta}_t = \mathbf{0}. \tag{12}$$

The parameter gradient for a token-wise linear map W is the sum of outer products over the sequence length. Substituting the decomposed terms yields:

$$\nabla_W \mathcal{L} \;=\; \sum_{t=1}^{T} \delta_t y_t^\top \;=\; \sum_{t=1}^{T} \big(\bar{\delta} + \tilde{\delta}_t\big)\big(\bar{y} + \tilde{y}_t\big)^\top. \tag{13}$$

Expanding the outer product gives four summation terms:

$$\nabla_W \mathcal{L} \;=\; \sum_{t=1}^{T} \bar{\delta}\bar{y}^\top \;+\; \sum_{t=1}^{T} \bar{\delta}\tilde{y}_t^\top \;+\; \sum_{t=1}^{T} \tilde{\delta}_t\bar{y}^\top \;+\; \sum_{t=1}^{T} \tilde{\delta}_t\tilde{y}_t^\top. \tag{14}$$

Because the sequence means $\bar{y}$ and $\bar{\delta}$ are constant across tokens, they can be factored out of the summations for the cross-terms:

$$\sum_{t=1}^{T} \bar{\delta}\tilde{y}_t^\top = \bar{\delta}\underbrace{\left(\sum_{t=1}^{T} \tilde{y}_t^\top\right)}_{=\mathbf{0}^\top} = \mathbf{0}, \tag{15}$$

$$\sum_{t=1}^{T} \tilde{\delta}_t\bar{y}^\top = \underbrace{\left(\sum_{t=1}^{T} \tilde{\delta}_t\right)}_{=\mathbf{0}}\bar{y}^\top = \mathbf{0}. \tag{16}$$

The cross-terms evaluate identically to zero matrices, and the first term sums to $T\,\bar{\delta}\,\bar{y}^\top$. The gradient therefore admits an exact additive decomposition into a mean-coherent rank-1 component (ΔW_μ) and a centered component (ΔW_c):

$$\nabla_W \mathcal{L} \;=\; \underbrace{T\,\bar{\delta}\,\bar{y}^\top}_{\Delta W_\mu} \;+\; \underbrace{\sum_{t=1}^{T} \tilde{\delta}_t\tilde{y}_t^\top}_{\Delta W_c}. \tag{17}$$

This recovers Equation 5 as an algebraic identity rather than an approximation.

C.2 Derivation of the Alignment-Amplification Law

We derive Eq. (6) and its equal-magnitude specialization.

Exact expansion and diagonal–off-diagonal split. Let W be a token-wise linear map with gradient

$$\nabla_W \mathcal{L} \;=\; \sum_{t=1}^{T} \delta_t y_t^\top \;\in\; \mathbb{R}^{m\times n}, \tag{18}$$

where $y_t \in \mathbb{R}^n$ is the forward input and $\delta_t \in \mathbb{R}^m$ is the corresponding backward gradient for token t. Using the Frobenius inner-product identity for rank-1 matrices,

$$\langle ab^\top,\; cd^\top \rangle_F = \langle a, c\rangle \, \langle b, d\rangle,$$

we obtain

$$\|\nabla_W \mathcal{L}\|_F^2 \;=\; \Big\langle \sum_{t=1}^{T} \delta_t y_t^\top,\; \sum_{s=1}^{T} \delta_s y_s^\top \Big\rangle_F \;=\; \sum_{t=1}^{T}\sum_{s=1}^{T} \langle y_t, y_s\rangle \, \langle \delta_t, \delta_s\rangle. \tag{19}$$

Separating diagonal and off-diagonal terms gives

$$\|\nabla_W \mathcal{L}\|_F^2 \;=\; \underbrace{\sum_{t=1}^{T} \|y_t\|^2 \|\delta_t\|^2}_{S} \;+\; \underbrace{\sum_{s\neq t} \langle y_t, y_s\rangle \, \langle \delta_t, \delta_s\rangle}_{C}. \tag{20}$$

Here S is the diagonal, no-interaction baseline, while C collects the cross-token interference terms.

Rayleigh-quotient form. Define the per-token magnitude

$$w_t \triangleq \|\delta_t\| \, \|y_t\|, \qquad w \in \mathbb{R}^T_{\geq 0}, \tag{21}$$

and the pairwise alignment matrix

$$M_{ts} \triangleq \cos(y_t, y_s) \, \cos(\delta_t, \delta_s). \tag{22}$$

Then Eq. (19) becomes

$$\|\nabla_W \mathcal{L}\|_F^2 \;=\; w^\top M w, \tag{23}$$

while the diagonal baseline is

$$S = \sum_{t=1}^{T} w_t^2 = \|w\|_2^2, \tag{24}$$

since $M_{tt} = 1$. Therefore the alignment amplification (cf. main text Eq. 6) is

$$\mathcal{A} \;\triangleq\; \frac{\|\nabla_W \mathcal{L}\|_F^2}{S} \;=\; \frac{w^\top M w}{\|w\|_2^2}, \tag{25}$$

and subtracting the diagonal baseline yields

$$\mathcal{A} - 1 \;=\; \frac{w^\top (M - I) w}{\|w\|_2^2} \;=\; \frac{\sum_{s\neq t} w_t w_s M_{ts}}{\sum_{t=1}^{T} w_t^2}. \tag{26}$$

This is exactly Eq. (6) in the main text.

Moreover, M is positive semidefinite. Indeed, it is the Hadamard product of the cosine Gram matrix of $\{y_t\}$ and that of $\{\delta_t\}$; both are positive semidefinite, and the Schur Product Theorem preserves positive semidefiniteness. Thus $\mathcal{A}$ is a Rayleigh quotient of a PSD matrix.

Equal-magnitude specialization. Suppose the per-token magnitude is approximately constant,

$$w_t \equiv w_0. \tag{27}$$

Then Eq. (26) reduces to

$$\mathcal{A} - 1 \;=\; \frac{1}{T} \sum_{s\neq t} M_{ts}. \tag{28}$$

Define

$$\kappa \triangleq \mathbb{E}_{s\neq t}[M_{ts}] \;=\; \mathbb{E}_{s\neq t}\big[\cos(y_t, y_s)\cos(\delta_t, \delta_s)\big], \tag{29}$$

where $\mathbb{E}_{s\neq t}$ denotes the uniform average over ordered pairs (s, t) with $s \neq t$. Then

$$\mathcal{A} - 1 \;=\; (T-1)\,\kappa. \tag{30}$$

When $\kappa \approx 0$, the cross-terms cancel on average and

$$\|\nabla_W \mathcal{L}\|_F^2 = \mathcal{O}(T), \qquad \|\nabla_W \mathcal{L}\|_F = \mathcal{O}(\sqrt{T}).$$

When $\kappa \to 1$, the accumulation becomes coherent and

$$\|\nabla_W \mathcal{L}\|_F^2 = \mathcal{O}(T^2), \qquad \|\nabla_W \mathcal{L}\|_F = \mathcal{O}(T).$$

Absolute-coherence upper bound. From Eq. (26),

$$|\mathcal{A} - 1| \;\le\; \frac{\sum_{s\neq t} w_t w_s |M_{ts}|}{\sum_{t=1}^{T} w_t^2}. \tag{31}$$

Under the equal-magnitude approximation, this becomes

$$|\mathcal{A} - 1| \;\le\; (T-1)\,\hat{\kappa}, \qquad \hat{\kappa} \triangleq \mathbb{E}_{s\neq t}\big[|\cos(y_t, y_s)|\,|\cos(\delta_t, \delta_s)|\big]. \tag{32}$$

In the main-text experiment, $\hat{\kappa}$ is used as an absolute-coherence proxy. The gap between $(T-1)\hat{\kappa}$ and the signed quantity $\mathcal{A} - 1$ reflects signed cancellation, together with any looseness introduced by the absolute-value relaxation.

C.3 Proof of Lemma 1

Let $a_i \in \mathbb{R}^T$ denote the i-th attention row written as a column vector, so that

$$Y_i \;=\; \sum_{j=1}^{T} a_{ij} V_j, \qquad a_i \;=\; \mathrm{softmax}(S_i). \tag{33}$$

By the chain rule,

$$\frac{\partial \mathcal{L}}{\partial a_{ij}} \;=\; \Big\langle \frac{\partial \mathcal{L}}{\partial Y_i},\, V_j \Big\rangle. \tag{34}$$

If $V_j = \bar{v}$ for all j, then the right-hand side is independent of j. Hence

$$\frac{\partial \mathcal{L}}{\partial a_i} \;=\; \gamma_i \mathbf{1}, \qquad \gamma_i \triangleq \Big\langle \frac{\partial \mathcal{L}}{\partial Y_i},\, \bar{v} \Big\rangle. \tag{35}$$

The softmax Jacobian at a_i is

$$J_{\mathrm{sm}}(a_i) \;=\; \mathrm{diag}(a_i) - a_i a_i^{\top}, \tag{36}$$

and satisfies

$$J_{\mathrm{sm}}(a_i)\mathbf{1} \;=\; a_i - a_i(\mathbf{1}^{\top} a_i) \;=\; \mathbf{0}, \tag{37}$$

since $\mathbf{1}^{\top} a_i = 1$. Exploiting the symmetry of the Softmax Jacobian,

$$\frac{\partial \mathcal{L}}{\partial S_i} \;=\; J_{\mathrm{sm}}(a_i)^{\top} \frac{\partial \mathcal{L}}{\partial a_i} \;=\; J_{\mathrm{sm}}(a_i) \frac{\partial \mathcal{L}}{\partial a_i} \;=\; \gamma_i\, J_{\mathrm{sm}}(a_i)\mathbf{1} \;=\; \mathbf{0}. \tag{38}$$

Since this holds for every row i, the logit gradient vanishes identically, which proves Lemma 1.

Residual-writer gradients bypass the Softmax null space. The null-space argument above concerns only the gradient through the attention logits S_i, and therefore the Q/K pathway. It does not zero the attention output projection. If $H_i = \sum_j a_{ij} V_j$ denotes the pre-W_O attention output and g_i is the upstream adjoint at the output projection, then

$$\nabla_{W_O} \mathcal{L} = \sum_i g_i H_i^{\top}, \tag{39}$$

which bypasses the Softmax Jacobian. Under value homogenization, $H_i \to \bar{H}$ becomes approximately token-constant, so the writer gradient becomes mean-coherent rather than zero. Gradients to the value pathway can also remain nonzero: the strict null-space extinction applies to the logit, and hence to the Q/K pathway, only.

D Detailed Comparison: MV-Split vs. LayerScale and ReZero

LayerScale (Touvron et al., 2021) parameterizes the merge as $Z_l^{\mathrm{LS}} = X_l + \lambda_l \odot F_l$ with $\lambda_l \in \mathbb{R}^D$ a per-channel learnable vector; ReZero (Bachlechner et al., 2020) is the single-scalar special case $\lambda_l \in \mathbb{R}$ initialized at zero. We show that both differ from MV-Split in three structural respects, each of which corresponds to a specific failure mode in Sec. 4.

1. Open-loop vs. leaky-integrator mean dynamics. Projecting the two merges into the mean subspace via J:

$$JZ_l^{\mathrm{LS}} = JX_l + (\lambda_l \odot JF_l), \qquad JZ_l^{\mathrm{MV}} = (1-\alpha) \odot JX_l + \alpha \odot JF_l. \tag{40}$$

LayerScale leaves the trunk's mean component *untouched at every layer* (the coefficient of JX_l is identically 1): it does not damp the carried trunk mean and only scales newly injected branch updates. MV-Split contracts the trunk's mean by $1-\alpha$ before each injection, which is a leaky integrator whenever $\alpha_d \in (0,1)$. The two cannot be made equivalent by any choice of λ_l: taking $F_l \equiv 0$ gives $JZ_l^{\mathrm{LS}} = JX_l$ while $JZ_l^{\mathrm{MV}} = (1-\alpha) \odot JX_l$, so the dynamics differ whenever $\alpha \neq 0$, irrespective of λ_l.

2. Isotropic vs. anisotropic gain on the residual branch. By Eq. 5 the gradient decomposes as $\nabla_W \mathcal{L} = \Delta W_\mu + \Delta W_c$ with $\|\Delta W_\mu\|_F \sim T\hat{\kappa}$ in the coherent regime and $\|\Delta W_c\|_F$ scaling diffusively under weak centered alignment. In the scalar-gain simplification, both modes are scaled by the same gain:

$$\frac{\|\Delta W_\mu^{\mathrm{LS}}\|_F}{\|\Delta W_c^{\mathrm{LS}}\|_F} \propto \sqrt{T}\hat{\kappa}. \tag{41}$$

For scalar gates like ReZero, the ratio is exactly invariant. LayerScale uses a per-channel diagonal operator $\lambda_l \in \mathbb{R}^D$; while feature-wise anisotropy can incidentally alter the mean/centered ratio, it provides no structural token-subspace filter. MV-Split scales the two modes by α and β independently. In the scalar-gain simplification:

$$\frac{\|\Delta W_\mu^{\mathrm{MV}}\|_F}{\|\Delta W_c^{\mathrm{MV}}\|_F} \propto \frac{\alpha}{\beta}\sqrt{T}\hat{\kappa}, \tag{42}$$

allowing the unstable-to-stable ratio α/β to be reduced without coupling to the absolute centered gain β. If the leaky mean replacement term were removed (i.e. replacing $\alpha \odot J(F_l - X_l)$ in Eq. 7 by $\alpha \odot JF_l$), the special case $\alpha = \beta$ would reduce to a LayerScale-like token-space isotropic branch gain. With the leaky term in Eq. 7, however, MV-Split remains dynamically distinct from LayerScale even when $\alpha = \beta$, because $JZ_l^{\mathrm{MV}} = (1-\alpha) \odot JX_l + \alpha \odot JF_l$ contracts the trunk's mean component at every layer (cf. §1 above), whereas LayerScale leaves it untouched.

3. Independent gain on the centered path. Whatever absolute gain on the centered branch is needed for stability at a given depth, MV-Split treats it as a free parameter set independently of α (Eq. 9). LayerScale ties the two paths to the same token-independent per-channel gain, so any reduction in the mean-coherent contribution unavoidably reduces centered replenishment by the same factor.

Sec. 4 describes a self-reinforcing failure: a gradient spike injects mean-mode content into the trunk, the trunk's mean direction aligns with the residual branch's coherent direction, and this alignment amplifies the next mean-coherent update. The two MV-Split gates intervene at two different points: α contracts the trunk's mean component at every layer (Eq. 9), bounding mean accumulation; the α/β gap damps the mean-coherent update relative to the centered update (Eq. 10), interrupting the alignment-amplification step. A scalar gate scales both modes equally, while LayerScale applies no explicit token-subspace filter; neither contracts the carried trunk mean.

E SEGMENT-WISE PROJECTORS FOR MULTIMODAL SEQUENCES

For a sequence partitioned into image tokens $\mathcal{I}$ and text tokens $\mathcal{T}$, we define group-mean projectors that average within each segment only:

$$J_{\text{seg}} = \text{blkdiag}(J_{\mathcal{I}}, J_{\mathcal{T}}), \qquad P_{\text{seg}} = I - J_{\text{seg}}, \tag{43}$$

where $J_{\mathcal{I}} = \frac{1}{|\mathcal{I}|}\mathbf{1}_{\mathcal{I}}\mathbf{1}_{\mathcal{I}}^{\top}$ and similarly for $J_{\mathcal{T}}$. The projector does not average across modalities; it prevents the residual merge from directly mixing image and text means through the mean operator, preserving modality-specific mean scales in the residual control path. The diagnostic and mechanistic derivations in the main text use the global sequence-mean projector; the segment-wise projector is used only in the multimodal MV-Split residual merge.

Segment-wise control still acts on the global mean mode. Let $J_g = \frac{1}{T}\mathbf{1}\mathbf{1}^{\top}$ denote the global projector. The global mean subspace is contained in the segment-wise mean subspace, so

$$J_g\, J_{\text{seg}} = J_g, \qquad J_g\, P_{\text{seg}} = \mathbf{0}. \tag{44}$$

Because the gates α, β are feature-wise and broadcast across tokens, they commute with the token projectors. Applying J_g to the pre-normalization MV-Split merge with J, P instantiated as $J_{\text{seg}}, P_{\text{seg}}$ (Eq. 7) gives

$$J_g Z_l \;=\; J_g X_l \;+\; \beta \odot J_g\, P_{\text{seg}} F_l \;+\; \alpha \odot J_g\, J_{\text{seg}}(F_l - X_l) \;=\; (1-\alpha) \odot J_g X_l + \alpha \odot J_g F_l. \tag{45}$$

Thus the segment-wise implementation applies the same leaky control to the global MMS mode that the main-text theory analyzes, while avoiding direct averaging of image and text means in the residual control path.

F STEP-LEVEL GRADIENT TRACE FOR FAILURE ATTRIBUTION

The main text analyzes MMS through token-space and gradient decompositions. This appendix describes the inline trace that audits the representative gradient spike before those subspace diagnostics are applied: it rules out data/loss-side artifacts, localizes where large gradients appear, and identifies which internal quantities to measure next.

Trace protocol. Figure 8 summarizes the pipeline. When the global gradient norm crosses the threshold, the training loop records (i) per-rank loss, loss weight, and output-gradient statistics, (ii) a `NaN`/`Inf` scan over stored parameters, (iii) distributed gradient norms grouped by layer and parameter family, and (iv) at instrumented residual writers, the mean-coherent and centered components from Eq. 5.

For a parameter family τ at layer l, the grouped norm aggregated across R distributed ranks is

$$G_{l,\tau}(t) = \left(\sum_{r=1}^{R} \sum_{\theta \in \Theta_{l,\tau}} \|\nabla_\theta \mathcal{L}_r(t)\|_F^2 \right)^{1/2}. \tag{46}$$

This top-K grouping localizes which parameter families receive large gradients at the detected step; it does not identify the responsible token-space mode.

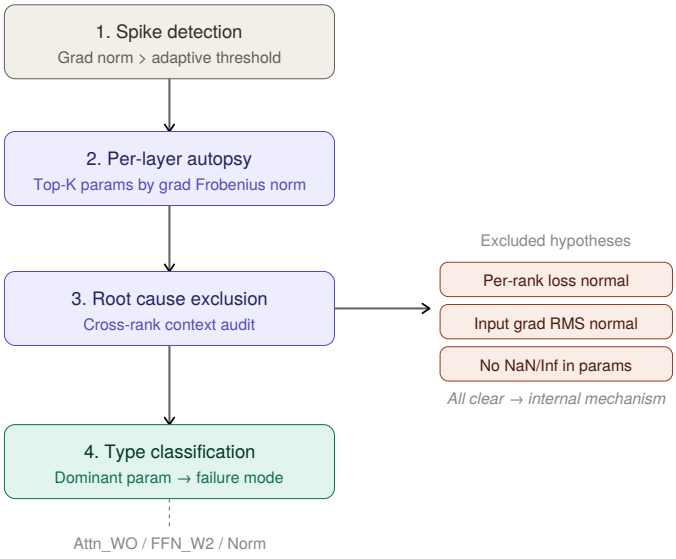

Figure 8: **Step-level gradient trace pipeline.** A global-norm threshold (1) triggers a per-family top-K ranking of distributed gradient norms (2) and a cross-rank exclusion audit (3) that checks per-rank loss agreement, final-output-gradient RMS, and `NaN`/`Inf` in stored parameters. When all three exclusions pass, the dominant top-K parameter family at the detected step (4) is recorded for the gradient-mode audit and subsequent paragraphs.

Data/loss-side checks. At the detected steps in the representative trace, per-rank losses remain clustered and the final-output-gradient statistics stay small at the printed precision. The maximum per-sample output-gradient norm is also nearly identical across ranks. This rules out two simple explanations: a single-rank data outlier and a global loss-weighting jump. The parameter scan finds no `NaN`/`Inf` values in stored parameters, so the event is not explained by persistent parameter corruption.

Parameter-family localization. We examine the representative trace in the lower-learning-rate baseline (Base $\eta/2$, no MV-Split, no LayerScale), where divergence at Step 26423 offers a longer pre-spike window than the default-LR run in Section 6.1 (Base η, $t^\star$=3400). The first warning snapshots

are mixed: their largest entries include embedding/final parameters, Q/K/V projections, FFN input weights, and residual output projections (Figure 9, left). We therefore do not interpret these early warnings as a fixed shallow-layer mechanism. At the escalation step, the largest printed entries shift toward residual output interfaces, with `Attn_WO` accounting for most of the top-K squared-norm mass and `FFN_W2` also appearing. This localization motivates auditing the residual writers directly rather than attributing the event to the output head, the batch, or a specific attention-logit pathology.

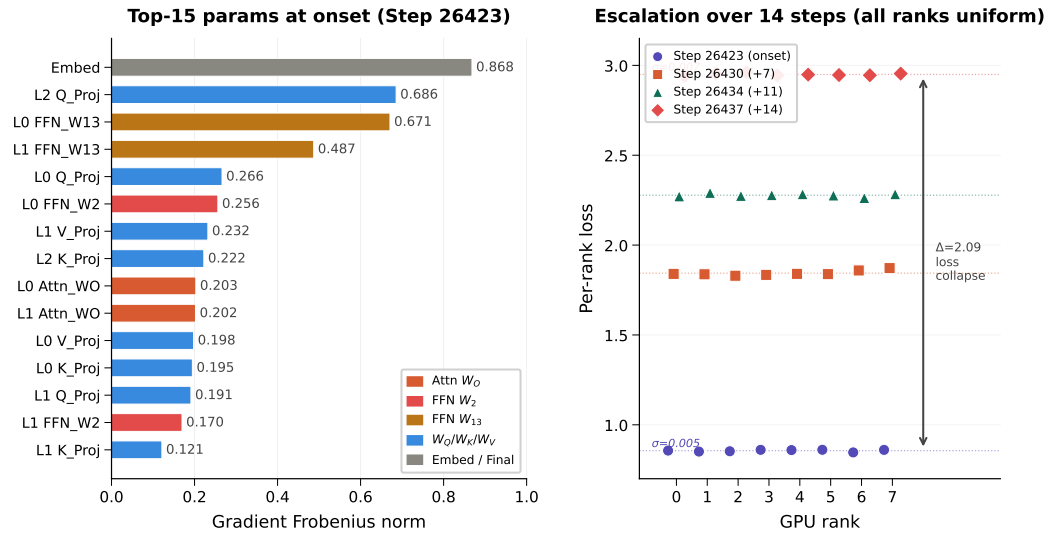

Figure 9: **Step-level gradient trace at a representative spike (400 layers, Base $\eta/2$). (Left)** Top parameter-family gradient norms $G_{l,\tau}$ at one detected step (Step 26423). The top-K entries span embedding/final parameters, Q/K/V projections, FFN input weights, and residual output projections. **(Right)** Per-rank loss across four snapshots (Steps 26423, 26430, 26434, 26437). The eight ranks stay tightly clustered ($\sigma \in [0.005, 0.007]$) and the maximum per-sample output-gradient norm $M_{\text{out}}^{(r)} \approx 1 \times 10^{-4}$ matches across ranks (inset).

Gradient-mode audit. For each instrumented writer, the trace caches the writer input y_t during the forward pass and the output adjoint δ_t during the backward pass. It then computes

$$\Delta W_\mu = T \bar{\delta} \bar{y}^\top, \qquad \Delta W_c = \sum_t \tilde{\delta}_t \tilde{y}_t^\top, \tag{47}$$

and reports $G_{\text{mean}} = \|\Delta W_\mu\|_F$ and $G_{\text{ctr}} = \|\Delta W_c\|_F$. This is the measurement that links the top-K localization to the mechanism in Section 4.1: at the spike, the writer update is amplified in the mean-coherent component, while the centered component does not show a comparable increase.

Attention-branch-only control. To test whether protecting the attention branch alone is sufficient, we apply MV-Split only to the attention output residual branch and leave the FFN residual branch unchanged. We tested this on a 1000-layer configuration. The training still spikes at Step 7415 (Global Norm ≈ 0.665), and the largest printed gradients move to the unprotected FFN branch: `Attn_WO` disappears from the top-K entirely, while `FFN_W2` accounts for 14 of the top 15 contributors and about 93% of the top-K squared-norm mass at this step (Figure 10). Depending on the step, these entries can include both the FFN input transformation and the FFN output projection.

The subsequent escalation is rapid. Within four steps (7415→7419), per-rank loss rises uniformly from $\sim$0.81 to $\sim$2.12 across all 16 ranks, with per-step standard deviation remaining low ($\sigma \in [0.011, 0.023]$), consistent with a global update event rather than a single-rank or single-batch fault.

The conclusion is therefore branch-level rather than weight-level: attention-only residual control is insufficient, and both attention and FFN residual branches require the mean/centered split.

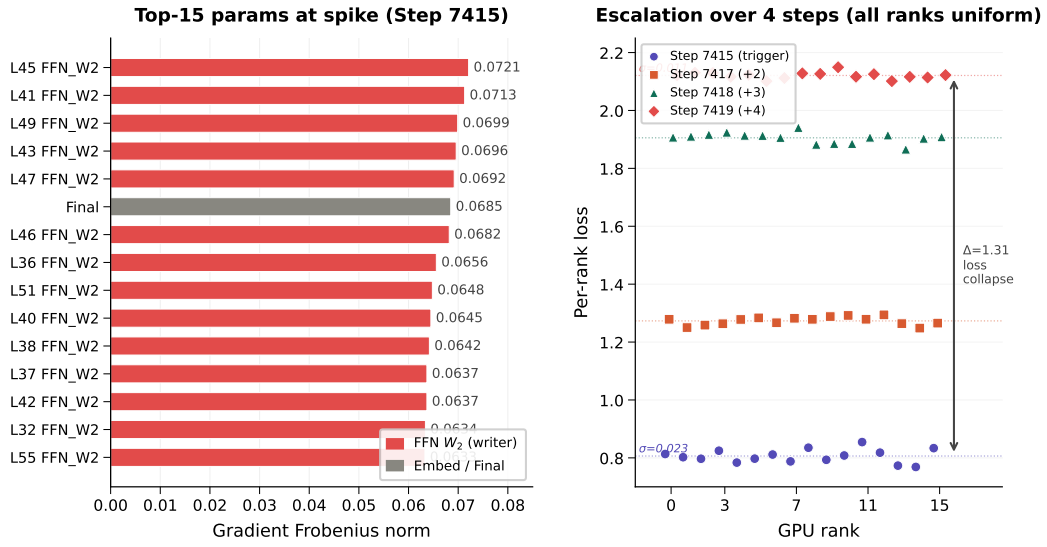

Figure 10: **Attention-branch-only MV-Split control (1000 layers). (Left)** Top parameter-family gradient norms at the detected step (Step 7415). With the attention output branch protected, no `Attn_WO` entries appear in the top-K; the largest printed entries are `FFN_W2`. **(Right)** Per-rank loss over four consecutive steps (7415–7419). Cross-rank losses stay tightly clustered ($\sigma \in [0.011, 0.023]$) while the loss rises uniformly from $\sim$0.81 to $\sim$2.12.

G TRAINING CONFIGURATION DETAILS

Table 3 summarizes the architecture and optimization hyperparameters for the four DiT configurations reported in this paper: **DiT-400L-Baseline** (400-layer Post-Norm without residual gating, used to characterize the Mean-Mode Screaming failure), **DiT-400L-LayerScale** (the matched 400-layer LayerScale control), **DiT-400L-MVSplit** (the matched 400-layer model with MV-Split residuals), and **DiT-1000L-MVSplit** (the 1000-layer text-to-image scale-up demonstration). The base learning rate for each run is obtained from the target value via μP (Yang et al., 2022) width-scaling, base $=$ target$/(0.2\sqrt{d_{\text{model}}})$ with target $= 10^{-3}$ and $d_{\text{model}} = 1024$, yielding base $= 1.5625\times10^{-4}$.

Table 3: **Architecture and training hyperparameters for the four DiT runs reported in this paper.** The three 400-layer controls share the same backbone, optimizer, batch size, and non-residual primitives, and vary in residual stabilization and initialization protocol. The 1000-layer run uses the same backbone family and MV-Split residual design, but differs in depth, hardware scale, and post-training pipeline. In MV-Split, α and β are unconstrained learnable vectors; empirically, α remains in $[0, 1]$ throughout training in all reported MV-Split runs. For the 1000-layer run, β_{init} is set to $0.03 \approx 1/\sqrt{L}$ following standard depth-variance scaling; the MMS protection itself stems from the anisotropic split ($\alpha_{\text{init}}{=}0 < \beta_{\text{init}}$) rather than from isotropic shrinkage.

Field	DiT-400L-Baseline	DiT-400L-LayerScale	DiT-400L-MVSplit	DiT-1000L-MVSplit
Pretraining Dataset	ImageNet-2012			
Image Autoencoder	Frozen FLUX.2 VAE			
Text Encoder	Frozen Qwen3-0.6B			
Trainable Components	DiT backbone only			
Training Hardware	8×H100	8×H100	8×H100	16×H100
Post-training Dataset	—	—	—	∼50k curated images
DiT Params	5.45 B	5.45 B	5.45 B	13.64 B
Layers	400	400	400	1000
Residual Mode	None	LayerScale	Mean-Variance Split	Mean-Variance Split
Residual Gates	—	learnable λ, $\lambda_{\text{init}} \in \{10^{-2}, \ldots, 10^{-5}\}$	learnable α, β; $\alpha_{\text{init}}{=}0,\ \beta_{\text{init}}{=}1$	learnable α, β; $\alpha_{\text{init}}{=}0,\ \beta_{\text{init}}{=}0.03$
Learning Rate	1.5625×10^{-4}, 7.8125×10^{-5}	1.5625×10^{-4}	1.5625×10^{-4}	1.5625×10^{-4}
Initialization Method	zero init W_O, W_2; standard init others	standard init, $\mathcal{N}(0, 0.02^2)$	zero init W_O, W_2; standard init others	zero init W_O, W_2; standard init others
Dimension (d_{model})	1024			
FFN Dimension	3072			
FFN Type	SwiGLU			
Attention Heads	8			
Attention Head Dim	128			
KV Heads	8			
Attention Type	MHA			
Position Embedding	2D RoPE			
RoPE θ	10000			
Layer Norm Type	RMSNorm, non-affine			
RMSNorm Affine Gain	disabled			
RMSNorm ϵ	10^{-6}			
QK-Norm	✓, non-affine			
LR Scheduler	warmup → constant			
Warmup Steps	1000			
Global Batch Size	1024			
Optimizer	AdamW (Loshchilov & Hutter, 2017)			
AdamW Betas	$(0.9,\ 0.999)$			
AdamW ϵ	10^{-8}			
Weight Decay	0.1 (2D Weights Only)			
Training Steps	crashed	100 k	100 k	100 k
Gradient Clipping	1.0			

H THE TOKEN MEAN AS AN IMPLICIT TIMESTEP CARRIER

Our backbone has no AdaLN and no explicit timestep embedding, so the model must infer the continuous rectified-flow time t from the noisy latent itself. In this appendix, we use "timestep" to refer to this continuous interpolation coordinate; equivalently, it is the noise-level coordinate controlling $z_t = (1-t)x_0 + tx_1$ from data latent to Gaussian noise.

We run a post-hoc linear probe on ImageNet-2012 validation images. Each image is encoded into the same VAE latent space used during training. We sample $x_1 \sim \mathcal{N}(0, I)$ and $t \sim U[0, 1]$, form $z_t = (1-t)x_0 + tx_1$, and record hidden states from the trained 400-layer MV-Split checkpoint. For each probed layer, we fit ridge regressors to predict t from the image-token mean m_l^{img}, a centered image-token RMS summary $\text{rms}(c_l^{\text{img}})$, and the text-token mean m_l^{txt}. Train/test splits are grouped by image id; scalar input-statistic, shuffled-label, and untrained-model controls are included. For panel (b), we report the fraction of residual squared error left by scalar input statistics that is removed by adding a hidden-state feature, $1 - \text{SSE}_{\text{input}+h}/\text{SSE}_{\text{input}}$.

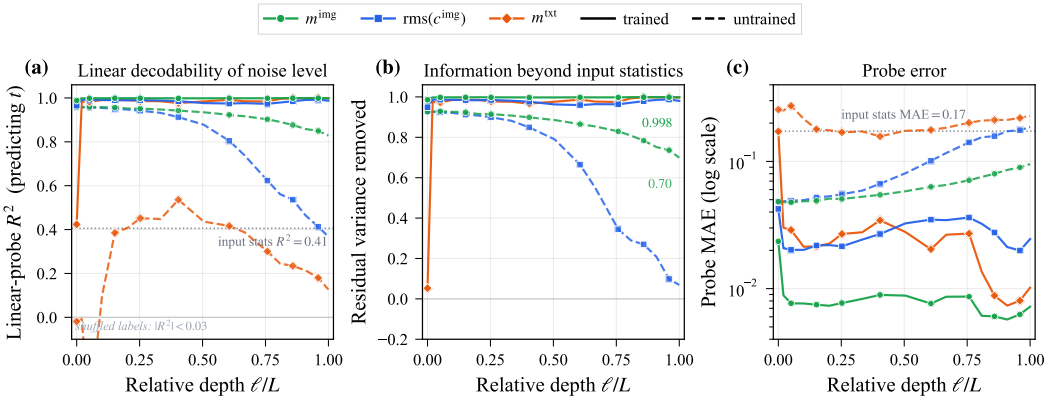

Figure 11: **Real-image timestep linear probe.** The backbone has no explicit timestep embedding or AdaLN modulation. **(a)** The trained image-token mean m^{img} predicts t with near-perfect linear R^2 across depth. The text-token mean m^{txt} becomes predictive after a few single-stream layers, indicating that the trained model routes image-derived timestep information into the text-token side. **(b)** Adding hidden-state summaries removes nearly all residual error left by scalar input statistics. **(c)** Probe MAE (mean absolute error) shows the same pattern on a log scale. The result shows that the token mean is not merely a collapse-prone direction: it is also a useful global timestep carrier. The same coordinate is also decodable from centered-energy summaries, so the claim is not uniqueness of the mean subspace but its usefulness and stability as a global-state path.

Figure 11 shows that the trained image-token mean predicts t with near-perfect linear R^2 across depth. The trained text-token mean becomes predictive after only a few single-stream layers, indicating that image-derived timestep information is routed into the shared multimodal sequence. A randomly initialized network already exposes substantial decodability in early image-token states, showing that the timestep is structurally available from the input latent; training preserves this signal through depth and routes it to the text-token side.

The token mean is therefore not only a collapse-prone direction but also a useful global-state carrier for the timestep. MV-Split preserves the trunk mean while gain-limiting new mean-path residual writes, controlling the dangerous mean-coherent writer channel without erasing this useful global state.

I System Implementation: Triton Fusion of RoPE, QK-Norm, SwiGLU, and MV-Split+RMSNorm

At the 8–16 H100 scale used in this work, training the ultra-deep DiT requires activation checkpointing (Chen et al., 2016) to fit in GPU memory. Checkpointing reduces activation memory, but it also replays checkpointed blocks during the backward pass. As a result, lightweight per-block operators such as RoPE, QK-Norm (Henry et al., 2020; Dehghani et al., 2023), SwiGLU (Shazeer, 2020), and MV-Split+RMSNorm are executed far more frequently than in a forward-only view of the model, making their memory-bound overhead non-negligible in ultra-deep training (Dao et al., 2022).

Fused operators. We implement these operators in Triton (Tillet et al., 2019). For MV-Split+RMSNorm, an eager implementation materializes the pre-normalized residual state and launches separate kernels for segment-wise correction, residual merging, and RMS normalization. Given precomputed segment means, the fused kernel applies the segment-wise MV-Split update and the subsequent non-affine RMS normalization without materializing the pre-normalized intermediate residual. Its backward uses pointwise recomputation together with compact segment-wise sufficient statistics, rather than caching the full pre-normalized state across the checkpoint boundary.

Two-pass backward recomputation. Let $F_l = f_l(X_l)$ denote the residual branch output and let Z_l be the pre-normalized MV-Split merge:

$$Z_l = X_l + \beta \odot P_{\text{seg}} F_l + \alpha \odot J_{\text{seg}}(F_l - X_l), \qquad X_{l+1} = \text{RMSNorm}(Z_l). \tag{48}$$

The backward does not require caching Z_l across the checkpoint boundary. Pass A recomputes Z_l on chip, evaluates the RMSNorm adjoint, and accumulates the segment-wise sufficient statistics needed for the input and gate gradients; Pass B then applies the closed-form gradients to X_l and F_l using the aggregated statistics. The full derivation is provided in Section I.1 below.

In-situ profiling. We evaluate the fused backend inside the distributed training loop rather than with isolated microbenchmarks. In our 8-GPU, 400-block profiling setup with activation checkpointing applied to three out of every four blocks, 300 blocks are replayed during backward. Consequently, RoPE, QK-Norm, and SwiGLU are each executed 700 times per active optimizer step, while MV-Split+RMSNorm is executed 1400 times.

Relative to a matched eager PyTorch baseline, the fused Triton backend reduces the aggregated self-CUDA time of these operators from 1697.4 ms to 614.0 ms per active optimizer step ($2.76\times$). The individual reductions are from 359.7 ms to 105.6 ms for RoPE ($3.41\times$), 255.2 ms to 101.2 ms for QK-Norm ($2.52\times$), 118.3 ms to 27.9 ms for SwiGLU ($4.24\times$), and 964.2 ms to 379.3 ms for MV-Split+RMSNorm ($2.54\times$). The explicit DiT forward range decreases from 1889.8 ms to 1553.9 ms, and the in-loop optimizer-step wall-clock decreases by 22.0%, from 5.87 s to 4.58 s, excluding dataloader wait. QKV projection and SDPA remain within a few percent under the same instrumentation, localizing the speedup to repeated normalization, activation, and residual-merge paths rather than to the main attention kernels.

I.1 Closed-form Backward of MV-Split+RMSNorm

For token i in segment $s(i)$, the pre-normalized merge can be written as

$$Z_{l,i} = X_{l,i} + \beta \odot (F_{l,i} - \bar{F}_l^{(s)}) + \alpha \odot (\bar{F}_l^{(s)} - \bar{X}_l^{(s)}). \tag{49}$$

Let $G_i = \partial \mathcal{L} / \partial X_{l+1,i}$ be the incoming gradient after RMSNorm, and let

$$r_i = \left(\tfrac{1}{D}\|Z_{l,i}\|_2^2 + \epsilon\right)^{-1/2}$$

be the inverse-RMS factor. The pre-normalization adjoint $\Delta_i = \partial \mathcal{L} / \partial Z_{l,i}$ is

$$\Delta_i = r_i G_i - Z_{l,i} \left(\frac{r_i^3}{D} \langle G_i, Z_{l,i} \rangle \right). \tag{50}$$

Define the segment-wise mean adjoint

$$\bar{\Delta}^{(s)} = \frac{1}{|s|} \sum_{i \in s} \Delta_i.$$

Then the merge gradients are

$$\frac{\partial \mathcal{L}}{\partial X_{l,i}} = \Delta_i - \alpha \odot \bar{\Delta}^{(s(i))}, \tag{51}$$

$$\frac{\partial \mathcal{L}}{\partial F_{l,i}} = \beta \odot \Delta_i + (\alpha - \beta) \odot \bar{\Delta}^{(s(i))}. \tag{52}$$

The gate gradients are

$$\frac{\partial \mathcal{L}}{\partial \alpha} = \sum_s \sum_{i \in s} \Delta_i \odot (\bar{F}_l^{(s)} - \bar{X}_l^{(s)}), \tag{53}$$

$$\frac{\partial \mathcal{L}}{\partial \beta} = \sum_s \sum_{i \in s} \Delta_i \odot (F_{l,i} - \bar{F}_l^{(s)}). \tag{54}$$

These expressions require only pointwise recomputation and segment-wise sums, which is what the two-pass kernel above evaluates.

J METHODS WE TRY BUT FAILED TO PREVENT MMS

We tested several interventions that target related objects: token means, attention mixing, attention-output gating, scalar gradient-norm control, and optimizer-side update geometry. None of these controls removed the mean-dominated failure in this backbone; some additionally degraded optimization. Their common limitation is that they do not combine local mean/centered branch-gradient control with the forward leaky mean replacement used by MV-Split.

Hard centering and attention reparameterizations. Explicit centering, $X \leftarrow PX$, removes the token mean rather than gain-limiting new mean writes. This degraded optimization in our runs and also removes useful global information, including image-level context and the implicit timestep signal discussed in Appendix H. Attention-matrix modifications such as $A - I$, $I - A$, or $(1 - \lambda)I + \lambda A$ change the attention branch but do not protect the FFN branch or the residual merge. Moreover, row-stochastic interpolations still preserve pure-mean states, and in multimodal sequences global centering does not remove segment-wise mean modes (image and text groups may each become internally homogeneous while their global average remains zero). None of these implement the local branch-gradient split

$$G \mapsto \alpha JG + \beta PG, \qquad \alpha \ll \beta,$$

nor the forward leaky mean replacement $JX_l \mapsto (1 - \alpha)JX_l + \alpha JF_l$ that defines MV-Split.

Gated attention. We also tested attention-output gates of the form

$$Y_i = g_i(X_i) \odot \mathrm{SDPA}(Q, K, V)_i,$$

where g_i is computed token-locally and acts along the **head** or **feature** dimension. It is not a sequence-level **token-space** projector and does not form JY or PY. Such gates (Qiu et al., 2025) can reduce attention-output magnitude and have been reported to mitigate attention-sink behavior (Xiao et al., 2023), but they do not compute JY and PY or apply different gains to them. In the mean-dominated regime, tokens are already aligned, so $g_i \approx \bar{g}$ tends to be similar across tokens; the gate then scales the mean and centered components together rather than separating them. More generally, attention-only controls leave the FFN W_2 uncontrolled. Our attention-only MV-Split trace shows that, once the attention branch is protected, the spike can relocate to the remaining ungated FFN branch (Appendix F).

The gradient-clipping paradox. All runs in the main comparison use global gradient clipping with threshold 1.0 (Appendix G), yet this does not remove MMS. The reason is that MMS is not only a large-norm event; it is a directional collapse of the writer update into the token-mean subspace. For a writer gradient decomposed as $G = G_\mu + G_c$, global norm clipping applies a scalar multiplier $\mathrm{clip}_\tau(G) = sG$ with $s = \min(1, \tau/\|G\|)$, so

$$\mathrm{clip}_\tau(G) = sG_\mu + sG_c, \qquad \frac{\|sG_\mu\|}{\|sG_c\|} = \frac{\|G_\mu\|}{\|G_c\|}.$$

Clipping can reduce the step length, but it cannot rotate a mean-coherent writer update back into the centered subspace. When G_μ dominates, the same scalar shrinkage also suppresses the already-small centered update, leaving the feature-learning path starved. A global-norm safety check is also blind to subspace structure: it can remain quiet while the residual writer direction has already become structurally unsafe. MV-Split addresses this failure mode at the residual interface by applying different gains to JG and PG, rather than by thresholding the scalar norm of G.

Muon optimizer. Muon (Jordan et al., 2024; Liu et al., 2025) orthogonalizes the momentum/update matrix using Newton–Schulz iterations. This can remove singular-value scale from a matrix update, but it acts after token gradients have already been summed into $G = \sum_t \delta_t y_t^\top = G_\mu + G_c$. If the momentum is dominated by an isolated mean-coherent term $G_\mu = T\bar{\delta}\bar{y}^\top = \sigma uv^\top$, the orthogonalized update removes σ but keeps the direction $uv^\top$. For homogeneous token inputs, this direction still produces the same update for every token and therefore remains mean-coherent. More generally, Muon reshapes singular values in parameter space; it does not implement the token-space split $G \mapsto \alpha JG + \beta PG$ or the forward leaky mean replacement $JX_l \mapsto (1 - \alpha)JX_l + \alpha JF_l$ used by MV-Split. In our runs, Muon could reduce update magnitude but did not remove the mean-dominated trajectory, consistent with MMS being a residual-subspace failure rather than only an optimizer-magnitude artifact.

K Additional Results on the MV-Split Runs

K.1 Text-Conditioned Evaluation of the 1000-Layer Checkpoint

We report GenEval and DPG-Bench measurements for the post-trained 1000-layer MV-Split checkpoint as a calibration of text-conditioned generation ability. These numbers are not intended as a controlled comparison to large public text-to-image systems: our model is trained on substantially smaller and differently sourced data (ImageNet-2012 pretraining followed by SFT and DPO (Rafailov et al., 2023; Wallace et al., 2023) on ∼50k curated images), uses a shorter training schedule, and uses a simpler post-training pipeline. The purpose is only to confirm that the 1000-layer scale-validation run remains usable as a text-conditioned generator, not to claim state-of-the-art text-to-image performance.

Table 4: **Text-conditioned evaluation of the 1000-layer MV-Split checkpoint.** Reported for calibration only and not used for state-of-the-art comparison.

Metric	Score
GenEval overall (avg. over tasks)	0.534
GenEval correct images	52.44%
GenEval correct prompts	67.63%
DPG-Bench overall	74.91

Table 5: **GenEval task breakdown (1000-layer MV-Split checkpoint).**

Task	Accuracy
single_object	92.81%
two_object	63.64%
counting	33.75%
colors	72.61%
position	25.75%
color_attr	31.75%

K.2 Full-Horizon Training Loss Curve for the MV-Split Runs

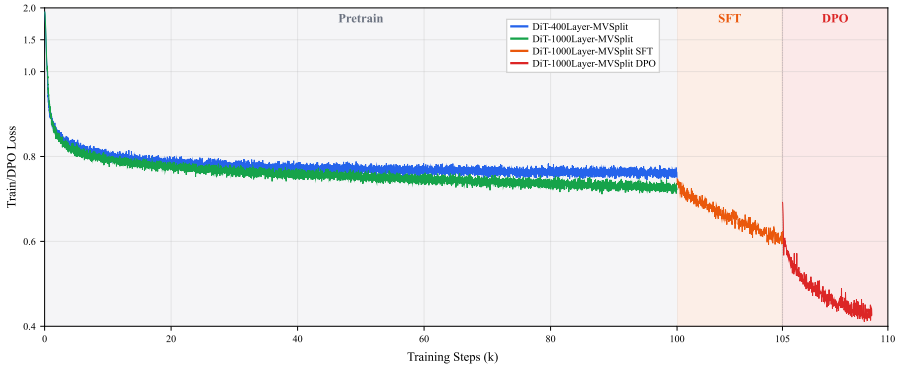

Figure 12: **Full-horizon training loss for the MV-Split 400-layer and 1000-layer runs.** Note that the SFT and DPO stages use a separately curated ∼50k image set rather than the ImageNet-2012 pre-training distribution; since loss values are data-dependent, the curves are shown for reference only.

L Limitations and Future Work

Our analysis identifies a residual-subspace failure pathway in ultra-deep Diffusion Transformers and shows that MV-Split stabilizes this pathway in the studied setting. The following boundary conditions define natural extensions rather than contradictions of the mechanism.

Predicting the exact onset time of MMS. The alignment-amplification law in Eq. 6 characterizes when token-wise writer gradients stop canceling and enter a coherent accumulation regime. This provides a mechanistic diagnostic for the MMS transition, but it does not by itself predict the exact training step $t^\star$ at which an un-stabilized run will cross the critical regime before the run is observed. The onset time depends on the coupled evolution of token representations, backward adjoints, optimizer momentum, data ordering, and mini-batch statistics. We therefore view exact onset prediction as a separate problem from architectural stabilization: MV-Split removes the unstable residual interface by controlling the mean and centered writer-gradient components directly, while deriving a closed-form scaling law for $t^\star$ remains an interesting direction for predictive theories of deep-network training dynamics.

Architectures beyond Softmax attention. Several parts of our analysis use Transformer-specific structure. In particular, row-stochastic attention preserves pure-mean token states (Proposition 1), and value homogenization suppresses Q/K logit gradients through the null space of the Softmax Jacobian (Lemma 1). These arguments do not directly transfer to attention-free sequence mixers such as convolutional diffusers or state-space models such as Mamba (Gu & Dao, 2023). At the same time, the writer-gradient decomposition in Eq. 5 only assumes a token-wise residual writer and is not specific to Softmax attention. This suggests a broader question: which components of the mean-dominated collapse mechanism are consequences of attention, and which are more general consequences of ultra-deep residual streams with token-wise writers? Testing this distinction in convolutional, hybrid, and state-space diffusion backbones is a natural next step.

Extreme-context spatiotemporal generation. Our scale validation focuses on image and text-to-image diffusion. Video, 3D, and other spatiotemporal generators often operate with substantially longer token sequences and additional structure across time, views, or modalities. In the coherent-alignment regime, the mean-coherent writer component can scale with sequence length as in Eq. 5, so these settings may place even stronger pressure on the residual interface. MV-Split is designed to decouple this mean-mode accumulation from the centered feature-learning path, but validating and possibly adapting the mechanism for ultra-long-context spatiotemporal DiTs remains an important direction for future large-scale generative modeling.

M MORE VISUAL RESULTS

We present additional uncurated samples from our 1000-layer MV-Split DiT to demonstrate the breadth and fidelity of the model across diverse semantic categories. All images are generated at 256×256 resolution using a Euler sampler (Karras et al., 2022) with 35 NFE steps and classifier-free guidance (Ho & Salimans, 2022) scale $w = 2.0$.

Text-Conditioned Generation. Unlike class-conditional DiTs that condition on a one-hot class label, our model is a **text-to-image** generator. Each sample is conditioned on a natural-language caption drawn from the ImageNet-2012 validation set, where captions were generated by a modern large language model and describe the scene content in 10–25 words (e.g., *"A colorful green jacamar on a branch with an insect in its beak, set against a blurred natural background"* or *"Snow-covered mountains under a dramatic cloudy sky, with sunlit huts and long shadows across the landscape"*). The captions vary in viewpoint, lighting, composition, and context, providing a diverse conditioning signal that goes beyond categorical labels. Within each grid below, the 12 images correspond to 12 *distinct* captions from the same ImageNet class, showcasing the model's ability to faithfully render varied scene descriptions. In each grid, the top row displays 4 images at $2\times$ magnification for detail inspection, and the bottom row shows 8 additional samples at $1\times$ scale.



Figure 13: Class "Alligator lizard" (044). Euler sampler, 35 NFE, CFG $w = 2.0$.



Figure 14: Class "Scorpion" (071). Euler sampler, 35 NFE, CFG $w = 2.0$.



Figure 15: Class “Jacamar” (095). Euler sampler, 35 NFE, CFG $w = 2.0$.



Figure 16: Class “Rhodesian ridgeback” (159). Euler sampler, 35 NFE, CFG $w = 2.0$.



Figure 17: Class “Bloodhound” (163). Euler sampler, 35 NFE, CFG $w = 2.0$.



Figure 18: Class “Bouvier des Flandres” (233). Euler sampler, 35 NFE, CFG $w = 2.0$.



Figure 19: Class “White wolf” (270). Euler sampler, 35 NFE, CFG $w = 2.0$.



Figure 20: Class “Chimpanzee” (367). Euler sampler, 35 NFE, CFG $w = 2.0$.



Figure 21: Class “Giant panda” (388). Euler sampler, 35 NFE, CFG $w = 2.0$.



Figure 22: Class “Beaker” (438). Euler sampler, 35 NFE, CFG $w = 2.0$.



Figure 23: Class “Caldron” (469). Euler sampler, 35 NFE, CFG $w = 2.0$.



Figure 24: Class “Candle” (470). Euler sampler, 35 NFE, CFG $w = 2.0$.



Figure 25: Class “Car wheel” (479). Euler sampler, 35 NFE, CFG $w = 2.0$.



Figure 26: Class “Coffeepot” (505). Euler sampler, 35 NFE, CFG $w = 2.0$.



Figure 27: Class “Convertible” (511). Euler sampler, 35 NFE, CFG $w = 2.0$.



Figure 28: Class “Crock Pot” (521). Euler sampler, 35 NFE, CFG $w = 2.0$.



Figure 29: Class “Drum” (541). Euler sampler, 35 NFE, CFG $w = 2.0$.



Figure 30: Class “Envelope” (549). Euler sampler, 35 NFE, CFG $w = 2.0$.



Figure 31: Class “Flute” (558). Euler sampler, 35 NFE, CFG $w = 2.0$.



Figure 32: Class “Freight car” (565). Euler sampler, 35 NFE, CFG $w = 2.0$.



Figure 33: Class “French horn” (566). Euler sampler, 35 NFE, CFG $w = 2.0$.



Figure 34: Class “Greenhouse” (580). Euler sampler, 35 NFE, CFG $w = 2.0$.



Figure 35: Class “Horse cart” (603). Euler sampler, 35 NFE, CFG $w = 2.0$.



Figure 36: Class “Knot” (616). Euler sampler, 35 NFE, CFG $w = 2.0$.



Figure 37: Class “Loupe” (633). Euler sampler, 35 NFE, CFG $w = 2.0$.



Figure 38: Class “Mask” (643). Euler sampler, 35 NFE, CFG $w = 2.0$.



Figure 39: Class “Minivan” (656). Euler sampler, 35 NFE, CFG $w = 2.0$.



Figure 40: Class “Mitten” (658). Euler sampler, 35 NFE, CFG $w = 2.0$.



Figure 41: Class “Monastery” (663). Euler sampler, 35 NFE, CFG $w = 2.0$.



Figure 42: Class “Mountain bike” (671). Euler sampler, 35 NFE, CFG $w = 2.0$.



Figure 43: Class “Pool table” (736). Euler sampler, 35 NFE, CFG $w = 2.0$.



Figure 44: Class “Pot” (738). Euler sampler, 35 NFE, CFG $w = 2.0$.



Figure 45: Class “Rugby ball” (768). Euler sampler, 35 NFE, CFG $w = 2.0$.



Figure 46: Class “Scoreboard” (781). Euler sampler, 35 NFE, CFG $w = 2.0$.



Figure 47: Class “Sweatshirt” (841). Euler sampler, 35 NFE, CFG $w = 2.0$.



Figure 48: Class “Teapot” (849). Euler sampler, 35 NFE, CFG $w = 2.0$.



Figure 49: Class “Trombone” (875). Euler sampler, 35 NFE, CFG $w = 2.0$.



Figure 50: Class “Windsor tie” (906). Euler sampler, 35 NFE, CFG $w = 2.0$.



Figure 51: Class “Alp” (970). Euler sampler, 35 NFE, CFG $w = 2.0$.



Figure 52: Class “Groom” (982). Euler sampler, 35 NFE, CFG $w = 2.0$.