EXPERT-DATA ALIGNMENT GOVERNS GENERATION QUALITY IN DECENTRALIZED DIFFUSION MODELS

Marcos Villagra, Bidhan Roy, Raihan Seraj & Zhiying Jiang
Bagel Labs (www.bagel.com), USA
{marcos,bidhan,raihan,gin}@bagel.com

ABSTRACT

Decentralized Diffusion Models (DDMs) route denoising through experts trained independently on disjoint data clusters, which can strongly disagree in their predictions. What governs the quality of generations in such systems? We present the first ever systematic investigation of this question. A priori, the expectation is that minimizing denoising trajectory sensitivity—minimizing how perturbations amplify during sampling—should govern generation quality. We demonstrate this hypothesis is incorrect: a stability–quality dissociation. Full ensemble routing, which combines all expert predictions at each step, achieves the most stable sampling dynamics and best numerical convergence while producing the worst generation quality (FID 47.9 vs. 22.6 for sparse Top-2 routing). Instead, we identify expert-data alignment as the governing principle: generation quality depends on routing inputs to experts whose training distribution covers the current denoising state. Across two distinct DDM systems, we validate expert-data alignment using (i) data-cluster distance analysis, confirming sparse routing selects experts with data clusters closest to the current denoising state, and (ii) per-expert analysis, showing selected experts produce more accurate predictions than non-selected ones, and (iii) expert disagreement analysis, showing quality degrades when experts disagree. For DDM deployment, our findings establish that routing should prioritize expert-data alignment over numerical stability metrics.

1 INTRODUCTION

Decentralized Diffusion Models (DDMs) McAllister et al. (2025) combine independently trained diffusion experts Ho et al. (2020) via an inference-time router. Because experts are trained on disjoint data clusters and can strongly disagree in their predictions, understanding what governs generation quality becomes crucial—yet this question has not been systematically studied.

A natural hypothesis is that numerical stability determines quality Yang et al. (2023), that is, routing strategies that minimize trajectory sensitivity should produce superior samples. We demonstrate that this hypothesis is incorrect. In this work, we observe that full ensemble routing, in which all expert predictions are combined, achieves the lowest trajectory sensitivity and best numerical convergence, yet produces the worst generation quality. These results rule out trajectory sensitivity as the primary determinant of generation quality.

We identify *expert-data alignment* (routing inputs to experts trained on similar data) as the governing principle. When sparse routing (e.g., Top-2) selects experts whose training distribution covers the current denoising state, each expert produces coherent velocity predictions that combine meaningfully. Full ensemble routing, by contrast, forces all experts to process every input; since each expert is trained on only a subset of the data, most of them are processing out-of-distribution data at any given time. The averaged velocity field may be smooth, but it points toward an incoherent compromise rather than the data manifold.

We provide direct experimental validation of this principle across two distinct DDM systems. Data-cluster distance analysis confirms that sparse routing selects experts with data clusters closest to the input embedding. Per-expert prediction quality analysis shows that selected experts produce velocity

predictions with higher alignment to the blended output. Expert disagreement analysis demonstrates that disagreement under full ensemble correlates with quality degradation.

Although numerical stability does not govern quality, understanding when DDM sampling converges remains valuable. Classical stability analysis suggests DDMs should fail: Lipschitz constants of deep networks grow exponentially with depth Fazlyab et al. (2019); Virmaux & Scaman (2018); Yang et al. (2023), and Grönwall's inequality implies small perturbations amplify over integration Hairer et al. (1993). We also examine *trajectory-local sensitivity*, denoted $\widehat{L}_{\text{eff}}^{(h)}$, which formalizes the idea that Jacobian spectral norms remain bounded along realized sampling paths. Empirically, we observe that trajectories exhibit moderate sensitivity compared to worst-case global bounds. While $\widehat{L}_{\text{eff}}^{(h)}$ does not predict quality across routing strategies, it may serve as a within-strategy diagnostic for identifying numerically sensitive samples.

The contributions of this work are as follows.

- **Expert-data alignment principle.** We identify expert-data alignment as the primary determinant of generation quality in DDMs. We provide direct experimental validation via (i) data-cluster distance analysis showing sparse routing selects experts with data clusters closest to the input, (ii) per-expert analysis demonstrating that selected experts produce superior velocity predictions, and (iii) expert disagreement analysis showing quality degrades when experts disagree.
- **Stability–quality dissociation.** We demonstrate that trajectory sensitivity does not govern generation quality: full ensemble routing achieves the lowest $\widehat{L}_{\text{eff}}^{(h)}$ and step-refinement disagreement, yet produces the worst Fréchet Inception Distance (FID) Heusel et al. (2017). This rules out numerical stability as the primary quality determinant. We additionally explore trajectory-local sensitivity as a within-strategy diagnostic, finding weak predictive power.

This paper is organized as follows. Section 2 introduces background on DDMs. Section 3 establishes the stability–quality dissociation. Section 4 identifies expert-data alignment as the governing principle. Section 5 presents trajectory sensitivity analysis. Sections 6 and 7 discuss implications and conclude. Related work is in Appendix A.

2 Background

2.1 Notation

For a differentiable scalar function $f : \mathbb{R}^d \times [0, 1] \to \mathbb{R}$, we let $\nabla_x f(x, t) \in \mathbb{R}^d$ denote the gradient with respect to x. We use $J_x f(x, t) \in \mathbb{R}^{d\times d}$ to denote the Jacobian matrix. We use $\|\cdot\|$ for the matrix spectral norm and to denote the ℓ_2-norm over $\mathbb{R}^d$.

2.2 Diffusion sampling

Diffusion sampling generates data by integrating an ordinary differential equation (ODE) $\frac{dx_t}{dt} = v_t(x_t)$ from noise to data Lipman et al. (2022). We call $v = \{v_t\}_{t=0}^1$ a *flow* and the solution $\{x_t\}$ a *trajectory*. The ODE has a unique solution when v is Lipschitz continuous Coddington et al. (1956).

2.3 Decentralized Diffusion

Decentralized Diffusion Models (DDMs) McAllister et al. (2025) train K expert diffusion models in complete isolation on disjoint data partitions. Unlike traditional MoE (which routes tokens to different FFN layers within a shared backbone), DDM routes entire inputs to separate full models at each denoising step.

At inference, a lightweight router predicts weights $w_t^{(k)}(x_t) \geq 0$ with $\sum_{k=1}^K w_t^{(k)}(x_t) = 1$ at each denoising step. The routed velocity field is given by

$$v_t(x_t) = \sum_{k=1}^{K} w_t^{(k)}(x_t)\, v_t^{(k)}(x_t). \tag{1}$$

Sampling integrates $\frac{dx_t}{dt} = v_t(x_t)$ from noise $x_0 \sim \mathcal{N}(0, I)$.

The router outputs probabilities $p_t(k|x_t)$ for each expert k. Full ensemble mode uses $w_t^{(k)}(x_t) = p_t(k|x_t)$ for all k. Top-k routing selects the k experts with highest probability and renormalizes their weights Shazeer et al. (2017). Top-1 selects a single expert: $v_{\text{Top-1}}(x_t) = v_{k^*}(x_t)$ with $k^* = \arg\max_k p_t(k|x_t)$. This yields a piecewise-smooth vector field with non-differentiability on measure-zero switching surfaces.

In this work, when we say that the DDM router converges we mean numerical convergence of the sampler, formalized below.

Definition 2.1 (DDM sampling convergence)**.** Fix trained experts and a router with a routed flow $v = \{v_t\}_{t=0}^1$. Let $x_1 \sim \mathcal{N}(0, I)$ be an initial condition at time $t = 1$ and let x_t denote the ODE solution at time $t < 1$ (note that t is decreasing). Let $\tilde{x}_t^{(h)}$ denote the output of a numerical ODE solver with step size h, coupled to the same initial noise x_1. We say the DDM sampler *converges in probability* if for every $\varepsilon > 0$, we have $P\big(\|\tilde{x}_0^{(h)} - x_0\| > \varepsilon\big) \to 0 \quad \text{as } h \to 0$, where the probability is over the initial noise x_1. This notion implies convergence in distribution of $\tilde{x}_0^{(h)}$ to x_0.

3 The Stability–Quality Dissociation

A natural hypothesis is that numerical stability governs generation quality in DDMs, that is, routing strategies that minimize trajectory sensitivity should produce superior samples. We establish that this hypothesis is incorrect.

Sections 4 and 5 will demonstrate the dissociation: full ensemble routing achieves the lowest trajectory sensitivity and lowest step-refinement disagreement, yet produces the worst generation quality—even worse generation quality is observed in Jiang et al. (2025). Top-2 achieves the best generation quality despite higher trajectory sensitivity. This rules out numerical stability as the quality determinant.

4 Expert-Data Alignment

The preceding section ruled out numerical stability as the quality determinant. We now establish *expert-data alignment*—routing inputs to experts trained on similar data—as the governing principle and provide direct experimental validation.

4.1 The Expert-Data Alignment Hypothesis

Full ensemble averaging produces a smoother velocity field because: (1) averaging over all experts reduces variance in velocity predictions, (2) this variance reduction directly lowers $\|J_x v(x_t, t)\|$, and (3) smoother trajectories also exhibit lower step-refinement disagreement. However, this smoothing forces experts to process out-of-distribution data. Each expert is trained on only small part of the data (one cluster); when all experts contribute to every input, most of them process data outside their training distribution. Therefore, the averaged velocity field may be smooth but it points toward an incoherent compromise rather than the data manifold.

Top-2 routing selects the two experts whose training data most closely matches the current input, keeping each expert producing coherent velocity predictions that combine meaningfully. The dominant factor for sample quality is whether experts process data similar to their training distribution (what we call *expert-data alignment*) rather than minimizing trajectory sensitivity.

If expert-data alignment governs quality, we expect that

1. sparse routing should achieve higher alignment (selected experts have lower cluster distance) than full ensemble;
2. selected experts should produce superior velocity predictions compared to non-selected experts; and,
3. expert disagreement should correlate with quality degradation under full ensemble.

Table 1: **Cluster distance analysis validates expert-data alignment.** Sparse routing selects experts whose training clusters match the input. Lower mean rank indicates better alignment (rank 1 = closest). Results averaged over $n = 500$ samples at $t \in \{0.3, 0.5, 0.7\}$. We also include the results for full ensemble as a baseline reference.

Routing	Mean Cluster Rank ↓	Top-2 Match Rate ↑
Top-1	1.54 ± 0.28	90.2%
Top-2	1.96 ± 0.26	83.9%
Full (8)	4.50 ± 0.00	25.0%

4.2 Cluster Distance Analysis

We use the pretrained DDM Paris model Jiang et al. (2025) via released pretrained checkpoints. The model consists of $K = 8$ experts trained on a subset of LAION-Aesthetics Schuhmann et al. (2022). The dataset was partitioned into 8 semantic clusters via two-stage hierarchical k-means on DINOv2-ViT-L/14 embeddings. The model uses a DiT-B/2 router (∼129M parameters) and 8 DiT-XL/2 experts (a modified version of the DiT-XL/2 experts with ∼606M parameters each, ∼5B total). The router was trained *post-hoc* on the full dataset, effectively learning to route inputs to the expert trained on the most similar data. We will use Paris DDM to test the Expert-Data Alignment Hypothesis.

Let $\mathcal{C}_k$ denote the training data cluster for expert k, and let $d(x, \mathcal{C}_k)$ denote the Euclidean distance from the DINOv2 embedding of input x to the centroid of cluster $\mathcal{C}_k$. We define *high expert-data alignment* as the condition where the selected experts have low $d(x, \mathcal{C}_k)$ relative to non-selected experts.

We test whether sparse routing selects experts whose training clusters match the input distribution. For $n = 500$ samples, we extract DINOv2-ViT-L/14 embeddings at timesteps $t \in \{0.3, 0.5, 0.7\}$ during sampling. For each state (x_t, t), we compute: (1) the Euclidean distance from the embedding to each of the 8 cluster centroids used during expert training; (2) the experts selected by each routing strategy at that timestep; (3) the rank of the selected expert(s)' cluster(s) among all 8 clusters, ordered by distance.

We emphasize that this embedding distance is used only for relative expert ranking; all conclusions depend solely on rank comparisons, not metric fidelity. Moreover, the training clusters themselves were defined via k-means in DINOv2 space, making this the canonical embedding for cluster proximity.

We report two metrics. (i) the average rank (1 = closest, 8 = farthest) of the selected expert's training cluster refered to as *Mean cluster rank*, and (ii) the percentage of timesteps where at least one selected expert's cluster is among the two closest to the current input refered to as *Top-2 Match Rate*. Table 1 shows the results.

Top-1 and Top-2 achieve mean cluster ranks of 1.54 and 1.96, far below the 4.5 random baseline, with Top-2 match rates exceeding 83%.

4.3 Per-Expert Prediction Quality

Now we want to test whether selected experts produce superior velocity predictions compared to non-selected experts.

For $n = 200$ samples generated with Top-2 routing, we record at each timestep (1) the blended velocity $v_t(x_t) = \sum_{k \in \mathcal{S}} w_t^{(k)} v_t^{(k)}(x_t)$, (2) the individual velocity predictions $v_t^{(k)}(x_t)$ for all 8 experts, and (3) the routing weights and selected set $\mathcal{S}$.

For each expert k at each timestep, we compute the velocity alignment score define as

$$a^{(k)}(x_t) = \frac{v_t^{(k)}(x_t)^\top \cdot v_t(x_t)}{\|v_t^{(k)}(x_t)\| \cdot \|v_t(x_t)\|},$$

Table 2: **Selected experts produce better-aligned predictions.** Angular deviation (degrees) from blended velocity for selected versus non-selected experts under Top-2 routing. Smaller is better. Results averaged over $n = 200$ samples. Statistical significance via independent samples t-test.

Expert Status	Angular Dev. ↓	Std Dev
Selected (Top-2)	3.6°	±1°
Non-selected	5.1°	±1°
Reduction	1.5° (29%)	$p < 0.001$

where $v_t^{(k)}(x_t)^\top$ denotes the transpose of $v_t^{(k)}(x_t)$. This cosine similarity measures how well expert k's prediction aligns with the routed velocity used for successful generation. We report angular deviation $\theta_k = \arccos(a_k)$ in degrees for interpretability. See Table 2 for the results.

Selected experts achieve smaller angular deviation from the blended velocity (3.6° vs. 5.1°; independent samples t-test, $p < 0.001$), a 29% reduction confirming systematic identification of coherent experts. The statistical significance demonstrates that routing systematically identifies the most coherent experts rather than selecting arbitrarily.

4.4 Expert Disagreement and Sample Quality

In this section, we want to test whether expert disagreement explains the poor quality of full ensemble routing.

For $n = 500$ samples generated with full ensemble, we measure (1) mean pairwise expert disagreement $D(x_t) = \frac{1}{\binom{K}{2}} \sum_{i<j} \|v_i(x_t) - v_j(x_t)\|_2$, (2) trajectory-integrated disagreement: $D_{\text{int}} = \int_0^1 D(x_t)\, dt$, and (3) perceptual quality as LPIPS distance Zhang et al. (2018) to the corresponding Top-2 output (matched initial noise). See Figure 1 for the results.

The monotonic increase in LPIPS across disagreement quartiles confirms that expert disagreement drives quality degradation in full ensemble routing.

We also validate the expert-data alignment hypothesis on a separate MNIST-based DDM with 10 UNet experts, where alignment effects are even more pronounced due to stronger expert specialization (see Appendix B for details).

4.5 Summary of Results

The experiments provide evidence for expert-data alignment as the governing principle of sample quality in DDMs. Cluster distance analysis confirms that sparse routing selects experts whose training data matches the input (Table 1). Per-expert prediction quality shows that selected experts produce superior velocity predictions (Table 2). Expert disagreement analysis explains why full ensemble fails, that is, averaging across disagreeing experts produces incoherent velocity fields that point off-manifold (Figure 1). Critically, this correlation provides causal evidence: high disagreement represents states where alignment naturally breaks down (multiple experts are forced to process inputs outside their training distribution), and we observe that this misalignment degrades actual sample quality (LPIPS), not merely step-refinement error.

We also provide independent validation of these findings on a MNIST-based DDM with 10 specialized expert, where the alignment effects are even more pronounced.

5 Trajectory Sensitivity Analysis

Having established that expert-data alignment governs sample quality, we now analyze trajectory sensitivity to understand when numerical convergence holds and to develop within-strategy diagnostics. We present a conditional convergence argument linking trajectory-local sensitivity to DDM convergence, followed by empirical validation.

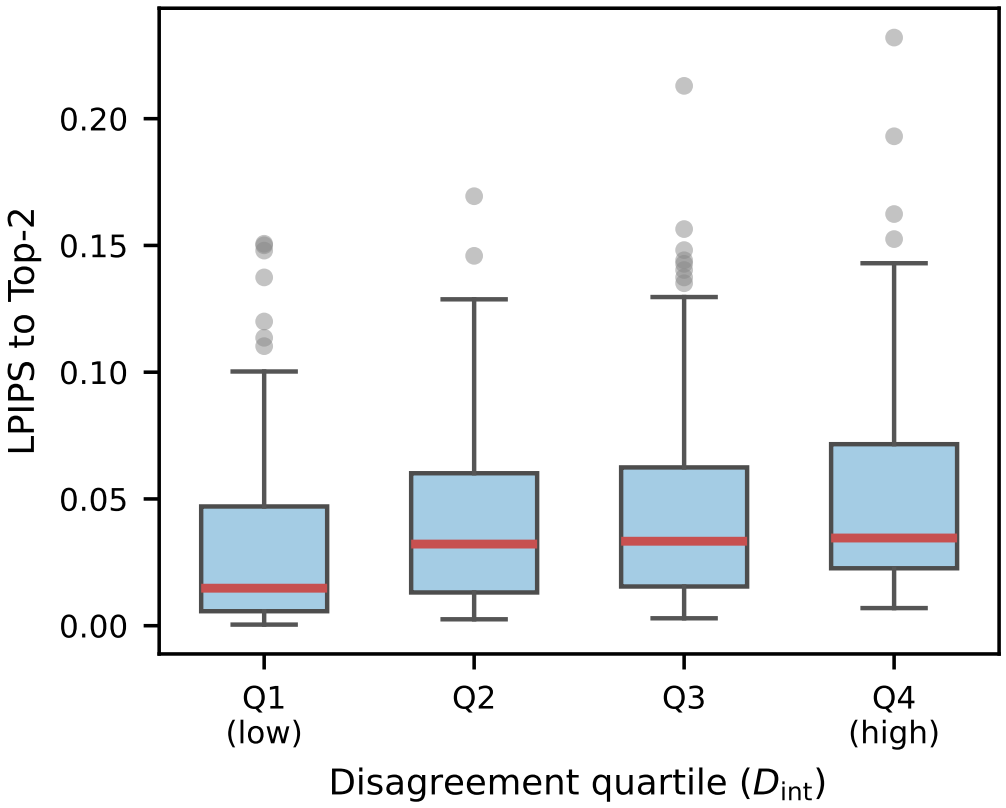

Figure 1: **Higher expert disagreement degrades perceptual quality in full ensemble routing.** Samples binned by trajectory-integrated disagreement (Q1=lowest, Q4=highest). Higher disagreement quartiles show greater perceptual distance (LPIPS) from the Top-2 reference, explaining why full ensemble underperforms sparse routing.

5.1 Trajectory-Local Sensitivity

We define the *effective Lipschitz constant* $L_{\text{eff}}(x_1) = \sup_{t\in[0,1]} \|J_x v_t(x_t)\|$ as the maximum Jacobian spectral norm along the exact ODE trajectory from initial condition x_1. Since exact trajectories are inaccessible, we use the *empirical effective Lipschitz constant* $\widehat{L}^{(h)}_{\text{eff}}(x_1) := \max_n \|J_x v(\tilde{x}^{(h)}_{t_n}, t_n)\|$ computed along numerical trajectories as a proxy (see Appendix C for formal definitions).

A standard Grönwall-based argument shows that when $L_{\text{eff}}(x_1) \leq L$, the global error satisfies $e_N \leq Ch^p e^L$, yielding convergence in probability as $h \to 0$ (see Appendix E for the formal statement and proof). We use $\widehat{L}^{(h)}_{\text{eff}}$ as a retrospective diagnostic for understanding sensitivity differences across routing strategies.

5.2 Empirical Validation

We design three experiments targeting each step of the convergence argument: (1) bounded local error, (2) bounded trajectory sensitivity, and (3) step-size refinement convergence. We use the pretrained DDM Paris model Jiang et al. (2025) with 8 experts on LAION-Aesthetics. Appendices F and G provide full details.

Local error (Exp. 1). Local truncation error is essentially identical across routing strategies, confirming routing does not affect single-step accuracy (Appendix G.2).

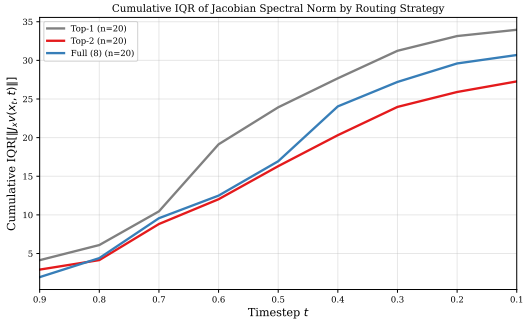

Figure 2: Cumulative IQR of the Jacobian spectral norm $\|\nabla_x v(x_t, t)\|$ as a measure of variability across sampling trajectories. The gap between Top-2 and other strategies widens as denoising progresses. Mid-trajectory timesteps ($t \in [0.1, 0.9]$); n=20 samples per strategy.

Table 3: **Routing strategy comparison.** FID from Jiang et al. (2025); $\widehat{L}_{\text{eff}}^{(h)}$ and Δ_{refine} measured on n=1000 samples. Full ensemble achieves the lowest trajectory sensitivity and step-refinement disagreement yet produces the worst FID.

Strategy	**FID** $\downarrow$	$\widehat{\boldsymbol{L}}_{\text{eff}}^{(\boldsymbol{h})}$	$\boldsymbol{\Delta}_{\text{refine}}$ $\downarrow$
Monolithic	*29.64*	–	–
Top-1	30.60	18.81 $\pm$7.15	0.075 $\pm$0.107
Top-2	22.60	17.48 $\pm$6.07	0.051 $\pm$0.070
Full (8)	47.89	17.07 $\pm$6.33	0.020 $\pm$0.020

Trajectory sensitivity (Exp. 2). We compute $\widehat{L}_{\text{eff}}^{(h)}$ along sampling trajectories using power iteration for Jacobian spectral norms. Figure 2 shows the temporal profile.

Step-size refinement (Exp. 3). We measure step-refinement disagreement

$$\Delta_{\text{refine}}(x_1) := \text{LPIPS}(D(\tilde{x}_0^{(h)}), D(\tilde{x}_0^{(h/2)})), \tag{2}$$

comparing samples at N and $2N$ steps from the same initial noise Zhang et al. (2018). This provides a Jacobian-independent error measure. Results are in Figure 3 and Appendix G.4.

5.3 Summary of Results

Table 3 shows a summary of the observed results. Across all routing strategies, correlation between $\widehat{L}_{\text{eff}}^{(h)}$ and step-refinement disagreement Δ_{refine} is low ($\rho < 0.08$; Figure 3)[1]. This further supports our finding that numerical stability metrics do not govern generation quality, aligning with the observed dissociation between sensitivity and FID. Factors beyond worst-case Jacobian norms—such as directional alignment of perturbations with the flow or cancellation effects across timesteps—likely govern actual error accumulation.

6 Discussion

The experimental results reveal a fundamental tradeoff: strategies that maximize numerical stability (full ensemble) necessarily sacrifice expert-data alignment, and vice versa. Cluster distance analysis quantifies this directly—sparse routing achieves mean cluster ranks of 1.54–1.96 versus 4.50 for full ensemble (Table 1)—while per-expert analysis shows selected experts produce 29% lower angular

[1]Preliminary analysis of the router Jacobian term $|\sum_k v_k \nabla_x w_k|$ also shows similarly low correlation with Δ_{refine} (Spearman $\rho = -0.07$, $p = 0.62$ for Top-2 routing, n=50), reinforcing that Jacobian-based sensitivity metrics are not aligned with discretization error.

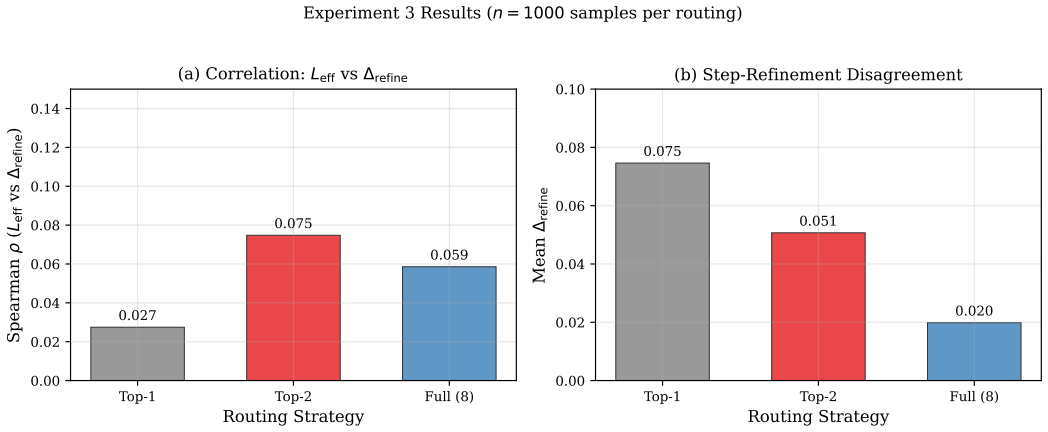

Figure 3: Correlation between trajectory sensitivity and step-refinement disagreement from Experiment 3 (n=1000 samples per routing strategy). **(a)** Spearman correlation $\rho(\widehat{L}_{\text{eff}}^{(h)}, \Delta_{\text{refine}})$ is weak across all strategies ($\rho < 0.08$), indicating that L_{eff} is not a tight predictor of discretization error. **(b)** Mean step-refinement disagreement Δ_{refine} shows clear ordering: Full ensemble achieves the lowest discretization error (0.020), followed by Top-2 (0.051) and Top-1 (0.075).

deviation from the blended velocity (Table 2). The disagreement-quality correlation provides shows that when alignment breaks down, then sample quality degrades proportionally.

Furthermore, we saw that L_{eff} is not a cross-strategy quality predictor, but remains useful for within-strategy diagnostics to identify numerically sensitive trajectories. The weak correlation $\rho(L_{\text{eff}}, \Delta_{\text{refine}}) \approx 0.03$–$0.08$ suggests factors beyond worst-case Jacobian norms affect actual error accumulation. The step-refinement ordering confirms that smoother velocity fields yield better numerical convergence, yet this convergence advantage does not translate to better generation quality.

7 Conclusions

We investigated what governs generation quality in Decentralized Diffusion Models (DDMs), where independently trained experts are combined via inference-time routing.

Our central finding is that **expert-data alignment governs generation quality**: routing inputs to experts trained on similar data is the primary determinant of quality, not numerical stability. We provide direct experimental validation through cluster distance analysis (showing sparse routing selects in-distribution experts), per-expert prediction quality (showing selected experts produce superior velocities), and disagreement analysis (explaining why full ensemble fails).

For practitioners, our findings demonstrate that when deploying DDMs with independently trained experts, routing that maintains expert-data alignment is more important than optimizing for numerical stability metrics. Future work should explore training objectives that improve expert robustness to out-of-distribution inputs. In Appendix H we discuss limitations and future directions.

References

Kun Cheng, Xiao He, Lei Yu, Zhijun Tu, Mingrui Zhu, Nannan Wang, Xinbo Gao, and Jie Hu. Diff-MoE: Diffusion transformer with time-aware and space-adaptive experts. In *Forty-second International Conference on Machine Learning*, 2025.

Earl A Coddington, Norman Levinson, and T Teichmann. Theory of ordinary differential equations, 1956.

Damai Dai, Li Dong, Shuming Ma, Bo Zheng, Zhifang Sui, Baobao Chang, and Furu Wei. StableMoE: Stable routing strategy for mixture of experts. *arXiv preprint arXiv:2204.08396*, 2022.

Mahyar Fazlyab, Alexander Robey, Hamed Hassani, Manfred Morari, and George Pappas. Efficient and accurate estimation of lipschitz constants for deep neural networks. *Advances in neural information processing systems*, 32, 2019.

William Fedus, Barret Zoph, and Noam Shazeer. Switch transformers: Scaling to trillion parameter models with simple and efficient sparsity. *Journal of Machine Learning Research*, 23(120):1–39, 2022.

Zhengcong Fei, Mingyuan Fan, Changqian Yu, Debang Li, and Junshi Huang. Scaling diffusion transformers to 16 billion parameters. *arXiv preprint arXiv:2407.11633*, 2024.

Ernst Hairer, Gerhard Wanner, and Syvert P Nørsett. *Solving ordinary differential equations I: Nonstiff problems*. Springer, 1993.

Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30, 2017.

Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*, 2022.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020.

Zhiying Jiang, Raihan Seraj, Marcos Villagra, and Bidhan Roy. Paris: A decentralized trained open-weight diffusion model. *arXiv preprint arXiv:2510.03434*, 2025.

Matt Jordan and Alexandros G Dimakis. Exactly computing the local lipschitz constant of relu networks. *Advances in Neural Information Processing Systems*, 33:7344–7353, 2020.

Tero Karras, Miika Aittala, Timo Aila, and Samuli Laine. Elucidating the design space of diffusion-based generative models. *Advances in neural information processing systems*, 35:26565–26577, 2022.

Tian Li, Anit Kumar Sahu, Manzil Zaheer, Maziar Sanjabi, Ameet Talwalkar, and Virginia Smith. Federated optimization in heterogeneous networks. *Proceedings of Machine learning and systems*, 2:429–450, 2020.

Yaron Lipman, Ricky TQ Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow matching for generative modeling. *arXiv preprint arXiv:2210.02747*, 2022.

Yahui Liu, Yang Yue, Jingyuan Zhang, Chenxi Sun, Yang Zhou, Wencong Zeng, Ruiming Tang, and Guorui Zhou. Efficient training of diffusion mixture-of-experts models: A practical recipe. *arXiv preprint arXiv:2512.01252*, 2025.

David McAllister, Matthew Tancik, Jiaming Song, and Angjoo Kanazawa. Decentralized diffusion models. *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 23323–23333, 2025.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023.

William Peebles and Saining Xie. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 4195–4205, 2023.

Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, et al. Laion-5b: An open large-scale dataset for training next generation image-text models. *Advances in neural information processing systems*, 35:25278–25294, 2022.

Noam Shazeer, Azalia Mirhoseini, Krzysztof Maziarz, Andy Davis, Quoc Le, Geoffrey Hinton, and Jeff Dean. Outrageously large neural networks: The sparsely-gated mixture-of-experts layer. *arXiv preprint arXiv:1701.06538*, 2017.

Minglei Shi, Ziyang Yuan, Haotian Yang, Xintao Wang, Mingwu Zheng, Xin Tao, Wenliang Zhao, Wenzhao Zheng, Jie Zhou, Jiwen Lu, et al. DiffMoE: Dynamic token selection for scalable diffusion transformers. *arXiv preprint arXiv:2503.14487*, 2025.

Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. *arXiv preprint arXiv:2011.13456*, 2020.

Haotian Sun, Tao Lei, Bowen Zhang, Yanghao Li, Haoshuo Huang, Ruoming Pang, Bo Dai, and Nan Du. Ec-dit: Scaling diffusion transformers with adaptive expert-choice routing. *arXiv preprint arXiv:2410.02098*, 2024.

Zheng Tan, Weizhen Wang, Andrea L. Bertozzi, and Ernest K. Ryu. Stork: Faster diffusion and flow matching sampling by resolving both stiffness and structure-dependence. *arXiv:2505.24210*, 2025.

Aladin Virmaux and Kevin Scaman. Lipschitz regularity of deep neural networks: analysis and efficient estimation. *Advances in Neural Information Processing Systems*, 31, 2018.

Zhantao Yang, Ruili Feng, Han Zhang, Yujun Shen, Kai Zhu, Lianghua Huang, Yifei Zhang, Yu Liu, Deli Zhao, Jingren Zhou, et al. Lipschitz singularities in diffusion models. In *The Twelfth International Conference on Learning Representations*, 2023.

Yike Yuan, Ziyu Wang, Zihao Huang, Defa Zhu, Xun Zhou, Jingyi Yu, and Qiyang Min. Expert race: A flexible routing strategy for scaling diffusion transformer with mixture of experts. *arXiv preprint arXiv:2503.16057*, 2025.

Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 586–595, 2018.

Youwei Zheng, Yuxi Ren, Xin Xia, Xuefeng Xiao, and Xiaohua Xie. Dense2MoE: Restructuring diffusion transformer to moe for efficient text-to-image generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 18661–18670, 2025.

Yanqi Zhou, Tao Lei, Hanxiao Liu, Nan Du, Yanping Huang, Vincent Zhao, Andrew M Dai, Quoc V Le, James Laudon, et al. Mixture-of-experts with expert choice routing. *Advances in Neural Information Processing Systems*, 35:7103–7114, 2022.

APPENDIX

A RELATED WORK

Diffusion as ODE/SDE and numerical stability. Diffusion sampling involves Lipschitz-constrained ODEs where discretization error affects accuracy. Sampling can be expressed as a probability-flow Ordinary Differential Equation (ODE) Song et al. (2020), where Lipschitz constants and discretization error determine solver accuracy. Recent work by Tan et al. (2025) addresses temporal stiffness via a specialized solver called STORK based on stabilized Runge-Kutta methods for stiff diffusion ODEs. STORK's stabilized solvers could, in principle, be combined with our approach (using stable solvers for temporal stiffness while using sparse routing to control spatial sensitivity). We use the released Euler and Heun solvers to isolate the effect of routing strategies, but STORK-style solvers may provide additional gains in decentralized settings.

Local Lipschitz analysis. Local Lipschitz bounds have been extensively studied in the neural network literature. Jordan & Dimakis (2020) developed methods for exactly computing local Lipschitz constants, enabling input-specific sensitivity analysis rather than global worst-case bounds. Our work applies this trajectory-local perspective to diffusion sampling. The key novelty is not the local Lipschitz concept itself, but its application to understanding decentralized diffusion dynamics and the discovery that routing implicitly stabilizes the sampling dynamics.

Decentralized diffusion. DDMs show that decentralized experts, when routed, can match a monolithic diffusion objective McAllister et al. (2025). Our work investigates why such combination succeeds despite expert disagreement, providing a stability-based explanation complementary to the original capacity arguments.

DDM vs. traditional Mixture-of-Experts. DDMs are *ensembles of independently trained models*, not traditional Mixture-of-Experts (MoE). In standard MoE architectures experts are FFN layers within a shared backbone, trained jointly with load balancing losses, and routed at the token level Shazeer et al. (2017); Fedus et al. (2022). In DDM, each "expert" is a *complete diffusion model* trained in isolation on a disjoint data partition (no shared parameters, no gradient communication, no joint training). Routing occurs at the *input level* (entire noisy images) rather than token level, and experts are combined only at inference time. Concurrent work applies traditional MoE architectures *within* diffusion models Fei et al. (2024); Sun et al. (2024); Shi et al. (2025); Yuan et al. (2025); Cheng et al. (2025); Liu et al. (2025); Zheng et al. (2025); DDM instead combines complete, independently trained models. This distinction matters for stability analysis: DDM experts can produce arbitrarily different outputs for the same input (having never coordinated during training), whereas MoE experts share a representational backbone that constrains their disagreement. Our analysis specifically addresses the stability challenges arising from DDM's decentralized training.

Data-aware routing and expert specialization. The importance of matching inputs to appropriately trained experts is well-established in MoE systems. Sparsely-gated MoE architectures rely on learned routing to direct inputs to relevant experts Shazeer et al. (2017), with subsequent work analyzing expert utilization and load balancing Dai et al. (2022); Zhou et al. (2022). In federated learning, data heterogeneity across clients creates analogous challenges: models trained on non-IID partitions may produce poor predictions on out-of-distribution inputs Li et al. (2020). Our work provides direct experimental evidence that this principle governs sample quality in DDMs: sparse routing succeeds precisely because it maintains alignment between inputs and expert training distributions.

B MNIST Validation of Expert-Data Alignment

We validate the expert-data alignment hypothesis on a separate MNIST-based DDM with 10 UNet experts and a CNN router. This controlled setting provides independent confirmation of our main findings on a simpler domain with more specialized experts.

The MNIST DDM consists of 10 independently trained UNet experts (each $\sim$10M parameters) and a lightweight CNN router. Each expert was trained on a digit-specific subset of MNIST, creating strong expert specialization. The router was trained separately to predict which expert best matches each input.

We run two experiments mirroring the Paris DDM analysis: (1) per-expert prediction quality comparing selected vs. non-selected experts, and (2) expert disagreement correlation with quality degradation. All experiments use the Heun solver with 50 steps and $n = 500$ samples per configuration.

Table 4: **MNIST: Selected experts produce better-aligned predictions.** Angular deviation (degrees) from blended velocity for selected versus non-selected experts under Top-2 routing. Smaller is better. Results averaged over $n = 500$ samples.

Expert Status	**Angular Dev. ↓**	**Std Dev**
Selected (Top-2)	6.4°	±1°
Non-selected	11.3°	±1°
Reduction	4.9° (43%)	$p \approx 0$

Table 4 shows that selected experts produce smaller angular deviation from the blended velocity than non-selected experts. Selected experts achieve smaller angular deviation from the blended velocity (6.4° vs. 11.3°, $p \approx 0$), a 43% reduction confirming systematic identification of coherent experts. The angular deviation gap (4.9°) is substantially larger than Paris DDM (1.5°), reflecting the stronger specialization of MNIST experts trained on digit-specific subsets.

Under full ensemble routing, we measure the correlation between trajectory-integrated expert disagreement and output quality degradation (MSE and LPIPS distance to Top-2 reference outputs). MNIST exhibits a substantially stronger disagreement-quality correlation than Paris DDM. This stronger correlation confirms that expert disagreement is a robust predictor of quality degradation, with the effect amplified in settings with stronger expert specialization.

C Formal Definitions for Trajectory Sensitivity

Definition C.1 (Trajectory-local sensitivity)**.** Given a flow v and an initial condition x_1, let $\{x_t\}_{t\in[0,1]}$ denote the (exact) ODE solution for $t \in [0,1]$. If the solution does not exist over $[0,1]$, we let $L_{\text{eff}}(x_1) = +\infty$. Otherwise, we define the *effective Lipschitz constant* at x_1 as $L_{\text{eff}}(x_1) = \sup_{t\in[0,1]} \|J_x v_t(x_t)\|$. Here, $\|\cdot\|$ denotes the operator norm induced by the Euclidean norm (i.e., the Jacobian spectral norm). We call the trajectory *locally stable* if $L_{\text{eff}}(x_1) < \infty$.

Definition C.2 (Empirical effective Lipschitz constant)**.** Given a numerical solver with step size h producing a discrete trajectory $\{\tilde{x}_{t_n}^{(h)}\}_{n=0}^{N}$, we define the *empirical effective Lipschitz constant* as $\widehat{L}_{\text{eff}}^{(h)}(x_1) := \max_{n\in\{0,\dots,N\}} \|J_x v(\tilde{x}_{t_n}^{(h)}, t_n)\|$.

Definition C.3 (Sampler sensitivity)**.** A sampler is (L,δ)-*trajectory-locally sensitive* if $P_{x_1\sim q}[L_{\text{eff}}(x_1) \le L] \ge 1-\delta$, for some noise distribution q, initial condition x_1, and constants L and δ.

Remark C.4 (Circularity of the diagnostic). Computing $\widehat{L}_{\text{eff}}^{(h)}$ requires the numerical trajectory, which is only available *after* sampling completes. This makes $\widehat{L}_{\text{eff}}^{(h)}$ a retrospective diagnostic rather than an a priori predictor.

D When $\widehat{L}_{\text{EFF}}^{(h)}$ approximates L_{EFF}

Assume the exact ODE solution $\{x_t\}_{t\in[0,1]}$ exists and remains in a set K, and that for all $t \in [0,1]$, $J_x v(\cdot, t)$ is L_J-Lipschitz on K, that is,

$$\|J_x v(x,t) - J_x v(y,t)\| \le L_J \|x - y\|$$

for all $x, y \in K$. Let $\tilde{x}^{(h)}(\cdot)$ be any continuous-time interpolation of the numerical trajectory such that $\sup_{t\in[0,1]} \|x_t - \tilde{x}^{(h)}(t)\| \le \eta(h)$ with $\eta(h) \to 0$ as $h \to 0$, and define

$$\widehat{L}_{\text{eff,cont}}^{(h)}(x_1) := \sup_{t\in[0,1]} \|J_x v(\tilde{x}^{(h)}(t), t)\|.$$

Then

$$\left|L_{\text{eff}}(x_1) - \widehat{L}_{\text{eff,cont}}^{(h)}(x_1)\right| \le L_J\, \eta(h),$$

so $\widehat{L}_{\text{eff,cont}}^{(h)}(x_1) \to L_{\text{eff}}(x_1)$ as $h \to 0$. Moreover, if $t \mapsto \|J_x v(\tilde{x}^{(h)}(t), t)\|$ is L_t-Lipschitz on $[0,1]$, then the grid maximum in Definition C.2 satisfies

$$0 \le \widehat{L}_{\text{eff,cont}}^{(h)}(x_1) - \widehat{L}_{\text{eff}}^{(h)}(x_1) \le L_t h,$$

so $\widehat{L}_{\text{eff}}^{(h)}(x_1)$ is a consistent proxy for $L_{\text{eff}}(x_1)$ under refinement.

This shows that the gap between this grid maximum and a continuous-time supremum is $O(h)$.

E Probabilistic Convergence under Empirical Stability

This appendix provides the formal convergence argument sketched in Section 5.1. We use a standard conditioning approach combined with deterministic ODE error bounds.

Proposition E.1 (Conditional convergence)**.** *Let v be a velocity field and q a noise distribution. Suppose that for $x_1 \sim q$, the effective Lipschitz constant $L_{\text{eff}}(x_1)$ (Definition C.1) satisfies $P(L_{\text{eff}}(x_1) < \infty) = 1$. Let $\tilde{x}_{t_N}^{(h)}$ denote the numerical solution at the final step using step size h, and let $e_N := \|x_{t_N} - \tilde{x}_{t_N}^{(h)}\|$ be the global error. Then for any $\varepsilon > 0$,*

$$\lim_{h\to 0} P(e_N > \varepsilon) = 0.$$

Proof. Fix $\varepsilon > 0$ and $\eta > 0$. Define the events $A_L := \{L_{\text{eff}}(x_1) \le L\}$ and $E := \{e_N > \varepsilon\}$. By the law of total probability,

$$P(E) \le P(E \mid A_L) + P(A_L^c).$$

Step 1 (Deterministic bound on A_L). Standard Grönwall-based global error analysis for one-step methods (see, e.g., Hairer et al. (1993)) shows that if the velocity field has Lipschitz constant L along the trajectory, then the global error satisfies

$$e_N \le Ch^p e^L$$

for constants $C > 0$ and $p \ge 1$ depending on the solver order and local truncation error bounds. On the event A_L, the trajectory-local Lipschitz constant is bounded by L, so this deterministic bound applies. Choosing $h^* = h^*(\varepsilon, L) := (\varepsilon/(Ce^L))^{1/p}$ ensures $e_N \le \varepsilon$ on A_L for all $h \le h^*$. Thus $P(E \mid A_L) = 0$ for $h \le h^*$.

Step 2 (Choosing L). Since $P(L_{\text{eff}}(x_1) < \infty) = 1$, for any $\eta > 0$ there exists $L = L(\eta)$ such that $P(A_L^c) = P(L_{\text{eff}}(x_1) > L) < \eta$.

Step 3 (Two-parameter limit). Given $\varepsilon, \eta > 0$: (i) choose $L = L(\eta)$ so that $P(A_L^c) < \eta$; (ii) choose $h \le h^*(\varepsilon, L)$ so that $P(E \mid A_L) = 0$. Then $P(E) < \eta$. Since $\eta > 0$ was arbitrary, $\lim_{h\to 0} P(e_N > \varepsilon) = 0$. □

Remark E.2. The key assumption is $P(L_{\text{eff}}(x_1) < \infty) = 1$, i.e., that almost all trajectories have finite effective Lipschitz constant. This is an empirical regularity condition that we validate by measuring $\widehat{L}_{\text{eff}}^{(h)}$ along numerical trajectories. The (L, δ)-trajectory-locally sensitive condition (Definition C.3) connects directly to this result: since the definition requires the bound to hold for *some* L and δ, we are free to choose any $\delta > 0$ and set $L = L(\delta)$ as the corresponding quantile of L_{eff}. For this $(L(\delta), \delta)$ pair and sufficiently small $h = h(\varepsilon, L(\delta))$, we obtain $P(e_N > \varepsilon) \le \delta$. Since δ can be made arbitrarily small, this recovers full convergence in probability.

F Expert-Data Alignment Experiment Details

This section provides implementation details for the alignment experiments described in Section 4.

DINOv2 embedding extraction. We use DINOv2-ViT-L/14 Oquab et al. (2023) to extract embeddings for both training data cluster centroids and intermediate sampling states. For intermediate states x_t during sampling, we first decode through the VAE to obtain pixel-space images, then extract DINOv2 embeddings. The 8 cluster centroids were computed during DDM training using hierarchical k-means on DINOv2 embeddings of the training set.

Cluster distance computation. For each routing decision at timestep t, we compute the Euclidean distance from the current sample's DINOv2 embedding to each of the 8 cluster centroids. The cluster rank is determined by sorting these distances (rank 1 = closest). For Top-k routing, we report the minimum rank among selected experts.

Velocity alignment computation. For each expert k at each timestep, we compute the velocity prediction $v_k(x_t)$ and measure its cosine similarity with the blended velocity $v_{\text{blend}}(x_t) = \sum_{j\in\mathcal{S}} w_j v_j(x_t)$. This requires evaluating all 8 experts at each recorded timestep, increasing computational cost by approximately $4\times$ compared to standard Top-2 sampling.

Statistical testing. Differences between selected and non-selected expert alignment scores are tested using paired t-tests, pairing by sample and timestep. Correlations are reported as Spearman's ρ with two-tailed p-values.

G Sensitivity Analysis Details

This appendix provides additional details for the trajectory sensitivity analysis in Section 5.

Table 5: **Local truncation error is routing-invariant.** One-step Heun local error $\epsilon_n^{\text{local}}$ at h=0.01 and h=0.005, with empirical scaling. Mean $\pm$ std. over n=1000 trajectories per routing strategy.

Routing	$\boldsymbol{\epsilon}^{\textbf{local}}$ (h=0.01)	$\boldsymbol{\epsilon}^{\textbf{local}}$ (h=0.005)	**Scaling**
Top-1	0.539 $\pm$0.045	0.304 $\pm$0.034	1.776$\times$
Top-2	0.542 $\pm$0.046	0.306 $\pm$0.034	1.773$\times$
Full (8)	0.543 $\pm$0.048	0.307 $\pm$0.035	1.771$\times$

Table 6 reveals that the router term $\sum_k v_t^{(k)} \nabla_x w_t^{(k)}$ dominates the expert term by 2–4 orders of magnitude across all routing strategies. This dominance reflects the inherent sensitivity of softmax routing to input perturbations, a property shared by Top-2 and full ensembling alike. However, when *comparing* routing strategies, the router term provides limited discriminative signal: it is uniformly large regardless of how many experts are selected. The expert term $\sum_k w_t^{(k)} \nabla_x v_t^{(k)}$, by contrast, captures how the *weighted mixture of expert outputs* changes with input. For this reason, our trajectory-local sensitivity traces (Figure 2) report the expert-only Jacobian $\| \sum_k w_t^{(k)} \nabla_x v_t^{(k)} \|$, isolating the component that explains sensitivity differences between routing strategies. The router gradient dominance is documented separately in Figure 4 and Table 6.

G.1 Experimental Setup

We use the pretrained DDM Paris model Jiang et al. (2025) via released pretrained checkpoints. The model consists of $K = 8$ experts trained on a subset of LAION-Aesthetics Schuhmann et al. (2022). The dataset was partitioned into 8 semantic clusters via two-stage hierarchical k-means on DINOv2-ViT-L/14 embeddings. The model uses a DiT-B/2 router ($\sim$129M parameters) and 8 DiT-XL/2 experts Peebles & Xie (2023) (a modified version of the DiT-XL/2 experts with $\sim$606M parameters each, $\sim$5B total).

We compare Top-1, Top-2, and full-ensemble routing strategies on the DDM architecture McAllister et al. (2025). To avoid circularity, that is defining failure via L_{eff} and then claiming L_{eff} predicts failure, we use Eq.(2) as a methodologically independent error signal.

We set τ_{refine} as the 99th percentile of Δ_{refine} on Top-2 runs, then apply this fixed threshold across all methods. This avoids tuning thresholds to match $\widehat{L}_{\text{eff}}^{(h)}$ predictions.

Appendix I presents additional experiments that test the robustness of our main findings.

G.2 Experiment 1: Local Truncation Error

Following Section 5.2, we estimate one-step local truncation error by comparing a single Heun step of size h Karras et al. (2022) against a higher-precision reference obtained by subdividing the same interval into 10 Heun sub-steps of size $h/10$. For each routing strategy, we sample 1,000 trajectories and evaluate at randomly selected trajectory points (x_t, t) the maximum local truncation error defined as $\epsilon^{\text{local}} = \max_n \|\epsilon_n^{\text{local}}(t_n)\| = Ch^{p+1}$ for some constant C. Table 5 reports these measurements.

Local error is essentially identical across routing strategies, confirming that routing does not affect single-step numerical accuracy.

G.3 Experiment 2: Trajectory-Local Sensitivity

We track $\|J_x v(\tilde{x}_{t_n}^{(h)}, t_n)\|$ across time for Top-1, Top-2, and full ensemble, and separate norms for selected vs. suppressed experts.

For a routed vector field $v_t(x_t) = \sum_k w_t^{(k)}(x_t) v_t^{(k)}(x_t)$, the Jacobian satisfies

$$\nabla_x v_t = \sum_k w_t^{(k)} \nabla_x v_t^{(k)} + \sum_k v_t^{(k)} \nabla_x w_t^{(k)}.$$

This decomposition separates sensitivity from experts ($\sum_k w_t^{(k)} \nabla_x v_t^{(k)}$) from sensitivity from routing ($\sum_k v_t^{(k)} \nabla_x w_t^{(k)}$). Since $\sum_k w_k = 1$ implies $\sum_k \nabla_x w_k = 0$, the router term can be rewritten as

Table 6: Jacobian decomposition at t=0.5 (n=100 samples). Both terms are measured along sampling trajectories. The router gradient term dominates for both strategies. Full ensemble shows nominally higher router term means, though the difference is small relative to variance.

Strategy	$\|J_{\text{expert}}\|$	$\|J_{\text{router}}\|$	**Dominant**
Top-2 routing	7.58 ±1.53	923 ±1.4K	Router
Full ensemble (8)	7.58 ±1.59	1161 ±2.1K	Router

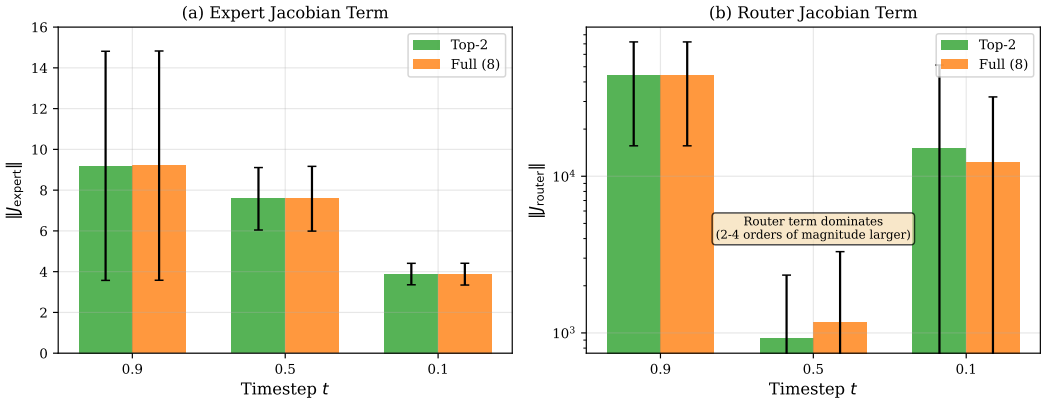

Figure 4: Temporal profile of the two Jacobian terms under Top-2 and full ensembling. Left: Expert term $\|\sum_k w_k \nabla_x v_k\|$ (linear scale). Right: Router term $\|\sum_k v_k \nabla_x w_k\|$ (log scale). The router term dominates by 2–4 orders of magnitude, but both routing strategies show similar router contributions.

$\sum_k (v_k - v_t)\nabla_x w_k$, showing that its magnitude is governed by inter-expert disagreement times router sensitivity—large $\|\nabla_x w_k\|$ alone does not inflate this term if experts agree. We report norms of each term separately as a diagnostic; by the triangle inequality, the full Jacobian norm satisfies $\|\nabla_x v_t\| \leq \|J_{\text{expert}}\| + \|J_{\text{router}}\|$, but these bounds need not be tight. Table 6 reports the two terms evaluated at t=0.5 (mean ± std. over n=100 trajectories).

Table 6 reveals the router term dominates by 2–4 orders of magnitude, but is similar across strategies. Since this shared dominance cannot explain the quality differences observed between routing strategies, the expert term $\sum_k w_k \nabla_x v_k$—which captures how the weighted mixture of expert outputs responds to input perturbations—is the relevant quantity for understanding routing-quality relationships. We therefore report the expert-only Jacobian in Figure 2.

G.4 Experiment 3: Step-Size Refinement

We compute Δ_{refine} by running N=50 and $2N$=100 steps from the same initial noise x_1 and prompt, decoding both final latents with $D(\cdot)$, and measuring LPIPS in image space. Figure 3 shows the correlation between $\widehat{L}_{\text{eff}}^{(h)}$ and Δ_{refine}.

H Limitations

Correlation strength. The weak correlations between L_{eff} and Δ_{refine} in our main in-distribution setting ($\rho < 0.1$) limit the predictive utility of trajectory Jacobian analysis for identifying failure cases. Future work should explore whether alternative sensitivity metrics (e.g., integrated sensitivity, switching frequency) provide stronger predictive signals.

Unexplained typical-set attractivity. Our convergence argument relies on a trajectory-local boundedness condition ($L_{\text{eff}}(x_1) < \infty$), and our empirical results are consistent with trajectories remaining in moderate-sensitivity regions. However, we do not provide a proof that the *exact* probability flow

dynamics must enter and remain in such low-sensitivity regions, nor do we characterize basins of attraction for the routed field. Developing a dynamical explanation is an important direction for future work.

Scope. This paper makes a *mechanistic* claim about decentralized expert systems: routing trades off numerical sensitivity and expert-data alignment, and alignment can dominate quality. Because this claim is established by controlled comparisons that hold the expert pool and router fixed, it does not require external diffusion baselines. Adding external models without matched training would primarily answer a different question, and could obscure the routing mechanism due to unavoidable training/data confounds.

I Extra Experiments

This subsection presents additional experiments that test the robustness of our main findings.

I.1 Full-ensemble tuning variants.

We test whether the full ensemble's sensitivity metrics can be improved via inference-only modifications that keep the expert pool fixed and require no retraining: (i) temperature scaling of router logits sweeping $T \in \{0.1, 0.25, 0.5, 1.0, 2.0, 4.0\}$, and (ii) top-$p$ truncation of the router distribution, keeping the smallest set of experts whose cumulative mass exceeds p and renormalizing within this set. Table 7 summarizes the best variants from each sweep. Neither modification substantially changes $\widehat{L}_{\text{eff}}^{(h)}$ or Δ_{refine} compared to the baseline full ensemble.

I.2 Counterfactual Routing

We evaluate two inference-only counterfactuals for the full ensemble (Table 8): (i) weight clipping that suppresses experts whose Jacobian norms are above the median at the current state, and (ii) **Misaligned Top-2** (random expert selection), which preserves sparsity while explicitly breaking proximity-based alignment.

I.3 Failure Mode Analysis

To understand how different routing strategies fail, we categorize samples by three failure indicators: high routing uncertainty, poor numerical convergence, and high effective Lipschitz constant. Table 13 reports the frequency of each failure mode across routing strategies.

I.4 Full-Ensemble Rescue Attempts

We test inference-time modifications to the full ensemble to investigate whether its poor sample quality (despite superior numerical stability) can be improved without retraining. As discussed in Section 5, full ensemble has the lowest $\widehat{L}_{\text{eff}}^{(h)}$ and Δ_{refine} but worst FID due to expert-data misalignment: experts process out-of-distribution inputs from data partitions they were not trained on.

We test inference-time temperature scaling of router logits: $w_T(x,t) = \text{softmax}(z(x,t)/T)$. We sweep $T \in \{0.1, 0.25, 0.5, 1.0, 2.0, 4.0\}$ for the *full ensemble* strategy ($v(x,t) = \sum_{k=1}^{K} w_{T,k}(x,t) v_k(x,t)$), holding the expert pool and solver fixed. Table 9 reports the results of the temperature sweep.

We also sweep $p \in \{0.8, 0.9, 1.0\}$. Table 10 reports the results. Truncation has minimal effect on sensitivity metrics.

I.5 Generalization Tests Without Retraining

To test whether the sensitivity ordering is specific to the training distribution, we evaluate the same frozen experts and router on a held-out prompt set. We generate n=100 samples per routing strategy and report the same sensitivity diagnostics. Table 11 shows that the Δ_{refine} ordering (Full lowest) persists under this prompt shift, while $\widehat{L}_{\text{eff}}^{(h)}$ values are comparable across strategies.

Table 7: **Full-ensemble rescue attempts (no retraining).** Temperature scaling uses $w_T = \mathrm{softmax}(z/T)$; we report the best T from a sweep over $\{0.1, 0.25, 0.5, 1.0, 2.0, 4.0\}$. Top-$p$ truncation keeps the smallest set of experts whose cumulative probability exceeds p and renormalizes. Results from n=50 samples per configuration.

Variant	$\widehat{L}_{\text{eff}}^{(h)} \downarrow$	$\Delta_{\text{refine}} \downarrow$
Full ensemble (baseline)	15.97 ±5.51	0.043 ±0.052
Temp scaling (T=4.0)	**15.00** ±4.71	0.044 ±0.056
Top-p truncation ($p^\star$=0.9)	15.85 ±5.30	**0.043** ±0.052

Table 8: **Counterfactual routing interventions.** Mean Δ_{refine} from Eq. 2 (computed without Jacobian metrics). Results from n=50 samples.

Condition	$\Delta_{\text{refine}} \downarrow$
Top-2 (base)	0.043 ±0.054
Full ensemble (8)	0.039 ±0.052
Full + weight clip ($\|\nabla_x v_k\| <$ median)	**0.037** ±0.029
Misaligned Top-2 (random)	0.040 ±0.035

We also evaluate two forms of distribution shift: (i) out-of-distribution prompts and (ii) stressed numerical regimes.

We evaluate the fixed trained experts and router on COCO captions as prompts. We sample n=100 prompts uniformly at random from the caption set, generate one sample per prompt with fixed seeds, and compute $\widehat{L}_{\text{eff}}^{(h)}$ and Δ_{refine} as in the main text. We additionally report the AUC of $\widehat{L}_{\text{eff}}^{(h)}$ for predicting high Δ_{refine} events using the same percentile thresholding protocol. We repeat the same evaluation under a harder numerical regime without retraining: (i) fewer solver steps (Heun-25 instead of Heun-50), and (ii) higher CFG Ho & Salimans (2022) (7.5 instead of 4.0). The goal is not quality, but whether sensitivity rankings and predictiveness persist. Table 14 compares baseline and stressed regimes.

I.6 Switching Sensitivity: Margin and Vector-Field Gap

We analyze nonsmooth expert switching by combining (i) proximity to the switching surface (router margin) and (ii) the jump size in the routed vector field (vector-field gap). This addresses the limitation that for hard Top-1 the Jacobian $J_x v$ is defined only inside routing regions and does not capture discontinuities at switches.

Let $z_k(x,t)$ be router logits, $p(k \mid x,t) = \mathrm{softmax}(z(x,t))_k$, and let $k_{(1)}, k_{(2)}$ index the top-2 logits. We report: (1) probability margin $m_p = p_{(1)} - p_{(2)}$, (2) logit margin $m_z = z_{(1)} - z_{(2)}$, (3) vector-field gap $g = \|v_{k_{(1)}} - v_{k_{(2)}}\|_2$, (4) switching score $S_{\text{switch}} = g/(m_z + \epsilon_{\text{sw}})$ with $\epsilon_{\text{sw}} = 10^{-3}$ for numerical stability. For each trajectory we summarize by $S_{\text{eff}} = \max_t S_{\text{switch}}(x(t),t)$ and an integrated variant $S_{\text{int}} = \int_0^1 S_{\text{switch}}(x(t),t)\,dt$.

For each saved trajectory state (x_t, t) we compute (m_p, m_z) from the router, then evaluate the two corresponding experts $v_{k_{(1)}}, v_{k_{(2)}}$ to obtain g and S_{switch}. For Top-1, this requires one extra expert evaluation at analysis time. We subsample timesteps identically to the L_{eff} pipeline.

We first reproduce margin statistics as a proxy for distance to switching surfaces. See Table 15.

Low-margin segments concentrate in failures, but margin alone does not measure the impact of switching. We evaluate how different predictors perform at identifying failures within Top-1 routing in Table 12.

The key findings are: (1) the vector-field gap g alone is the best single predictor for high Δ_{refine}, (2) combining L_{eff} with switching features does not improve over g alone.

Table 9: Temperature scaling sweep for full ensemble routing. Lower temperatures sharpen the weight distribution toward top experts; higher temperatures flatten it. All metrics measured on n=100 samples. T=4.0 achieves the lowest Δ_{refine} in this sweep.

T	**Entropy**	$\widehat{L}_{\text{eff}}^{(h)}$	Δ_{refine}
0.10	0.17	18.84 ±7.43	0.060 ±0.047
0.25	0.43	18.68 ±7.07	0.054 ±0.038
0.50	0.79	18.06 ±5.65	0.054 ±0.051
1.00	1.24	17.32 ±5.48	0.050 ±0.049
2.00	1.70	16.35 ±4.56	0.047 ±0.045
4.00	1.96	16.17 ±4.88	0.044 ±0.034

Table 10: Top-p truncation sweep for full ensemble routing (T=1.0). Lower p values exclude low-probability experts. All metrics measured on n=50 samples. Truncation has minimal effect on $\widehat{L}_{\text{eff}}^{(h)}$ or Δ_{refine}.

p	$\widehat{L}_{\text{eff}}^{(h)}$	Δ_{refine}
0.8	16.51 ±5.64	0.043 ±0.052
0.9	15.85 ±5.30	0.043 ±0.052
1.0	15.97 ±5.51	0.043 ±0.052

Table 11: **Generalization test on COCO captions (out-of-distribution prompts).** Frozen experts/router, new prompt distribution. Results from n=100 samples with Heun solver (50 steps). Δ_{refine} ordering matches Table 3 (Full lowest), while $\widehat{L}_{\text{eff}}^{(h)}$ values are comparable across strategies.

Strategy	$\widehat{L}_{\text{eff}}^{(h)} \downarrow$	$\Delta_{\text{refine}} \downarrow$	**Spearman** $\rho(\widehat{L}_{\text{eff}}^{(h)}, \Delta_{\text{refine}})$
Top-1	27.17 ±12.08	0.114 ±0.109	0.11
Top-2	26.32 ±11.95	0.083 ±0.089	0.01
Full (8)	26.37 ±11.83	**0.048** ±0.067	−0.01

Table 12: Comparison of failure predictors for Top-1 routing. S_{eff} (switching score) combines margin and vector-field gap. The vector-field gap g alone achieves the best prediction of high Δ_{refine} (AUC 0.63).

Top-1 predictor	**AUC (high** Δ_{refine}**)**	**Spearman vs.** Δ_{refine}
m_p only	0.50	0.07
g only	**0.63**	**0.20**
S_{eff} (margin+gap)	0.55	0.08
L_{eff} only	0.58	0.08
$L_{\text{eff}} + S_{\text{eff}}$	0.58	0.11

Table 13: Failure mode analysis across routing strategies (n=100 samples). Thresholds: Routing uncert. = max entropy >1.5 nats; Poor conv. = Δ_{refine}>0.1; High L_{eff} = $\widehat{L}_{\text{eff}}^{(h)}$>50. Poor convergence decreases with more experts. High L_{eff} events are rare. Note: Routing uncertainty naturally increases with ensemble size since more experts contribute non-zero weights.

Strategy	$\widehat{L}_{\text{eff}}^{(h)}$	Δ_{refine}	**Routing uncert.**	**Poor conv.**	**High** L_{eff}
Top-1	24.21 ±7.44	0.112 ±0.114	60%	39%	1%
Top-2	23.68 ±8.07	0.094 ±0.074	64%	33%	1%
Top-4	23.56 ±6.74	0.049 ±0.059	88%	11%	0%
Full (8)	23.72 ±6.97	**0.038** ±0.053	94%	5%	1%

Table 14: Generalization under baseline vs. stressed numerical regimes (COCO captions, n=100). Stressed regime uses fewer solver steps (Heun-25) and higher CFG (7.5). Sensitivity rankings persist across regimes.

Strategy	**Regime**	$\widehat{L}_{\text{eff}}^{(h)}$	Δ_{refine}	**Spearman**	**AUC**
Top-1	Baseline	20.09 $\pm$12.61	0.131 $\pm$0.128	0.09	0.56
Top-2	Baseline	18.39 $\pm$5.57	0.078 $\pm$0.079	**0.31**	**0.69**
Full (8)	Baseline	**17.29** $\pm$5.82	**0.054** $\pm$0.073	0.05	0.52
Top-1	Stressed	18.16 $\pm$5.99	0.206 $\pm$0.134	0.20	0.54
Top-2	Stressed	17.27 $\pm$5.53	0.190 $\pm$0.148	0.03	0.59
Full (8)	Stressed	**16.07** $\pm$4.68	**0.136** $\pm$0.135	-0.06	0.48

Table 15: Router probability margin statistics for stable vs. unstable samples (unstable = Δ_{refine} above 75th percentile). Unstable samples exhibit lower margins and more frequent near-switching events, but margin alone is insufficient for prediction.

Routing	**Median** m_p	**% Steps** $m_p < 0.05$
Top-1 (stable samples)	0.42 ± 0.18	8.3%
Top-1 (unstable samples)	0.21 ± 0.14	24.7%
Top-2 (stable samples)	0.38 ± 0.16	9.1%
Top-2 (unstable samples)	0.19 ± 0.12	27.2%