Stability and Diversity of Networked Self-Consuming Generative Ecosystems

Xiukun Wei[1], **Yang Zhang**[1], **Xueru Zhang**[1]
[1]The Ohio State University
{wei.1418, zhang.15751, zhang.12807}@osu.edu

Abstract

The widespread deployment of generative AI has made it increasingly difficult to distinguish synthetic content from real data. Consequently, synthetic data is inevitably incorporated into the training pipelines of future model generations, forming a *self-consuming training loop*. Prior work has studied the effects of such recursive self-consuming training, but analyses have largely been limited to *isolated* models, where a model consumes only its own synthetic data, or to simplified interactions between two models. This paper takes a first step toward understanding *networked* self-consuming generative models, in which multiple models consume synthetic data generated by one another through complex interaction pathways. We introduce a theoretical framework representing models as nodes in a directed, weighted graph, with edge weights governing the flow of synthetic data among models. Using this framework, we analyze the long-term behavior of networked models under retraining dynamics, establishing conditions for convergence and characterizing the resulting fixed points. We further investigate how the system's long-term stability and diversity are shaped by each model's access to real data, cross-model data consumption, and the structure of the interaction graph.

1 Introduction

The proliferation of generative models has fundamentally transformed the nature of data used for model training. As synthetic data increasingly permeates public repositories, online platforms, and downstream training pipelines [37], maintaining datasets that are entirely real or purely human-generated is becoming increasingly challenging. This has motivated a rapidly growing line of research on *self-consuming models*, in which models are iteratively updated using mixtures of real and synthetic data generated by the models themselves. Prior work has shown that models trained under such self-consuming loops often exhibit distributional drift [14, 22, 27, 29], collapse [2, 4, 16], loss of tail coverage [11, 28], and potentially bias amplification [32, 38, 42, 43].

However, existing work on self-consuming models has largely focused on **isolated** models, where a single model is trained independently and consumes only synthetic data generated by itself. In practice, real-world generative ecosystems are inherently multi-model and interactive: online platforms are populated with outputs from many models, and downstream training pipelines inevitably incorporate synthetic samples produced by peer models [30]. For example, in the domain of large language models (LLMs), *Orca* is fine-tuned using large-scale synthetic instruction-response pairs generated by ChatGPT and GPT-4 [25]. Similarly, *Phi-3* relies heavily on synthetic data generated by larger LLMs to achieve strong performance despite its smaller parameter scale [1]. This multi-model interaction is also central to modern text-to-image systems: *DALL·E 3* employs a dedicated captioning model to rewrite noisy prompts into high-quality synthetic captions, improving text-image alignment during training [5]; *PixArt-*α adopts a similar strategy by utilizing LLaVA to refine its training data [8].

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

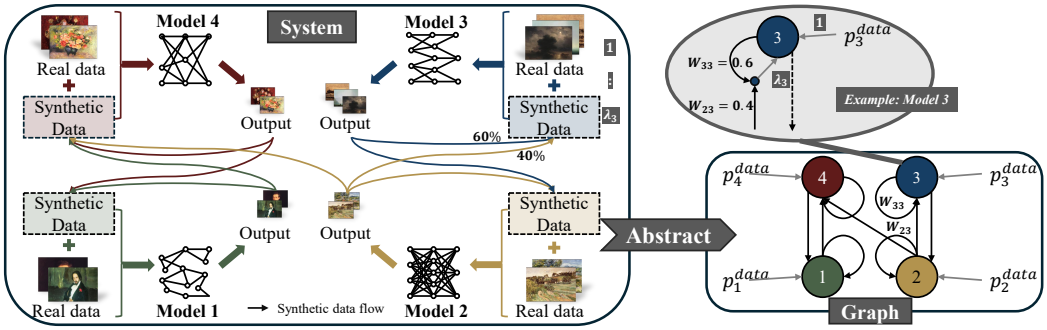

Figure 1: From networked self-consuming models to graph abstraction. (1) Left: A generative ecosystem in which each model is trained on its own real-data distribution and a weighted mixture of synthetic data produced by other models. (2) Right: Abstraction of the system as a directed, weighted interaction graph that captures effective synthetic-data propagation among models. (3) Top right: An example local retraining rule, where *Model* 3 is trained at iteration t on $p_3^{\text{data}} + \lambda_3 \sum_{j=1}^{K} w_{3j} p_{\boldsymbol{\theta}_j^{t-1}}$.

Although a few recent works have recognized cross-model interactions in self-consuming training [15, 36, 45], showing that such interactions can accelerate homogenization or model collapse, these studies are still highly limited. Most are restricted to *two-model* coupled systems [15, 45] or rely on highly *simplified interactions* and *linearized dynamics assumptions* [36], leaving the behavior of large, heterogeneous networks of interacting self-consuming models largely unexplored.

To fully understand real-world generative ecosystems, it is necessary to move beyond these simplified settings and explicitly account for the complex, networked interactions among many models. Unfortunately, existing analyses fail to explicitly capture the *networked, directed, and weighted* flows of synthetic data among heterogeneous models, nor do they characterize the asymptotic impact of graph structure on stability and diversity. As a result, a general theoretical framework is still lacking, one that (i) accommodates arbitrary interaction graphs, (ii) enables systematic analysis of how interaction topology and synthetic-data mixing jointly affect system stability, and (iii) enables quantification of how network connectivity and mixing strength influence system diversity.

In this paper, we initiate a theoretical study of **networked self-consuming models** by modeling the system as a directed interaction graph (Fig. 1). Each node represents a generative model, and weighted edges capture the probability that synthetic data produced by one model is consumed by another. Within this framework, we analyze the behavior of networked self-consuming models by characterizing (i) the evolution of networked models under self-consuming iterative training; (ii) the long-term stability of retraining dynamics; (iii) the effects of each model's access to real data, cross-model data consumption, and graph topology on the system's long-term stability and diversity.

Specifically, we show that under standard regularity conditions for iterative retraining, the retraining dynamics can be locally analyzed by the Jacobian around any fixed point of the system. Assuming such a fixed point exists, we derive sufficient conditions for the stability and convergence of the fixed points that separate *local amplification factors*—determined by each model's synthetic-data mixing weight and local curvature—from *global propagation effects* governed by the interaction graph. This condition strictly generalizes the stability threshold of prior single-model analyses [4, 13], accommodating arbitrary directed and weighted graphs.

We also study how the system's long-term diversity is affected by cross-model data consumption, especially when models have access to varying real data distributions. We first derive an upper bound on system diversity at the fixed point, showing that synthetic-data consumption among models induces a *topology-dependent smoothing effect* that contracts inter-model heterogeneity. More related work is discussed in Appx. C, and our contributions are summarized below.

- **A unified theoretical framework for networked self-consuming models**. We introduce a framework that captures directed, weighted interactions among heterogeneous models with arbitrary cross-model synthetic-data flows (Sec. 2).
- **Graph-structured convergence & stability analysis.** We derive explicit conditions for convergence and stability of the networked self-consuming models, showing that how the synthetic-data mixing weight and the interaction graph topology jointly affect long-term stability (Sec. 3).

- **Topology-dependent diversity analysis.** We characterize the long-term diversity at the fixed point and show that synthetic-data consumption induces a topology-dependent smoothing effect that contracts inter-model heterogeneity (Sec. 4).
- **Empirical validation.** We validate our theoretical predictions on a synthetic 8-Gaussian dataset and a real image dataset CIFAR-10 [1]. Across multiple generative model classes and a range of interaction topologies, the experimental observations align with our analysis (Sec. 5).

2 Problem Formulation

Consider a set $\mathcal{V} = \{1, \dots, K\}$ of generative models. Each model $i \in \mathcal{V}$ has access to a real data distribution $p_i^{\text{data}}(x)$ over a shared data space $\mathcal{X}$ and is parameterized by $\boldsymbol{\theta}_i \in \Theta_i$, inducing a generated data distribution $p_{\boldsymbol{\theta}_i}$. We allow heterogeneous model architectures by permitting $\Theta_i \neq \Theta_j$ for $i \neq j$ and heterogeneous real data where p_i^{data} varies across models.

Data flow among networked models. If models are trained in isolation using only their own real data sources, each model i converges to its own maximum likelihood estimator (MLE), denoted by $\boldsymbol{\theta}_i^\star := \arg\max_{\boldsymbol{\theta}} \mathbb{E}_{x\sim p_i^{\text{data}}}[\log p_{\boldsymbol{\theta}}(x)]$. We refer to the collection of these local optima, $\Theta^\star = (\boldsymbol{\theta}_1^\star, \dots, \boldsymbol{\theta}_K^\star)$, as the *uncoupled solution*.

However, as synthetic data proliferates and becomes difficult to distinguish from real data, it is often inevitably used to train future generations of models, including both the originating model itself and other models. We formalize such interactions using a directed, weighted graph $(\mathcal{V}, \mathcal{E}, \mathbf{W})$, where each node in $\mathcal{V}$ represents a generative model. A directed edge $j \to i \in \mathcal{E}$ indicates that synthetic data generated by model j may be used to retrain model i. We refer to models with edges pointing to i as its *parents*. The adjacency matrix $\mathbf{W} = [w_{ij}] \in \mathbb{R}^{K\times K}$ assigns a weight $w_{ij} \geq 0$ to each edge $j \to i$, representing the probability that any given synthetic sample consumed by model i during retraining is generated by model j. Consequently, $\mathbf{W}$ is *row-stochastic*, i.e., $\sum_{j=1}^K w_{ij} = 1$ for each $i \in \mathcal{V}$. *Self-loops* ($w_{ii} \geq 0$) are allowed, meaning that a model may reuse its own generated data.

Iterative training loop. As the environment evolves, models are updated repeatedly to maintain high performance. We formalize this iterative training process as follows. At $t = 0$, each model is initialized from a pre-trained state obtained on its own real data p_i^{data}. In subsequent rounds ($t \geq 1$), each model i updates its parameters $\boldsymbol{\theta}_i^t$ using a mixture of its real data and synthetic data generated by itself and its parents in the previous round, i.e.,

$$\boldsymbol{\theta}_i^t \in \arg\max_{\boldsymbol{\theta}\in\Theta_i} \mathbb{E}_{x\sim p_i^{\text{data}}}[\log p_{\boldsymbol{\theta}}(x)] + \lambda_i \sum_{j=1}^K w_{ij} \mathbb{E}_{x\sim p_{\boldsymbol{\theta}_j^{t-1}}}[\log p_{\boldsymbol{\theta}}(x)] \tag{1}$$

where $\lambda_i \geq 0$ controls the relative contribution of synthetic data.[2] Stacking the parameters across all models, we denote the global parameter vector at round t as $\Theta^t := (\boldsymbol{\theta}_i^t, \dots, \boldsymbol{\theta}_K^t) \in \prod_{i=1}^K \Theta_i$. We then define the global iteration map $\mathcal{G} : \prod_{i=1}^K \Theta_i \to \prod_{i=1}^K \Theta_i$ by

$$\Theta^{t+1} = \mathcal{G}(\Theta^t) := \big(\mathcal{G}_1(\Theta^t), \dots, \mathcal{G}_K(\Theta^t)\big) \tag{2}$$

where $\mathcal{G}_i$ is the update operator for model i as in Eq. (1).

Objectives. Under the dynamics (2), the network of generative models evolves over time, yet its long-term behavior remains poorly understood. While prior work has investigated the effects of self-consuming retraining paradigms, existing studies are largely limited to a single model in isolation [2, 4, 29] or to two-model settings with highly simplified interaction structures [15]. In contrast, real-world deployments involve multiple generative models interacting through complex and heterogeneous pathways, in which synthetic data produced by one model can propagate throughout the ecosystem and influence the training and behavior of many others.

This motivates a system-level investigation of how synthetic data and self-consuming iterative training loops shape the long-term dynamics of a network of generative models. In this paper, we seek to address the following fundamental questions regarding the asymptotic behavior of the network:

[1] Please see our code at https://github.com/osu-srml/Gen-System

[2] When the maximizer is not unique, we select the solution from the set of maximizers $\widehat{\Theta}_i^t$ that is closest to the previous iterate, i.e., $\boldsymbol{\theta}_i^t := \arg\min_{\boldsymbol{\theta}'\in\widehat{\Theta}_i^t} \|\boldsymbol{\theta}' - \boldsymbol{\theta}_i^{t-1}\|$, reflecting common practice in which models are *fine-tuned* from their prior state using gradient-based training paradigms rather than retrained from scratch.

- How do networked generative models evolve under self-consuming iterative training loops, and under what conditions do the dynamics converge and admit fixed points?
- How do structural properties of the interaction graph $\mathbf{W}$ (e.g., connectivity, length and configuration of feedback cycles) and the strength of synthetic-data contributions $\{\lambda_i\}_{i=1}^K$ affect the long-term stability and diversity of individual models and of the network as a whole?
- How does the diversity of model outputs evolve under cross-model interactions, and to what extent is it preserved at equilibrium?

Sec. 3 addresses the first two questions through a local stability analysis around an arbitrary fixed point. Sec. 4 addresses the third question by analyzing how synthetic data interactions and graph topology shape inter-model diversity at the system's fixed point.

3 Convergence & Stability Analysis

3.1 Fixed Point and Technical Assumptions

We begin by formalizing the notion of *fixed point*, a state in which the models reach equilibrium and remain unchanged under iterative training loop.

Definition 3.1 (Fixed point)**.** The global parameter vector $\bar{\Theta}$ is a fixed point of the dynamics (2) if $\bar{\Theta} = \mathcal{G}(\bar{\Theta})$.

As shown in Appx. D.1, any fixed point $\bar{\Theta}$ satisfies the first-order optimality condition:

$$\nabla_{\boldsymbol{\theta}} \mathbb{E}_{x\sim p_i^{\text{data}}}[\log p_{\bar{\boldsymbol{\theta}}_i}(x)] + \lambda_i \sum_{j=1}^{K} w_{ij} \nabla_{\boldsymbol{\theta}} \mathbb{E}_{x\sim p_{\bar{\boldsymbol{\theta}}_j}}[\log p_{\bar{\boldsymbol{\theta}}_i}(x)] = 0, \forall i \in \mathcal{V} \tag{3}$$

In the single-model setting ($|\mathcal{V}| = 1$), the MLE $\boldsymbol{\theta}_i^\star$ satisfies Eq. (3) under Assumptions 3.3-3.4 and is therefore a fixed point [4]. In the networked multi-model setting, however, the uncoupled solution $\Theta^\star = (\boldsymbol{\theta}_1^\star, \ldots, \boldsymbol{\theta}_K^\star)$ is generally *not* a fixed point of $\mathcal{G}$; Appx. D.3 gives a two-model example exhibiting a fixed point distinct from $\Theta^\star$. Nevertheless, we can identify conditions under which $\Theta^\star$ remains a fixed point in multi-model systems, as established in Prop. 3.2 (see Appx. D.2).

Proposition 3.2 (Uncoupled solution as a fixed point)**.** *When* (i) *the real data distributions are homogeneous ($p_i^{\text{data}} \equiv p^{\text{data}}$ for all i) and* (ii) *the per-model MLEs induce a common generative distribution ($p_{\boldsymbol{\theta}_i^\star} \equiv p^\star$ for all $i \in \mathcal{V}$, for some $p^\star$), uncoupled solution $\Theta^\star$ is a fixed point.*

In general multi-model settings, establishing the existence of fixed points satisfying Eq. (3) and characterizing them is substantially more challenging than in the single-model case, due to the high-dimensional coupled structure of $\mathcal{G}$ and the flexibility of the interaction graph $\mathbf{W}$. Their existence typically requires global conditions ensuring that the retraining map is a continuous self-map on a compact parameter space. Instead of pursuing such a global fixed-point analysis, we study the local dynamics around an arbitrary fixed point $\bar{\Theta}$ whenever one exists. In particular, we aim to characterize the stability of a given equilibrium and how this stability is influenced by synthetic-data mixing and the interaction graph.

We now introduce the assumptions imposed at every fixed point $\bar{\Theta}$ under consideration. Let U_i denote a neighborhood of $\bar{\boldsymbol{\theta}}_i$, and $\mathcal{U} := \prod_i U_i$. These assumptions are standard and follow those of [4].

Assumption 3.3 (Local strong concavity)**.** For each model i, the map $\boldsymbol{\theta} \mapsto \mathbb{E}_{x\sim p_i^{\text{data}}}[\log p_{\boldsymbol{\theta}}(x)]$ is twice continuously differentiable on U_i, and there exists $\alpha_i > 0$ such that $-\nabla^2_{\boldsymbol{\theta}} \mathbb{E}_{x\sim p_i^{\text{data}}}[\log p_{\boldsymbol{\theta}}(x)] \succeq \alpha_i I_{d_i}$ for all $\boldsymbol{\theta} \in U_i$, where $d_i = \dim(\Theta_i)$.

Assumption 3.4 (Hessian stability under distribution shift)**.** Denote $d_{\mathcal{W}}(\cdot,\cdot)$ as 1-Wasserstein distance [35]. For each model i, there exists $L_i \geq 0$ such that for all $\boldsymbol{\theta} \in U_i$ and any probability measures q, q' over $\mathcal{X}$, $\left\|\mathbb{E}_{x\sim q} \nabla^2_{\boldsymbol{\theta}} \log p_{\boldsymbol{\theta}}(x) - \mathbb{E}_{x\sim q'} \nabla^2_{\boldsymbol{\theta}} \log p_{\boldsymbol{\theta}}(x)\right\| \leq L_i d_{\mathcal{W}}(q, q')$.

The above assumptions characterize the local regularity of each model around the fixed point $\bar{\Theta}$. To analyze the networked dynamics, we further introduce two quantities.

Definition 3.5 (Distributional gap)**.** For each $i \in \mathcal{V}$, the *distributional gap* between model i's real data distribution and its parent output distribution at $\bar{\Theta}$ is $\varepsilon_i := \max_{j:w_{ij}>0} d_W\left(p_{\bar{\boldsymbol{\theta}}_j}, p_i^{\text{data}}\right)$.

Definition 3.6 (Local amplification factor)**.** The *local amplification factor* at $\bar{\Theta}$ for model $i \in \mathcal{V}$ is

$$\beta_i := \frac{\lambda_i M_i}{\alpha_i + \lambda_i(\alpha_i - L_i\varepsilon_i)}.$$

where $M_i := \max_{j:w_{ij}>0} \left\| \mathbb{E}_{x\sim p_{\bar{\boldsymbol{\theta}}_j}} \left[\nabla_{\boldsymbol{\theta}} \log p_{\bar{\boldsymbol{\theta}}_i}(x) \, \nabla_{\boldsymbol{\theta}_j} \log p_{\bar{\boldsymbol{\theta}}_j}(x)^\top \right] \right\|_2$ captures the strongest cross-model sensitivity between the model i and its parent models at the fixed point.

Throughout the paper, $\|\cdot\|$ denotes the Euclidean norm for vectors, and $\|\cdot\|_2$ denotes the induced operator (spectral) norm for matrices. Local amplification factor β_i quantifies how strongly a perturbation arriving at model i through synthetic data is amplified by its local retraining dynamics. Stacking these factors with the interaction graph, we define the nonnegative *comparison matrix*

$$\mathbf{D} := \mathrm{diag}(\beta_1, \ldots, \beta_K)\mathbf{W}.$$

3.2 Convergence of networked self-consuming models.

Next, we analyze the evolution of networked models and identify conditions under which the dynamics converge. Prior studies have shown that a *single* self-consuming model can fail to converge when the synthetic-data contribution λ_i is sufficiently large [4]. When multiple models coexist and interact over a network, analyzing the stability of the entire system becomes substantially more challenging. In this case, stability is no longer solely a local property determined by λ_i, rather, it emerges as a *system-level* property that also depends on the structure of the interaction network encoded by $\mathbf{W}$.

Theorem 3.7 (Local asymptotic stability)**.** *Suppose $\alpha_i > L_i\varepsilon_i$, $\forall i \in \mathcal{V}$. If $\rho(\mathbf{D}) < 1$, then $\bar{\Theta}$ is locally asymptotically stable. Moreover, there exists a neighborhood $\mathcal{U}$ of $\bar{\Theta}$ such that for any initialization $\Theta^0 \in \mathcal{U}$, the iterates $\Theta^{t+1} = \mathcal{G}(\Theta^t)$ converge to $\bar{\Theta}$ at asymptotic linear rate*

$$\limsup_{t\to\infty} \|\Theta^t - \bar{\Theta}\|_2^{1/t} \le \rho(\mathbf{J}) \le \rho(\mathbf{D}). \tag{4}$$

where $\mathbf{J} := \nabla_\Theta \mathcal{G}(\bar{\Theta})$ denotes the Jacobian of $\mathcal{G}$ at $\bar{\Theta}$, and $\rho(\cdot)$ is the spectral radius of a matrix.

The proof is in Appx. D.4. Thm. 3.7 establishes conditions for the convergence of networked self-consuming models. Since $\mathbf{D}$ is nonnegative and encodes the interaction graph $\mathbf{W}$ directly, $\rho(\mathbf{D})$ serves as a topology-aware upper bound on the spectral radius of $\mathbf{J}$. The stability holds whenever $\rho(\mathbf{D}) < 1$, i.e., when local amplification and network propagation together do not compound to an expanding fixed point. Since $\rho(\mathbf{D}) \le \|\mathbf{D}\|_2 \le \beta_{\max}\|\mathbf{W}\|_2$ with $\beta_{\max} := \max_{i\in\mathcal{V}} \beta_i$, we have the following corollary with its proof in Appx. D.5.

Corollary 3.8 (A norm-based sufficient condition)**.** *If $\beta_{\max}\|\mathbf{W}\|_2 < 1$, then $\rho(\mathbf{D}) < 1$, and $\bar{\Theta}$ is locally asymptotically stable. Equivalently, when $M_i\|\mathbf{W}\|_2 > \alpha_i - L_i\varepsilon_i$ for all $i \in \mathcal{V}$, a sufficient condition for local asymptotic stability is*

$$\lambda_i < \frac{\alpha_i}{M_i\|\mathbf{W}\|_2 - (\alpha_i - L_i\varepsilon_i)}, \quad \forall i \in \mathcal{V}. \tag{5}$$

As in isolated model settings, stability is guaranteed when the synthetic-data contribution λ_i is sufficiently small. In the presence of cross-model interactions, the key difference is that the threshold now depends on two network-level quantities: the cross-model sensitivity M_i which measures how parameter perturbations in parent models propagate through their synthetic data to model i's gradient signal; and the spectral norm of the interaction graph $\|\mathbf{W}\|_2$, which measures the graph's worst-case amplification of perturbations across nodes. As $\|\mathbf{W}\|_2$ and M_i increases, local instabilities propagate and compound across models, and the maximum allowable synthetic-data weight λ_i decreases.

While this norm-based condition is convenient and analytically tractable, it relies on a global worst-case characterization of the interaction graph and can therefore be conservative. In particular, the spectral norm aggregates all interactions across the network into a single scalar quantity, which can obscure the specific topological features that contribute to instability. As a result, different interaction graphs may share the same value of $\|\mathbf{W}\|_2$ yet exhibit qualitatively different dynamical behaviors.

Impacts of graph topology. To move beyond global norm-based characterizations and identify which structural features of the interaction graph drive instability, we analyze how $\rho(\mathbf{D})$ is shaped by the graph's cycle structure.

Definition 3.9 (Directed cycle and cycle gain)**.** A *simple directed cycle* is an ordered sequence of distinct model nodes $C = (i_\ell \to \cdots \to i_2 \to i_1 \to i_\ell)$ with $i_r \in \mathcal{V}$ for all r. The *cycle gain* of C is defined as the geometric mean of the corresponding entries in $\mathbf{D}$:

$$\gamma(C) := \left(\prod_{r=1}^{\ell} D_{i_r i_{r+1}}\right)^{1/\ell} = \left(\prod_{r=1}^{\ell} \beta_{i_r} w_{i_r i_{r+1}}\right)^{1/\ell}$$

where $i_{\ell+1} = i_1$.

Intuitively, cycle gain $\gamma(C)$ captures per-iteration amplification accumulated along cycle C, where moving along each edge $i_{r+1} \to i_r$ contributes a factor $\beta_{i_r} w_{i_r i_{r+1}}$, and geometric mean reflects the average amplification per retraining step around the cycle.

Proposition 3.10 (Cycle gain as a lower bound of $\rho(\mathbf{D})$)**.** *Let $\mathbf{D} \geq 0$ be a nonnegative matrix and let $\{C_m\}_{m=1}^{M}$ denote the set of all simple directed cycles in the corresponding interaction graph. Then the spectral radius of $\mathbf{D}$ satisfies*

$$\rho(\mathbf{D}) \geq \max_{m \in [M]} \gamma(C_m)$$

Prop. 3.10 indicates that in a network with multiple cycles, $\rho(D)$ is primarily governed by the cycle with the largest geometric mean gain. In other words, a single "dominant cycle" can bottleneck the entire system, regardless of the total number of cycles or their lengths. Cycle length affects stability only through the geometric mean: if every model along a cycle contributes the same per-step gain (i.e., $\beta_i w_{ij} = b$), then the overall cycle gain $\gamma(C) = b$ is independent of the cycle's length. Combining this with Thm. 3.7 yields the following necessary condition on cycle gains; see Appx. D.6 for proofs.

Corollary 3.11 (Necessary condition on cycles)**.** *If $\rho(\mathbf{D}) < 1$, then every simple directed cycle C satisfies $\prod_{(j \to i) \in C} \beta_i w_{ij} < 1$. In particular, every self-loop must satisfy $\beta_i w_{ii} < 1$.*

4 Diversity Analysis

When p_i^{data} differ across models, it is natural to expect the resulting generative models to produce diverse outputs. A central question, however, is how such diversity evolves in the long run under cross-model interactions. In the following, we assume that models converge to a fixed point $\bar{\Theta}$ under the iterative retraining dynamics $\mathcal{G}$, and we study how model diversity manifests at $\bar{\Theta}$.

At any fixed point $\bar{\Theta}$, each model i is trained on a composite dataset drawn from the mixture distribution $\bar{p}_i \propto p_i^{\text{data}} + \lambda_i \sum_j w_{ij} p_{\bar{\boldsymbol{\theta}}_j}$. Because models may differ in architecture and parameter dimensionality, directly comparing parameters is not meaningful. We therefore measure diversity in a common representation space.

Definition 4.1 (Output representations)**.** Let $\phi : \mathcal{X} \to \mathbb{R}^D$ be a fixed 1-Lipschitz feature map[3]. For each model i, we define its *output representation* at a fixed point $\bar{\Theta}$ as $\bar{\boldsymbol{f}}_i := \mathbb{E}_{x \sim p_{\bar{\boldsymbol{\theta}}_i}}[\phi(x)]$. The *real-data representation centroid* for model i is defined as $\boldsymbol{f}_i^{\text{data}} := \mathbb{E}_{x \sim p_i^{\text{data}}}[\phi(x)]$.

Definition 4.2 (System diversity)**.** Define *system output diversity* $\mathcal{D}_{\text{out}}$ and *system real-data diversity* $\mathcal{D}_{\text{data}}$ as the empirical variances of $\{\bar{\boldsymbol{f}}_i\}_{i \in \mathcal{V}}$ and of $\{\boldsymbol{f}_i^{\text{data}}\}_{i \in \mathcal{V}}$ across the network, respectively:

$$\mathcal{D}_{\text{out}} := \text{Var}_{i \in \mathcal{V}}\left[\bar{\boldsymbol{f}}_i\right] = \tfrac{1}{|\mathcal{V}|} \textstyle\sum_{i \in \mathcal{V}} \|\bar{\boldsymbol{f}}_i - \bar{\boldsymbol{f}}_{\text{avg}}\|^2, \text{ with } \bar{\boldsymbol{f}}_{\text{avg}} := \tfrac{1}{|\mathcal{V}|} \textstyle\sum_{i \in \mathcal{V}} \bar{\boldsymbol{f}}_i$$

$$\mathcal{D}_{\text{data}} := \text{Var}_{i \in \mathcal{V}}\left[\boldsymbol{f}_i^{\text{data}}\right] = \tfrac{1}{|\mathcal{V}|} \textstyle\sum_{i \in \mathcal{V}} \|\boldsymbol{f}_i^{\text{data}} - \boldsymbol{f}_{\text{avg}}^{\text{data}}\|^2, \text{ with } \boldsymbol{f}_{\text{avg}}^{\text{data}} := \tfrac{1}{|\mathcal{V}|} \textstyle\sum_{i \in \mathcal{V}} \boldsymbol{f}_i^{\text{data}}$$

Definition 4.3 (Fixed-point feature discrepancy)**.** For each model $i \in \mathcal{V}$, define the *fixed-point feature discrepancy* as $\bar{\boldsymbol{r}}_i := \bar{\boldsymbol{f}}_i - \mathbb{E}_{x \sim \bar{p}_i}[\phi(x)] \in \mathbb{R}^D$, which measures the mismatch between the output representation of model i and the representation of its training mixture distribution $\bar{p}_i$.

A nonzero $\bar{\boldsymbol{r}}_i$ arises when the model cannot exactly realize the mixture distribution, due to limited model expressivity, approximation error, or imperfect optimization during training.

With these definitions, we can characterize the fixed-point output representations $\{\bar{\boldsymbol{f}}_i\}_i$ as the solution of a graph-structured linear system, which sets the stage for analyzing system diversity.

[3] ϕ could be the identity function or a mapping that captures semantic or perceptual properties of the data.

Proposition 4.4 (Fixed-point output representation of the system)**.** *Let* $\bar{\mathbf{F}} = [\bar{\boldsymbol{f}}_1, \ldots, \bar{\boldsymbol{f}}_K]^\top$ *and* $\mathbf{F}_{\text{data}} = [\boldsymbol{f}_1^{\text{data}}, \ldots, \boldsymbol{f}_K^{\text{data}}]^\top$ *denote the fixed-point output representations and the real-data representations of all models, respectively. Let* $\boldsymbol{\Lambda} = \text{diag}(\lambda_1, \ldots, \lambda_K)$ *and* $\mathbf{L} := \mathbf{I} - \mathbf{W}$*. We have*

$$\bar{\mathbf{F}} = (\mathbf{I} + \boldsymbol{\Lambda}\mathbf{L})^{-1}\big[\mathbf{F}_{\text{data}} + (\mathbf{I} + \boldsymbol{\Lambda})\bar{\mathbf{R}}\big], \tag{6}$$

where $\bar{\mathbf{R}} := [\bar{\boldsymbol{r}}_1, \ldots, \bar{\boldsymbol{r}}_K]^\top$ *and* $\bar{\boldsymbol{r}}_i$ *satisfies* $\|\bar{\boldsymbol{r}}_i\| \le \frac{1+2\lambda_i}{1+\lambda_i}\,\varepsilon_i$*.*

The proof is given in Appx. D.7. With this solution, we can analyze how the system's output diversity at the fixed point deviates from the real-data diversity, and the following theorem provides insight into how iterative self-consuming training dynamics shape system diversity.

Theorem 4.5 (Upper bound on system output diversity)**.** *Under Prop. 4.4, the system output diversity at any fixed point* $\bar{\Theta}$ *satisfies*

$$\mathcal{D}_{\text{out}} \le 2\big\|(\mathbf{I} + \boldsymbol{\Lambda}\mathbf{L})^{-1}\big\|_2^2\left[\mathcal{D}_{\text{data}} + \frac{1}{|\mathcal{V}|}\sum_{i\in\mathcal{V}}(1+\lambda_i)^2\,\|\bar{\boldsymbol{r}}_i\|^2\right]. \tag{7}$$

Eq. (6) shows that the output representations at the fixed point are obtained by applying a *linear smoothing* operator $(\mathbf{I} + \Lambda\mathbf{L})^{-1}$ to real-data centroids. The diversity bound in Thm. 4.5 (see proof in Appx. D.8) quantifies the worst-case attenuation of heterogeneity along the non-consensus subspace, i.e., the subspace orthogonal to the all-ones vector where models disagree. The factor $\|(\mathbf{I}+\Lambda\mathbf{L})^{-1}\|_2^2$ acts as a *collapse coefficient*, mapping the initial data diversity $\mathcal{D}_{\text{data}}$ to the final output diversity $\mathcal{D}_{\text{out}}$. This bound highlights three key insights:

- *No cross-model interaction preserves diversity.* If there is no synthetic data ($\Lambda = \mathbf{0}$), or interaction graph has no cross edges ($w_{ij} = 0, \forall i \neq j$), then operator $(\mathbf{I} + \Lambda\mathbf{L})^{-1}$ does not average across models. Moreover, when each model's output matches its training mixture in feature space ($\bar{\boldsymbol{r}}_i = \mathbf{0}$ for all i), the bound reduces to $\mathcal{D}_{\text{out}} = \mathcal{D}_{\text{data}}$.
- *Stronger synthetic-data mixing has competing effects.* Larger λ_i increases the averaging effect of $(\mathbf{I} + \Lambda\mathbf{L})^{-1}$, pulling model representations toward a common value and reducing differences. However, larger λ_i also amplifies the term $(1+\lambda_i)^2\,\|\bar{\boldsymbol{r}}_i\|^2$, since each model relies more heavily on imperfect synthetic training data.
- *Graph topology shapes which disagreements are reduced.* The operator $(\mathbf{I} + \Lambda\mathbf{L})^{-1}$ averages out disagreements that change rapidly across strongly connected regions of the graph, while differences across weakly connected regions are diminished much less.
- *Imperfect representation introduces a topology-independent diversity term.* The discrepancy term $\frac{1}{|\mathcal{V}|}\sum_i(1+\lambda_i)^2\|\bar{\boldsymbol{r}}_i\|^2$ captures, in the feature space, the gap between each model's fixed-point output and its training mixture; it contributes to $\mathcal{D}_{\text{out}}$ regardless of $\mathbf{W}$. Such diversity reflects model error rather than meaningful inter-model disagreement.

Special Case: undirected graph. While Thm. 4.5 provides a general bound on diversity collapse, it is expressed in terms of the global operator norm $\big\|(\mathbf{I}+\Lambda\mathbf{L})^{-1}\big\|_2^2$, which obscures how specific structural features of the interaction graph affect diversity. To gain more interpretable, topology-dependent insights, we consider the instructive case of symmetric (i.e., undirected) interactions.

Corollary 4.6 (Diversity Bound for Symmetric Interaction)**.** *Suppose the interaction graph is undirected (*$\mathbf{W} = \mathbf{W}^\top$*), the synthetic weight is homogeneous (*$\lambda_i \equiv \lambda$*) and the feature discrepancy vanishes (*$\bar{\boldsymbol{r}}_i = \mathbf{0}$*). Let* μ_2 *denote the algebraic connectivity (the second smallest eigenvalue of* $\mathbf{L}$*). Then:*

$$\mathcal{D}_{\text{out}} \le \frac{1}{(1+\lambda\mu_2)^2}\mathcal{D}_{\text{data}} \tag{8}$$

Cor. 4.6 provides a clean, topology-dependent characterization of diversity collapse, explicitly linking it to the algebraic connectivity μ_2. Intuitively, μ_2 captures the "bottleneckedness" of the network: a small μ_2 indicates weak connections or community structures, whereas a large μ_2 corresponds to strong connectivity with few bottlenecks. The corollary implies that higher connectivity (larger μ_2) leads to a smaller upper bound and thus stronger collapse of diversity. Conversely, even in a densely connected network, a small μ_2 (e.g., when communities are only weakly connected) can significantly slow down diversity collapse.

5 Experiments

Our theoretical analysis is under simplifying assumptions and abstracts away sources of error arising from finite data and imperfect optimization. To understand how self-consuming retraining operates beyond these idealized settings, we complement our theory with experiments, evaluating whether the qualitative behaviors predicted by the theory persist under realistic training conditions.

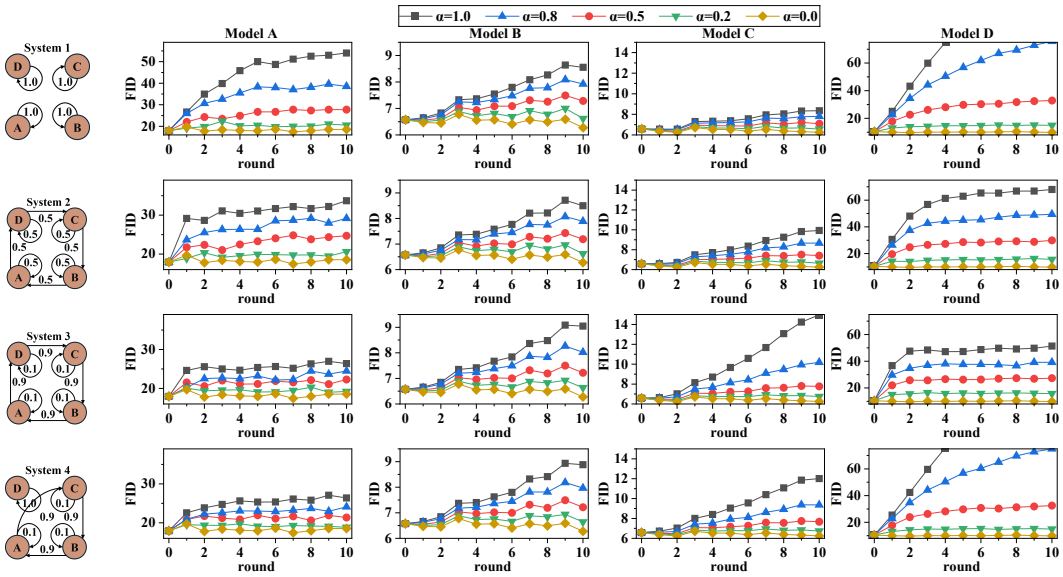

Figure 2: Stability experiments when all models trained on full CIFAR-10. Each row corresponds to a different interaction graph. For each row, the left figure illustrates the direction and strength of synthetic data flow, while the right figures report the FID between generated samples and real data for each model. Results are shown under different synthetic mixing strengths, with $\alpha_i = \alpha$ shared across models.

Datasets. We conduct experiments on two datasets: (i) **Synthetic Gaussian**: A dataset following 8-mode Gaussian mixture model, the details are in Appx. F. (ii) **CIFAR-10** [23]: It contains 60,000 images from 10 classes $\{\text{airplane} := 0, \text{automobile} := 1, \text{bird} := 2, \text{cat} := 3, \text{deer} := 4, \text{dog} := 5, \text{frog} := 6, \text{horse} := 7, \text{ship} := 8, \text{truck} := 9\}$.

Models. For real-data experiments, we consider two representative classes of generative models: denoising diffusion probabilistic models (DDPM) [20] and continuous flow matching models (CFM) [33]. We instantiate four generative models from two paradigms, with two variants per paradigm as shown in Table 1. All models are adopted from public GitHub repositories and are used with their default hyperparameters. Each model is initialized from its released pretrained checkpoint, keeping architectures and optimization settings unchanged. For the synthetic experiments, we additionally include a GAN-based [18] model; details are in Appx. F.

Table 1: Models and real-data assignments. The same data for all models is used for stability; the heterogeneous assignment is used for diversity experiments.

Model	Model Type	Stability	Diversity
A	VLB Diffusion	$\{0, \cdots, 9\}$	$\{0, \cdots, 9\}$
B	OT-CFM		$\{3, 4\}$
C	iCFM		$\{5, 6\}$
D	Hybrid Diffusion		$\{7, 8, 9\}$

Iterative Retraining. At each iteration, each model generates $10,000$ samples and is fine-tuned for a fixed number of steps on its mixed dataset of size $10,000$. This dataset consists of real samples and synthetic samples generated by other models. The contribution of synthetic data is governed by the interaction matrix $\mathbf{W}$ and the synthetic mixing ratio $\alpha_i = \frac{\lambda_i}{\lambda_i + 1}$. The procedure is in Alg. 1.

Interaction Graphs. We design four interaction graphs (Fig. 2, left) to isolate the effects of graph structure, edge weights, and cycle configuration through controlled pairwise comparisons. **System 1** is an isolated-retraining baseline. **Systems 1** and **2** share the same spectral radius $\|\mathbf{W}\| = 1$ but

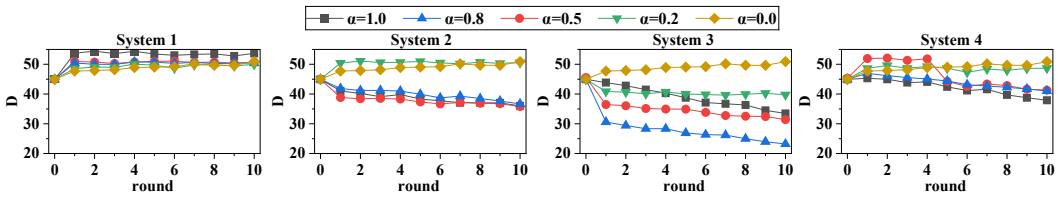

Figure 3: Diversity across models under heterogeneous data with different interaction graphs. *Systems* are defined in Fig. 2. Each figure reports the evolution of diversity $\mathcal{D}$ across retraining rounds under different synthetic mixing ratios $\alpha_i = \alpha$ shared across models.

differ in graph structure. **Systems 2** and **3** share the same graph structure but differ in edge weights. **System 4** has a shorter directed cycles. Additional graphs are explored in Appx. F.1.

Evaluation Metrics. To assess distributional quality, we mainly compute the FID [26] between generated samples and real data using standard Inception feature embeddings. As a supplement, Appx. G reports precision and recall. To quantify diversity across models, we use 512-dimensional features from an ImageNet-pretrained [10] ResNet-18 [19] with the final classification layer removed. System diversity $\mathcal{D}$ is computed following Def. 4.2. Since a single run of iterative retraining is computationally expensive, we introduce stochasticity across rounds by varying the random seed. Specifically, in round r for model i, we use a seed of the form $\text{seed} + 10{,}000 \cdot r + 97 \cdot i$. For synthetic data experiments, all reported results are averaged over three independent runs with different seeds. We present the results of CIFAR-10 data here and the results of Gaussian data are in Appx. F.

Stability experiments. We first verify the stability analysis, training all models on the full CIFAR-10 ($p_i^{\text{data}} \equiv p^{\text{data}}$). The synthetic-data mixing ratio is shared across models ($\alpha_i \equiv \alpha$). Fig. 2 illustrates several interaction graph structures with varying patterns of synthetic data flow; additional graph configurations are reported in Appx. F.1. **System 1** corresponds to isolated retraining as a baseline. Across all systems, increasing α leads to higher FID, indicating degraded distributional quality. The four models show different sensitivities, reflecting architecture and training dynamics rather than data access. Crucially, neither the interaction norm $\|\mathbf{W}\|$ nor graph structure alone determines stability: **Systems 1** and **2** share $\|\mathbf{W}\|$ but exhibit different FID trends, while **Systems 2** and **3** share graph structure but differ in interaction strength. This is consistent with our spectral-radius criterion in Thm. 3.7, which combines both ingredients. Comparing **Systems 3** and **4**, shorter interaction cycles can mitigate degradation when real data are present. This *contrasts* with observations in the synthetic settings (Appx. F), where short cycles may instead accelerate instability, indicating that instability depends on which models participate in dominant feedback paths rather than cycle length itself. Additional synthetic-data experiments appear in Appx. F.1; precision and recall are in Appx. G.

Diversity experiments. We next verify the diversity analysis as shown in Fig. 3, training models on disjoint or partially overlapping subsets of CIFAR-10 (Table 1); this captures the practical scenario where different model providers have access to different data sources. Overall, increasing α and using denser interaction graphs both reduce diversity, consistent with stronger synthetic-data reliance encouraging alignment. **System 1** (isolated retraining) is an exception: its diversity is driven by model instability rather than meaningful differentiation, and does not decrease with α. **Systems 2** and **3** have similar topology but different interaction weights, where **System 3** exhibits a faster and more pronounced decline in $\mathcal{D}$, indicating that interaction strength matters beyond just the presence of α. **Systems 2** and **3** also feature denser interactions than **System 4** and exhibit faster diversity decay across most $\alpha > 0$, aligning with our prediction (Thm. 4.5, Cor. 4.6) that higher graph connectivity accelerates homogenization. Appx. F.2 reports synthetic experiments with similar qualitative behavior. Note that increases in $\mathcal{D}$ may also result from instability, which is difficult to quantify since the fixed point is generally unknown; sample visualizations in Appx. G aid interpretation.

6 Conclusion

This work takes a first step toward studying networked self-consuming generative models through a graph-based dynamical framework, revealing how synthetic-data mixing and interaction topology jointly govern system stability and diversity. Our results show that cross-model feedback can induce topology-dependent amplification and diversity contraction, offering a principled understanding of the long-term behavior of multi-model generative ecosystems.

Acknowledgments and Disclosure of Funding

This material is based upon work supported by the U.S. National Science Foundation under Grant Nos. IIS-2202699, IIS-2416895, IIS-2301599, CMMI-2301601, DMS-2529302, and OIA-2616279. The authors declare no competing interests.

References

[1] Marah Abdin, Jyoti Aneja, Hany Awadalla, Ahmed Awadallah, Ammar Ahmad Awan, Nguyen Bach, Amit Bahree, Arash Bakhtiari, Jianmin Bao, Harkirat Behl, Alon Benhaim, Misha Bilenko, Johan Bjorck, Sébastien Bubeck, Martin Cai, Qin Cai, Vishrav Chaudhary, Dong Chen, Dongdong Chen, Weizhu Chen, Yen-Chun Chen, Yi-Ling Chen, Hao Cheng, Parul Chopra, Xiyang Dai, Matthew Dixon, Ronen Eldan, Victor Fragoso, Jianfeng Gao, Mei Gao, Min Gao, Amit Garg, Allie Del Giorno, Abhishek Goswami, Suriya Gunasekar, Emman Haider, Junheng Hao, Russell J. Hewett, Wenxiang Hu, Jamie Huynh, Dan Iter, Sam Ade Jacobs, Mojan Javaheripi, Xin Jin, Nikos Karampatziakis, Piero Kauffmann, Mahoud Khademi, Dongwoo Kim, Young Jin Kim, Lev Kurilenko, James R. Lee, Yin Tat Lee, Yuanzhi Li, Yunsheng Li, Chen Liang, Lars Liden, Xihui Lin, Zeqi Lin, Ce Liu, Liyuan Liu, Mengchen Liu, Weishung Liu, Xiaodong Liu, Chong Luo, Piyush Madan, Ali Mahmoudzadeh, David Majercak, Matt Mazzola, Caio César Teodoro Mendes, Arindam Mitra, Hardik Modi, Anh Nguyen, Brandon Norick, Barun Patra, Daniel Perez-Becker, Thomas Portet, Reid Pryzant, Heyang Qin, Marko Radmilac, Liliang Ren, Gustavo de Rosa, Corby Rosset, Sambudha Roy, Olatunji Ruwase, Olli Saarikivi, Amin Saied, Adil Salim, Michael Santacroce, Shital Shah, Ning Shang, Hiteshi Sharma, Yelong Shen, Swadheen Shukla, Xia Song, Masahiro Tanaka, Andrea Tupini, Praneetha Vaddamanu, Chunyu Wang, Guanhua Wang, Lijuan Wang, Shuohang Wang, Xin Wang, Yu Wang, Rachel Ward, Wen Wen, Philipp Witte, Haiping Wu, Xiaoxia Wu, Michael Wyatt, Bin Xiao, Can Xu, Jiahang Xu, Weijian Xu, Jilong Xue, Sonali Yadav, Fan Yang, Jianwei Yang, Yifan Yang, Ziyi Yang, Donghan Yu, Lu Yuan, Chenruidong Zhang, Cyril Zhang, Jianwen Zhang, Li Lyna Zhang, Yi Zhang, Yue Zhang, Yunan Zhang, and Xiren Zhou. Phi-3 technical report: A highly capable language model locally on your phone, 2024. URL https://arxiv.org/abs/2404.14219.

[2] Sina Alemohammad, Josue Casco-Rodriguez, Lorenzo Luzi, Ahmed Imtiaz Humayun, Hossein Babaei, Daniel LeJeune, Ali Siahkoohi, and Richard Baraniuk. Self-consuming generative models go MAD. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=ShjMHfmPs0.

[3] Sina Alemohammad, Ahmed Imtiaz Humayun, Shruti Agarwal, John Collomosse, and Richard Baraniuk. Self-improving diffusion models with synthetic data, 2024. URL https://arxiv.org/abs/2408.16333.

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

[5] James Betker, Gabriel Goh, Li Jing, Tim Brooks, Jianfeng Wang, Linjie Li, Long Ouyang, Juntang Zhuang, Joyce Lee, Yufei Guo, Wesam Manassra, Dhariwal Prafulla, Aditya Ramesh, Tyna Eloundou, Alec Glaese, and Sandhini Bhatia. Improving image generation with better captions, 2023. URL https://cdn.openai.com/papers/dall-e-3.pdf. OpenAI Technical Report.

[6] Rajendra Bhatia. *Matrix Analysis*, volume 169. Springer, New York, NY, 1997. doi: 10.1007/978-1-4612-0653-8.

[7] Zhongteng Cai, Yaxuan Wang, Yang Liu, and Xueru Zhang. Stabilizing self-consuming diffusion models with latent space filtering, 2025. URL https://arxiv.org/abs/2511.12742.

[8] Junsong Chen, Jincheng YU, Chongjian GE, Lewei Yao, Enze Xie, Zhongdao Wang, James Kwok, Ping Luo, Huchuan Lu, and Zhenguo Li. Pixart-α: Fast training of diffusion transformer for photorealistic text-to-image synthesis. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=eAKmQPe3m1.

[9] Krzysztof Ciesielski. On Stefan Banach and some of his results. *Banach Journal of Mathematical Analysis*, 1(1):1 – 10, 2007. doi: 10.15352/bjma/1240321550. URL https://doi.org/10.15352/bjma/1240321550.

[10] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. ImageNet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255, 2009. doi: 10.1109/CVPR.2009.5206848.

[11] Elvis Dohmatob, Yunzhen Feng, Pu Yang, Francois Charton, and Julia Kempe. A tale of tails: Model collapse as a change of scaling laws. In *Forty-first International Conference on Machine Learning*, 2024. URL https://openreview.net/forum?id=KVvku47shW.

[12] Yunzhen Feng, Elvis Dohmatob, Pu Yang, Francois Charton, and Julia Kempe. Beyond model collapse: Scaling up with synthesized data requires verification, 2024. URL https://arxiv.org/abs/2406.07515.

[13] Damien Ferbach, Quentin Bertrand, Avishek Joey Bose, and Gauthier Gidel. Self-consuming generative models with curated data provably optimize human preferences, 2024. URL https://arxiv.org/abs/2407.09499.

[14] Shi Fu, Yingjie Wang, Yuzhu Chen, Xinmei Tian, and Dacheng Tao. A theoretical perspective: How to prevent model collapse in self-consuming training loops. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=WttfQGwpES.

[15] Weiguo Gao and Ming Li. Convergence dynamics and stabilization strategies of co-evolving generative models, 2025. URL https://arxiv.org/abs/2503.08117.

[16] Matthias Gerstgrasser, Rylan Schaeffer, Apratim Dey, Rafael Rafailov, Tomasz Korbak, Henry Sleight, Rajashree Agrawal, John Hughes, Dhruv Bhandarkar Pai, Andrey Gromov, Dan Roberts, Diyi Yang, David L. Donoho, and Sanmi Koyejo. Is model collapse inevitable? Breaking the curse of recursion by accumulating real and synthetic data. In *First Conference on Language Modeling*, 2024. URL https://openreview.net/forum?id=5B2K4LRgmz.

[17] Nate Gillman, Michael Freeman, Daksh Aggarwal, Chia-Hong Hsu, Calvin Luo, Yonglong Tian, and Chen Sun. Self-correcting self-consuming loops for generative model training, 2024. URL https://arxiv.org/abs/2402.07087.

[18] Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Proceedings of the 28th International Conference on Neural Information Processing Systems - Volume 2*, NIPS'14, page 2672–2680, Cambridge, MA, USA, 2014. MIT Press.

[19] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2016. doi: 10.1109/CVPR.2016.90.

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

[21] Zizhao Hu, Mohammad Rostami, and Jesse Thomason. Multimodal synthetic data finetuning and model collapse: Insights from VLMs and diffusion models. In *Proceedings of the 27th International Conference on Multimodal Interaction*, ICMI '25, page 588–599, New York, NY, USA, 2025. Association for Computing Machinery. ISBN 9798400714993. doi: 10.1145/3716553.3750806. URL https://doi.org/10.1145/3716553.3750806.

[22] Grgur Kovač, Jérémy Perez, Rémy Portelas, Peter Ford Dominey, and Pierre-Yves Oudeyer. Recursive training loops in LLMs: How training data properties modulate distribution shift in generated data? In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 32278–32297, Suzhou, China, November 2025. Association for Computational Linguistics. URL https://aclanthology.org/2025.emnlp-main.1643/.

[23] Alex Krizhevsky. Learning multiple layers of features from tiny images. 2009. URL https://api.semanticscholar.org/CorpusID:18268744.

[24] Serge Lang. *Fundamentals of differential geometry*, volume 191. Springer Science & Business Media, 1999.

[25] Subhabrata Mukherjee, Arindam Mitra, Ganesh Jawahar, Sahaj Agarwal, Hamid Palangi, and Ahmed Awadallah. Orca: Progressive learning from complex explanation traces of GPT-4, 2023. URL https://arxiv.org/abs/2306.02707.

[26] Anton Obukhov, Maximilian Seitzer, Po-Wei Wu, Semen Zhydenko, Jonathan Kyl, and Elvis Yu-Jing Lin. High-fidelity performance metrics for generative models in pytorch, 2020. URL https://github.com/toshas/torch-fidelity. Version: 0.3.0, DOI: 10.5281/zenodo.4957738.

[27] Mohamed El Amine Seddik, Suei-Wen Chen, Soufiane Hayou, Pierre Youssef, and Merouane Abdelkader DEBBAH. How bad is training on synthetic data? A statistical analysis of language model collapse. In *First Conference on Language Modeling*, 2024. URL https://openreview.net/forum?id=t3z6UlV09o.

[28] Ilia Shumailov, Zakhar Shumaylov, Yiren Zhao, Yarin Gal, Nicolas Papernot, and Ross Anderson. The curse of recursion: Training on generated data makes models forget, 2024. URL https://arxiv.org/abs/2305.17493.

[29] Ilia Shumailov, Zakhar Shumaylov, Yiren Zhao, Nicolas Papernot, Ross Anderson, and Yarin Gal. AI models collapse when trained on recursively generated data. *Nature*, 631(8022):755–759, 2024.

[30] Zhen Sun, Zongmin Zhang, Xinyue Shen, Ziyi Zhang, Yule Liu, Michael Backes, Yang Zhang, and Xinlei He. Are we in the AI-generated text world already? Quantifying and monitoring AIGT on social media. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 22975–23005, July 2025. URL https://aclanthology.org/2025.acl-long.1120/.

[31] Ananda Theertha Suresh, Andrew Thangaraj, and Aditya Nanda Kishore Khandavally. Rate of model collapse in recursive training, 2024. URL https://arxiv.org/abs/2412.17646.

[32] Rohan Taori and Tatsunori B. Hashimoto. Data feedback loops: Model-driven amplification of dataset biases. In *Proceedings of the 40th International Conference on Machine Learning*, ICML'23. JMLR.org, 2023.

[33] Alexander Tong, Kilian FATRAS, Nikolay Malkin, Guillaume Huguet, Yanlei Zhang, Jarrid Rector-Brooks, Guy Wolf, and Yoshua Bengio. Improving and generalizing flow-based generative models with minibatch optimal transport. *Transactions on Machine Learning Research*, 2024. ISSN 2835-8856. URL https://openreview.net/forum?id=CD9Snc73AW. Expert Certification.

[34] Cédric Villani. Optimal transport: Old and new. 2008. doi: https://doi.org/10.1007/978-3-540-71050-9.

[35] Cédric Villani. The Wasserstein distances. In *Optimal Transport: Old and New*, pages 93–111. Springer Berlin Heidelberg, Berlin, Heidelberg, 2009.

[36] Hung Anh Vu, Galen Reeves, and Emily Wenger. What happens when generative AI models train recursively on each others' outputs?, 2025. URL https://arxiv.org/abs/2505.21677.

[37] Maorong Wang, Nicolas Michel, Jiafeng Mao, and Toshihiko Yamasaki. Dealing with synthetic data contamination in online continual learning. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024. URL https://openreview.net/forum?id=Lc8gemv97Y.

[38] Yaxuan Wang, Zhongteng Cai, Yujia Bao, Xueru Zhang, and Yang Liu. Observations and remedies for large language model bias in self-consuming performative loop. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 33862–33882, 2026.

[39] Xiukun Wei and Xueru Zhang. Self-consuming generative models with adversarially curated data, 2025. URL https://arxiv.org/abs/2505.09768.

[40] Xiukun Wei, Min Shi, and Xueru Zhang. Market games for generative models: Equilibria, welfare, and strategic entry. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=T7c5LXuebY.

[41] Xiukun Wei, Tian Xie, Ding Zhu, and Xueru Zhang. Self-consuming generative models with co-evolving human preferences. In *The Fortieth Annual Conference on Neural Information Processing Systems*, 2026. URL https://openreview.net/forum?id=tsXMIKjzrO.

[42] Sierra Wyllie, Ilia Shumailov, and Nicolas Papernot. Fairness feedback loops: Training on synthetic data amplifies bias. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency*, FAccT'24, page 2113–2147, New York, NY, USA, 2024. Association for Computing Machinery. URL https://doi.org/10.1145/3630106.3659029.

[43] Tian Xie and Xueru Zhang. Automating data annotation under strategic human agents: Risks and potential solutions. *Advances in Neural Information Processing Systems*, 37:127436–127482, 2024.

[44] Bingji Yi, Qiyuan Liu, Yuwei Cheng, and Haifeng Xu. Escaping model collapse via synthetic data verification: Near-term improvements and long-term convergence, 2025. URL https://arxiv.org/abs/2510.16657.

[45] Yang Zhang, Xiukun Wei, and Xueru Zhang. When and how human curation backfires: Preference alignment under multi-model self-consuming loop, 2026. URL https://arxiv.org/abs/2605.29267.

Appendix Contents

A Limitations and Future Directions **14**

B Broader Impact **14**

C Related Work **15**

C.1 Single-model Self-Consuming Loop 15

C.2 Multi-Model Interactive Ecosystems 15

C.3 Mitigation Strategies for Model Collapse 16

D Proofs **17**

D.1 Proofs of Characterization of Fixed Points 17

D.2 Proof of the Uncoupled Solution as a Fixed Point 17

D.3 Example: Heterogeneous Fixed Point Different from the Uncoupled Solution 18

D.4 Proofs of Theorem 3.7 19

D.5 Proofs of Corollary 3.8 21

D.6 Proofs of Proposition 3.10 and Corollary 3.11 22

D.7 Proofs of Proposition 4.4 23

D.8 Proofs of Theorem 4.5 and Corollary 4.6 25

E Closed-Form Spectral Validation of the Networked Dynamics **27**

F Experiments with Synthetic Gaussian Data **29**

F.1 Stability Experimental Results 31

F.2 Diversity Experimental Results 31

G Additional Results on CIFAR-10 Datasets **37**

A Limitations and Future Directions

This work provides a first theoretical step toward understanding networked self-consuming generative ecosystems. Our main results take a local stability perspective around an arbitrary fixed point whenever such a fixed point exists. While this view is standard in dynamical-systems analysis, a natural future direction is to establish broader global existence and uniqueness guarantees for heterogeneous, nonconvex retraining systems.

Our analysis builds on the standard local regularity assumptions adopted in prior single-model self-consuming [4, 13] analyses, and extends this style of local stability reasoning to networked retraining dynamics. These assumptions make the coupled system analytically tractable, but practical deep generative models involve finite-sample data, stochastic optimization, and nonconvex objectives. Extending the theory to approximate updates and weaker regularity conditions is an important direction for further bridging theory and large-scale generative-model training.

Finally, our diversity result gives an upper bound that reveals a graph-smoothing effect at equilibrium. Future work could derive sharper diversity characterizations, separate meaningful specialization from instability-induced diversity, and study adaptive or time-varying interaction graphs. The framework also suggests mitigation strategies, such as controlling high-gain feedback cycles or allocating real data to nodes with large local amplification.

B Broader Impact

This work develops a theoretical framework for understanding stability and diversity in networked self-consuming generative ecosystems. The framework maynhelp model providers and platforms diagnose long-term risks such as distributional drift, instability, homogenization, and loss of diversity when synthetic data is reused across models. It may also inform mitigation strategies, such as limiting high-gain feedback cycles, preserving data diversity, or allocating real data to parts of the system that are most vulnerable to amplification.

At the same time, understanding how graph structure and synthetic-data mixing shape long-term model behavior could be misused to deliberately steer model ecosystems toward narrower or biased output distributions. For this reason, practical use of such analyses should be paired with transparent data tracing, auditing of synthetic-data flows, and careful evaluation of downstream effects on different user groups and content domains.

C Related Work

C.1 Single-model Self-Consuming Loop

High-quality synthetic data has been widely used for model pretraining, fine-tuning, and data augmentation. However, when models are repeatedly trained on their own or other models' generated outputs, the training distribution may drift away from the true data distribution.

Through extensive experimentation, early studies such as [2, 16, 29] demonstrated that this retraining loop leads to *model collapse*, a degenerative process where models progressively lose their ability to represent the underlying data distribution. Earlier work by Taori and Hashimoto [32] showed that feedback loops in model-generated data can amplify existing dataset biases, laying the foundation for understanding recursive error reinforcement. Specifically, Shumailov et al. [29] identified that statistical approximation errors accumulate across generations, causing the model to lose information about the distribution's tails. Simultaneously, Alemohammad et al. [2] and Gerstgrasser et al. [16] observed that without sufficient fresh real data, the loop leads to a complete divergence from the ground truth. Additional large-scale experiments by Kovač et al. [22] highlighted how properties of the training data, such as redundancy, diversity, and prompt type, modulate the degree of distribution shift in recursive LLM training.

Building on these empirical findings, subsequent work focused on theoretically characterizing the conditions under which collapse occurs. Bertrand et al. [4] provided conditions for stability, emphasizing the importance of retaining a non-negligible fraction of real data. Similarly, Fu et al. [14] modeled self-consuming loops as dynamical systems and showed that outputs lose variance and converge to trivial fixed points unless synthetic data is carefully scheduled. From a different angle, Dohmatob et al. [11] analyzed how recursive training alters scaling laws, leading to diminished tail coverage and faster error decay on rare examples. Complementary to this, Seddik et al. [27] provided a statistical treatment of model collapse, revealing monotonic increases in KL divergence and sharp variance decay during recursive fine-tuning. Suresh et al. [31] further modeled the rate at which collapse emerges, quantifying error accumulation speed under recursive sampling and establishing bounds for early-stage divergence.

These works reveal how recursive training can cause models to drift, lose diversity, and collapse. They offer strong theoretical and empirical support. However, most of this work focuses on single-model behavior. Our framework subsumes these single-model settings as a special case. In particular, when $w_{ii} = 1$ and $w_{ij} = 0$ for all $i \neq j$, each model retrains exclusively on its own generated data, recovering the single-model self-consuming dynamics analyzed.

C.2 Multi-Model Interactive Ecosystems

Sun et al. [30] found that a large and growing fraction of online content is now generated by AI. This implies that training data used by one model often includes outputs from others, making cross-model data consumption common in practice. Motivated by this, recent work has begun to explore multi-model generative ecosystems, where models are explicitly trained on each other's outputs.

Gao and Li [15] studied the stability of a co-evolving system involving one image model and one text model. Through simulations, they showed that when models learn from each other, small distribution shifts can compound and lead to joint instability. Zhang et al. [45] further studied interacting self-consuming models under human curation, showing that cross-model interactions can substantially alter the long-term effect of preference-based curation. However, their analysis is limited to a two-model setup with fixed data domains, and may not generalize to larger or more diverse systems. Vu et al. [36] approached the problem from a theoretical angle. They analyzed how one model behaves when trained on data that mixes outputs from multiple other models. Their theoretical results rely on simplifying assumptions, such as linearized dynamics assumptions. Their experiments also focus on just two interacting models, limiting the scope of their findings. Hu et al. [21] studied repeated fine-tuning of multi-modal generative systems composed of two coupled models using self-generated synthetic data. They showed that such closed feedback loops can lead to performance degradation over time, despite occasional short-term improvements in task metrics. Wei and Zhang [39] studied a setup where two models curate data for each other with adversarial intent. Their work focuses on how harmful interactions between models affect training dynamics.

While these works [15, 21, 36, 40, 45] provide early insights into multi-model collapse, they are subject to several constraints. Most existing studies focus on small model pairs. The systems considered are relatively narrow in scope and not designed to represent general-purpose model ecosystems. In addition, many works assume predefined interaction structures rather than allowing interactions to emerge naturally.

In contrast, our work presents a more general framework that captures real-world interaction patterns. We explicitly model the flow of synthetic data across models and analyze the resulting network from a graph-theoretic perspective. This allows us to identify key factors that affect system stability, such as synthetic data

ratio, topology density, and model heterogeneity. We further validate our analysis through simulations involving multiple models and diverse interaction structures.

C.3 Mitigation Strategies for Model Collapse

Model collapse is driven by loss of diversity and distributional drift. To counter this, recent research has explored several directions.

One widely supported strategy is to keep real data in the training loop. Alemohammad et al. [2] and Bertrand et al. [4] show that even a small reserve of real data can stabilize learning. Instead of discarding past data, cumulative synthetic samples without replacing original data have proven more robust [16, 36].

Another approach introduces external selection. Ferbach et al. [13] and Wei et al. [41] provide a theoretical analysis of human-curated training loops. They show that when models are retrained on their own outputs filtered by human feedback or a reward model, the process implicitly optimizes for those preferences. This curation mechanism can prevent collapse and guide models toward desirable behavior over time. Building on this, Feng et al. [12] emphasize that scaling up with synthetic data requires explicit verification mechanisms. They argue that without such checks, recursive training amplifies model biases and misalignment. Yi et al. [44] propose using a verifier model to score generated samples, admitting only high-quality outputs into training.

Some defenses are built into the model loop itself. Alemohammad et al. [3] proposed a method called SIMS for diffusion models. It adds a negative guidance signal during generation to reduce drift and improve stability. Gillman et al. [17] designed a training loop where generated samples are first corrected using expert rules before being reused. This helps keep the synthetic data high-quality and reduces the chance of reinforcing errors. Cai et al. [7] introduce latent space filtering to diffusion models. By removing low-entropy samples during retraining, their method preserves diversity and delays degradation.

Consistent with prior findings, our study also highlights the importance of real data during retraining. We observe that real data provides a stabilizing signal against distributional drift. Moreover, several existing mitigation strategies discussed above are conceptually orthogonal to our approach and could be incorporated into our framework.

D Proofs

Throughout this appendix, we work under Assumptions. For model i, recall that $\varepsilon_i := \max_{j:w_{ij}>0} d_{\mathcal{W}}\big(p_{\bar{\boldsymbol{\theta}}_j}, p_i^{\text{data}}\big)$. The local objective of model i is denoted

$$\mathcal{H}_i(\boldsymbol{\theta}; \Theta) := \mathbb{E}_{x\sim p_i^{\text{data}}}[\log p_{\boldsymbol{\theta}}(x)] + \lambda_i \sum_{j=1}^{K} w_{ij}\, \mathbb{E}_{x\sim p_{\boldsymbol{\theta}_j}}[\log p_{\boldsymbol{\theta}}(x)],$$

so that $\mathcal{G}_i(\Theta) \in \arg\max_{\boldsymbol{\theta}\in\Theta_i} \mathcal{H}_i(\boldsymbol{\theta}; \Theta)$.

D.1 Proofs of Characterization of Fixed Points

Lemma D.1 (The local maximum likelihood solution is unique)**.** *Under assumptions and $\alpha_i > L_i\varepsilon_i$, for each model i there exists a neighborhood $\mathcal{U}$ of $\bar{\Theta}$ such that for any $\Theta \in \mathcal{U}$, the local objective $\boldsymbol{\theta}_i' \mapsto H_i(\boldsymbol{\theta}_i'; \Theta)$ admits a unique local maximizer on U_i. Consequently, $\mathcal{G}_i$ is single-valued on $\mathcal{U}$.*

Proof. Fix a model i. The Hessian of H_i with respect to the optimizing variable $\boldsymbol{\theta}_i'$ is

$$\nabla^2_{\boldsymbol{\theta}_i'} H_i(\boldsymbol{\theta}_i'; \Theta) = (1+\lambda_i) \underbrace{\mathbb{E}_{p_i^{\text{data}}}[\nabla^2 \log p_{\boldsymbol{\theta}_i'}]}_{\preceq -\alpha_i I} + \lambda_i \sum_j w_{ij} \underbrace{\left(\mathbb{E}_{p_{\boldsymbol{\theta}_j}}[\nabla^2 \log p_{\boldsymbol{\theta}_i'}] - \mathbb{E}_{p_i^{\text{data}}}[\nabla^2 \log p_{\boldsymbol{\theta}_i'}]\right)}_{=:\,\Delta_j^i}.$$

By Assumption 3.4, $\|\Delta_j^i\|_2 \le L_i\, d_{\mathcal{W}}(p_{\boldsymbol{\theta}_j}, p_i^{\text{data}})$. When evaluated at $\Theta = \bar{\Theta}$ and any j with $w_{ij} > 0$, this is at most $L_i\varepsilon_i$ by definition.

By continuity of $\boldsymbol{\theta}_j \mapsto d_{\mathcal{W}}(p_{\boldsymbol{\theta}_j}, p_i^{\text{data}})$, the bound extends to a neighborhood of $\bar{\Theta}$ with a slightly relaxed constant $\varepsilon_i + \delta$ for any $\delta > 0$.

Since $\alpha_i > L_i\varepsilon_i$ strictly, we may choose $\delta > 0$ small enough that $\alpha_i > L_i(\varepsilon_i + \delta)$.

This gives a neighborhood $\mathcal{U}$ of $\bar{\Theta}$ on which, for all $\boldsymbol{\theta}_i' \in U_i$ and $\Theta \in \mathcal{U}$,

$$\nabla^2_{\boldsymbol{\theta}_i'} H_i(\boldsymbol{\theta}_i'; \Theta) \preceq -\big((1+\lambda_i)\alpha_i - \lambda_i L_i(\varepsilon_i + \delta)\big) I \prec 0.$$

Hence $\boldsymbol{\theta}_i' \mapsto H_i(\boldsymbol{\theta}_i'; \Theta)$ is strongly concave on U_i uniformly in $\Theta \in \mathcal{U}$, so it admits a unique local maximizer and $\mathcal{G}_i$ is single-valued on $\mathcal{U}$. □

Lemma D.2 (Characterization of fixed points)**.** *Under assumptions and $\alpha_i > L_i\varepsilon_i$, a global parameter vector $\bar{\Theta}$ is a fixed point of $\mathcal{G}$ if and only if, for every $i \in \mathcal{V}$,*

$$\nabla_{\boldsymbol{\theta}} \mathbb{E}_{x\sim p_i^{\text{data}}}[\log p_{\bar{\boldsymbol{\theta}}_i}(x)] + \lambda_i \sum_{j=1}^{K} w_{ij}\, \nabla_{\boldsymbol{\theta}} \mathbb{E}_{x\sim p_{\bar{\boldsymbol{\theta}}_j}}[\log p_{\bar{\boldsymbol{\theta}}_i}(x)] = 0.$$

Proof. ($\Rightarrow$) Suppose $\bar{\Theta} = \mathcal{G}(\bar{\Theta})$. Then for each i, $\bar{\boldsymbol{\theta}}_i = \mathcal{G}_i(\bar{\Theta}) \in \arg\max_{\boldsymbol{\theta}\in\Theta_i} H_i(\boldsymbol{\theta}; \bar{\Theta})$.

Since $\mathcal{H}_i$ is twice continuously differentiable, and $\bar{\boldsymbol{\theta}}_i$ is an interior local maximizer (strong concavity on U_i from Lemma D.1 places the maximum in the interior of U_i), Fermat's rule gives $\nabla_{\boldsymbol{\theta}} H_i(\bar{\boldsymbol{\theta}}_i; \bar{\Theta}) = 0$.

($\Leftarrow$) Suppose $\nabla_{\boldsymbol{\theta}} H_i(\bar{\boldsymbol{\theta}}_i; \bar{\Theta}) = 0$ holds for every i. By Lemma D.1, $\boldsymbol{\theta} \mapsto H_i(\boldsymbol{\theta}; \bar{\Theta})$ is strongly concave on U_i, so any stationary point in U_i is the unique local maximizer on U_i.

Evaluating the tie-breaking rule at $\boldsymbol{\theta}_i^{t-1} = \bar{\boldsymbol{\theta}}_i$ gives distance zero, so $\mathcal{G}_i(\bar{\Theta}) = \bar{\boldsymbol{\theta}}_i$. Stacking over i: $\mathcal{G}(\bar{\Theta}) = \bar{\Theta}$. □

D.2 Proof of the Uncoupled Solution as a Fixed Point

Proposition D.3 (Uncoupled solution as a fixed point)**.** *Suppose the real data distributions are homogeneous, $p_i^{\text{data}} \equiv p^{\text{data}}$ for all $i \in \mathcal{V}$, and there exists a distribution $p^\star$ such that $p_{\boldsymbol{\theta}_i^\star} \equiv p^\star$ for every $i \in \mathcal{V}$. Then $\Theta^\star = (\boldsymbol{\theta}_1^\star, \ldots, \boldsymbol{\theta}_K^\star)$ is a fixed point of $\mathcal{G}$ for any $\lambda_i \ge 0$ and any row-stochastic $\mathbf{W}$.*

Proof. Fix any $i \in \mathcal{V}$. The local objective of model i evaluated at $\Theta = \Theta^\star$ is

$$\begin{aligned} H_i(\boldsymbol{\theta}; \Theta^\star) &= \mathbb{E}_{x\sim p^{\text{data}}}[\log p_{\boldsymbol{\theta}}(x)] + \lambda_i \sum_{j=1}^{K} w_{ij}\, \mathbb{E}_{x\sim p_{\boldsymbol{\theta}_j^\star}}[\log p_{\boldsymbol{\theta}}(x)] \\ &= \mathbb{E}_{x\sim p^{\text{data}}}[\log p_{\boldsymbol{\theta}}(x)] + \lambda_i \sum_{j=1}^{K} w_{ij}\, \mathbb{E}_{x\sim p^\star}[\log p_{\boldsymbol{\theta}}(x)] \\ &= \mathbb{E}_{x\sim p^{\text{data}}}[\log p_{\boldsymbol{\theta}}(x)] + \lambda_i\, \mathbb{E}_{x\sim p^\star}[\log p_{\boldsymbol{\theta}}(x)], \end{aligned} \tag{9}$$

where the second equality uses $p_{\boldsymbol{\theta}_j^\star} \equiv p^\star$ for all j, and the third uses row-stochasticity $\sum_j w_{ij} = 1$.

We show $\boldsymbol{\theta}_i^\star$ maximizes each term on the right-hand side of (9) over Θ_i.

For the first term, by definition of $\boldsymbol{\theta}_i^\star$ as the MLE on p^{data}:

$$\mathbb{E}_{x\sim p^{\text{data}}}[\log p_{\boldsymbol{\theta}}(x)] \le \mathbb{E}_{x\sim p^{\text{data}}}[\log p_{\boldsymbol{\theta}_i^\star}(x)], \qquad \forall \boldsymbol{\theta} \in \Theta_i. \tag{10}$$

For the second term, since $p_{\boldsymbol{\theta}_i^\star} \equiv p^\star$, for any $\boldsymbol{\theta} \in \Theta_i$:

$$\begin{aligned}\mathbb{E}_{x\sim p^\star}[\log p_{\boldsymbol{\theta}}(x)] &= \mathbb{E}_{x\sim p^\star}[\log p_{\boldsymbol{\theta}_i^\star}(x)] - D_{\text{KL}}(p^\star \,\|\, p_{\boldsymbol{\theta}}) \\ &\le \mathbb{E}_{x\sim p^\star}[\log p_{\boldsymbol{\theta}_i^\star}(x)],\end{aligned} \tag{11}$$

using non-negativity of KL divergence.

Combining (9)–(11) yields

$$H_i(\boldsymbol{\theta}; \Theta^\star) \le H_i(\boldsymbol{\theta}_i^\star; \Theta^\star), \qquad \forall \boldsymbol{\theta} \in \Theta_i.$$

Hence $\boldsymbol{\theta}_i^\star$ is a maximizer of $H_i(\cdot; \Theta^\star)$.

By the nearest-local-maximizer tie-breaking rule, $\boldsymbol{\theta}_i^{t-1} = \boldsymbol{\theta}_i^\star$ gives distance zero, so $\mathcal{G}_i(\Theta^\star) = \boldsymbol{\theta}_i^\star$.

Stacking over i: $\mathcal{G}(\Theta^\star) = \Theta^\star$. □

D.3 Example: Heterogeneous Fixed Point Different from the Uncoupled Solution

We present a simple two-model example illustrating that, under heterogeneous data, the coupled retraining dynamics can admit a stable fixed point distinct from the uncoupled solution $\Theta^\star$.

Consider two models $K = 2$ with scalar parameters $\theta_1, \theta_2 \in \mathbb{R}$.

Each model observes a different real-data distribution, let the log-likelihood be given by

$$\log p_{\boldsymbol{\theta}_i}(x) = -\tfrac{1}{2}(\theta_i - a_i)^2$$

so that in the absence of interactions, the uncoupled solution is $\theta_i^\star = a_i$, $\Theta^\star = (a_1, a_2)$.

Introduce symmetric synthetic interactions $\lambda_i = \lambda_j = \lambda$ and $w_{ij} = w_{ji} = 1$ by letting each model update according to

$$H_i(\theta_i; \theta_j) = \log p_{\boldsymbol{\theta}_i}(x) - \tfrac{\lambda}{2}(\theta_i - \theta_j)^2$$

For each i, the objective H_i is strictly concave in θ_i, so the best-response map $\mathcal{G}_i$ is uniquely defined and smooth.

To compute the update map $\mathcal{G}_i$, we take the first-order condition with respect to θ_i:

$$\frac{\partial H_i}{\partial \theta_i} = -(\theta_i - a_i) - \lambda(\theta_i - \theta_j)$$

Setting this derivative to zero yields

$$(1+\lambda)\theta_i = a_i + \lambda\theta_j$$

and hence the unique best response

$$\mathcal{G}_i(\theta_j) = \frac{a_i + \lambda\theta_j}{1+\lambda}$$

A fixed point $\bar{\Theta} = (\bar{\theta}_1, \bar{\theta}_2)$ therefore satisfies

$$\bar{\theta}_1 = \frac{a_1 + \lambda\bar{\theta}_2}{1+\lambda} \qquad \bar{\theta}_2 = \frac{a_2 + \lambda\bar{\theta}_1}{1+\lambda}$$

So,

$$\bar{\theta}_1 = \frac{(1+\lambda)a_1 + \lambda a_2}{1+2\lambda} \qquad \bar{\theta}_2 = \frac{\lambda a_1 + (1+\lambda)a_2}{1+2\lambda}$$

For example, setting $a_1 = 1$, $a_2 = -1$, and $\lambda = 1$ gives the unique fixed point

$$\bar{\Theta} = \left(\tfrac{1}{3}, -\tfrac{1}{3}\right)$$

whereas the uncoupled solution is $\Theta^\star = (1, -1)$.

Local stability. The Jacobian of the update map $\mathcal{G}$ is

$$\mathbf{J} = \begin{pmatrix} 0 & \frac{\lambda}{1+\lambda} \\ \frac{\lambda}{1+\lambda} & 0 \end{pmatrix}$$

Its spectral radius is $\rho(\mathbf{J}) = \frac{\lambda}{1+\lambda} < 1$, confirming that the fixed point $\bar{\Theta}$ is locally stable.

Thus, under heterogeneous data, the fixed point of the coupled retraining dynamics exists and is locally stable, but differs from the uncoupled solution.

D.4 Proofs of Theorem 3.7

Proposition D.4 (Block Jacobian bound)**.** *Let $\bar{\Theta}$ be a fixed point of $\mathcal{G}$. Under assumptions and $\alpha_i > L_i\varepsilon_i$, $\mathcal{G}$ is C^1 in a neighborhood of $\bar{\Theta}$, the Jacobian $\mathbf{J} := \nabla_\Theta \mathcal{G}(\bar{\Theta})$ is well-defined, and its (i,j)-th block satisfies*

$$\|J_{ij}\|_2 \le \beta_i w_{ij}, \qquad \beta_i := \frac{\lambda_i M_i}{\alpha_i + \lambda_i(\alpha_i - L_i\varepsilon_i)}. \tag{12}$$

Proof. *C^1 smoothness and Jacobian decomposition.* Define the stationarity map

$$F_i(\boldsymbol{\theta}_i'; \Theta) := \nabla_{\boldsymbol{\theta}_i'} H_i(\boldsymbol{\theta}_i'; \Theta).$$

By Lemma D.2, $F_i(\bar{\boldsymbol{\theta}}_i; \bar{\Theta}) = 0$. Moreover, by Lemma D.1, $H_{ii} := \nabla^2_{\boldsymbol{\theta}_i'} H_i(\bar{\boldsymbol{\theta}}_i; \bar{\Theta}) \prec 0$, hence invertible.

The Implicit Function Theorem [24] applied to $F_i = 0$ yields a C^1 selection $\Theta \mapsto \mathcal{G}_i(\Theta)$ in a neighborhood of $\bar{\Theta}$ with $\mathcal{G}_i(\bar{\Theta}) = \bar{\boldsymbol{\theta}}_i$ and

$$\nabla_\Theta \mathcal{G}_i(\bar{\Theta}) = -H_{ii}^{-1}\, \nabla_\Theta F_i(\bar{\boldsymbol{\theta}}_i; \bar{\Theta}).$$

Stacking over i, $\mathcal{G}$ is C^1 and $\mathbf{J} := \nabla_\Theta \mathcal{G}(\bar{\Theta})$ is well-defined.

Differentiating F_i with respect to $\boldsymbol{\theta}_j$:

$$\frac{\partial F_i}{\partial \boldsymbol{\theta}_j}(\bar{\boldsymbol{\theta}}_i) = \lambda_i w_{ij} \frac{\partial}{\partial \boldsymbol{\theta}_j}\Big(\mathbb{E}_{p_{\boldsymbol{\theta}_j}}\left[\nabla_\theta \log p_{\bar{\boldsymbol{\theta}}_i}\right]\Big)\Big|_{\boldsymbol{\theta}_j = \bar{\boldsymbol{\theta}}_j} = \lambda_i w_{ij} B_{ij}$$

Hence the (i,j)-th block of $\mathbf{J}$ is

$$J_{ij} = -H_{ii}^{-1}\, \lambda_i w_{ij}\, B_{ij}.$$

Taking spectral norms,

$$\|J_{ij}\|_2 \le \|H_{ii}^{-1}\|_2 \cdot \lambda_i w_{ij} \cdot \|B_{ij}\|_2. \tag{13}$$

Bounding $\|H_{ii}^{-1}\|_2$. From the proof of Lemma D.1, $H_{ii} \preceq -\big((1+\lambda_i)\alpha_i - \lambda_i L_i \varepsilon_i\big) I$.

Since $\alpha_i > L_i\varepsilon_i$, the bracketed quantity is positive, so

$$\|H_{ii}^{-1}\|_2 \le \frac{1}{(1+\lambda_i)\alpha_i - \lambda_i L_i \varepsilon_i}. \tag{14}$$

Bounding $\|B_{ij}\|_2$. By assumption, $\|B_{ij}\|_2 \le M_i$.

Combining (13), (14), and $\|B_{ij}\|_2 \le M_i$:

$$\|J_{ij}\|_2 \le \frac{\lambda_i w_{ij} M_i}{(1+\lambda_i)\alpha_i - \lambda_i L_i \varepsilon_i} = \frac{\lambda_i M_i}{\alpha_i + \lambda_i(\alpha_i - L_i\varepsilon_i)} \cdot w_{ij} = \beta_i w_{ij}.$$

□

Lemma D.5 (Comparison principle $\rho(\mathbf{J}) \le \rho(\mathbf{D})$)**.** *Let $\mathbf{J} = \nabla_\Theta \mathcal{G}(\bar{\Theta})$ with block decomposition satisfying $\|J_{ij}\|_2 \le \beta_i w_{ij}$, as established in Proposition D.4, and let $\mathbf{D} := \operatorname{diag}(\beta_1, \dots, \beta_K)\mathbf{W}$. Then $\rho(\mathbf{J}) \le \rho(\mathbf{D})$.*

Proof. We first show by induction that for all $t \ge 1$,

$$\|(\mathbf{J}^t)_{ij}\|_2 \le (\mathbf{D}^t)_{ij}, \qquad \forall i, j \in \mathcal{V}. \tag{15}$$

Base case ($t = 1$): $\|J_{ij}\|_2 \le \beta_i w_{ij} = D_{ij}$ by Proposition D.4.

Inductive step: assume (15) holds for some $t \ge 1$. Then for the (i,j)-th block of $\mathbf{J}^{t+1} = \mathbf{J} \cdot \mathbf{J}^t$,

$$\begin{aligned}\|(\mathbf{J}^{t+1})_{ij}\|_2 &= \Big\|\sum_{k \in \mathcal{V}} J_{ik}\, (\mathbf{J}^t)_{kj}\Big\|_2 \le \sum_k \|J_{ik}\|_2\, \|(\mathbf{J}^t)_{kj}\|_2 \\ &\le \sum_k D_{ik}\, (\mathbf{D}^t)_{kj} = (\mathbf{D}^{t+1})_{ij},\end{aligned}$$

where the first inequality uses triangle inequality and submultiplicativity of the spectral norm, and the seconduses the inductive hypothesis.

To convert the entrywise block bound to a spectral norm bound on $\mathbf{J}^t$, using the Frobenius norm $\|\cdot\|_F$:

$$\|\mathbf{J}^t\|_2 \le \|\mathbf{J}^t\|_F = \sqrt{\sum_{i,j} \|(\mathbf{J}^t)_{ij}\|_F^2}$$

$d_{\max} := \max_i d_i$ accounts for block dimensions, using $\|M\|_F \le \sqrt{d}\,\|M\|_2$ for any $d \times d$ matrix M:

$$\sqrt{\sum_{i,j} \|(\mathbf{J}^t)_{ij}\|_F^2} \le \sqrt{d_{\max}} \sqrt{\sum_{i,j} \|(\mathbf{J}^t)_{ij}\|_2^2} \le \sqrt{d_{\max}} \sqrt{\sum_{i,j} (\mathbf{D}^t)_{ij}^2} = \sqrt{d_{\max}}\,\|\mathbf{D}^t\|_F$$

Combined with $\|\mathbf{D}^t\|_F \le \sqrt{K}\,\|\mathbf{D}^t\|_2$:

$$\|\mathbf{J}^t\|_2 \le \sqrt{K\, d_{\max}}\,\|\mathbf{D}^t\|_2.$$

Applying Gelfand's formula [6]:

$$\rho(\mathbf{J}) = \lim_{t\to\infty} \|\mathbf{J}^t\|_2^{1/t} \le \lim_{t\to\infty} (K\, d_{\max})^{1/(2t)}\, \|\mathbf{D}^t\|_2^{1/t} = \rho(\mathbf{D}),$$

since $(K\, d_{\max})^{1/(2t)} \to 1$. □

Theorem 3.7 (Local asymptotic stability)**.** *Suppose $\alpha_i > L_i\varepsilon_i$, $\forall i \in \mathcal{V}$. If $\rho(\mathbf{D}) < 1$, then $\bar{\Theta}$ is locally asymptotically stable. Moreover, there exists a neighborhood $\mathcal{U}$ of $\bar{\Theta}$ such that for any initialization $\Theta^0 \in \mathcal{U}$, the iterates $\Theta^{t+1} = \mathcal{G}(\Theta^t)$ converge to $\bar{\Theta}$ at asymptotic linear rate*

$$\limsup_{t\to\infty} \|\Theta^t - \bar{\Theta}\|_2^{1/t} \le \rho(\mathbf{J}) \le \rho(\mathbf{D}). \tag{4}$$

where $\mathbf{J} := \nabla_\Theta \mathcal{G}(\bar{\Theta})$ denotes the Jacobian of $\mathcal{G}$ at $\bar{\Theta}$, and $\rho(\cdot)$ is the spectral radius of a matrix.

Proof. By Proposition D.4, $\mathcal{G}$ is C^1 in a neighborhood of $\bar{\Theta}$ and the Jacobian $\mathbf{J} := \nabla_\Theta \mathcal{G}(\bar{\Theta})$ is well-defined.

By Lemma D.5, $\rho(\mathbf{J}) \le \rho(\mathbf{D})$.

Combined with the hypothesis $\rho(\mathbf{D}) < 1$, we have $\rho(\mathbf{J}) < 1$, which implies that $\bar{\Theta}$ is a locally asymptotically stable fixed point.

Choice of contraction norm. For any $\delta > 0$ with $\rho(\mathbf{J}) + 2\delta < 1$, there exists a matrix norm $|\!|\!| \cdot |\!|\!|$ on $\prod_i \Theta_i$ such that $|\!|\!|\mathbf{J}|\!|\!| \le \rho(\mathbf{J}) + \delta < 1$. (This is a standard consequence of Gelfand's formula [6]: such a norm exists because $\rho(\mathbf{J})$ is the infimum over all sub-multiplicative matrix norms of $\mathbf{J}$.)

Contraction neighborhood. Since $\nabla\mathcal{G}$ is continuous at $\bar{\Theta}$ with $\nabla\mathcal{G}(\bar{\Theta}) = \mathbf{J}$, there exists a neighborhood $\mathcal{U}$ of $\bar{\Theta}$ such that

$$|\!|\!|\nabla\mathcal{G}(\Theta) - \mathbf{J}|\!|\!| \le \delta, \qquad \forall \Theta \in \mathcal{U}. \tag{16}$$

Combining $|\!|\!|\mathbf{J}|\!|\!| \le \rho(\mathbf{J}) + \delta$ and (16) via the triangle inequality,

$$|\!|\!|\nabla\mathcal{G}(\Theta)|\!|\!| \le |\!|\!|\mathbf{J}|\!|\!| + |\!|\!|\nabla\mathcal{G}(\Theta) - \mathbf{J}|\!|\!| \le \rho(\mathbf{J}) + 2\delta =: \rho_{\mathcal{G}} < 1, \qquad \forall \Theta \in \mathcal{U}.$$

By the mean-value inequality, for all $\Theta, \Theta' \in \mathcal{U}$ with the segment $[\Theta, \Theta'] \subset \mathcal{U}$,

$$|\!|\!|\mathcal{G}(\Theta) - \mathcal{G}(\Theta')|\!|\!| \le \rho_{\mathcal{G}} |\!|\!|\Theta - \Theta'|\!|\!|.$$

Taking $\mathcal{U}$ to be a convex ball in the $|\!|\!| \cdot |\!|\!|$ metric, and shrinking if needed so $\mathcal{G}(\mathcal{U}) \subseteq \mathcal{U}$ (which is possible since $\mathcal{G}(\bar{\Theta}) = \bar{\Theta}$ and $\mathcal{G}$ is continuous), $\mathcal{G}$ is a contraction self-map on $\mathcal{U}$.

Convergence. By Banach's fixed-point theorem [9], for any $\Theta^0 \in \mathcal{U}$, the iterates $\Theta^{t+1} = \mathcal{G}(\Theta^t)$ converge to the unique fixed point of $\mathcal{G}$ in $\mathcal{U}$, which is $\bar{\Theta}$.

Moreover,

$$|\!|\!|\Theta^t - \bar{\Theta}|\!|\!| \le \rho_{\mathcal{G}}^t |\!|\!|\Theta^0 - \bar{\Theta}|\!|\!|, \quad \forall t \ge 0.$$

Since all norms on finite-dimensional $\prod_i \Theta_i$ are equivalent, there exist constants $c_1, c_2 > 0$ with $c_1\|\cdot\|_2 \le |\!|\!|\cdot|\!|\!| \le c_2\|\cdot\|_2$. Hence,

$$\|\Theta^t - \bar{\Theta}\|_2 \le \frac{c_2}{c_1}\, \rho_{\mathcal{G}}^t \|\Theta^0 - \bar{\Theta}\|_2,$$

establishing geometric convergence in $\|\cdot\|_2$ and local asymptotic stability of $\bar{\Theta}$.

Asymptotic linear rate. From the above bound,

$$\|\Theta^t - \bar{\Theta}\|_2^{1/t} \le \left(\frac{c_2}{c_1}\right)^{1/t} \rho_{\mathcal{G}}.$$

Taking $t \to \infty$, $\left(\frac{c_2}{c_1}\right)^{1/t} \to 1$, so

$$\limsup_{t\to\infty} \|\Theta^t - \bar{\Theta}\|_2^{1/t} \le \rho_{\mathcal{G}} = \rho(\mathbf{J}) + 2\delta.$$

Since $\delta > 0$ was arbitrary (subject to $\rho(\mathbf{J}) + 2\delta < 1$), letting $\delta \to 0$:

$$\limsup_{t\to\infty} \|\Theta^t - \bar{\Theta}\|_2^{1/t} \le \rho(\mathbf{J}) \le \rho(\mathbf{D}),$$

where the last inequality uses Lemma D.5. □

D.5 Proofs of Corollary 3.8

Corollary 3.8 (A norm-based sufficient condition)**.** *If $\beta_{\max}\|\mathbf{W}\|_2 < 1$, then $\rho(\mathbf{D}) < 1$, and $\bar{\Theta}$ is locally asymptotically stable. Equivalently, when $M_i\|\mathbf{W}\|_2 > \alpha_i - L_i\varepsilon_i$ for all $i \in \mathcal{V}$, a sufficient condition for local asymptotic stability is*

$$\lambda_i < \frac{\alpha_i}{M_i\|\mathbf{W}\|_2 - (\alpha_i - L_i\varepsilon_i)}, \quad \forall i \in \mathcal{V}. \tag{5}$$

Proof. We prove two parts.

Part 1: $\beta_{\max}\|\mathbf{W}\|_2 < 1 \Rightarrow \rho(\mathbf{D}) < 1$. Since $\mathbf{D} = \mathrm{diag}(\beta_1, \ldots, \beta_K)\mathbf{W}$ with $\beta_i \geq 0$,

$$\|\mathbf{D}\|_2 \leq \|\mathrm{diag}(\beta_1, \ldots, \beta_K)\|_2 \cdot \|\mathbf{W}\|_2 = \beta_{\max}\|\mathbf{W}\|_2.$$

Combined with $\rho(\mathbf{D}) \leq \|\mathbf{D}\|_2$:

$$\rho(\mathbf{D}) \leq \|\mathbf{D}\|_2 \leq \beta_{\max}\|\mathbf{W}\|_2 < 1.$$

Local asymptotic stability of $\bar{\Theta}$ then follows from Theorem 3.7.

Part 2: Equivalence with the λ_i condition.

The condition $\beta_{\max}\|\mathbf{W}\|_2 < 1$ is equivalent to $\beta_i\|\mathbf{W}\|_2 < 1$ for all $i \in \mathcal{V}$.

Substituting $\beta_i = \frac{\lambda_i M_i}{(\alpha_i + \lambda_i(\alpha_i - L_i\varepsilon_i))}$:

$$\frac{\lambda_i M_i}{\alpha_i + \lambda_i(\alpha_i - L_i\varepsilon_i)}\|\mathbf{W}\|_2 < 1.$$

Since $\alpha_i > L_i\varepsilon_i$ implies the denominator is positive, cross-multiplying gives

$$\lambda_i M_i\|\mathbf{W}\|_2 < \alpha_i + \lambda_i(\alpha_i - L_i\varepsilon_i).$$

Isolating λ_i on the left:

$$\lambda_i\big[M_i\|\mathbf{W}\|_2 - (\alpha_i - L_i\varepsilon_i)\big] < \alpha_i.$$

Since $M_i > 0$ and $M_i\|\mathbf{W}\|_2 > \alpha_i - L_i\varepsilon_i$

$$\lambda_i < \frac{\alpha_i}{M_i\|\mathbf{W}\|_2 - (\alpha_i - L_i\varepsilon_i)}.$$

□

Proposition D.6 (Single-model recovery)**.** *Suppose $K = 1$, $w_{11} = 1$. Under assumptions with the regularity margin $\alpha_1 > L_1\varepsilon_1$, the normalized self-coupling in the single-model Jacobian calculation is bounded by $M_1 := \alpha_1 + L_1\varepsilon_1$, and the single-model stability condition reduces to $\lambda_1 < \frac{\alpha_1}{2L_1\varepsilon_1}$.*

Proof. When $K = 1$ and $w_{11} = 1$, the networked retraining map reduces to the standard single-model self-consuming update

$$G_\lambda(\theta) \in \text{local-argmax}_{\theta'}\Big[H_1(\theta') + \lambda H_2(\theta, \theta')\Big],$$

where

$$H_1(\theta') = \mathbb{E}_{x\sim p^{\mathrm{data}}}[\log p_{\theta'}(x)], \qquad H_2(\theta, \theta') = \mathbb{E}_{x\sim p_\theta}[\log p_{\theta'}(x)].$$

Let

$$A := \nabla^2_{\theta',\theta'} H_1(\bar{\theta}_1), \qquad B := \nabla^2_{\theta,\theta'} H_2(\bar{\theta}_1, \bar{\theta}_1).$$

By the probability-conservation identity used in single-model self-consuming analyses,

$$\nabla^2_{\theta',\theta} H_2(\bar{\theta}_1, \bar{\theta}_1) = -\nabla^2_{\theta',\theta'} H_2(\bar{\theta}_1, \bar{\theta}_1).$$

Therefore, applying the implicit-function calculation for the local update branch gives

$$J_G(\bar{\theta}_1) = (I + \lambda_1 A^{-1}B)^{-1}\lambda_1 A^{-1}B.$$

Next, Assumption 3.3 gives

$$\|A^{-1}\|_2 \leq \frac{1}{\alpha_1}.$$

Moreover, by Assumption 3.4,

$$\|B - A\|_2 \leq L_1 d_{\mathcal{W}}(p_{\bar{\theta}_1}, p_1^{\mathrm{data}}) = L_1\varepsilon_1.$$

Hence

$$\|A^{-1}B - I\|_2 = \|A^{-1}(B - A)\|_2 \leq \frac{L_1\varepsilon_1}{\alpha_1},$$

and

$$\|A^{-1}B\|_2 \leq 1 + \frac{L_1\varepsilon_1}{\alpha_1}.$$

Equivalently, the normalized coupling can be bounded by

$$\alpha_1\|A^{-1}B\|_2 \leq \alpha_1 + L_1\varepsilon_1.$$

Thus, in the single-model self-loop case, the coupling term appearing in the normalized single-model Jacobian calculation is controlled by $M_1 := \alpha_1 + L_1\varepsilon_1$.

Since $\alpha_1 > L_1\varepsilon_1$, we have $\|A^{-1}B - I\|_2 \leq \frac{L_1\varepsilon_1}{\alpha_1} < 1$. Moreover, $\frac{\lambda_1}{(1+\lambda_1)} < 1$ for $\lambda_1 \geq 0$. Hence

$$\frac{\lambda_1}{1+\lambda_1}\|A^{-1}B - I\|_2 < 1.$$

Using

$$J_G(\bar{\theta}_1) = (I + \lambda_1 A^{-1}B)^{-1}\lambda_1 A^{-1}B,$$

we bound

$$\|J_G(\bar{\theta}_1)\|_2 \leq \|(I + \lambda_1 A^{-1}B)^{-1}\|_2 \cdot \lambda_1\|A^{-1}B\|_2.$$

Since

$$I + \lambda_1 A^{-1}B = (1+\lambda_1)I + \lambda_1(A^{-1}B - I),$$

we have, whenever

$$\frac{\lambda_1}{1+\lambda_1}\|A^{-1}B - I\|_2 < 1,$$

that

$$\|(I + \lambda_1 A^{-1}B)^{-1}\|_2 \leq \frac{1}{1 + \lambda_1 - \lambda_1\frac{L_1\varepsilon_1}{\alpha_1}}.$$

Therefore,

$$\|J_G(\bar{\theta}_1)\|_2 \leq \frac{\lambda_1(1 + \frac{L_1\varepsilon_1}{\alpha_1})}{1 + \lambda_1 - \lambda_1\frac{L_1\varepsilon_1}{\alpha_1}}.$$

A sufficient condition for the right-hand side to be smaller than one is

$$\lambda_1\left(1 + \frac{L_1\varepsilon_1}{\alpha_1}\right) < 1 + \lambda_1 - \lambda_1\frac{L_1\varepsilon_1}{\alpha_1},$$

which is equivalent to

$$2\lambda_1\frac{L_1\varepsilon_1}{\alpha_1} < 1.$$

Thus

$$\lambda_1 < \frac{\alpha_1}{2L_1\varepsilon_1}.$$

This recovers the single-model stability threshold. □

D.6 Proofs of Proposition 3.10 and Corollary 3.11

Proposition 3.10 (Cycle gain as a lower bound of $\rho(\mathbf{D})$)**.** *Let $\mathbf{D} \geq 0$ be a nonnegative matrix and let $\{C_m\}_{m=1}^{M}$ denote the set of all simple directed cycles in the corresponding interaction graph. Then the spectral radius of $\mathbf{D}$ satisfies*

$$\rho(\mathbf{D}) \geq \max_{m\in[M]} \gamma(C_m)$$

Proof. Under the convention w_{ij} = weight of edge $j \to i$ (i.e., row i of $\mathbf{W}$ represents incoming signals to model i), we have $\mathbf{D}_{ij} = \beta_i w_{ij}$ associated with edge $j \to i$.

For a nonnegative matrix $\mathbf{D}$, the (a, b)-th entry of $\mathbf{D}^t$ admits the path-sum interpretation

$$(\mathbf{D}^t)_{ab} = \sum_{a\leftarrow k_1\leftarrow\cdots\leftarrow k_{t-1}\leftarrow b} \prod (\text{edge weights along the path}),$$

where the sum is over all directed paths from b to a of length t.

Consider a simple directed cycle $C = (i_\ell \to i_{\ell-1} \to \cdots \to i_1 \to i_\ell)$. Traversing C exactly m times yields a directed sequence $(i_\ell, \ldots, i_1, i_\ell, \ldots, i_1, \ldots, i_\ell)$ of length $m\ell + 1$, constituting a valid path from i_1 to i_1 of length $m\ell$.

Since $\mathbf{D}$ is nonnegative, the entry $(\mathbf{D}^{m\ell})_{i_1,i_1}$ is lower-bounded by the weight of this specific path:

$$(\mathbf{D}^{m\ell})_{i_1,i_1} \geq \left(\prod_{r=1}^{\ell} D_{i_r,i_{r+1}}\right)^m,$$

where $i_{\ell+1} = i_1$.

The induced ℓ_∞ norm satisfies $\|\mathbf{D}^{m\ell}\|_\infty = \max_a \sum_b (\mathbf{D}^{m\ell})_{ab}$, which is at least any single entry:

$$\|\mathbf{D}^{m\ell}\|_\infty \geq (\mathbf{D}^{m\ell})_{i_1,i_1} \geq \left(\prod_{r=1}^{\ell} D_{i_r,i_{r+1}}\right)^m.$$

Taking the $(m\ell)$-th root:

$$\|\mathbf{D}^{m\ell}\|_\infty^{1/(m\ell)} \geq \left(\prod_{r=1}^{\ell} D_{i_r,i_{r+1}}\right)^{1/\ell} = \gamma(C),$$

using the definition of cycle gain.

By Gelfand's formula, $\rho(\mathbf{D}) = \lim_{t\to\infty} \|\mathbf{D}^t\|_\infty^{1/t}$. Taking the limit $m \to \infty$ along the subsequence $t = m\ell$:

$$\rho(\mathbf{D}) = \lim_{m\to\infty} \|\mathbf{D}^{m\ell}\|_\infty^{1/(m\ell)} \geq \gamma(C).$$

Since C was arbitrary, $\rho(\mathbf{D}) \geq \max_{m\in[M]} \gamma(C_m)$. □

Corollary 3.11 (Necessary condition on cycles)**.** *If $\rho(\mathbf{D}) < 1$, then every simple directed cycle C satisfies $\prod_{(j\to i)\in C} \beta_i w_{ij} < 1$. In particular, every self-loop must satisfy $\beta_i w_{ii} < 1$.*

Proof. We prove the contrapositive. Suppose there exists a simple directed cycle

$$C = (i_\ell \to i_{\ell-1} \to \cdots \to i_1 \to i_\ell)$$

such that

$$\prod_{r=1}^{\ell} D_{i_r i_{r+1}} \geq 1, \qquad i_{\ell+1} = i_1.$$

Equivalently, $\gamma(C) \geq 1$. By Proposition 3.10,

$$\rho(\mathbf{D}) \geq \gamma(C) \geq 1.$$

Therefore $\rho(\mathbf{D}) < 1$ cannot hold. Taking the contrapositive, if $\rho(\mathbf{D}) < 1$, then every simple directed cycle C must satisfy

$$\prod_{r=1}^{\ell} D_{i_r i_{r+1}} < 1.$$

Using $D_{ij} = \beta_i w_{ij}$, this is

$$\prod_{r=1}^{\ell} \beta_{i_r} w_{i_r i_{r+1}} < 1.$$

In particular, for a self-loop $C = (i \to i)$, the condition becomes

$$\beta_i w_{ii} < 1.$$

□

D.7 Proofs of Proposition 4.4

Proposition D.7 (Fixed-point representation mixture identity)**.** *Suppose assumptions hold with $\alpha_i > L_i\varepsilon_i$ for all $i \in \mathcal{V}$. Then at any fixed point $\bar{\Theta}$ of $\mathcal{G}$, the representation $\bar{\boldsymbol{f}}_i$ admits the decomposition*

$$\bar{\boldsymbol{f}}_i = \frac{1}{1+\lambda_i} \boldsymbol{f}_i^{\text{data}} + \frac{\lambda_i}{1+\lambda_i} \sum_{j\in\mathcal{V}} w_{ij}\, \bar{\boldsymbol{f}}_j + \bar{\boldsymbol{r}}_i, \tag{17}$$

where the residual $\bar{\boldsymbol{r}}_i \in \mathbb{R}^D$ satisfies

$$\|\bar{\boldsymbol{r}}_i\|_2 \leq \frac{1+2\lambda_i}{1+\lambda_i}\, \varepsilon_i. \tag{18}$$

Proof. Step 1: Mixture distribution and its representation. At the fixed point $\bar{\Theta}$, the synthetic-data training mixture for model i is

$$\bar{p}_i(x) := \frac{1}{1+\lambda_i} p_i^{\text{data}}(x) + \frac{\lambda_i}{1+\lambda_i} \sum_{j \in \mathcal{V}} w_{ij}\, p_{\bar{\boldsymbol{\theta}}_j}(x).$$

Applying the operator $\int \phi(x)(\cdot)dx$:

$$\mathbb{E}_{x\sim\bar{p}_i}[\phi(x)] = \int \phi(x)\bar{p}_i(x)dx = \int \phi(x) \left[\frac{1}{1+\lambda_i} p_i^{data}(x) + \frac{\lambda_i}{1+\lambda_i} \sum_{j=1}^{K} w_{ij} p_{\bar{\boldsymbol{\theta}}_j}(x) \right] dx$$

$$= \frac{1}{1+\lambda_i} \underbrace{\int \phi(x) p_i^{data}(x)dx}_{\boldsymbol{f}_i^{\text{data}}} + \frac{\lambda_i}{1+\lambda_i} \sum_{j=1}^{K} w_{ij} \underbrace{\int \phi(x) p_{\boldsymbol{\theta}_j^{t-1}} dx}_{\bar{\boldsymbol{f}}_j}$$

So,

$$\mathbb{E}_{x\sim\bar{p}_i}[\phi(x)] = \frac{1}{1+\lambda_i} \boldsymbol{f}_i^{\text{data}} + \frac{\lambda_i}{1+\lambda_i} \sum_{j \in \mathcal{V}} w_{ij}\, \bar{\boldsymbol{f}}_j. \tag{19}$$

Step 2: Residual definition. Define the representation residual

$$\bar{\boldsymbol{r}}_i := \bar{\boldsymbol{f}}_i - \mathbb{E}_{x\sim\bar{p}_i}[\phi(x)] = \mathbb{E}_{x\sim p_{\bar{\boldsymbol{\theta}}_i}}[\phi(x)] - \mathbb{E}_{x\sim\bar{p}_i}[\phi(x)]. \tag{20}$$

Substituting (19) into (20):

$$\bar{\boldsymbol{f}}_i = \frac{1}{1+\lambda_i} \boldsymbol{f}_i^{\text{data}} + \frac{\lambda_i}{1+\lambda_i} \sum_{j \in \mathcal{V}} w_{ij}\, \bar{\boldsymbol{f}}_j + \bar{\boldsymbol{r}}_i$$

Step 3: Bound the residual. For any 1-Lipschitz scalar function $g : \mathcal{X} \to \mathbb{R}$ and any pair of probability measures p, q on $\mathcal{X}$, the Kantorovich–Rubinstein duality [34] gives

$$\left|\mathbb{E}_p[g] - \mathbb{E}_q[g]\right| \le d_W(p, q).$$

Since $\phi : \mathcal{X} \to \mathbb{R}^D$ is 1-Lipschitz with respect to the Euclidean norm on $\mathbb{R}^D$, for any unit vector $u \in \mathbb{R}^D$, the scalar function $x \mapsto \langle u, \phi(x)\rangle$ is 1-Lipschitz on $\mathcal{X}$.

Applying Kantorovich–Rubinstein with $p = p_{\bar{\boldsymbol{\theta}}_i}$ and $q = \bar{p}_i$:

$$\begin{aligned}
\langle u, \bar{\boldsymbol{r}}_i\rangle &= \left\langle u,\ \mathbb{E}_{x\sim p_{\bar{\boldsymbol{\theta}}_i}}[\phi(x)] - \mathbb{E}_{x\sim\bar{p}_i}[\phi(x)]\right\rangle && \text{(residual definition)} \\
&= \left\langle u,\ \mathbb{E}_{x\sim p_{\bar{\boldsymbol{\theta}}_i}}[\phi(x)]\right\rangle - \left\langle u,\ \mathbb{E}_{x\sim\bar{p}_i}[\phi(x)]\right\rangle && \text{(linearity of inner product)} \\
&= \mathbb{E}_{x\sim p_{\bar{\boldsymbol{\theta}}_i}}[\langle u, \phi(x)\rangle] - \mathbb{E}_{x\sim\bar{p}_i}[\langle u, \phi(x)\rangle] && \text{(swap inner product and expectation)} \\
&= \mathbb{E}_{x\sim p_{\bar{\boldsymbol{\theta}}_i}}[g_u(x)] - \mathbb{E}_{x\sim\bar{p}_i}[g_u(x)] && \text{(definition of } g_u\text{)} \\
&\le d_W(p_{\bar{\boldsymbol{\theta}}_i}, \bar{p}_i) && \text{(Kantorovich–Rubinstein).}
\end{aligned}$$

Taking $u = \bar{\boldsymbol{r}}_i / \|\bar{\boldsymbol{r}}_i\|$,

$$\|\bar{\boldsymbol{r}}_i\| \le d_W(p_{\bar{\boldsymbol{\theta}}_i}, \bar{p}_i). \tag{21}$$

Step 4: Bound $d_W(p_{\bar{\boldsymbol{\theta}}_i}, \bar{p}_i)$. By the triangle inequality of Wasserstein distance:

$$d_W(p_{\bar{\boldsymbol{\theta}}_i}, \bar{p}_i) \le \underbrace{d_W(p_{\bar{\boldsymbol{\theta}}_i}, p_i^{\text{data}})}_{=:\,A} + \underbrace{d_W(p_i^{\text{data}}, \bar{p}_i)}_{=:\,B}. \tag{22}$$

Bounding A. By the definition of ε_i in Assumption 3.4, $A = d_W(p_{\bar{\boldsymbol{\theta}}_i}, p_i^{\text{data}}) \le \varepsilon_i$.

Bounding B. Using the convexity of Wasserstein distance in each argument and the definition of $\bar{p}_i$:

$$\begin{aligned}
B &= d_W\left(p_i^{\text{data}},\ \frac{1}{1+\lambda_i} p_i^{\text{data}} + \frac{\lambda_i}{1+\lambda_i} \sum_j w_{ij} p_{\bar{\boldsymbol{\theta}}_j}\right) \\
&\le \frac{1}{1+\lambda_i}\, d_W(p_i^{\text{data}}, p_i^{\text{data}}) + \frac{\lambda_i}{1+\lambda_i} \sum_j w_{ij}\, d_W(p_i^{\text{data}}, p_{\bar{\boldsymbol{\theta}}_j}) \\
&= \frac{\lambda_i}{1+\lambda_i} \sum_j w_{ij}\, d_W(p_i^{\text{data}}, p_{\bar{\boldsymbol{\theta}}_j}),
\end{aligned}$$

where the first term vanishes since $d_W(p_i^{\text{data}}, p_i^{\text{data}}) = 0$.

By the definition of $\varepsilon_i = \max_{j:w_{ij}>0} d_W(p_{\bar{\boldsymbol{\theta}}_j}, p_i^{\text{data}})$ and $\sum_j w_{ij} = 1$ (row-stochasticity of $\mathbf{W}$):

$$B \le \frac{\lambda_i}{1+\lambda_i} \cdot \varepsilon_i.$$

Combining. Substituting the bounds for A and B into (22):

$$d_W(p_{\bar{\boldsymbol{\theta}}_i}, \bar{p}_i) \le \varepsilon_i + \frac{\lambda_i}{1+\lambda_i}\,\varepsilon_i = \frac{1+2\lambda_i}{1+\lambda_i}\,\varepsilon_i.$$

Combined with (21),

$$\|\bar{\boldsymbol{r}}_i\| \le \frac{1+2\lambda_i}{1+\lambda_i}\,\varepsilon_i.$$

□

Proposition 4.4 (Fixed-point output representation of the system)**.** *Let* $\bar{\mathbf{F}} = [\bar{\boldsymbol{f}}_1, \dots, \bar{\boldsymbol{f}}_K]^\top$ *and* $\mathbf{F}_{\text{data}} = [\boldsymbol{f}_1^{\text{data}}, \dots, \boldsymbol{f}_K^{\text{data}}]^\top$ *denote the fixed-point output representations and the real-data representations of all models, respectively. Let* $\boldsymbol{\Lambda} = \operatorname{diag}(\lambda_1, \dots, \lambda_K)$ *and* $\mathbf{L} := \mathbf{I} - \mathbf{W}$*. We have*

$$\bar{\mathbf{F}} = (\mathbf{I} + \boldsymbol{\Lambda}\mathbf{L})^{-1}\big[\mathbf{F}_{\text{data}} + (\mathbf{I}+\boldsymbol{\Lambda})\bar{\mathbf{R}}\big], \tag{6}$$

where $\bar{\mathbf{R}} := [\bar{\boldsymbol{r}}_1, \dots, \bar{\boldsymbol{r}}_K]^\top$ *and* $\bar{\boldsymbol{r}}_i$ *satisfies* $\|\bar{\boldsymbol{r}}_i\| \le \frac{1+2\lambda_i}{1+\lambda_i}\,\varepsilon_i$.

Proof. By Proposition D.7, for each $i \in \mathcal{V}$,

$$\bar{\boldsymbol{f}}_i = \frac{1}{1+\lambda_i}\boldsymbol{f}_i^{\text{data}} + \frac{\lambda_i}{1+\lambda_i}\sum_{j\in\mathcal{V}} w_{ij}\,\bar{\boldsymbol{f}}_j + \bar{\boldsymbol{r}}_i.$$

Multiplying both sides by $(1+\lambda_i)$ and rearranging:

$$(1+\lambda_i)\bar{\boldsymbol{f}}_i - \lambda_i \sum_{j\in\mathcal{V}} w_{ij}\,\bar{\boldsymbol{f}}_j = \boldsymbol{f}_i^{\text{data}} + (1+\lambda_i)\bar{\boldsymbol{r}}_i.$$

Stacking over $i \in \mathcal{V}$ in matrix form:

$$(\mathbf{I}+\boldsymbol{\Lambda})\bar{\mathbf{F}} - \boldsymbol{\Lambda}\mathbf{W}\bar{\mathbf{F}} = \mathbf{F}_{\text{data}} + (\mathbf{I}+\boldsymbol{\Lambda})\bar{\mathbf{R}}.$$

Since $(\mathbf{I}+\boldsymbol{\Lambda}) - \boldsymbol{\Lambda}\mathbf{W} = \mathbf{I} + \boldsymbol{\Lambda}(\mathbf{I}-\mathbf{W}) = \mathbf{I} + \boldsymbol{\Lambda}\mathbf{L}$:

$$(\mathbf{I}+\boldsymbol{\Lambda}\mathbf{L})\bar{\mathbf{F}} = \mathbf{F}_{\text{data}} + (\mathbf{I}+\boldsymbol{\Lambda})\bar{\mathbf{R}}. \tag{23}$$

Let $\mathbf{M} := \mathbf{I} + \boldsymbol{\Lambda}\mathbf{L} = (\mathbf{I}+\boldsymbol{\Lambda}) - \boldsymbol{\Lambda}\mathbf{W}$. The entries of $\mathbf{M}$ are

$$M_{ii} = 1 + \lambda_i - \lambda_i w_{ii} = 1 + \lambda_i(1 - w_{ii}), \qquad M_{ij} = -\lambda_i w_{ij} \quad (j \ne i).$$

The off-diagonal absolute row sum is

$$\sum_{j\ne i} |M_{ij}| = \lambda_i \sum_{j\ne i} w_{ij} = \lambda_i(1-w_{ii}),$$

using $\sum_j w_{ij} = 1$ (row-stochasticity of $\mathbf{W}$). Comparing with the diagonal:

$$M_{ii} = 1 + \lambda_i(1-w_{ii}) > \lambda_i(1-w_{ii}) = \sum_{j\ne i} |M_{ij}|,$$

so $\mathbf{M}$ is strictly diagonally dominant, $\mathbf{M}$ is invertible.

Multiplying both sides of (23) by $\mathbf{M}^{-1} = (\mathbf{I}+\boldsymbol{\Lambda}\mathbf{L})^{-1}$:

$$\bar{\mathbf{F}} = (\mathbf{I}+\boldsymbol{\Lambda}\mathbf{L})^{-1}\big[\mathbf{F}_{\text{data}} + (\mathbf{I}+\boldsymbol{\Lambda})\bar{\mathbf{R}}\big].$$

□

D.8 Proofs of Theorem 4.5 and Corollary 4.6

Theorem 4.5 (Upper bound on system output diversity)**.** *Under Prop. 4.4, the system output diversity at any fixed point* $\bar{\Theta}$ *satisfies*

$$\mathcal{D}_{\text{out}} \le 2\big\|(\mathbf{I}+\boldsymbol{\Lambda}\mathbf{L})^{-1}\big\|_2^2\left[\mathcal{D}_{\text{data}} + \frac{1}{|\mathcal{V}|}\sum_{i\in\mathcal{V}}(1+\lambda_i)^2\,\|\bar{\boldsymbol{r}}_i\|^2\right]. \tag{7}$$

Proof. Let $\mathbf{M} := \mathbf{I} + \mathbf{\Lambda}\mathbf{L}$ and $\mathbf{P} := \mathbf{I}_K - \frac{1}{|\mathcal{V}|}\mathbf{1}\mathbf{1}^\top$ be the centering matrix.

For $\mathbf{A} \in \mathbb{R}^{K\times D}$ with rows $\boldsymbol{a}_i$ and column-mean $\boldsymbol{a}_{\text{avg}} = \frac{1}{|\mathcal{V}|}\sum_i \boldsymbol{a}_i$, the rows of $\mathbf{PA}$ are $\boldsymbol{a}_i - \boldsymbol{a}_{\text{avg}}$, so

$$\mathcal{D}_{\text{out}} = \frac{1}{|\mathcal{V}|}\|\mathbf{P}\bar{\mathbf{F}}\|_F^2, \qquad \mathcal{D}_{\text{data}} = \frac{1}{|\mathcal{V}|}\|\mathbf{P}\mathbf{F}_{\text{data}}\|_F^2.$$

By Proposition 4.4,

$$\bar{\mathbf{F}} = \mathbf{M}^{-1}\mathbf{F}_{\text{data}} + \mathbf{M}^{-1}(\mathbf{I}+\mathbf{\Lambda})\bar{\mathbf{R}}.$$

Since $\mathbf{W}$ is row-stochastic, $\mathbf{M}\mathbf{1} = \mathbf{1}$, hence $\mathbf{M}^{-1}\mathbf{1} = \mathbf{1}$.

Decomposing $\mathbf{F}_{\text{data}} = \mathbf{P}\mathbf{F}_{\text{data}} + \frac{1}{|\mathcal{V}|}\mathbf{1}\mathbf{1}^\top\mathbf{F}_{\text{data}}$ and applying $\mathbf{M}^{-1}$:

$$\begin{aligned}
\mathbf{P}\mathbf{M}^{-1}\mathbf{F}_{\text{data}} &= \mathbf{P}\mathbf{M}^{-1}\mathbf{P}\mathbf{F}_{\text{data}} + \tfrac{1}{|\mathcal{V}|}\mathbf{P}\mathbf{M}^{-1}\mathbf{1}\mathbf{1}^\top\mathbf{F}_{\text{data}} \\
&= \mathbf{P}\mathbf{M}^{-1}\mathbf{P}\mathbf{F}_{\text{data}} + \tfrac{1}{|\mathcal{V}|}\mathbf{P}\mathbf{1}\mathbf{1}^\top\mathbf{F}_{\text{data}} && (\mathbf{M}^{-1}\mathbf{1} = \mathbf{1}) \\
&= \mathbf{P}\mathbf{M}^{-1}\mathbf{P}\mathbf{F}_{\text{data}} && (\mathbf{P}\mathbf{1} = 0).
\end{aligned}$$

Hence

$$\mathbf{P}\bar{\mathbf{F}} = \mathbf{P}\mathbf{M}^{-1}\mathbf{P}\mathbf{F}_{\text{data}} + \mathbf{P}\mathbf{M}^{-1}(\mathbf{I}+\mathbf{\Lambda})\bar{\mathbf{R}}.$$

Apply $\|a+b\|_F^2 \le 2\|a\|_F^2 + 2\|b\|_F^2$.

$$\|\mathbf{P}\bar{\mathbf{F}}\|_F^2 \le 2\|\mathbf{P}\mathbf{M}^{-1}\mathbf{P}\mathbf{F}_{\text{data}}\|_F^2 + 2\|\mathbf{P}\mathbf{M}^{-1}(\mathbf{I}+\mathbf{\Lambda})\bar{\mathbf{R}}\|_F^2.$$

Using $\|\mathbf{A}\mathbf{B}\|_F \le \|\mathbf{A}\|_2\|\mathbf{B}\|_F$ and $\|\mathbf{P}\|_2 = 1$:

$$\|\mathbf{P}\mathbf{M}^{-1}\mathbf{P}\mathbf{F}_{\text{data}}\|_F^2 \le \|\mathbf{M}^{-1}\|_2^2 \cdot \|\mathbf{P}\mathbf{F}_{\text{data}}\|_F^2 = |\mathcal{V}|\,\|\mathbf{M}^{-1}\|_2^2 \cdot \mathcal{D}_{\text{data}}.$$

Similarly,

$$\|\mathbf{P}\mathbf{M}^{-1}(\mathbf{I}+\mathbf{\Lambda})\bar{\mathbf{R}}\|_F^2 \le \|\mathbf{M}^{-1}\|_2^2 \cdot \|(\mathbf{I}+\mathbf{\Lambda})\bar{\mathbf{R}}\|_F^2.$$

Since $(\mathbf{I}+\mathbf{\Lambda})\bar{\mathbf{R}}$ has rows $(1+\lambda_i)\bar{\boldsymbol{r}}_i$,

$$\|(\mathbf{I}+\mathbf{\Lambda})\bar{\mathbf{R}}\|_F^2 = \sum_{i\in\mathcal{V}}(1+\lambda_i)^2\|\bar{\boldsymbol{r}}_i\|_2^2.$$

By the residual bound from Proposition D.7, $\|\bar{\boldsymbol{r}}_i\| \le \frac{1+2\lambda_i}{1+\lambda_i}\varepsilon_i$. Hence

$$(1+\lambda_i)^2\|\bar{\boldsymbol{r}}_i\|^2 \le (1+\lambda_i)^2 \cdot \frac{(1+2\lambda_i)^2}{(1+\lambda_i)^2}\,\varepsilon_i^2 = (1+2\lambda_i)^2\,\varepsilon_i^2.$$

Therefore

$$\|(\mathbf{I}+\mathbf{\Lambda})\bar{\mathbf{R}}\|_F^2 \le \sum_{i\in\mathcal{V}}(1+2\lambda_i)^2\,\varepsilon_i^2.$$

Substituting into Step 3 and dividing by $|\mathcal{V}|$:

$$\mathcal{D}_{\text{out}} = \frac{1}{|\mathcal{V}|}\|\mathbf{P}\bar{\mathbf{F}}\|_F^2 \le 2\|\mathbf{M}^{-1}\|_2^2\left[\mathcal{D}_{\text{data}} + \frac{1}{|\mathcal{V}|}\sum_{i\in\mathcal{V}}(1+2\lambda_i)^2\varepsilon_i^2\right].$$

□

Corollary 4.6 (Diversity Bound for Symmetric Interaction)**.** *Suppose the interaction graph is undirected ($\mathbf{W} = \mathbf{W}^\top$), the synthetic weight is homogeneous ($\lambda_i \equiv \lambda$) and the feature discrepancy vanishes ($\bar{\boldsymbol{r}}_i = \mathbf{0}$). Let μ_2 denote the algebraic connectivity (the second smallest eigenvalue of $\mathbf{L}$). Then:*

$$\mathcal{D}_{\text{out}} \le \frac{1}{(1+\lambda\mu_2)^2}\mathcal{D}_{\text{data}} \tag{8}$$

Proof. Under the assumption $\bar{\boldsymbol{r}}_i = 0$, hence $\bar{\mathbf{R}} = \mathbf{0}$, and Proposition 4.4 reduces to

$$\bar{\mathbf{F}} = \mathbf{M}^{-1}\mathbf{F}_{\text{data}}, \qquad \mathbf{M} := \mathbf{I} + \lambda\mathbf{L}.$$

Step 1: Spectral analysis of $\mathbf{M}$. Since $\mathbf{W}$ is symmetric, $\mathbf{L} = \mathbf{I} - \mathbf{W}$ is symmetric.

Because $\mathbf{W}$ is row-stochastic and nonnegative, $\mathbf{L}$ is positive semi-definite with eigenvalues $0 = \mu_1 < \mu_2 \le \cdots \le \mu_K$ and orthonormal eigenvectors $\mathbf{u}_1 = \frac{1}{\sqrt{K}}\mathbf{1}, \mathbf{u}_2, \ldots, \mathbf{u}_K$.

Hence $\mathbf{M}^{-1}$ has the same eigenvectors with eigenvalues $\frac{1}{1+\lambda\mu_k}$. The function $g(\mu) = \frac{1}{1+\lambda\mu}$ is monotonically decreasing for $\lambda > 0$, so $\mathbf{M}^{-1}$'s largest eigenvalue is $\frac{1}{1+\lambda\mu_1} = 1$ (along $\mathbf{u}_1$), and its second largest is $\frac{1}{1+\lambda\mu_2}$ (along $\mathbf{u}_2$).

Step 2: Centering and restriction to $\mathbf{1}^\perp$. Let $\mathbf{P} := \mathbf{I}_K - \frac{1}{K}\mathbf{1}\mathbf{1}^\top$ be the centering matrix.

$$\mathcal{D}_{\text{out}} = \frac{1}{K}\|\mathbf{P}\bar{\mathbf{F}}\|_F^2, \qquad \mathcal{D}_{\text{data}} = \frac{1}{K}\|\mathbf{P}\mathbf{F}_{\text{data}}\|_F^2.$$

Since $\mathbf{M}^{-1}\mathbf{1} = \mathbf{1}$ (the eigenvector for $\mu_1 = 0$), $\mathbf{M}^{-1}$ commutes with $\mathbf{P}$ in this symmetric case, and $\mathbf{P}$ projects onto $\mathbf{1}^\perp$:

$$\mathbf{P}\bar{\mathbf{F}} = \mathbf{P}\mathbf{M}^{-1}\mathbf{F}_{\text{data}} = \mathbf{M}^{-1}\mathbf{P}\mathbf{F}_{\text{data}}.$$

Step 3: Bound via restricted spectral norm. On $\mathbf{1}^\perp$, $\mathbf{M}^{-1}$ has spectral norm

$$\|\mathbf{M}^{-1}|_{\mathbf{1}^\perp}\|_2 = \max_{k\geq 2}\frac{1}{1+\lambda\mu_k} = \frac{1}{1+\lambda\mu_2}.$$

Hence

$$\|\mathbf{P}\bar{\mathbf{F}}\|_F^2 = \|\mathbf{M}^{-1}\mathbf{P}\mathbf{F}_{\text{data}}\|_F^2 \leq \|\mathbf{M}^{-1}|_{\mathbf{1}^\perp}\|_2^2 \cdot \|\mathbf{P}\mathbf{F}_{\text{data}}\|_F^2 = \frac{1}{(1+\lambda\mu_2)^2}\|\mathbf{P}\mathbf{F}_{\text{data}}\|_F^2.$$

Dividing both sides by K:

$$\mathcal{D}_{\text{out}} \leq \frac{1}{(1+\lambda\mu_2)^2}\,\mathcal{D}_{\text{data}}.$$

□

E Closed-Form Spectral Validation of the Networked Dynamics

To complement the neural generative-model experiments, we consider a controlled networked dynamical system in which the fixed point, Jacobian, comparison matrix, and cycle gains are all available in closed form. While the experiments in Sec. 5 test whether the qualitative mechanisms predicted by our theory persist under finite data, imperfect optimization, and model misspecification, the purpose of this experiment is different: it provides a direct numerical validation.

Setup. We consider $K = 4$ interacting nodes, denoted by $\{A, B, C, D\}$, with each node represented by a two-dimensional state $\theta_i \in \mathbb{R}^2$. We assign the heterogeneous reference centroids

$$a_A = \begin{bmatrix}1\\0\end{bmatrix}, \qquad a_B = \begin{bmatrix}-1\\0\end{bmatrix}, \qquad a_C = \begin{bmatrix}0\\1\end{bmatrix}, \qquad a_D = \begin{bmatrix}0\\-1\end{bmatrix}$$

and let

$$A = \begin{bmatrix}a_A^\top\\a_B^\top\\a_C^\top\\a_D^\top\end{bmatrix} \in \mathbb{R}^{4\times 2}$$

To introduce controlled model heterogeneity, we associate each source node with a sensitivity coefficient $c = (1,\ 1,\ 1.8,\ 1)^\top$ and $\Gamma = \text{diag}(c)$. Thus, node C has a larger local sensitivity than the other nodes. Importantly, the same sensitivity vector c is used for every interaction topology; no topology-specific parameters are tuned.

For an interaction matrix W, the network evolves according to the affine update

$$\Theta^{t+1} = \frac{A + \lambda W\Gamma\Theta^t}{1+\lambda}$$

where $\Theta^t = [(\theta_A^t)^\top, \ldots, (\theta_D^t)^\top]^\top$ collects the four node states and $\lambda \geq 0$ controls the strength of the interaction term. As in the main experiments, we report this strength using the synthetic-data ratio $s := \frac{\lambda}{1+\lambda} \in [0, 1)$.

We evaluate the first four interaction systems used in the main paper, as shown in Fig. 4. System 1 corresponds to isolated self-consumption, Systems 2 and 3 share the same ring topology but use different interaction strengths, and System 4 is fully connected. These are the same four controlled topologies used in the main experiments.

Closed-form fixed point and exact Jacobian. Whenever $(1+\lambda)I - \lambda W\Gamma$ is nonsingular, Eq. (3) admits the unique fixed point $\bar{\Theta} = [(1+\lambda)I - \lambda W\Gamma]^{-1}A$. Unlike the realistic generative-model experiments, this controlled setting therefore allows us to directly measure the distance to the fixed point, $\|\Theta^t - \bar{\Theta}\|_F$. And its Jacobian is constant and available exactly:$J = s(W\Gamma) \otimes I_2 = sW\Gamma$. Consequently, the error $E^t := \Theta^t - \bar{\Theta}$ follows $E^{t+1} = JE^t$. Hence, for a generic perturbation with a nonzero component in the dominant eigenspace,

the asymptotic convergence factor is $\rho(J)$. All graph-level quantities, including the comparison matrix D and the cycle gain $\gamma(C)$, follow the definitions in Sec. 3.

Direct validation of the spectral criterion We first fix $\lambda = 1$, corresponding to a synthetic-data ratio $s = 0.5$. For each interaction topology, we initialize the dynamics in a small neighborhood of the exact fixed point $\Theta^0 = \bar{\Theta} + 10^{-3}Z$, where the entries of Z are independently sampled from a standard Gaussian distribution. We run 250 iterations and record $e_t := \|\Theta^t - \bar{\Theta}\|_F$. To estimate the asymptotic contraction rate, we fit

$$\log e_t \approx b + t \log r_{\text{emp}}$$

over the latter half of the trajectory before numerical saturation, and report the resulting r_{emp}.

Tbl. 2 reports the exact and empirical spectral quantities. The empirical convergence rate agrees with $\rho(J)$ to three decimal places in every system. Moreover, $\rho(J) \leq \rho(D)$ holds throughout, consistent with the comparison principle in Thm. 3.7.

Table 2: Closed-form spectral validation at $\lambda = 1$ ($s = 0.5$). The measured asymptotic rate r_{emp} matches the exact Jacobian spectral radius $\rho(J)$, while $\rho(D)$ upper-bounds $\rho(J)$. The last column reports the largest cycle gain and verifies the lower bound in Prop. 3.10.

System	$\|W\|_2$	$\rho(J)$	r_{emp}	$\rho(D)$	$\max_C \gamma(C)$
Sys. 1 (isolated)	1.00	0.900	0.900	0.900	0.900
Sys. 2 (weak ring)	1.00	0.606	0.606	0.700	0.450
Sys. 3 (strong ring)	1.00	0.582	0.582	0.674	0.604
Sys. 4 (dense)	1.00	0.600	0.600	0.900	0.225

All four systems have the same spectral norm $\|W\|_2 = 1$, yet their exact convergence factors range from 0.582 to 0.900. Thus, the interaction norm alone cannot distinguish their dynamics. In contrast, the topology-aware spectral quantities retain information about how heterogeneous local sensitivities are propagated through the network. This illustrates why the norm-based condition in Cor. 3.8 can be strictly more conservative than the spectral criterion of Thm. 3.7..In all systems,$\max_C \gamma(C) \leq \rho(D)$, as predicted by Prop. 3.10. The bound can be tight, as in the isolated system, or substantially loose in graphs containing many simultaneously interacting paths.

Critical synthetic-data ratio Because both J and D scale linearly with $s = \lambda/(1+\lambda)$, the two stability thresholds are available in closed form. For the systems considered here, $s_c^J = \frac{1}{\rho(W\Gamma)}$, $s_c^D = \frac{1}{\rho(\text{diag}(M_1,\ldots,M_K)W)}$. Table 3 reports these two thresholds. For every system, $s_c^D \leq s_c^J$, showing quantitatively that the D-based criterion lies on the conservative side of the exact Jacobian threshold.

Table 3: Critical synthetic-data ratio $s = \lambda/(1+\lambda)$. The comparison-matrix threshold s_c^D is a sufficient stability certificate and is therefore no larger than the exact Jacobian threshold s_c^J in all tested systems.

System	s_c^D	s_c^J
Sys. 1 (isolated)	0.556	0.556
Sys. 2 (weak ring)	0.714	0.825
Sys. 3 (strong ring)	0.742	0.860
Sys. 4 (dense)	0.556	0.833

The gap between the two thresholds quantifies the conservativeness introduced by replacing heterogeneous block sensitivities with the nonnegative comparison system. This gap depends on topology. For example, the bound is tight in the isolated System 1, whereas in the dense System 4 the sufficient threshold is 0.556, compared with the exact Jacobian threshold 0.833. Thus, the comparison matrix provides a tractable topology-aware certificate without requiring it to exactly recover the full Jacobian dynamics.

Summary. This controlled experiment complements the neural generative-model experiments in two respects. First, because the fixed point and Jacobian are available exactly, it directly verifies that $\rho(J)$ governs the asymptotic convergence rate. Second, it numerically illustrates both levels of graph-based comparison developed in the theory. Together with the realistic experiments in Sec. 5, these results distinguish two complementary forms of validation: the controlled system directly verifies the spectral predictions under an exactly analyzable update map, while the neural-model experiments test whether the same qualitative mechanisms remain visible under finite data, imperfect optimization, and heterogeneous model architectures.

Algorithm 1 Iterative Self-Consuming Retraining

Input: $\{\mathcal{D}_i^{\text{real}}\}_{i=1}^K$, $\boldsymbol{\theta}_i^0$ trained on $\mathcal{D}_i^{\text{real}}$, interaction weights $\mathbf{W}$, synthetic mixing ratio α_i
for round $t = 1$ to T **do**
  for each model $j = 1, \ldots, K$ **do**
    Generate synthetic samples $\mathcal{D}_j^t \sim p_{\boldsymbol{\theta}_j^{t-1}}$
  end for
  for each model $i = 1, \ldots, K$ **do**
    Construct training set $\mathcal{D}_i^t = (1-\alpha_i)\mathcal{D}_i^{\text{real}} \cup \alpha_i \bigcup_j w_{ij}\mathcal{D}_j^{(t)}$
    Update $\boldsymbol{\theta}_i^t$ by training on $\mathcal{D}_i^t$
  end for
end for
Output: $\{\boldsymbol{\theta}_i^T\}_{i=1}^K$

F Experiments with Synthetic Gaussian Data

In this section, we present synthetic data experiments to empirically validate the theoretical results of the main paper, providing a low-dimensional illustration that complements large-scale image experiments.

Table 4: Model architectures and data access.

#	Model	Architecture	Stability	Diveristy	Retrain Steps
A	DDPM	MLP (hidden dimensions 128)	$t \in \{0, \cdots, 7\}$	$t \in \{0, 4\}$	100
B	CFM	MLP	$t \in \{0, \cdots, 7\}$	$t \in \{2, 6\}$	50
C	GAN	MLP	$t \in \{0, \cdots, 7\}$	$t \in \{3, 5\}$	30
D	DDPM	MLP (hidden dimensions 256)	$t \in \{0, \cdots, 7\}$	$t \in \{1, 7\}$	100

Dataset. The synthetic dataset is a two-dimensional dataset constructed from mixtures of isotropic Gaussian distributions. The global data distribution follows a standard 8-mode Gaussian, where eight mode centers are uniformly distributed on a circle of radius 2. The coordinates of these centers are: $\mu_t = 2 \times \left(\cos(\frac{t\pi}{4}), \sin(\frac{t\pi}{4})\right)$ where $t = 0, 1, 2, \ldots, 7$. We use two real-data assignments to evaluate different aspects of the coupled system. The same data for all models is used for stability experiments, and the heterogeneous assignment is used for diversity experiments, where each model i is associated with a subset of these Gaussian modes, as specified in Table 4. Each data point is independently and uniformly sampled from one of the assigned mode centers, with isotropic Gaussian noise (mean of 0 and standard deviation of 0.02) added: $x = \mu_t + \epsilon$, $\epsilon \sim \mathcal{N}(0, 0.02^2 I_2)$

Models. Each model is implemented as a lightweight generative model operating in $\mathbb{R}^2$. We consider three standard generative paradigms: DDPM [20], CFM [33], GANs [18]. All models are implemented using multilayer perceptrons and are intentionally kept simple to minimize architectural inductive bias. For DDPM models, we use two network architectures with different capacities. All models are trained using identical optimization procedures within each paradigm.

Each model is first trained on its own real data distribution during a warm-up phase until convergence. This initialization ensures that all models start from a converged solution of their real-data objectives. A detailed summary of model architectures, data preferences is reported in Table 4.

Training Settings. All experiments follow a fixed iterative retraining protocol. In each retraining round, every model is trained on a mixture of real samples and synthetic samples generated in the previous round. The total training set size is fixed to 10,000 samples, with the proportion of synthetic samples determined by the mixing ratio $\alpha_i = \frac{\lambda_i}{\lambda_i + 1}$. In each round, $10,000$ samples are generated to calculate the metrics. All experiments are repeated with three different random seeds, and results are reported as the mean across runs.

Metrics. We measure the distributional discrepancy between generated samples and reference distributions using the 1-Wasserstein [35] distance $\mathcal{WD}$. To quantify diversity across models, we compute a system-level diversity metric $\mathcal{D}$ defined as the average deviation of each model's generated sample mean from the global mean of all models: $\mathcal{D} = \frac{1}{K}\sum_{i=1}^K \left\| \mu_i - \frac{1}{K}\sum_{j=1}^K \mu_j \right\|_2$

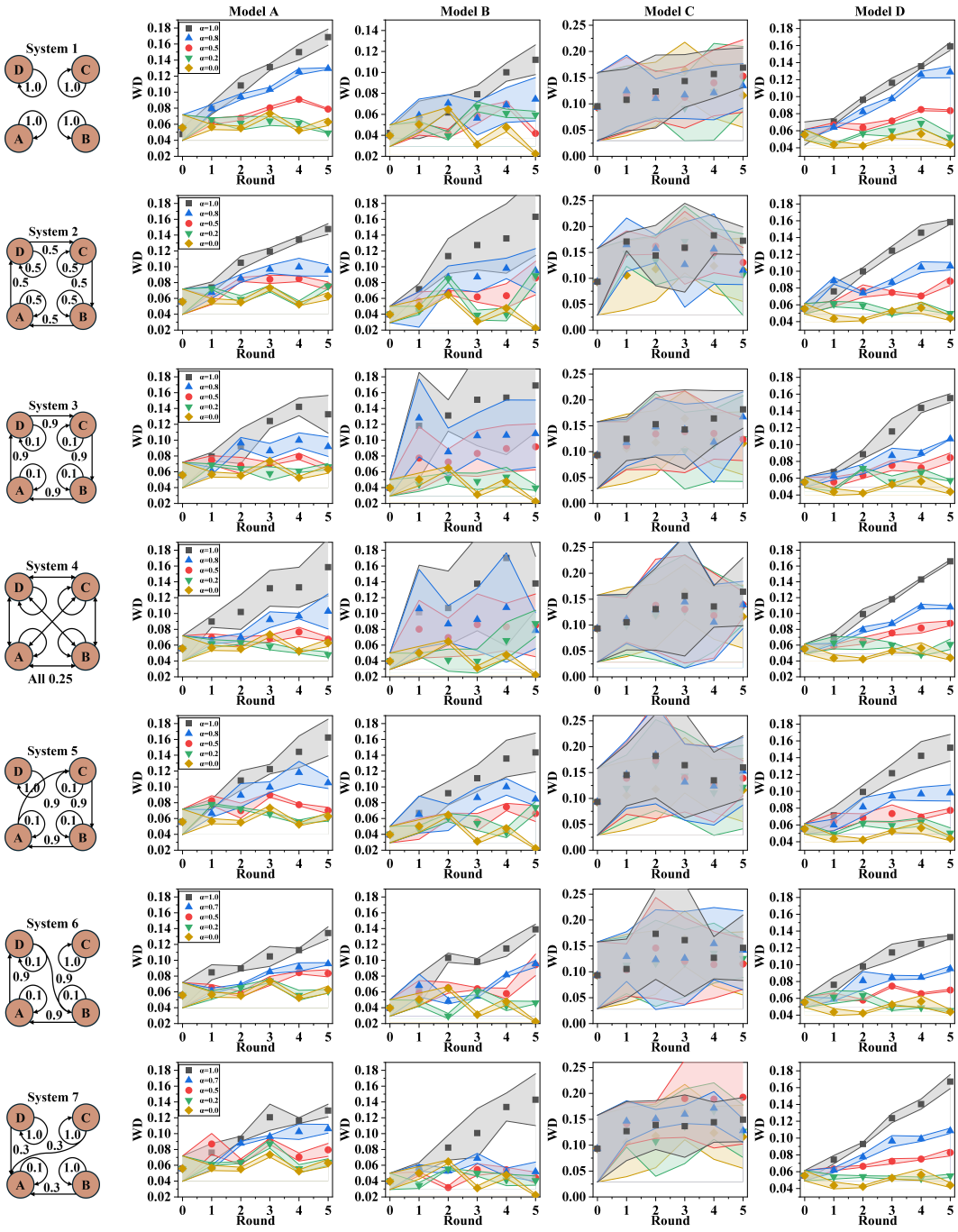

Figure 4: Stability experiments when all models trained on full dataset. Each row corresponds to a different interaction graph. For each row, the left figure illustrates the direction and strength of synthetic data flow, while the right figures report the Wasserstein distance ($\mathcal{WD}$) between generated samples and real data for each model. Results are shown under different synthetic mixing strengths, with $\alpha_i = \alpha$ shared across models.

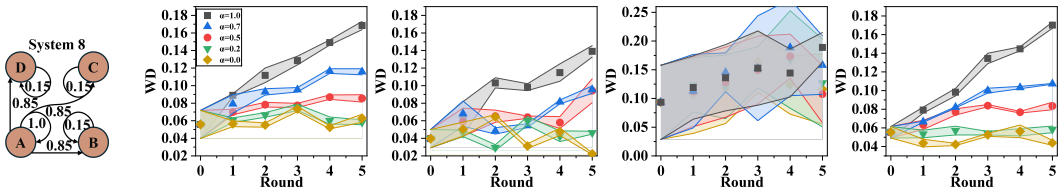

Figure 5: Stability experiments when all models trained on full dataset (continued).

F.1 Stability Experimental Results

We first present the results of stability, where all models access to the same real data distribution $t \in \{0, \cdots, 7\}$. We fix the synthetic mixing ratio to be identical $\alpha_i = \alpha_j = \alpha$. Fig. 4 (and its continuation Fig. 5) illustrates different interaction graph structures with varying synthetic data flows and reports the Wasserstein distance $\mathcal{WD}$ across three independent runs with different seeds between generated and real samples.

We analyze eight interaction graph structures across four models under the homogeneous data setting. Since all models have the same real-data target in this experiment, persistent growth in $\mathcal{WD}$ indicates instability of the retraining dynamics rather than disagreement in the desired data distribution. Although all models are trained on the same real data distribution and use an identical synthetic mixing strength α, their retraining behaviors differ substantially due to both model-specific stability and interaction topology.

Model C adopts a GAN-based architecture and exhibits weaker baseline stability compared to the other models. This instability is reflected throughout the experiments. Across nearly all graph structures, Model C shows larger WD fluctuations across retraining rounds.

System 1 serves as a baseline that approximates isolated self-consuming retraining, where models primarily train on their own generated samples. In this setting, increasing α consistently leads to higher $\mathcal{WD}$ for all models, with the effect being most pronounced for *Model C*. This behavior aligns with the theoretical expectation that stronger synthetic mixing amplifies local instability even in the absence of cross-model interactions.

We next examine the effect of cycles by comparing **System 1**, **System 3**, and **System 5**, which differ in their maximum cycle lengths. Under the same synthetic mixing strength, longer cycles do not necessarily result in more severe degradation. In particular, **System 5** exhibits the highest $\mathcal{WD}$ mean and variance across models, despite not having the longest cycle. This observation is consistent with the theoretical result that stability is governed by the dominant cycle gain rather than cycle length alone.

The influence of interaction strength is illustrated by comparing **System 2** and **System 3**. **System 2** satisfies the theoretical stability condition based on the interaction norm, while **System 3** exceeds this threshold. Empirically, **System 3** displays faster $\mathcal{WD}$ growth and increased variance across models under the same α. These results support the theoretical prediction that stronger interactions reduce the stability margin of self-consuming retraining. At the same time, differences among systems with similar interaction norms (**System 1**, **System 2** and **System 4**) indicate that norm-based conditions are conservative and do not fully capture structural effects.

Finally, comparing **System 5** and **System 6** highlights the role of model participation within the interaction graph. Although the two systems share similar topological structures, **System 5** involves the unstable *Model C* more directly in high-gain cycles. As a result, all models in **System 5** experience higher $\mathcal{WD}$ and larger variance, whereas the corresponding effect is attenuated in **System 6**. This comparison suggests that instability introduced by a single weak model can propagate through the network when it participates in dominant interaction cycles, thereby affecting the entire system.

Overall, the homogeneous data experiments confirm the main theoretical predictions. Stability depends jointly on synthetic mixing strength, interaction structure, and the presence of unstable models. Moreover, graph topology shapes how local instabilities are amplified or contained, beyond what can be explained by interaction norms alone.

We further visualize the generated samples in Fig. 6 ($\alpha = 1.0$) and Fig. 7 ($\alpha = 0.2$).

F.2 Diversity Experimental Results

We next study the diversity, where models are trained on different subsets of the underlying modes as shown in Table 4. In this setting, we focus on a system-level diversity metric $\mathcal{D}$, which captures the degree of differentiation among models without requiring a shared reference distribution as shown in Fig. 8. We further visualize the generated samples to qualitatively illustrate the stability and diversity of the system as shown in Fig. 9 and Fig. 10.

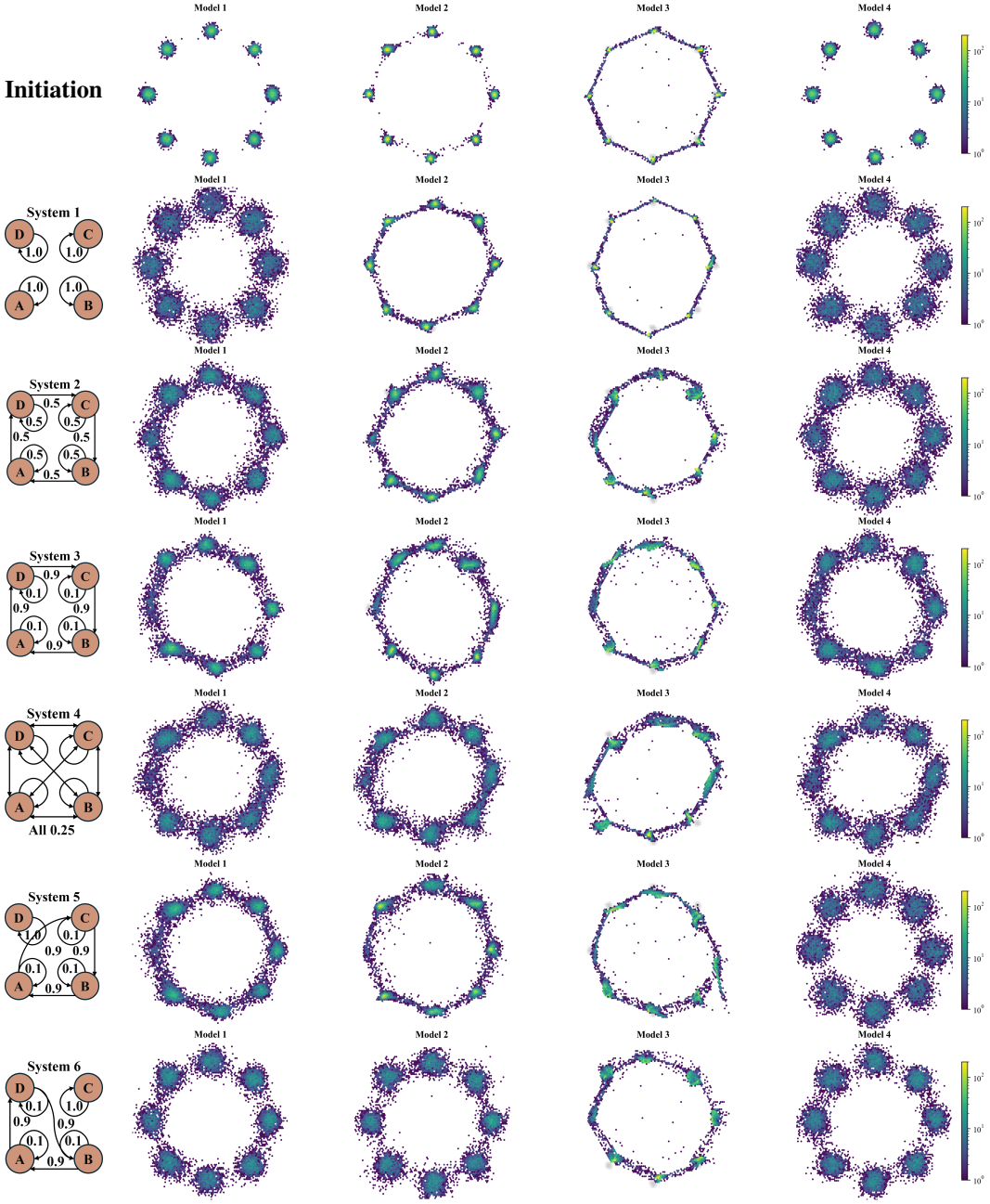

Figure 6: Generated samples when all models trained on full dataset with different interaction graphs. All models use $\alpha_i = \alpha = 1.0$.

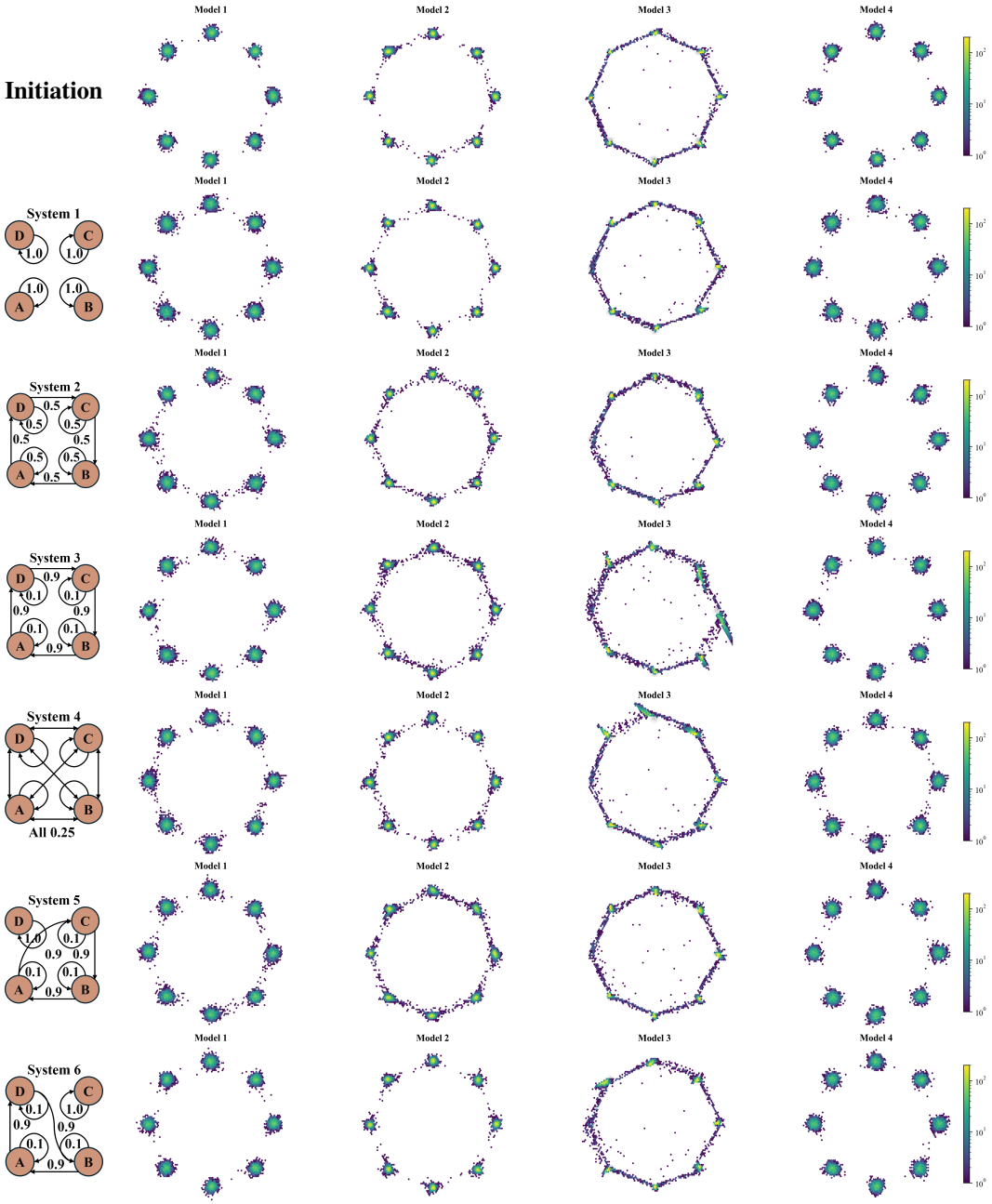

Figure 7: Generated samples when all models trained on full dataset. All models use $\alpha_i = \alpha = 0.2$.

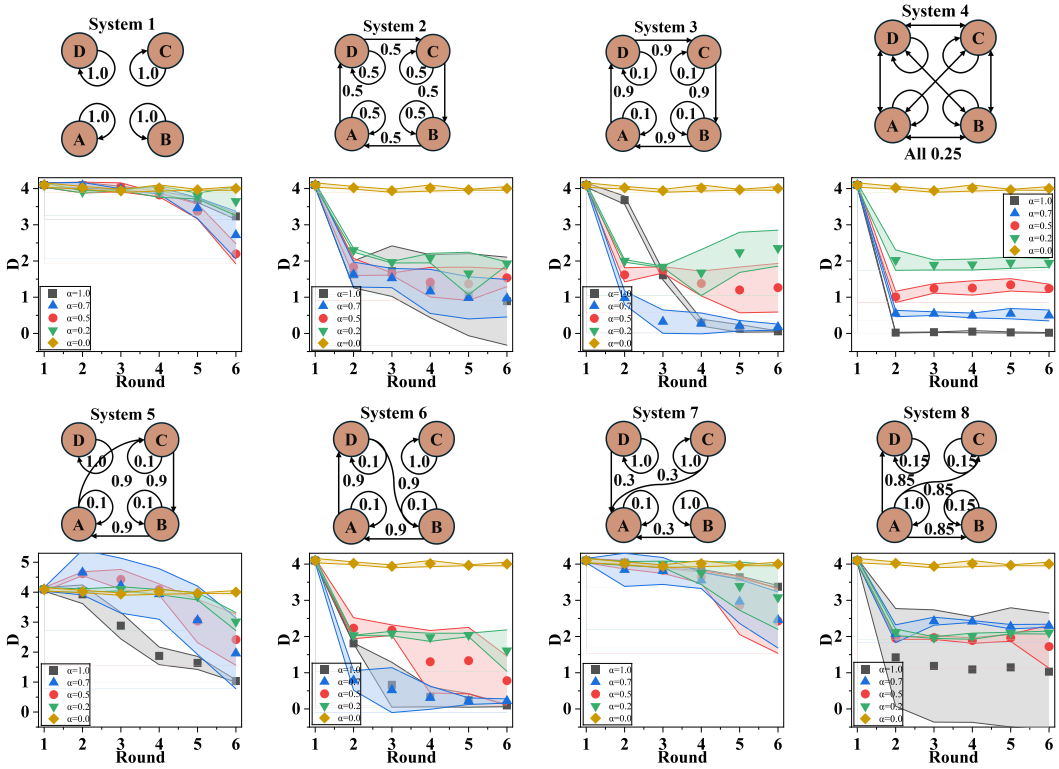

Figure 8: Diversity across models under heterogeneous data with different interaction graphs. Each figure reports the system-level diversity metric $\mathcal{D}$ across retraining rounds under different synthetic mixing strengths α.

Across all interaction graphs, increasing the synthetic mixing strength α generally leads to a reduction in diversity. When synthetic data dominates training, models increasingly align with one another despite heterogeneous real data access, resulting in homogenization.

System 1 provides a baseline corresponding to nearly isolated self-consuming retraining. In this case, large α can lead to elevated diversity. However, this increase does not reflect meaningful differentiation. Instead, it arises from independent model collapse, where unstable retraining causes different models to drift toward distinct degenerate distributions. This interpretation is supported by the large variance observed across runs and is consistent with the theory, which characterizes stability rather than the semantic quality of diversity.

Interaction topology also plays a decisive role in shaping heterogeneous results. This trend is strongest in densely connected systems such as **System 4**, where diversity collapses rapidly even for moderate α. This behavior is consistent with the theoretical prediction. On the other side, systems with comparable interaction magnitudes can exhibit qualitatively different behaviors. For example, **System 2** and **System 4** share similar interaction norms, yet **System 4** experiences rapid diversity loss. This contrast reflects structural effects that go beyond conditions, which are sufficient but conservative.

The influence of unstable models is most clearly illustrated in **System 5** and **System 6**. *Model C*, which adopts a GAN-based architecture and exhibits weaker baseline stability, participates directly in cycles in **System 5**. As a result, synthetic errors generated by *Model C* are repeatedly amplified and propagated, leading to elevated diversity accompanied by large variance. This behavior reflects system-level instability rather than stable heterogeneity. In contrast, systems where the same model is weakly coupled or excluded from dominant cycles show significantly attenuated effects. These observations align with the theoretical prediction that instability introduced by a single node can dominate global behavior when embedded in dominant interaction cycles.

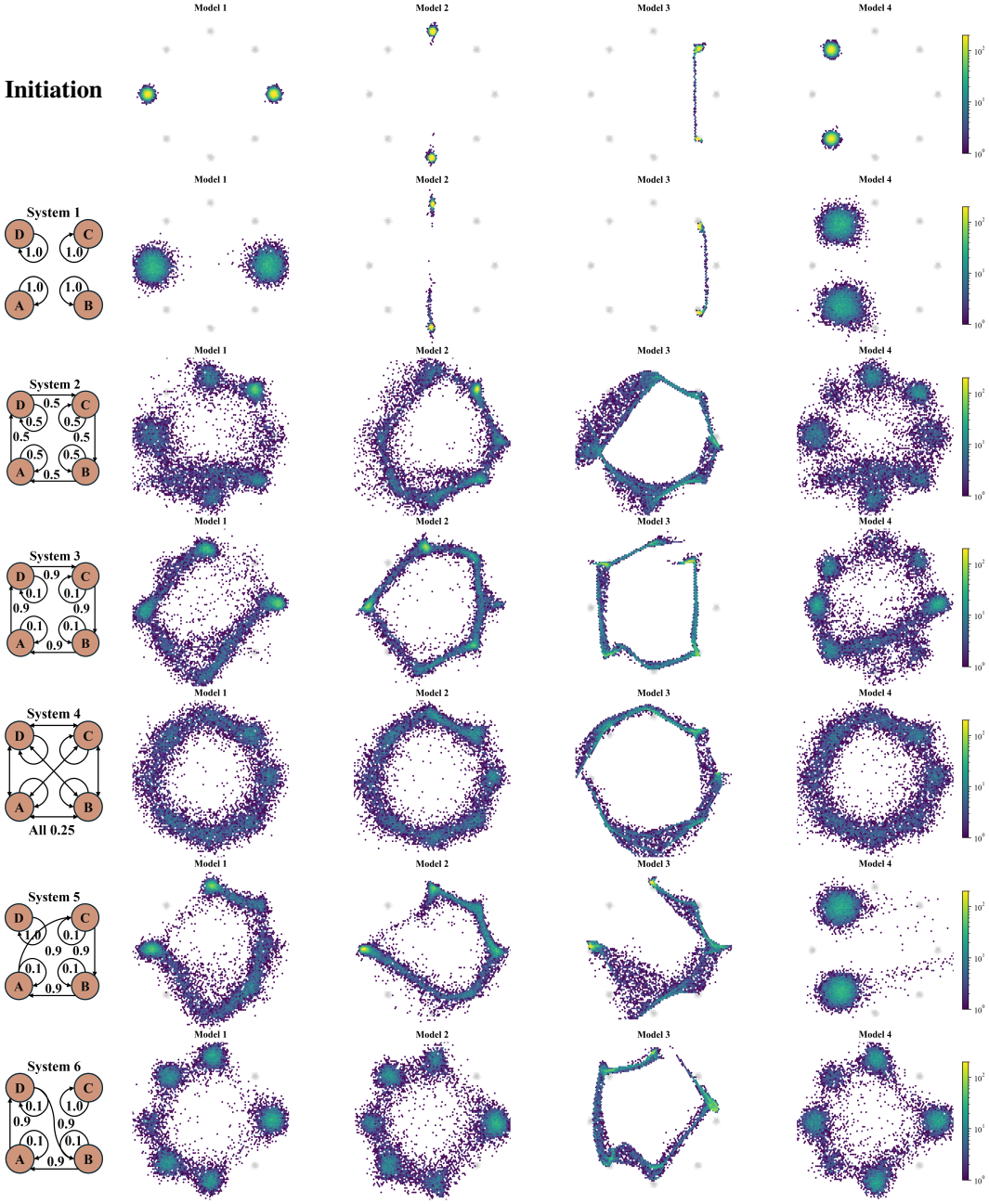

Figure 9: Generated samples under the heterogeneous data with different interaction graphs. All models use $\alpha_i = \alpha = 1.0$.

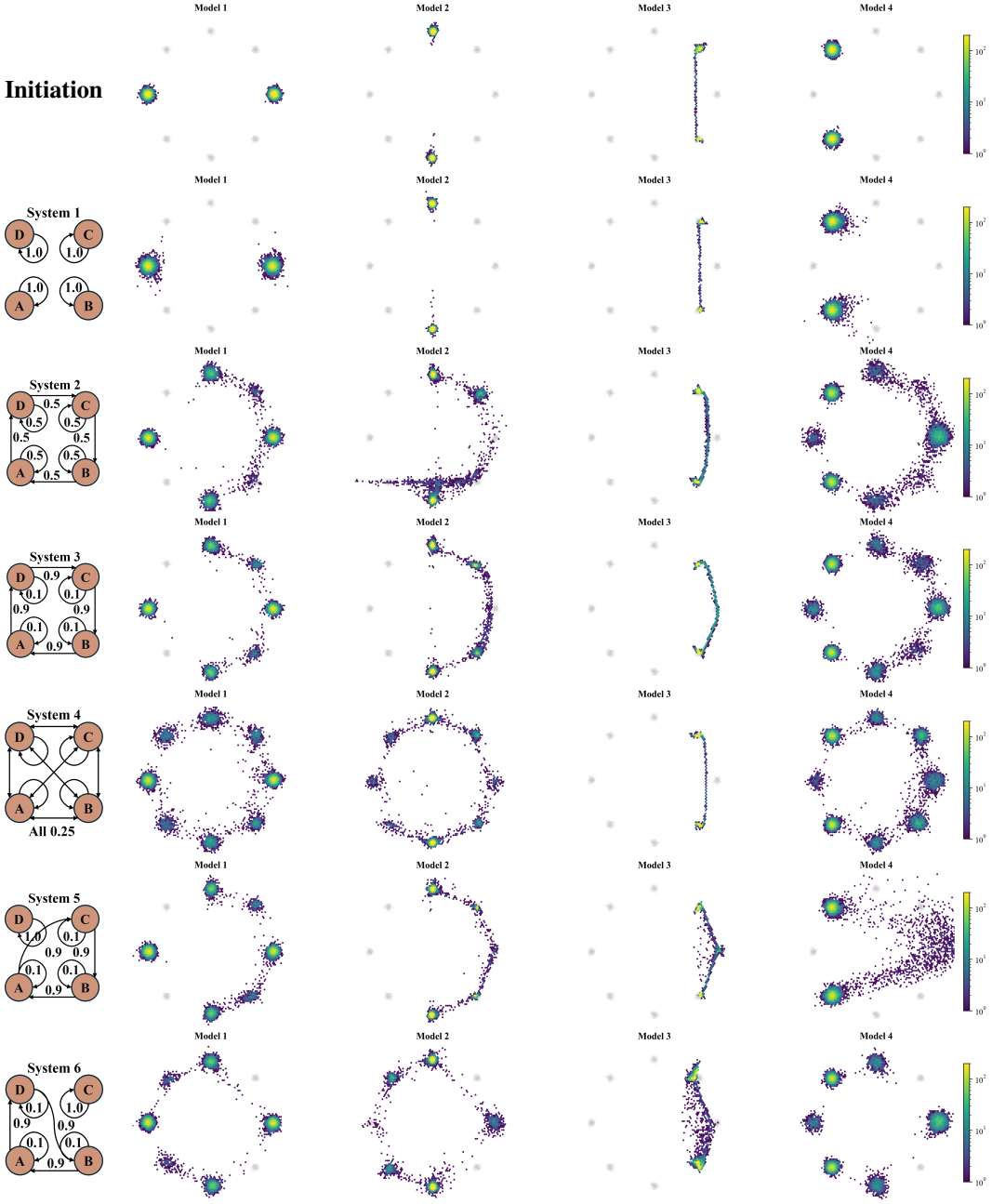

Figure 10: Generated samples under the heterogeneous data with different interaction graphs. All models use $\alpha_i = \alpha = 0.2$.

G Additional Results on CIFAR-10 Datasets

We conduct all the experiments on a server which has two Intel Xeon 6326 CPU and 6 Nvidia A6000 GPU.

While FID summarizes overall distributional similarity, it does not distinguish whether degradation comes primarily from reduced sample fidelity or from loss of coverage. We therefore report precision and recall, together with qualitative samples, to better understand how networked self-consuming retraining affects both image quality and distributional support.

Table 5: Model architectures and data access.

#	Model	Variants	Stability	Diversity	Retrain Steps
A	DDPM	VLB	class $\{0, \cdots, 9\}$	class $\{0, \cdots, 9\}$	100
B	CFM	OT-CFM	class $\{0, \cdots, 9\}$	class $\{3, 4\}$	600
C	CFM	iCFM	class $\{0, \cdots, 9\}$	class $\{5, 6\}$	600
D	DDPM	hybrid	class $\{0, \cdots, 9\}$	class $\{7, 8, 9\}$	100

Details in models. For real-data experiments, we use four generative models from two paradigms, with two variants per paradigm. Using multiple variants within the same broad paradigm allows us to separate paradigm-level effects from objective-level effects. For example, two diffusion models may share the same denoising backbone but respond differently to synthetic-data accumulation because their training objectives impose different likelihood and regression biases.

For **DDPM variants**, we include two diffusion models based on the improved diffusion framework with a UNet backbone. The first variant is a *VLB* diffusion model, which is trained using a variational lower bound objective with a learned noise schedule and likelihood-based parameterization. The second variant is a *Hybrid* diffusion model, which combines the likelihood-style objective with a simplified regression loss used in practical diffusion training. While both variants share the same diffusion forward process and network backbone, they differ in their training objectives and, consequently, in their optimization behavior and sample quality trade-offs.

For **CFM variants**, we include two flow-matching models implemented with neural ODE parameterizations. The first variant follows *OT-CFM*, where the training pairs are constructed using an optimal-transport-inspired coupling, yielding flow trajectories that encourage coherent mass transport between source and target distributions. The second variant follows *iCFM*, which uses an implicit coupling and a simplified construction of flow-matching targets, leading to a different inductive bias on the learned transport dynamics. Both variants optimize flow-matching objectives but differ in how the intermediate training targets are defined and how probability mass is coupled across time.

We initialize all models from publicly released pretrained checkpoints. To ensure a fair comparison, we follow the original training configurations for each model and modify only the number of training steps per round. Dataset specifications and step counts are provided in Table 5.

Additional Results for Stability. In addition to FID, we report precision as shown in Fig. 12 and recall curves as shown in Fig. 11 to provide a more fine-grained characterization of generation quality. These two metrics are particularly useful for diagnosing different failure modes. A drop in precision suggests that generated images become less realistic or less aligned with the data manifold, whereas a drop in recall indicates that the model covers fewer modes or classes of the target distribution.

Additional Results Diveristy. We additionally visualize generated samples at $r = 10$ under $\alpha = 1.0$ as shown in Fig. 13 and $\alpha = 0.2$ as shown in Fig. 14. These qualitative results illustrate how different levels of synthetic mixing affect sample quality in practice.

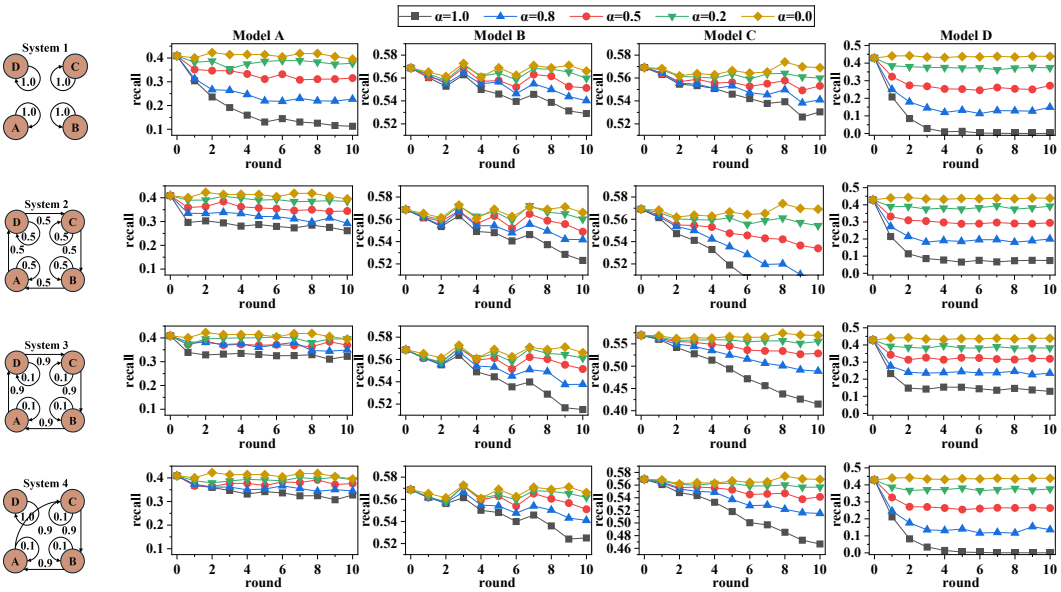

Figure 11: Stability experiments when all models trained on full dataset. Each row corresponds to a different interaction graph. For each row, the left figure illustrates the direction and strength of synthetic data flow, while the right figures report the recall between generated samples and real data for each model. Results are shown under different synthetic mixing strengths, with $\alpha_i = \alpha$ shared across models.

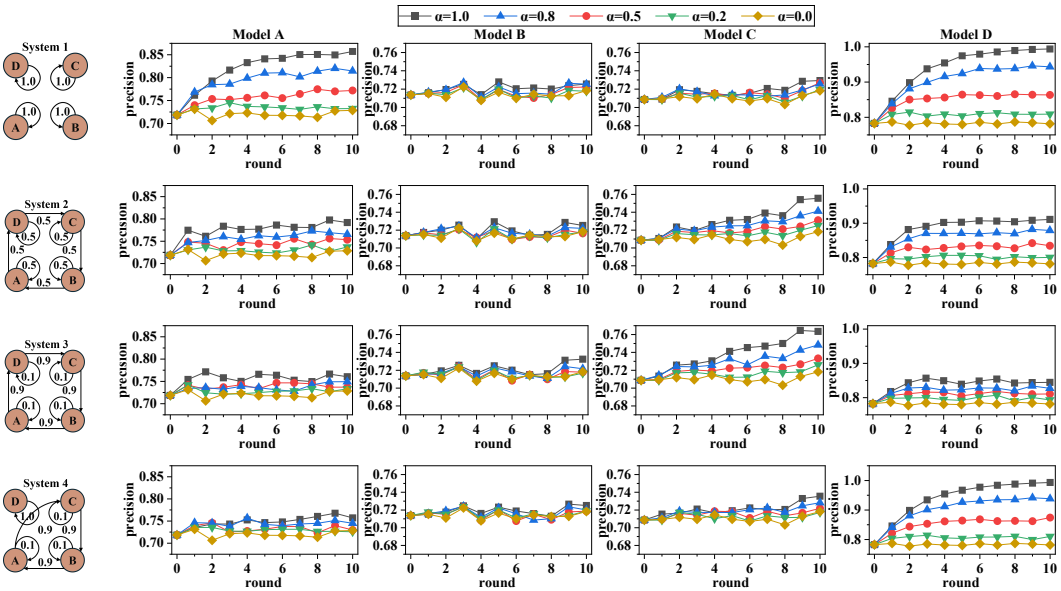

Figure 12: Stability experiments when all models trained on full dataset. Each row corresponds to a different interaction graph. For each row, the left figure illustrates the direction and strength of synthetic data flow, while the right figures report the precision between generated samples and real data for each model. Results are shown under different synthetic mixing strengths, with $\alpha_i = \alpha$ shared across models.

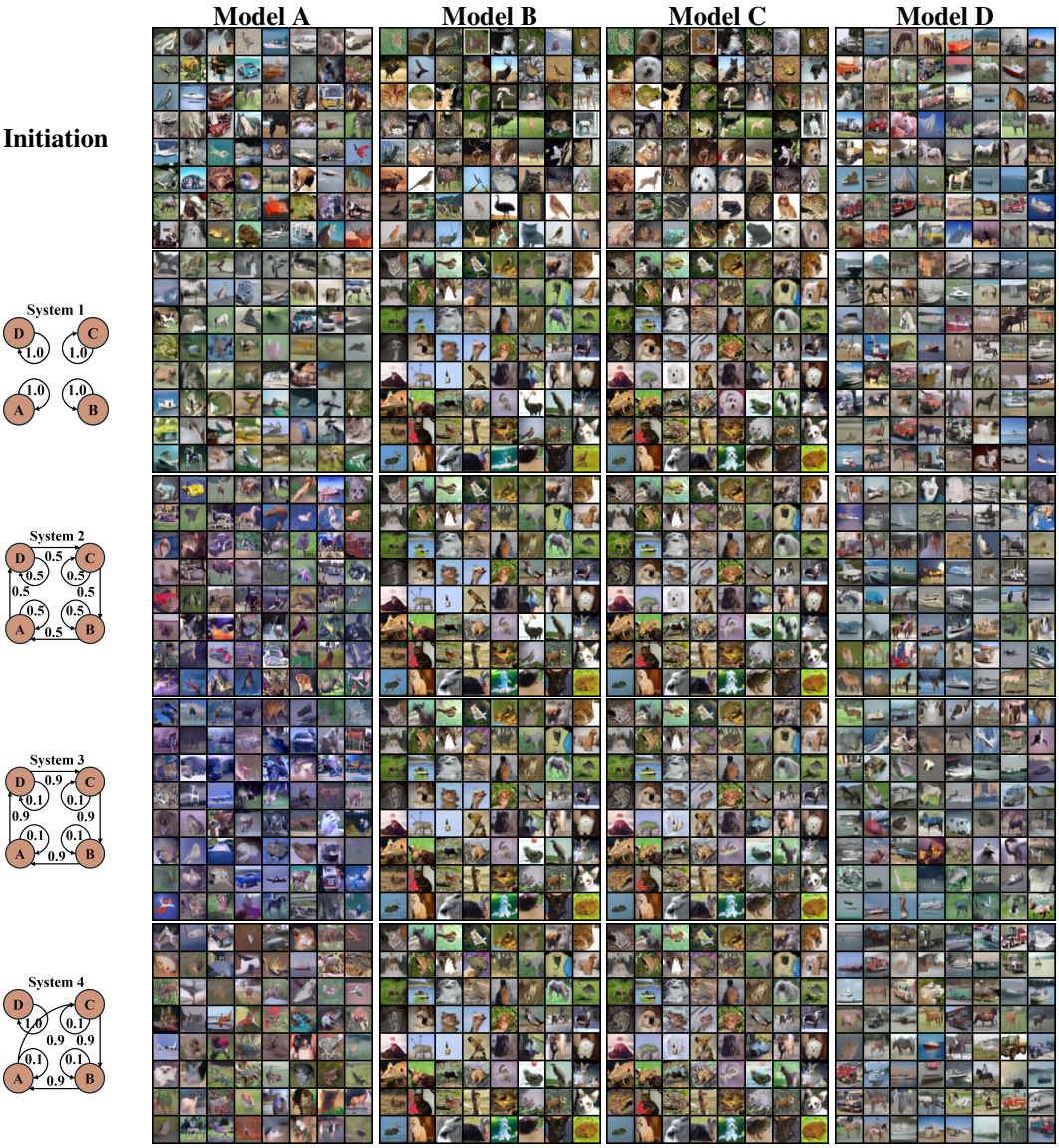

Figure 13: Generated samples under the heterogeneous real data setting with different interaction graphs. All models use $\alpha_i = \alpha = 1.0$.

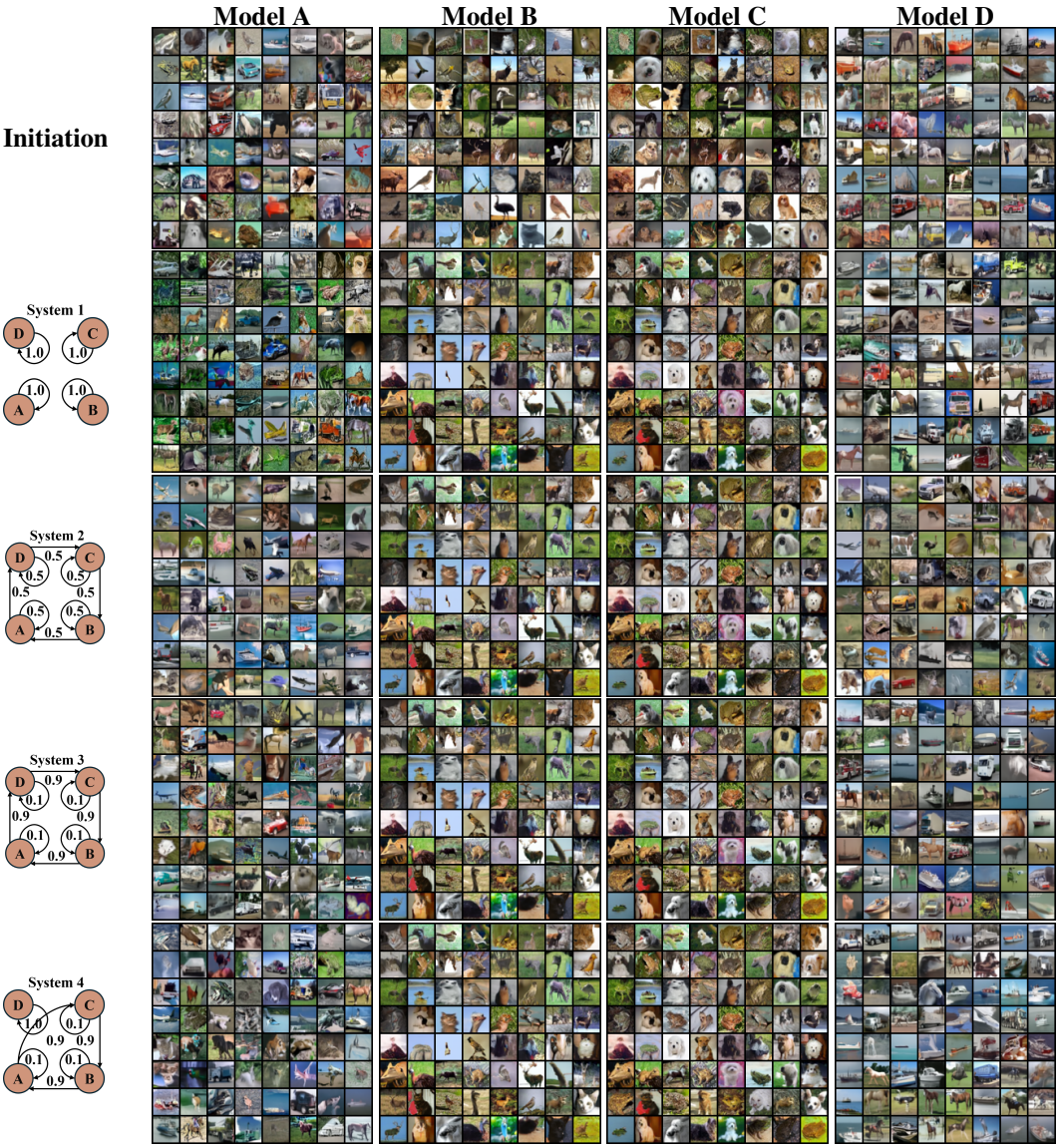

Figure 14: Generated samples under the heterogeneous real data setting with different interaction graphs. All models use $\alpha_i = \alpha = 0.2$.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and introduction state that the paper develops a graph-based theoretical framework for networked self-consuming generative models, with local stability and fixed-point diversity analyses. The scope is stated as a local, fixed-point-conditional analysis, and the experiments are presented as qualitative validation.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: The paper discusses limitations in Appx. A, including the fixed-point-conditional nature of the analysis, local regularity assumptions, finite-sample and nonconvex training effects, and directions for broader theoretical extensions.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

   Justification: The theoretical assumptions are stated in Sec. 3. The complete proofs of the main theoretical statements provided in Appx. D.

Guidelines:

- The answer [N/A] means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: Sec. 5 describes key experimental settings (e.g., datasets, models, interaction graphs, retraining procedure , and evaluation metrics). Additional experimental details are provided in Appx. E, F and G.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: The paper uses public datasets and publicly available model implementations. We provide experimental details sufficient to reproduce the reported results and provide the link to the code

   Guidelines:

   - The answer [N/A] means that paper does not include experiments requiring code.

- Please see the NeurIPS code and data submission guidelines (https://neurips.cc/public/guides/CodeSubmissionPolicy) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (https://neurips.cc/public/guides/CodeSubmissionPolicy) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: Sec. 5 describes the datasets, model types, retraining protocol, interaction graphs, synthetic mixing ratios, and evaluation metrics. Additional experimental details, including synthetic experiments and supplementary metrics, are provided in Appx. E, F and G.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: The synthetic-data experiments report means with error bars over three independent runs. For the CIFAR-10 experiments, full independent reruns are computationally expensive, so we present qualitative trends and vary the random seed across retraining rounds.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).
   - It should be clear whether the error bar is the standard deviation or the standard error of the mean.
   - It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
   - For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
   - If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

Answer: [Yes]

Justification: The appendix reports the compute environment used for the experiments.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
- The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
- The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

Answer: [Yes]

Justification: The work uses public datasets and pretrained models for theoretical and empirical analysis of generative-model feedback loops. There is no violation of the code of ethics.

Guidelines:

- The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
- If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
- The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

Answer: [Yes]

Justification: Appx. B discusses positive impacts, such as diagnosing instability, homogenization, and diversity loss in generative ecosystems, as well as possible misuse, such as steering model ecosystems toward narrower or biased output distributions.

Guidelines:

- The answer [N/A] means that there is no societal impact of the work performed.
- If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

Answer: [N/A]

Justification: The paper does not release new high-risk pretrained models, image generators, or scraped datasets. The experiments use public datasets and existing model implementations for analysis.

Guidelines:

- The answer [N/A] means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [Yes]

Justification: The paper cites the datasets, model architectures, pretrained classifiers, and public repositories used in the experiments.

Guidelines:

- The answer [N/A] means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

Answer: [N/A]

Justification: The paper does not introduce a new dataset, benchmark, or pretrained model as a primary asset. The contribution is a theoretical framework with empirical illustrations.

Guidelines:

- The answer [N/A] means that the paper does not release new assets.
- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

Answer: [N/A]

Justification: The paper does not involve crowdsourcing, user studies, or research with human subjects.

Guidelines:

- The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

Answer: [N/A].

Justification: The paper does not involve human subjects, crowdsourcing, or user studies, so IRB approval is not applicable.

Guidelines:

- The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

Answer: [N/A]

Justification: LLMs are not used as an important, original, or non-standard component of the core methodology. Any use for writing, editing, or formatting does not affect the scientific claims, theory, or experiments.

Guidelines:

- The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.