Novelty Assessment Report

Paper: Neon: Negative Extrapolation From Self-Training Improves Image Generation
PDF URL: https://openreview.net/pdf?id=kpLRYtPGt3
Venue: ICLR 2026 Conference Submission
Year: 2026
Report Generated: 2025-12-29

Abstract

Scaling generative AI models is bottlenecked by the scarcity of high-quality training data. The ease of synthesizing from a generative model suggests using (unverified) synthetic data to augment a limited corpus of real data for the purpose of fine-tuning in the hope of improving performance. Unfortunately, however, the resulting positive feedback loop leads to model autophagy disorder (MAD, aka model collapse) that results in a rapid degradation in sample quality and/or diversity. In this paper, we introduce Neon (for Negative Extrapolation frOm self-traiNing), a new learning method that turns the degradation from self-training into a powerful signal for self-improvement. Given a base model, Neon first fine-tunes it on its own self-synthesized data but then, counterintuitively, reverses its gradient updates to extrapolate away from the degraded weights. We prove that Neon works because typical inference samplers that favor high-probability regions create a predictable anti-alignment between the synthetic and real data population gradients, which negative extrapolation corrects to better align the model with the true data distribution. Neon is remarkably easy to implement via a simple post-hoc merge that requires no new real data, works effectively with as few as 1k synthetic samples, and typically uses less than 1\% additional training compute. We demonstrate Neon's universality across a range of architectures (diffusion, flow matching, autoregressive, and inductive moment matching models) and datasets (ImageNet, CIFAR-10, and FFHQ). In particular, on ImageNet 256x256, Neon elevates the xAR-L model to a new state-of-the-art FID of 1.02 with only 0.36\% additional training compute.

Core Task Landscape

This paper addresses: **Improving Generative Models Through Negative Extrapolation From Synthetic Data**

A total of **6 papers** were analyzed and organized into a taxonomy with **7 categories**.

Taxonomy Overview

The research landscape has been organized into the following main categories:

- **Self-Training and Synthetic Data Feedback Mechanisms**
- **Constraint-Guided Generation With Negative Examples**
- **Hybrid Generative-Discriminative Training Frameworks**

Complete Taxonomy Tree

- Improving Generative Models Through Negative Extrapolation From Synthetic Data Survey Taxonomy
- Self-Training and Synthetic Data Feedback Mechanisms
  - Negative Extrapolation and Gradient Reversal ★ (1 papers)
  - [0] Neon: Negative Extrapolation From Self-Training Improves Image Generation (Anon et al., 2026) View paper
  - Model Collapse Prevention Through Data Quality Control (1 papers)
  - [1] Self-improving diffusion models with synthetic data (Alemohammad, 2024) View paper
- Constraint-Guided Generation With Negative Examples
  - Engineering Design Constraint Satisfaction (1 papers)
  - [2] Constraining Generative Models for Engineering Design with Negative Data (Regenwetter, 2023) View paper
  - Synthetic Hard Negatives for Representation Learning (1 papers)
  - [5] Fake & Square: Training Self-Supervised Vision Transformers with Synthetic Data and Synthetic Hard Negatives (Floros, 2025) View paper
  - Synthetic Negatives in Recommendation Systems (1 papers)
  - [6] Time and Space Aggregation Recommendation Model Based on Synthetic Negative Samples (Yang Xingyao, 2022) View paper
- Hybrid Generative-Discriminative Training Frameworks
  - Discriminative Learning of Generative Distributions (1 papers)
  - [4] Learning generative models via discriminative approaches (Zhuowen Tu, 2007) View paper
  - Generative-Discriminative Fusion for Anomaly Detection (1 papers)
  - [3] Hybrid Open-Set Segmentation With Synthetic Negative Data (Matej GrciÄ, 2023) View paper

Narrative

Core task: improving generative models through negative extrapolation from synthetic data. This field explores how generative systems can be refined by leveraging not only positive examples but also synthetic negative or contrastive signals. The taxonomy reveals three main branches. Self-Training and Synthetic Data Feedback Mechanisms encompasses methods that iteratively refine models using their own outputs, often employing negative extrapolation or gradient reversal to steer away from undesirable generations—exemplified by approaches like Self-improving diffusion[1] and Neon[0]. Constraint-Guided Generation With Negative Examples focuses on incorporating explicit negative constraints or contrastive data to shape the generation process, as seen in Negative Data Constraints[2] and Synthetic Negative Samples[6]. Hybrid Generative-Discriminative Training Frameworks blends generative and discriminative objectives, drawing on ideas from earlier work such as Generative via Discriminative[4] and more recent studies like Hybrid Open-Set[3] and Fake and Square[5], to balance likelihood maximization with classification or rejection of poor samples.

Across these branches, a central theme is the trade-off between exploiting model-generated data for self-improvement and avoiding the pitfalls of reinforcing model biases or collapsing onto narrow modes. Many studies explore how to construct informative negative examples—whether by perturbing real data, sampling from the model's own distribution, or using auxiliary discriminators—and how to integrate these signals without destabilizing training. Neon[0] sits squarely within the Self-Training and Synthetic Data Feedback Mechanisms branch, emphasizing negative extrapolation and gradient reversal to push the model away from synthetic failure modes. Its approach contrasts with constraint-based methods like Negative Data Constraints[2], which impose hard rules on generation, and with hybrid frameworks such as Fake and Square[5], which interleave generative and discriminative updates. By focusing on gradient-level steering rather than explicit constraints or separate discriminators, Neon[0] offers a streamlined path to refining generative quality through self-generated negative signals.

Related Works in Same Category

No sibling papers were found in the same taxonomy leaf. A taxonomy-subtopic-level comparison will be produced instead.

Taxonomy-Level Summary

Both subtopics address the challenge of maintaining generative model quality when training on synthetic data, but employ fundamentally different strategies. The original leaf uses gradient reversal to actively extrapolate away from degraded model weights, while the sibling focuses on preventing degradation through careful control of synthetic data quality and distribution. These represent complementary approaches: one corrective (reversing damage) and one preventive (avoiding damage).

Similarities: - Both aim to prevent or mitigate model collapse when fine-tuning generative models on synthetic data - Both recognize that naive self-training on synthetic data can degrade model quality - Both are concerned with the autophagy/feedback loop problem in generative model training

Differences: - Original leaf uses gradient reversal as an active correction mechanism, while sibling uses data filtering/selection as a preventive measure - Original leaf operates at the optimization level (reversing gradient updates), while sibling operates at the data level (controlling input distribution) - Original leaf explicitly extrapolates away from degraded weights, while sibling maintains quality by avoiding exposure to degrading data - Original leaf excludes data filtering methods, which are central to the sibling category

Suggested Search Directions: - Hybrid approaches combining gradient reversal with data quality control - Comparative studies measuring when gradient-based correction outperforms data filtering - Methods that detect when to apply reversal vs. when to filter data

Sibling Subtopics

- **Model Collapse Prevention Through Data Quality Control** (leaves: 1, papers: 1)
- Scope: Approaches maintaining generative model quality by controlling synthetic data distribution to avoid autophagy disorder.
- Exclude: Excludes gradient-based extrapolation methods and discriminative constraint enforcement; see sibling and parent categories.

Contributions Analysis

Overall novelty summary. The paper introduces Neon, a method that fine-tunes generative models on self-synthesized data and then reverses gradient updates to extrapolate away from degraded weights. According to the taxonomy, this work resides in the 'Negative Extrapolation and Gradient Reversal' leaf under 'Self-Training and Synthetic Data Feedback Mechanisms'. Notably, this leaf contains only the original paper itself—no sibling papers are listed—suggesting this specific combination of gradient reversal and negative extrapolation from self-training represents a relatively unexplored niche within the broader self-training literature.

The taxonomy reveals that the broader parent category 'Self-Training and Synthetic Data Feedback Mechanisms' includes a sibling leaf focused on 'Model Collapse Prevention Through Data Quality Control', which addresses similar autophagy concerns but through filtering rather than gradient manipulation. Adjacent branches explore 'Constraint-Guided Generation With Negative Examples' (using explicit negative constraints in engineering, vision, and recommendation domains) and 'Hybrid Generative-Discriminative Training Frameworks' (combining generative and discriminative objectives). Neon diverges from these by operating purely at the gradient level without external discriminators or hard constraints, positioning it as a distinct approach within the self-improvement paradigm.

Among the three contributions analyzed, the core Neon method examined two candidates and found one potentially refutable prior work, indicating some overlap in the limited search scope of 22 papers. The theoretical proof of anti-alignment between synthetic and population gradients examined ten candidates with none clearly refuting it, suggesting this formalization may be relatively novel. The demonstration of universality across architectures and datasets also examined ten candidates without clear refutation. These statistics reflect a focused semantic search rather than exhaustive coverage, so the apparent novelty should be interpreted cautiously within this bounded exploration.

Given the limited search scope and the paper's placement in an otherwise-empty taxonomy leaf, Neon appears to occupy a distinct methodological position—combining self-training feedback with gradient reversal in a way not directly captured by the examined prior work. However, the single refutable candidate for the core method suggests that related ideas may exist in the broader literature beyond the 22 papers examined. The theoretical and empirical contributions show fewer direct overlaps within this search, though a more exhaustive review would be needed to assess their full novelty.

This paper presents **3 main contributions**, each analyzed against relevant prior work:

Contribution 1: Neon method for negative extrapolation from self-training

Description: Neon is a post-processing method that improves generative models by first fine-tuning them on self-synthesized data to obtain degraded weights, then reversing the gradient updates via negative extrapolation. This simple parameter merge requires no new real data, works with as few as 1k synthetic samples, and uses less than 1% additional training compute.

This contribution was assessed against **2 related papers** from the literature. Papers with potential prior art are analyzed in detail with textual evidence; others receive brief assessments.

---

1. Self-improving diffusion models with synthetic data

URL: View paper

Prior Art Analysis

Self-improving diffusion[1] demonstrates that a nearly identical approach was published prior to the ORIGINAL paper. Both methods fine-tune a base model on self-synthesized data to obtain degraded weights, then use negative extrapolation via parameter merging to improve performance. Self-improving diffusion[1] explicitly presents the formula $s\theta(xt, t) = (1 + \omega)s\theta r(xt, t) - \omega s\theta s(xt, t)$, which is mathematically equivalent to the ORIGINAL paper's $\theta neon = (1 + w)\theta r - w\theta s$. Both methods require no new real data, work with limited synthetic samples, and use minimal additional training compute. The candidate paper was published on arXiv in August 2024, predating the ORIGINAL submission to ICLR 2026.

Evidence

Evidence 1 - **Rationale**: Both papers present mathematically equivalent negative extrapolation formulas. The ORIGINAL uses $\theta neon = (1 + w)\theta r - w\theta s$ for parameters, while Self-improving diffusion[1] uses $s\theta(xt, t) = (1 + \omega)s\theta r(xt, t) - \omega s\theta s(xt, t)$ for score functions. The core innovation—reversing degradation from self-training via weighted parameter combination—is identical. - **Original**: neon (negative extrapolation from self-training) exploits this anti-alignment through a simple parameter merge. given a base model with parameters θr

trained on real data, we first apply the naïve self-training approach: we generate synthetic samples and briefly fine-tune to obtain the parameters θs... - **Candidate**: sims: extrapolating to self-improvement. let us unpack the sims steps outlined in algorithm 1 in the introduction. consider a base diffusion model characterized by the score function sθr (xt, t) that was trained on real data samples drawn from a target real data distribution pr. [...] this motivates...

Evidence 2 - **Rationale**: Self-improving diffusion[1]'s Algorithm 1 describes the same procedure: (1) train base model on real data, (2) generate synthetic data from base model, (3) fine-tune on synthetic data to get auxiliary model, (4) combine via negative extrapolation. Both methods require no new real data and use temporary synthetic datasets that are discarded after training. - **Original**: we introduceneon, a deceptively simple post-processing method that improves generative models by reversing their degradation on self-generated data (section 3). in contrast to existing methods for synthetic data augmentation, neon requires no additional real training data, no access to the original ... - **Candidate**: algorithm 1 sims procedure input: training dataset d hyperparameters: synthetic dataset size ns, guidance strength ω, training budget b 1: train base diffusion model: use dataset d to train the diffusion model using standard training, resulting in the score function sθr (xt, t). 2: generate auxiliar...

Evidence 3 - **Rationale**: Both papers identify the same core insight: self-training on synthetic data creates a systematic degradation signal that can be reversed. The ORIGINAL calls this 'anti-alignment with the real-data population gradient,' while Self-improving diffusion[1] describes it as steering 'away from the non-ideal synthetic data manifold.' The conceptual foundation is identical. - **Original**: our key insight is that the degradation due to self-training is not random noise but rather a power signal that is anti-aligned with the real-data population gradient. neon (negative extrapolation from self-training) exploits this anti-alignment through a simple parameter merge. - **Candidate**: our key insight is that, to most effectively exploit synthetic data in training a generative model, we need to change how we employ synthetic data. instead of naïvely training a model on synthetic data as though it were real, sims guides the model towards better performance but away from the pattern...

---

2. Improving Compound-Protein Interaction Prediction by Self-Training with Augmenting Negative Samples.

URL: View paper

Brief Assessment

Self-Training Augmenting Negatives[17] focuses on compound-protein interaction prediction using self-training with augmented negative samples, which is a completely different domain and methodology from Neon's image generation approach via negative extrapolation from degraded model weights.

Contribution 2: Theoretical proof of anti-alignment between synthetic and population gradients

Description: The authors rigorously prove that mode-seeking inference samplers induce a predictable anti-alignment between synthetic data gradients and real data population gradients. This theoretical result explains why reversing the degradation direction through negative extrapolation reduces the true data risk and guarantees Neon's effectiveness.

This contribution was assessed against **10 related papers** from the literature. Papers with potential prior art are analyzed in detail with textual evidence; others receive brief assessments.

---

1. Domain-adaptive neural networks improve supervised machine learning based on simulated population genetic data

URL: View paper

Brief Assessment

Domain-adaptive Neural Networks[23] addresses simulation mis-specification in population genetics using domain adaptation techniques, not theoretical analysis of gradient anti-alignment in generative model training.

---

2. iNeRF: Inverting Neural Radiance Fields for Pose Estimation

URL: View paper

Brief Assessment

iNeRF[22] focuses on pose estimation by inverting neural radiance fields through gradient-based optimization of camera poses, not on synthetic data gradients or population gradient alignment in generative model training.

---

3. Gradient projection Newton algorithm for sparse collaborative learning using synthetic and real datasets of applications

URL: View paper

Brief Assessment

Gradient Projection Newton[25] focuses on sparse collaborative learning with double-sparsity constraints for multi-dataset optimization, not on synthetic data generation or gradient anti-alignment in generative models.

---

4. Simulation and the reality gap: Moments in a prehistory of synthetic data

URL: View paper

Brief Assessment

Reality Gap[20] examines the historical and epistemological aspects of the reality gap in simulation technologies, not mathematical proofs about gradient alignment in machine learning optimization.

---

5. Mind the gap between synthetic and real: Utilizing transfer learning to probe the boundaries of stable diffusion generated data

URL: View paper

Brief Assessment

Mind the Gap[21] focuses on transfer learning with Stable Diffusion synthetic data for image classification, not on theoretical analysis of gradient alignment in self-training generative models.

---

6. iSDF: Real-Time Neural Signed Distance Fields for Robot Perception

URL: View paper

Brief Assessment

iSDF[24] focuses on real-time neural signed distance field reconstruction for robot perception using depth images, not on theoretical analysis of gradient alignment in generative model training with synthetic data.

---

7. Synthetic text generation for training large language models via gradient matching

URL: View paper

Brief Assessment

Synthetic Text Gradient[27] focuses on generating synthetic text for LLMs via gradient matching in discrete token space, not on proving anti-alignment between synthetic and population gradients in generative image models. The theoretical frameworks address fundamentally different problems.

---

8. Robin hood and matthew effects: Differential privacy has disparate impact on synthetic data

URL: View paper

Brief Assessment

Robin Hood Matthew[26] focuses on differential privacy's disparate impact on synthetic data quality across demographic subgroups, not on gradient anti-alignment mechanisms in self-training or negative extrapolation methods.

---

9. Gradient matching for domain generalization

URL: View paper

Brief Assessment

Gradient Matching[18] focuses on inter-domain gradient matching for domain generalization across multiple training domains, not on synthetic vs. real data gradients in generative model training.

---

10. Information Diffusion Modeling in Social Networks: A Comparative Analysis of Delay Mechanisms Using Population Dynamics

URL: View paper

Brief Assessment

Information Diffusion Modeling[19] focuses on information spread in social networks using delay mechanisms and network structures, not on synthetic data gradients or generative model training.

Contribution 3: Demonstration of Neon's universality across model architectures and datasets

Description: The authors empirically validate Neon across diverse generative model families including diffusion models, flow matching, autoregressive models, and few-step generators on multiple standard benchmarks. On ImageNet 256x256, Neon achieves state-of-the-art FID of 1.02 with only 0.36% additional training compute.

This contribution was assessed against **10 related papers** from the literature. Papers with potential prior art are analyzed in detail with textual evidence; others receive brief assessments.

---

1. Discrete diffusion vla: Bringing discrete diffusion to action decoding in vision-language-action policies

URL: View paper

Brief Assessment

Discrete Diffusion VLA[13] focuses on vision-language-action models for robotics manipulation, not on universal methods for improving generative models across diffusion, flow matching, and autoregressive architectures on image datasets like ImageNet.

---

2. Distilled decoding 1: One-step sampling of image auto-regressive models with flow matching

URL: View paper

Brief Assessment

Distilled Decoding[16] focuses on accelerating autoregressive image generation models through flow matching distillation, not on universal methods across diffusion, flow matching, and autoregressive models as claimed by the original paper.

---

3. D-AR: Diffusion via Autoregressive Models

URL: View paper

Brief Assessment

Diffusion via Autoregressive[10] focuses on bridging diffusion and autoregressive modeling for visual generation through a sequential diffusion tokenizer, not on universal training methods across different generative model families.

---

4. Flowar: Scale-wise autoregressive image generation meets flow matching

URL: View paper

Brief Assessment

Flowar[15] focuses on scale-wise autoregressive image generation with flow matching, not on universal methods across diffusion, flow matching, and autoregressive models. The candidate addresses a different technical problem (next-scale prediction) rather than demonstrating universality of a single method across architectures.

---

5. The Mathematics of Modern Generative Modeling: Normalizing Flows, Autoregressive and Diffusion Models

URL: View paper

Brief Assessment

Mathematics Generative Modeling[12] is a mathematical survey paper covering normalizing flows, autoregressive models, and diffusion models. It does not present empirical validation of any method across these architectures or report FID scores on ImageNet.

---

6. Pyramidal flow matching for efficient video generative modeling

URL: View paper

Brief Assessment

Pyramidal Flow Matching[7] focuses on efficient video generation through spatial and temporal pyramids in flow matching models, not on universal methods across diverse generative model families (diffusion, flow matching, autoregressive, few-step) on image datasets like the original paper demonstrates.

---

7. Unigenx: Unified generation of sequence and structure with autoregressive diffusion

URL: View paper

Brief Assessment

Unigenx[9] focuses on unified generation of sequences and structures across scientific domains (proteins, molecules, materials), not on universal training methods across diffusion/flow/autoregressive models for image generation benchmarks like ImageNet.

---

8. Ar-diffusion: Asynchronous video generation with auto-regressive diffusion

URL: View paper

Brief Assessment

AR-Diffusion[11] focuses on video generation using a hybrid auto-regressive diffusion approach for temporal sequences, not on universal training methods across diverse generative model families (diffusion, flow matching, autoregressive, few-step) for image generation as in the original paper.

---

9. Lumos-1: On autoregressive video generation from a unified model perspective

URL: View paper

Brief Assessment

Lumos[14] focuses on autoregressive video generation using LLM architectures with MM-RoPE and AR-DF techniques, not on universal methods across diffusion, flow matching, and autoregressive models for image generation benchmarks like ImageNet.

---

10. Beyond next-token: Next-x prediction for autoregressive visual generation

URL: View paper

Brief Assessment

Beyond Next-Token[8] focuses on autoregressive visual generation with next-x prediction entities (tokens, cells, scales), not on universal training methods across diffusion, flow matching, and autoregressive models as claimed in the original paper.

Appendix: Text Similarity Detection

No high-similarity text segments were detected across any compared papers.

References

- [0] Neon: Negative Extrapolation From Self-Training Improves Image Generation View paper
- [1] Self-improving diffusion models with synthetic data View paper
- [2] Constraining Generative Models for Engineering Design with Negative Data View paper
- [3] Hybrid Open-Set Segmentation With Synthetic Negative Data View paper
- [4] Learning generative models via discriminative approaches View paper
- [5] Fake & Square: Training Self-Supervised Vision Transformers with Synthetic Data and Synthetic Hard Negatives View paper
- [6] Time and Space Aggregation Recommendation Model Based on Synthetic Negative Samples View paper
- [7] Pyramidal flow matching for efficient video generative modeling View paper
- [8] Beyond next-token: Next-x prediction for autoregressive visual generation View paper
- [9] Unigenx: Unified generation of sequence and structure with autoregressive diffusion View paper
- [10] D-AR: Diffusion via Autoregressive Models View paper
- [11] Ar-diffusion: Asynchronous video generation with auto-regressive diffusion View paper
- [12] The Mathematics of Modern Generative Modeling: Normalizing Flows, Autoregressive and Diffusion Models View paper
- [13] Discrete diffusion vla: Bringing discrete diffusion to action decoding in vision-language-action policies View paper
- [14] Lumos-1: On autoregressive video generation from a unified model perspective View paper
- [15] Flowar: Scale-wise autoregressive image generation meets flow matching View paper
- [16] Distilled decoding 1: One-step sampling of image auto-regressive models with flow matching View paper
- [17] Improving Compound-Protein Interaction Prediction by Self-Training with Augmenting Negative Samples. View paper
- [18] Gradient matching for domain generalization View paper
- [19] Information Diffusion Modeling in Social Networks: A Comparative Analysis of Delay Mechanisms Using Population Dynamics View paper
- [20] Simulation and the reality gap: Moments in a prehistory of synthetic data View paper
- [21] Mind the gap between synthetic and real: Utilizing transfer learning to probe the boundaries of stable diffusion generated data View paper
- [22] iNeRF: Inverting Neural Radiance Fields for Pose Estimation View paper
- [23] Domain-adaptive neural networks improve supervised machine learning based on simulated population genetic data View paper
- [24] iSDF: Real-Time Neural Signed Distance Fields for Robot Perception View paper
- [25] Gradient projection Newton algorithm for sparse collaborative learning using synthetic and real datasets of applications View paper
- [26] Robin hood and matthew effects: Differential privacy has disparate impact on synthetic data View paper
- [27] Synthetic text generation for training large language models via gradient matching View paper