

Model Collapse in the Self-Consuming Chain of Diffusion Finetuning: A Novel Perspective from Quantitative Trait Modeling

Youngseok Yoon
UCSB
California, USA
youngseok_yoon@ucsb.edu

Dainong Hu
UCSB
California, USA
dainong@ucsb.edu

Iain Weissburg
UCSB
California, USA
ixw@ucsb.edu

Yao Qin
UCSB, Google Deepmind
California, USA
yaoqin@ucsb.edu

Haewon Jeong
UCSB
California, USA
haewon@ucsb.edu

ICLR 2026(International Conference on Learning Representations)

発表：藤原 颯太

背景：LLMの普及とモデル崩壊

LLMによる膨大な量のデータ生成

- 合成テキスト
- 画像データ

合成データによる学習

- 自己改善の有効性を示す研究も存在
- 大多数はモデル崩壊現象を報告

画像モデルでのモデル崩壊

- 画像劣化
- 多様性の喪失

目的：再利用可能な画像生成手法

ハイパーパラメータの影響の調査

- 各変数がモデル崩壊にどう影響するか調査

統計遺伝学を用いた理論的考察

- 遺伝モデルと機械学習モデルの対応

パラメータ調整不要手法ReDiFineの提案

- 最適値を探索せずに再利用可能な画像を生成

実験設定：モデルとデータセット

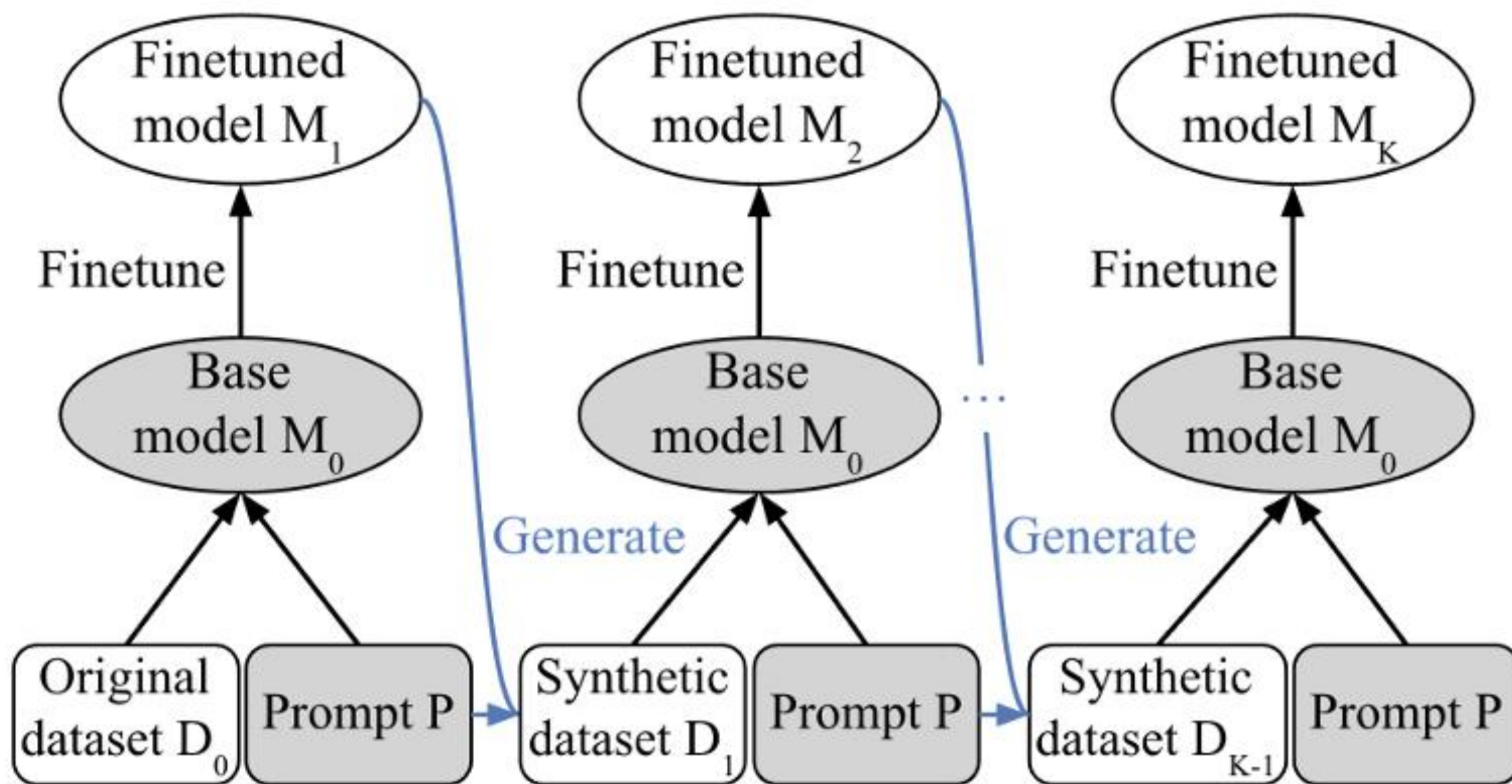
モデル

- Stable Diffusion v1.5

データセット

- ポケモン
- Kumapi
- Butterfly
- CelebA-1k

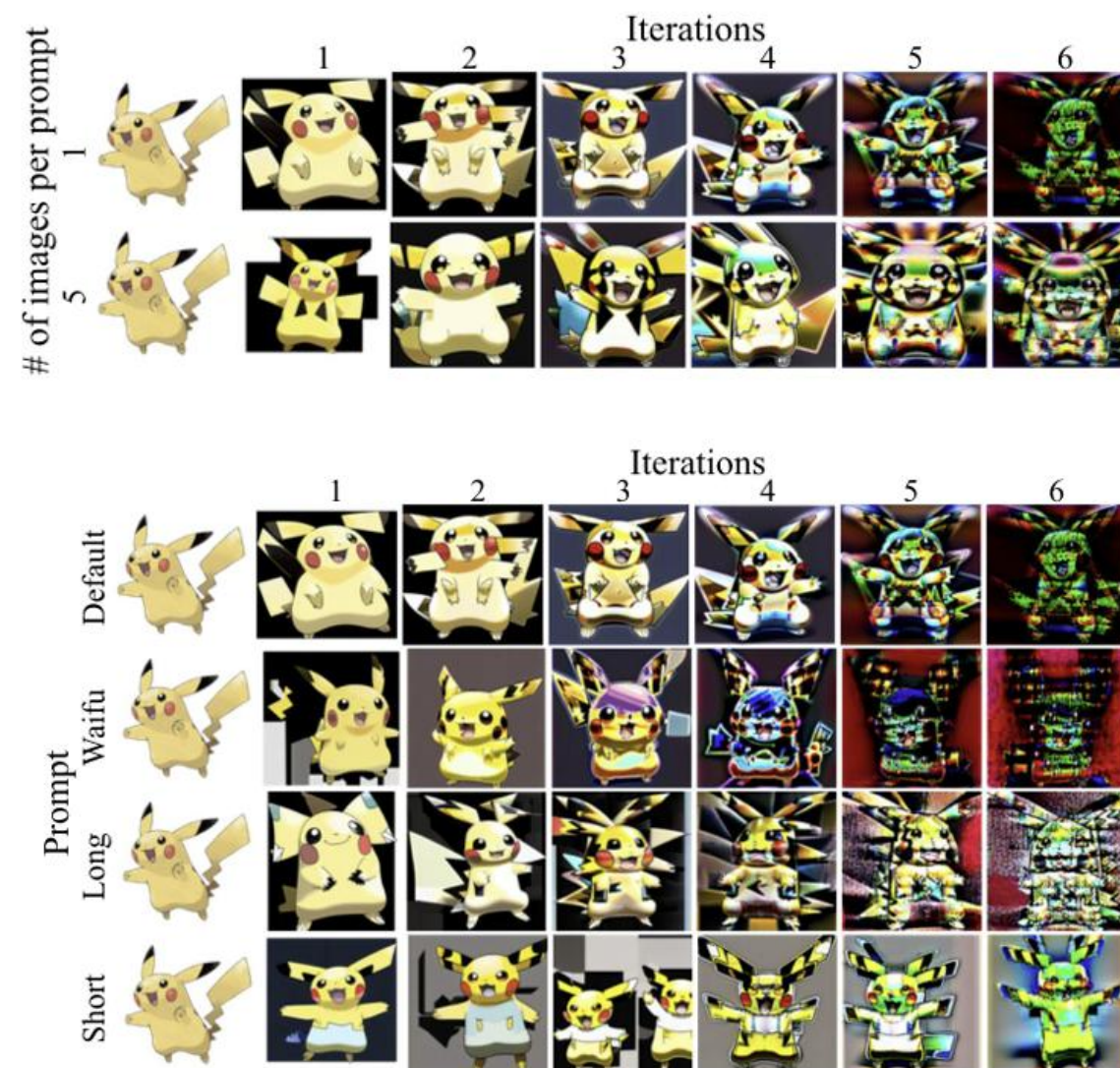
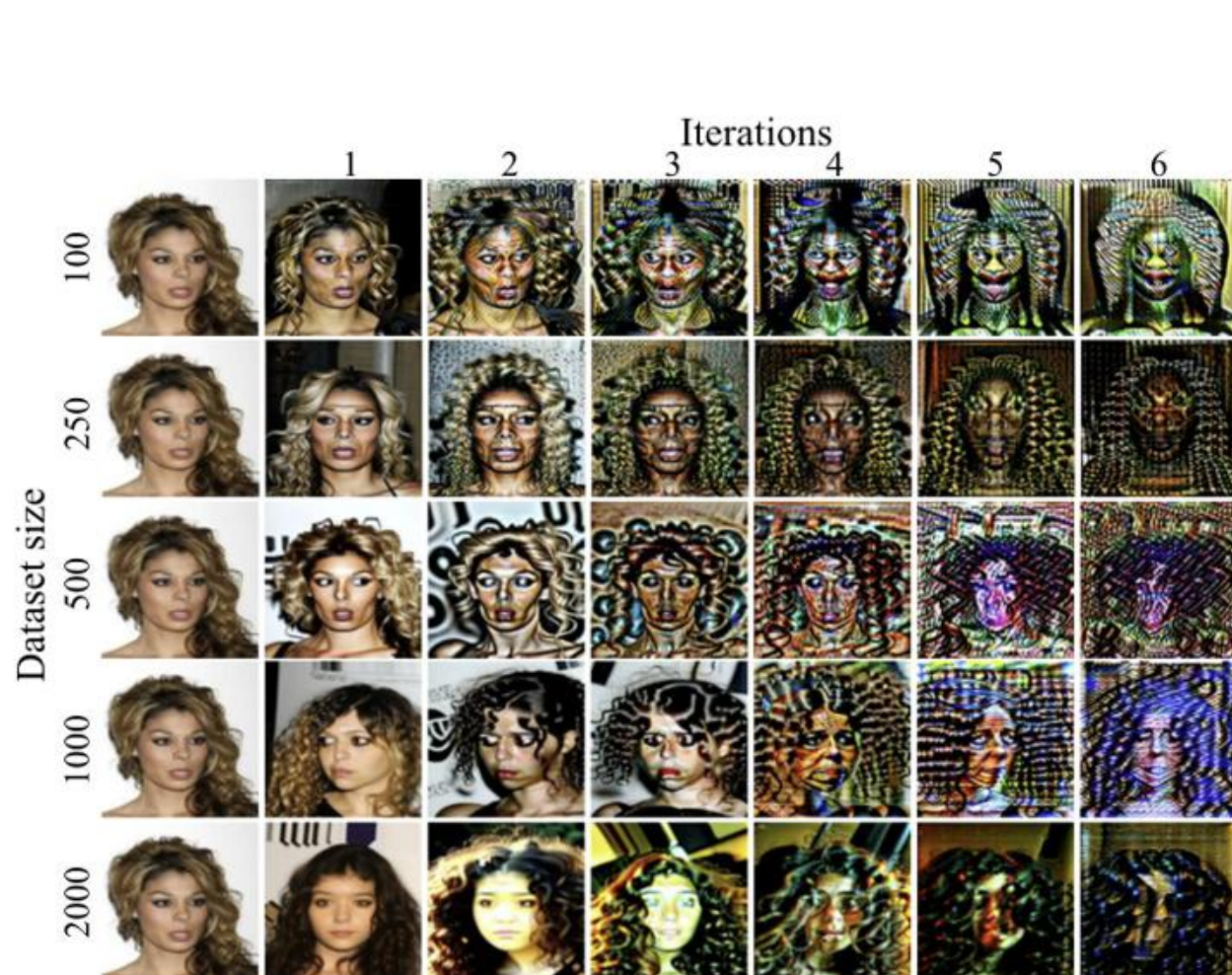
実験設定：拡散チェーン



実験結果：モデル崩壊現象の観測

普遍的発生

- 4つのデータセットすべてで崩壊現象を観測
- どのようにパラメータを変化させても回避できない



実験結果：主要パラメータの特定

CFGスケール(Classifier-Free Guidance)

- 性能劣化に大きな影響を与える
- データセットごとに最適値が異なる

CFGスケールとは

総スコア = 無条件スコア + CFG × (条件付きスコア - 無条件スコア)

- 平均的な画像らしさとプロンプトへの忠実度を制御する

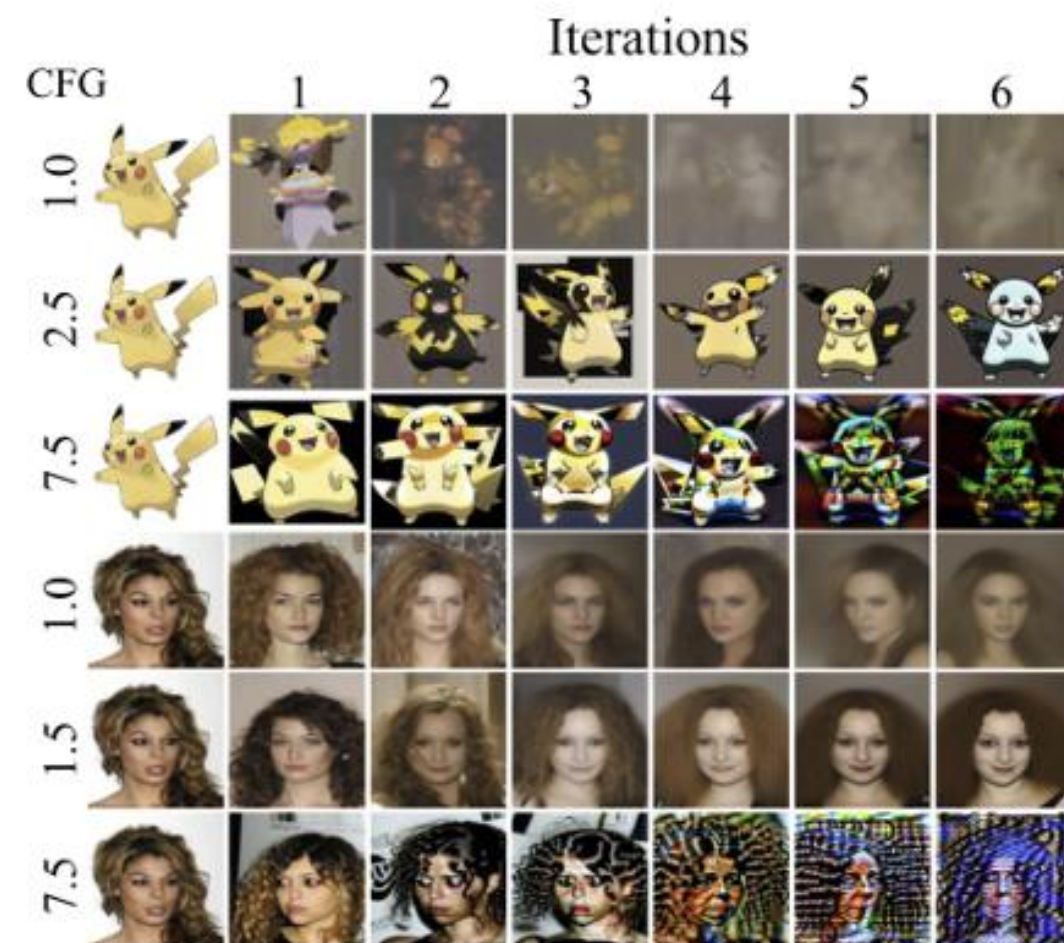
実験結果：高周波劣化と低周波劣化

高周波劣化

- CFGスケールを高くすることで発生
- 特徴が過剰に強調される
- 反復パターンや色飽和の発生

低周波劣化

- CFGスケールを低くすることで発生
- 画像が徐々にぼやける
- 最終的には構造を失ってしまう

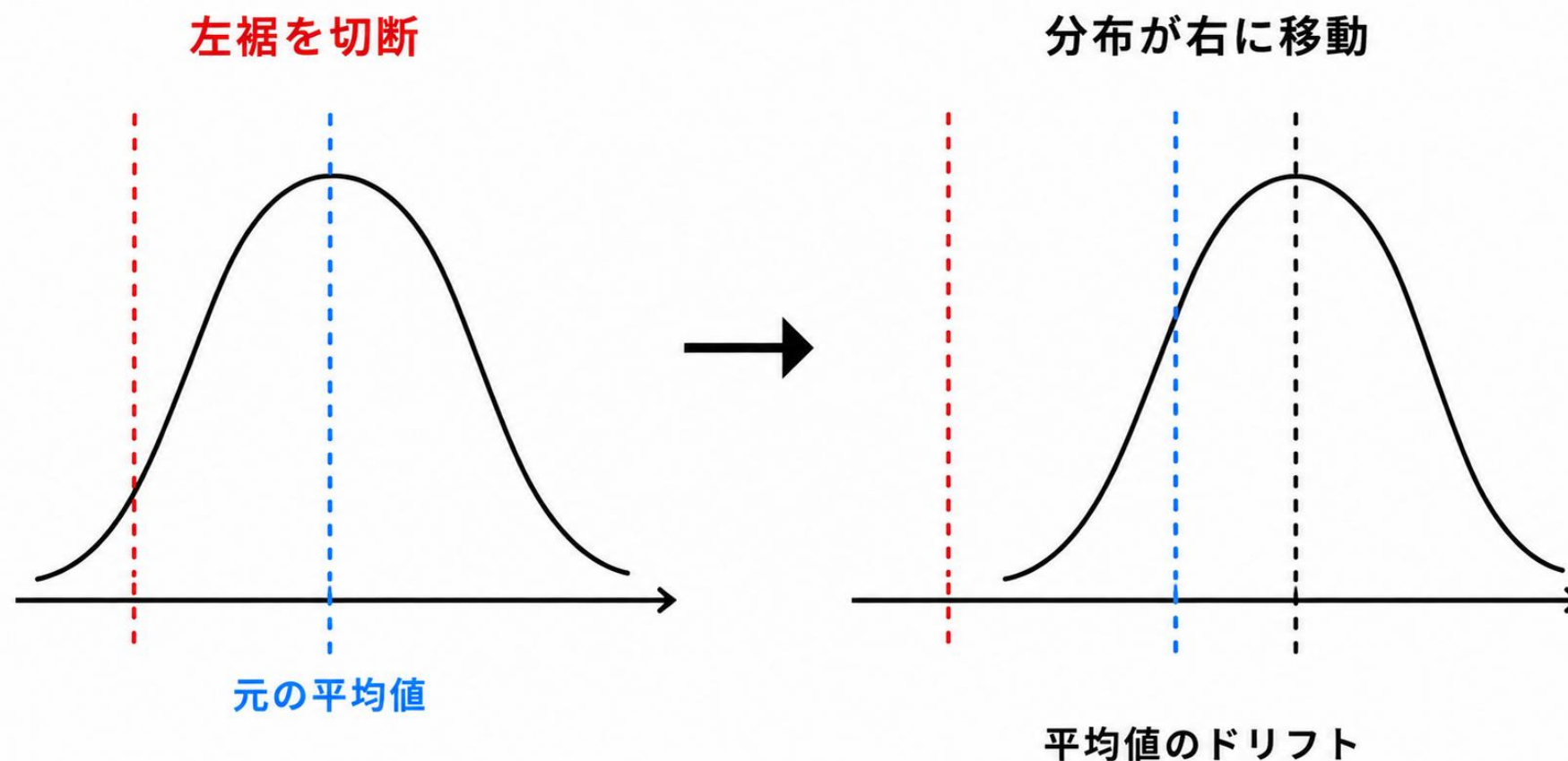


統計遺伝学を用いた考察

量的形質モデリング

- 人間の好みを反映した自然選択を組み込む
- ガウス分布の左右の裾を切断

⇒分布が変化し平均値のドリフトが発生



統計遺伝学を用いた考察

CFGスケールによる選択

- CFGスケールは選択メカニズムとして機能
- 高CFGは分布の右裾を選択
⇒画像の詳細な情報を強調する分布に変化
- 低CFGは分布の左裾を選択
⇒画像の大まかな情報を強調する分布に変化

提案手法：ReDiFine

再利用可能な画像生成

- 最適CFGなしでもモデル崩壊を抑制可能

- 二つの手法の組み合わせ

①条件ドロップファインチューニング

- 条件付きスコアと無条件スコアの誤差を小さく保つ
- ランダムにテキスト条件をドロップする

②CFGスケジューリング

- 条件付きスコアと無条件スコアの誤差の増幅率を軽減する
- 後期ステップのCFGによる選択圧を緩和

$$\text{総スコア} = \text{無条件スコア} + \text{CFG} \times (\text{条件付きスコア} - \text{無条件スコア})$$

CFGスケジューリング

拡散ステップ中にCFGスケールを減少させる

- 後期ステップでの条件付きスコアの過剰強調を緩和
- 反復を繰り返すことによる切断の累積を軽減
- 後期ステップでは分布がほとんど変化しない

条件ドロップファインチューニング

条件テキストをランダムで空文字列に置き換える

- 画像を教師なしで学習させることで無条件スコアを更新
- 無条件スコアと条件付きスコアの差を軽減
- 反復による劣化特徴の強調を緩和

条件ドロップファインチューニング

最初の世代

$$\text{条件付きスコア} - \text{無条件スコア} = \text{差分(小)}$$

既存手法

$$\begin{array}{ccc} \text{条件付きスコア} & - & \text{無条件スコア} \\ \text{更新} & & \text{更新なし} \end{array} = \text{差分(大)}$$

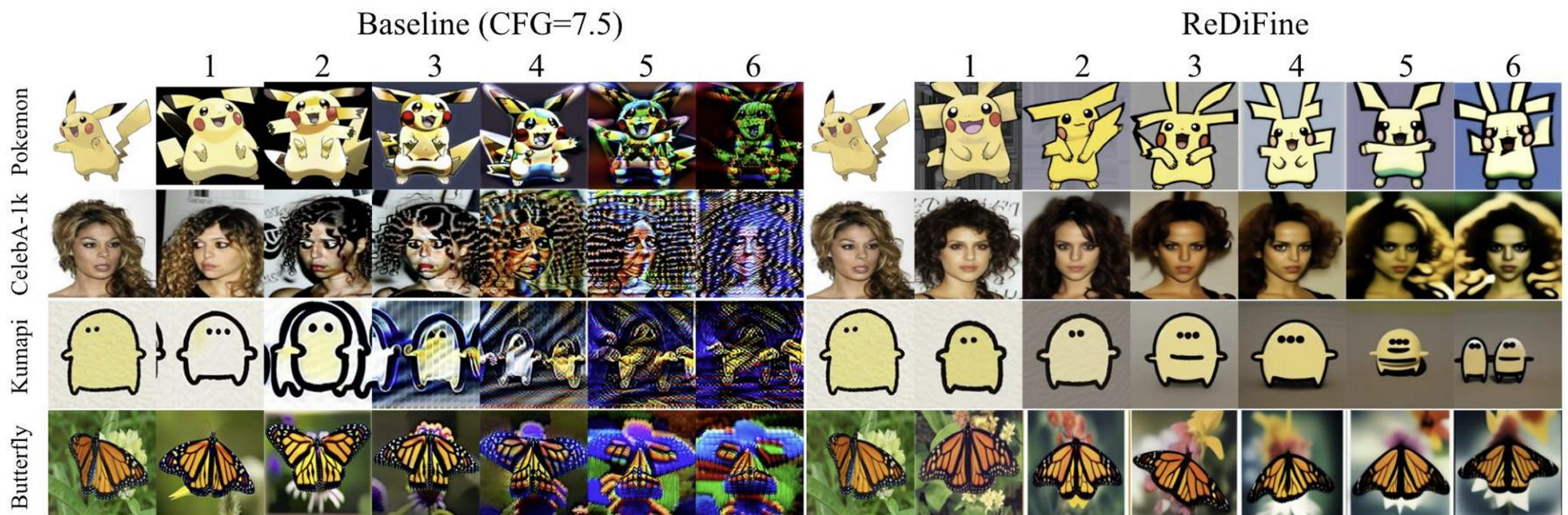
条件ドロップファインチューニング

$$\begin{array}{ccc} \text{条件付きスコア} & - & \text{無条件スコア} \\ \text{更新} & & \text{ランダムで更新} \end{array} = \text{差分(小)}$$

ReDiFineの結果

すべてのデータセットで良好な結果

- 最適CFGはデータセットによって異なるが,データセットに依らず安定した品質を維持できる



結論

- | 合成データの流入は不可避であり,合成データを学習データに含めることも避けられない
- | 現在のテキストから画像を生成するモデルでは合成データを自己改善に用いることは難しい
- | 実データと合成データを区別し異なる学習方法を設定するアルゴリズムの開発など多くの課題が残っている

所感

- | 遺伝学からモデル崩壊現象を説明しているのが斬新で面白い
- | CFGスケジューリングは画像生成特有の手段だが条件ドロップの手法はコード生成分野でも有効なのではないか
- | 拡散チェーンではベースとなるモデルが固定されているがそれでもモデル崩壊が発生しているのが驚きだった