Photorealistic Text-to-Image Diffusion Models with Deep Language Understanding

- 先行研究は、画像-テキストデータを使用してモデルを学習させていたが、Imagenでは、LLMのtext embeddingを使用した
- Imagenの大まかな概要
  - T5-XXLのエンコーダを使用してテキストをembeddingにする
  - 64×64の画像用の拡散モデルに続いて256x256, 1024×1024の画像を生成するために、超解像の拡散モデルも用意されている
  - 全ての拡散モデルはテキストのEmbeddingで条件付けされており、classifier free guidanceが使用されている
  - 先行研究と比べてguidanceの重みを大きくしても推論画像の品質を落とさずに生成することができた
  - contribution
    - テキストデータのみで学習された大規模言語モデルがテキストから画像の生成において効果的であり、このテキストエンコーダをスケールさせる方が画像生成する拡散モデルをスケールさせるよりもサンプルの品質が向上する
    - dynamic thresholdingを提案
      - guidanceの重みを大きくすることができ、写実的で詳細な画像生成を実現
    - 拡散モデルのアーキテクチャとしてEfficient U-Netを提案
      - シンプル、収束が早い、メモリ効率が良い

        - COCO FIDでSOTA、人の評価でも良い
        - text-to-imageの評価用ベンチマークDrawBenchを提案、DALLE-2よりも良い
- imagenのモデル構造

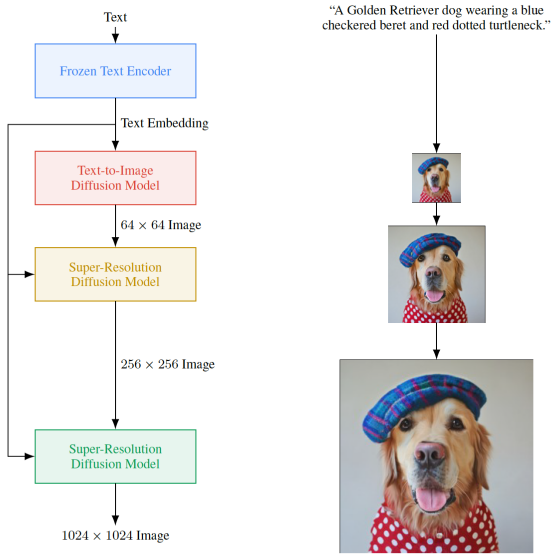


Figure A.4: Visualization of Imagen. Imagen uses a frozen text encoder to encode the input text into text embeddings. A conditional diffusion model maps the text embedding into a 64×64 image. Imagen further utilizes text-conditional super-resolution diffusion models to upsample the image, first $64 \times 64 \rightarrow 256 \times 256$, and then $256 \times 256 \rightarrow 1024 \times 1024$.

    - 学習済みのテキストエンコーダ
        - text-to-imageモデルのtext encoderでは、image-textのペアのデータで学習させたようなモデル(CLIPなど)を用いる方法と、BERT, GPT, T5などのLLMを使用する方法がある
        - 後者の方がテキストデータのみで学習しているためリッチなテキストの情報を保持していると思われる
        - Imagenではその有効性を検証するために、text encoderとして、BERT, T5, CLIPを用いた場合で比較する
        - また、実験によってtext encoderをスケールさせた方がtext-to-imageの生成画像の品質が良いことがわかった
        - MS-COCOのようなシンプルなベンチマークでは、T5-XXLがCLIPと同等の制度であったのに対して、DrawBench(難しいベンチマーク)では、人の評価によるとT5-XXLの方がCLIPよりも良い結果となった

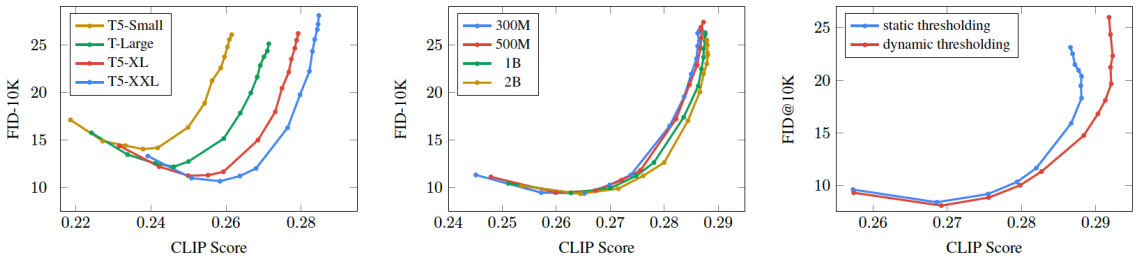
- classifier guidance
  - 学習済みモデル(画像分類器など)の勾配を利用して条件付けた拡散モデルの多様性？を軽減し、サンプリング品質を向上させる手法
- classifier free guidance (cfg)
  - 条件付けをしていない拡散モデルと条件付けをした拡散モデルを同時に学習させる
    - 条件付け(クラス、テキスト、低解像度画像）を10%の確率でドロップ(0埋め)させる
  - 推論時のノイズの予測は以下のようになる

$$\tilde{\boldsymbol{\epsilon}}_\theta(\mathbf{z}_t, \mathbf{c}) = w\boldsymbol{\epsilon}_\theta(\mathbf{z}_t, \mathbf{c}) + (1-w)\boldsymbol{\epsilon}_\theta(\mathbf{z}_t).$$

  - 先行研究から、cfgの重みwを大きくするとtext-imageのアライメントをよくするが、不自然な画像を生成してしまう
    - これは、train-testのミスマッチによって引き起こさせる
      - テスト時に時刻tのサンプリング結果x_tが[-1,1]の幅を超えてしまう
    - static thresholding
      - [-1, 1]にx_tの値をクリッピングする
        - 依然として不自然な画像を生成してしまう
    - dynamic thresholding
      - 各サンプリングステップで、予測されたピクセル値x_tの絶対値のあるパーセンタイルを設定し、それを閾値sとする。もしsが1を超える場合、予測されたピクセル値を[-s, s]の範囲にクリップし、その後、sで割る
      - cfgの重みが大きい際に良い画像生成を可能にした
- robust cascaded diffusion models
  - noise conditioning augmentationが有効

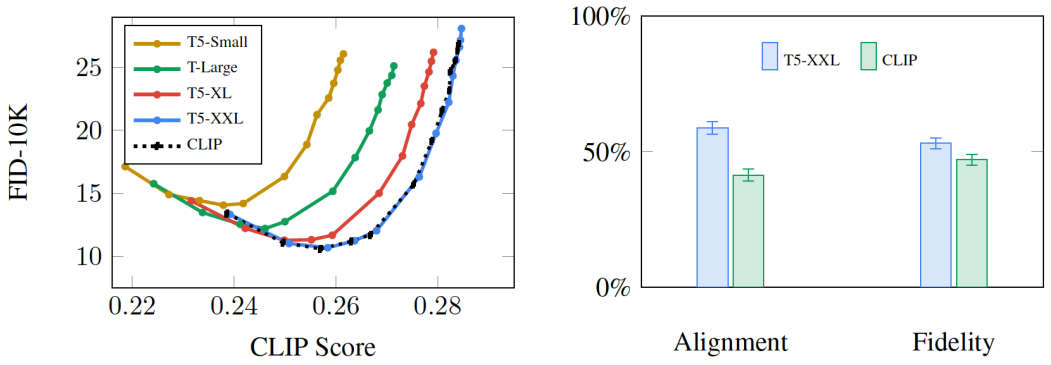
        - さらに超解像モデルにおいてnoise level conditioningがサンプリング効率を向上させる
        - 低解像度の画像とデータ拡張レベル(ガウスノイズやガウスぼかしの強さ)を受け取り、低解像度の画像に対してノイズを乗せる。それとノイズの強さが超解像モデルの入力となる
        - このノイズの強さが学習中はランダムに選択され、推論時には様々な値が探索されて一番サンプル効率が良いものが選択される
    - アーキテクチャ
        - ベースモデル
            - テキストのembeddingで条件付けされたU-Netを使用する
            - アテンション層やプーリング層におけるlayer normalization層におけるテキスト条件付けが良いことがわかった
        - 超解像モデル
            - テキストを用いたクロスアテンション層を導入
- モデルの評価
    - 評価指標
        - FID - 生成画像のクオリティ
        - CLIP scores - テキストと画像の類似度
        - 人の評価
            - モデルの生成画像とレファレンス画像を見せて、どっちがより写実的か評価してもらう - prefer model A, indifferent, prefer model B
            - また、生成画像がテキストを表現しているかを評価するために、それぞれの画像に対して三段階で人に評価してもらう - 0, 50, 100
    - 評価用データ
        - COCO
        - DrawBench : 提案
            - 様々な種類の画像が用意されており、テキストに対しても長文のものや珍しい単語を含むものやスペルミスがある文章などを用意した

- その他の分析
  - テキストエンコーダのサイズを大きくした方が良い
  - テキストエンコーダのサイズを大きくする方が拡散モデル(U-Net)のサイズを大きくするよりも有効
  - Dynamic thresholdingが有効



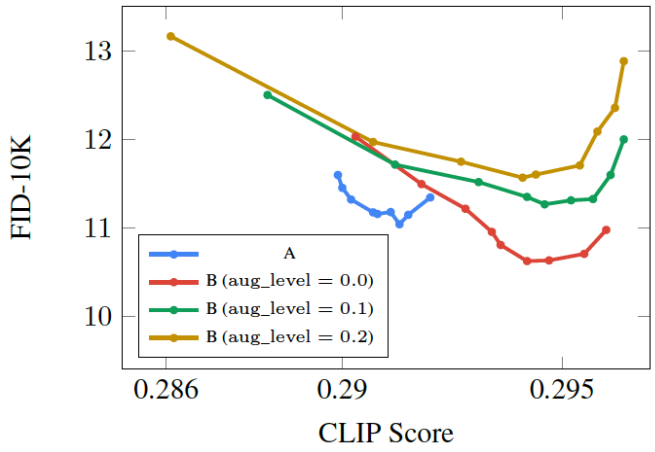
(a) Impact of encoder size. (b) Impact of U-Net size. (c) Impact of thresholding.

  - 人の評価ではテキストエンコーダでT5-XXLを使用する方がCLIPよりも良い

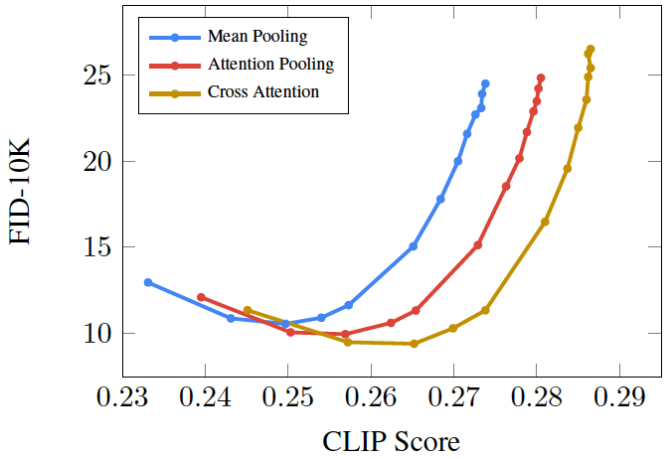


(a) Pareto curves comparing various text encoders. (b) Comparing T5-XXL and CLIP on DrawBench.

  - noise conditioning augmentationが有効



- テキストの条件付け手法の選択が影響を及ぼす



- Efficient U-Netが良い