一般社団法人　電子情報通信学会
THE INSTITUTE OF ELECTRONICS,
INFORMATION AND COMMUNICATION ENGINEERS

信学技報
IEICE Technical Report

CommonArt β: 国産大規模言語モデルによる透明性の高い画像生成用拡散トランスフォーマー

尾崎安範 (責任著者)†　三嶋　隆史††　冨平　準喜†

† 本部, 株式会社 AIdeaLab　〒100–6001 東京都千代田区霞が関三丁目 2 番 5 号霞が関ビル 13F
†† 本部, 株式会社 AI Picasso　〒107–0052 東京都港区赤坂 3–13–4 赤坂三河家ビル 5F
E-mail: †{ozaki,tomihira}@aidealab.com, ††mishima@aipicasso.app

あらまし　本研究では、著作権に配慮した透明性の高い画像生成モデルである CommonArt β を提案する。データセットには CC-0 や CC-BY といった改変可能な画像約 2500 万枚と合成キャプション 5000 万個を使い、アルゴリズムには拡散トランスフォーマーを国産 LLM で条件付けすることとした。 30000 L4 GPU 時間による学習の結果、FID といった画像品質や CLIP Score といった指示追従の観点から日本語と英語を総合して定量評価した場合、従来の手法よりも最も高い性能になることが示された。今後は動画生成モデルへの応用が考えられる。

キーワード　画像生成, 大規模言語モデル, 拡散モデル, 生成 AI

CommonArt β: Diffusion Transformer for Text-to-Image Generation by Japanese Large Language Model

Yasunori OZAKI (CORRESPONDING AUTHOR)†, Ryuji MISHIMA††, and Toshiki TOMIHIRA†

† Head Quarters, AIdeaLab, Inc.　13F KASUMIGASEKI BUILDING, 3–2–5 Kasumigaseki, Chiyoda, Tokyo, 100–6001 Japan
†† Head Quarters, AI Picasso, Inc.　5F 3–13–4 Akasaka, Minato, 107–0052 Japan
E-mail: †{ozaki,tomihira}@aidealab.com, ††mishima@aipicasso.app

Abstract　In this paper, we propose CommonArt β, a transparent image generation model that respects copyright. The dataset consists of approximately 25 million modifiable images under licenses such as CC-0 and CC-BY. For the algorithm, we used a diffusion transformer conditioned on a domestically developed LLM. After 30,000 L4 GPU hours of training, quantitative evaluation combining Japanese and English metrics in terms of image quality and instruction following showed that our method achieved the highest performance compared to conventional approaches. Future work may include applications to video generation models.

Key words　Text-to-Image Generation, Large Language Model, Diffusion Model, GenAI

1. はじめに

拡散モデルをはじめとした新しい世代の人工知能が、テキストなどの入力に応じてテキストや画像などのコンテンツを大量に生成している。これらの総称を内閣府の AI 戦略会議は「生成 AI」と呼んでいる [33]。生成 AI は産業革命時の内燃機関と同じく、我が国の労働力を補う可能性がある。特に我が国が力を入れるコンテンツ産業の基礎となる可能性を秘めている。しかし、著作権侵害をはじめとする負のリスクが指摘されている。

そこで、本研究の目的は、我が国のコンテンツ産業を支えながらも、負のリスクを抑えた、テキストからの画像生成モデル (以下、画像生成 AI)、CommonArt β を作ることとする。本研究の貢献は次のとおりである。

（1） 国産の大規模言語モデルにより、Stable Diffusion 2.1 [19] と英語に対する同等程度の画像品質と指示追従を持ち、日本語に対しては大幅に上回る、テキストからの画像生成モデルを初めて Apache-2.0 で公開した(注1)

（2） 画像生成 AI の透明性や再現性を担保するために、学習に使用した CC-0 と CC-BY の画像データセットをすべて公開した。また、画像テキストペアやプログラムは可能な限りすべて公開した。

（3） 既存の学習方法を効果的に組み合わせ、負のリスクで

(注1)：https://huggingface.co/aipicasso/commonart-beta

ある、学習時や生成時の著作権侵害の可能性を抑えられることを示した

2. 従来の画像生成 AI を作る方法

画像生成 AI を作るにはデータセットとアルゴリズムが必要である。ここではデータセットとアルゴリズムについて整理する。

2.1 画像生成 AI を作るためのデータセット

画像生成 AI を作るためのデータセットは複数あるが、最も議論を呼んだものは LAION らの LAION-5B [21] だろう。LAION-5B はインターネット上の画像リンクとそのキャプションを約 50 億枚分集めたデータセットである。これらは著作権により保護されているものもあるため、類似性がある画像を意図して生成すると著作権侵害になる可能性がある。LAION-5B を元にした Stable Diffusion でメンバーシップ推論攻撃を試みた Somepalli らの研究では LAION-5B の中にある画像と類似性の高い画像が出ることが指摘されている [24]。

そこで、これらの問題を解決するために提案されたのが Thomee らの YFCC100M [27] をベースとした Gokaslan らの CommonCatalog [8] がある。CommonCatalog は、コンテンツの再利用を前提としたライセンス Creative Commons の画像と合成キャプションで構成された約 7000 万ペアのデータセットである。Gokaslan らはこのうち CC-BY と CC–BY–SA の画像を使った CC–BY–SA のモデル CommonCanvas–S–C を作ることで問題の解決を試みた。しかし、CC-BY-SA のモデルが作り出す画像は CC–BY–SA にせざるを得ないのか、そうでないのかよく分かっていない。もし、CC–BY–SA にせざるを得ない場合、通常の著作物とは違い、著作権による保護を緩めなければならなくなる。このため、安心して利用することができない。

さらに、これらを解決するために Megalith–10M [3] が提案された。Megath–10M はパブリックドメインの画像を含む 1000 万枚の画像リンクで構成されるデータセットである。しかし、キャプションがないために画像生成 AI の学習にはそのままでは使えない。

以上から現状では安心して画像生成 AI に使えるデータセットがないことがわかる。

2.2 画像生成 AI を作るためのアルゴリズム

画像生成 AI を作るアルゴリズムとして、テキストエンコーダーと画像生成モデルを分離して開発するアルゴリズムとそれらを一体にして開発するアルゴリズムがある。前者の例として Stable Diffusion 2.1 [19] や PIXART-α [5] があり、後者の例として Chameleon [26] がある。ここでは前者を焦点に当てて説明する。Stable Diffusion 2.1 は OpenCLIP [6] のテキストエンコーダー (BERT [7] ベース) と Latent Diffusion Model [19] を分離して開発したアルゴリズムである。ただし、LAION-5B を元に作った内部データセットで作られているため、アルゴリズムを利用したモデルに再現性がない。また、同じアルゴリズムで作られた CommonCanvas-S-C はライセンスが CC-BY-SA なので安心して利用できない。Pixart-α は T5 [18] のエンコーダーと拡散トランスフォーマーを分離して開発したアルゴリズムである。これを利用したモデルもやはり内部データセットを使っており、再現性がない。しかし、この内部データセットには特徴がある。2500 万枚の画像を LLaVA [14] でキャプション付けしているところである。これは Stable Diffusion 2.1 の 20 億枚に比べて極めて少ない枚数である。また、このことにより学習コストも低いとしている。DALLE3 [2] も LLaVA のようなモデルを使い、キャプション付けしたり、Imagen 3 [11] も Gemini を使ってキャプション付けしていることから、このキャプション付けは様々なメリットがあると考えられる。

以上から既存の画像生成 AI の研究は再現性が低い。しかし、視覚言語モデルによる画像のキャプション付けは、テキストからの画像生成に有効である可能性が高い。

3. CommonArt β の作り方

従来の作り方を整理すると次の仮説が立てられる。

（1） テキストからの画像生成モデルを作るには LLaVA による日本語キャプションづけをして日本語 T5 と拡散トランスフォーマーを使えば、約 2500 万枚程度の画像で Stable Diffusion 2.1 のような日本語画像生成モデルが作れる

（2） T5 や BERT のようにテキストエンコーダーが変えられるならば、Llama のようなデコーダーオンリーのモデルでも、テキストからの画像生成モデルのテキストエンコーダーを代替できる

（3） Pixart-α ほど、効率よく作れるならば、学習用の GPU、例えば、H100 や A100 ではなく、推論用の GPU、例えば、NVIDIA L4 や RTX 4090 で作れる

これらから更なる仮説がたてられる。

CommonArt β により実証する仮説

テキストからの画像生成モデルを作るには、インターネット上にある約 2500 万枚程度の画像を LLaVA のようなモデルでキャプションづけをして Llama ベースの大規模言語モデルと拡散トランスフォーマーを使えば、Stable Diffusion 2.1 のような画像生成モデルが NVIDIA L4 で安全に作れる

以下ではこの仮説を実証するための実現方法を段階的に説明する。

3.1 データセットの構築

まず、安全に使えるデータセットがないため、データセットを構築する。この際、OpenAI や Google の利用規約に反して、競合するモデルを作らないように気をつけなければならない。

画像データセットとして、3 つのデータセットを用いる。

（1） CommonCatalog-CC-BY

（2） Megalith-10M

（3） ArtBench [13] (ただし、CC-0 のものに限る) (注2)

（4） Smithsonian Open Access (注3)

これにより、学習用画像は CC-0、CC-BY だけに限られる。このため、もし、これらの画像が出たとしても、著作権侵害と

(注2)：https://huggingface.co/datasets/alfredplpl/artbench-pd-256x256
(注3)：https://huggingface.co/datasets/aipicasso/soa-full-florence2

なることはない。

次にこれらの画像にキャプション付けを行う。学習画像が出ることへのリスク緩和をするため、画像にはすべて 2 つ以上のキャプションをつける [25]。また、同じ目的のため、全て合成キャプションを使う [25]。キャプション付けに使ったモデルは以下の通りである。

（1） BLIP-2 [12]
（2） Florence-2-large [30]
（3） Phi-3 Vision [1]
（4） Qwen2-VL-2B-Instruct [29]
（5） LLaVA-JP-Captioner

このうち、LLaVA-JP-Captioner は、今回の研究のために作られた日本語小型視覚言語モデルである。作り方は LLaVA と同じであるが、入力はキャプション付けしか受け付けないようになっている。データセットには OpenAI や Google の競合にならないように彼らのモデルの出力は一切入らないようにして、商用利用可能な状態にした。また、一部のデータセットは Phi-3 Medium [1] により英語から日本語に翻訳されている。

以上の作業から、画像数約 2500 万枚、画像テキストペア 5000 万ペアを用意した。

3.2 アルゴリズムの構築

今回は、Pixart-α の改良版 Pixart-Σ [4] をベースに新しい組み合わせに書き換える。基本となるフレームワークは、既存研究と同じであるテキストエンコーダーと拡散トランスフォーマーを組み合わせて作ることである。アルゴリズムは学習と推論に分かれる。

3.2.1 学習アルゴリズム

学習アルゴリズムでは入力テキストの文脈を意味するテンソル y を条件として、条件に合わせたノイズ除去方法を学習する。この疑似コードを Listing 1 に示す。

```
# Get embbedings
tokens = tokenizer(prompts)
h = text_encoder(tokens.input_ids)
y = h.hidden_states[-1]

# Sample a random timestep
timesteps = torch.randint()

# Initialize optimizer
optimizer.zero_grad()

# Calcurate loss value by Pixart Sigma loss function
loss = loss_function(model, clean_images, timesteps, y)

# Optimize
loss.backward()
optimizer.step()
```

Listing 1– 学習コード

今回テキストエンコーダーとしてデコーダーオンリーのアーキテクチャである Llama [28] を使った CALM2-7B(注4) を使った。その出力のうち、生成の条件付けには隠れ層の最終層の出力 y を使う。この出力 y は、次の単語を予測するための文脈を保持することに相当する情報を持っているとされている。

表 1 学習時のパラメータ

学習段階	最適化	バッチサイズ	ハードウェア
256 × 256 初期	CAME [17]	64	A6000 2 台
256 × 256 中期	AdamW8bit	1024	L4 8 台
256 × 256 後期	AdamW8bit	256	L4 32 台
512 × 512 全体	AdamW8bit	128	L4 32 台

3.2.2 推論アルゴリズム

推論アルゴリズムでは入力テキストの文脈を意味するテンソル y を条件として、条件に合わせたノイズ除去方法を推論する。この疑似コードを Listing 2 に示す。推論には DPM++ [16] を用いる。

```
# Get text embeddings
token = tokenizer(prompt)
h = text_encoder(token.input_ids)
y = h.hidden_states[-1]

# Generate Image
x = pixart_sigma_pipeline(y)
```

Listing 2– 推論コード

3.2.3 学 習 手 順

セクション 3.1 のデータセットとセクション 3.2 のアルゴリズムを使い、拡散トランスフォーマーを学習する。学習手順は Weak-Strong 法 [4] と呼ばれる、低解像度から高解像度へ学習する方法に従った。学習は学習の段階に合わせて最適化アルゴリズムやハードウェア、ハイパーパラメータを変えたため、表 1 にまとめる。なお、このバッチサイズのうち、1024 は Gradient Accumulation した結果である。最適化の学習率は一貫して 2×10^{-5} だった。その他のパラメータはデフォルトの値を用いた。このため、AdamW [15] の ϵ は 1×10^{-8} であった。また、なお、 CALM2-7B が VRAM を大量に消費するため、学習時には Quanto (注5) により 8bit 量子化をして使った。

ステップあたりの学習枚数が変わるため、参考程度しかならないが、最終的に 200 万ステップ程度となった。学習時間は 30000 L4 GPU 時間だった。

4. 評価方法と評価結果

今回、画像品質と指示追従 (Prompt Following, Text Alignment の仮訳) の観点から機械的に性能評価を行う。英語と日本語でそれぞれ計算する。

4.1 英語に対する画像品質と指示追従

英語に関するベースモデルは Stable Diffusion 2.1 – base、CommonCanvas–S–C、Japanese Stable Diffusion、提案手法とする。Japanese Stable Diffusion は Stable Diffusion 1.4 をベースに日本語でも使えるように LAION–5B から 1 億ペアを学習した Stable Diffusion の派生モデルである。このため、透明性が低い

(注4)：https://huggingface.co/cyberagent/calm2-7b

(注5)：https://huggingface.co/docs/transformers/main/en/quantization/quanto

が、参考として計測する。なお、これらのモデルを選んだ理由として、標準の解像度が 512×512 であり、比較しやすいからである。

画像品質は、Zero–shot FID–30K とする。これは COCO(注6)からランダムに 3 万個選んだ画像とテキストペアを選び、そのテキストペアから生成された画像の集合と 3 万個選んだ画像の集合との距離を FID [10] で測定する方法である。今回は COCO 2017 の train set の中からランダムに選んだ集合を用いた。NFE (Number of Function Evaluations) は提案手法で 20、その他でデフォルトの設定とした。

指示追従は CLIP Score [9] とする。これは MS COCO から画像品質と同様に 3 万個選んだテキストを使い、各モデルで画像を生成する。その生成された画像とテキストの類似度を CLIP を使い、コサイン類似度を見る。ここで使った CLIP の Visual Encoder は、ViT-B/32 である。今回は画像品質で抽出した集合と同じものを使った。なお、ここでは $w = 100$ とした。

画像品質と指示追従、それぞれを比較した結果が表が 2 である。

今回生成された定性的結果を図 1 に次のセクションの結果と合わせてはる。

4.2 日本語に対する画像品質と指示追従

日本語に関するベースモデルは Stable Diffusion 2.1 – base、CommonCanvas–S–C、Japanese Stable Diffusion [22] [20] 、提案手法とする。

画像品質は、日本語 Zero–shot FID–30K とする。これは STAIR Captions [32] からランダムに 3 万個選んだ画像とテキストペアを選び、そのテキストペアから生成された画像の集合と 3 万個選んだ画像の集合との距離を FID で測定する方法である。今回は train set の中からランダムに選んだ集合を用いた。NFE は提案手法で 20、その他でデフォルトの設定とした。

指示追従は日本語 CLIP Score とする。これは STAIR Captions から画像品質と同様に 3 万個選んだテキストを使い、各モデルで画像を生成する。その生成された画像とテキストの類似度を CLIP を使い、コサイン類似度を見る。ここで使った CLIP のモデルは、line–corporation/clip–japanese–base [31] である。今回は画像品質で抽出した集合と同じものを使った。なお、ここでは $w = 100$ とした。

画像品質と指示追従、それぞれを比較した結果が表が 2 である。

4.3 メンバーシップ推論攻撃への耐性

この章では学習したキャプションから学習した画像を抽出できるか試みる。これをメンバーシップ推論攻撃 [23] という。メンバーシップ推論攻撃に対して十分に強ければ、著作権侵害を起こすことは少ないと言える。

メンバーシップ推論攻撃した例を図 2 にのせる。

5. 実験結果から得られた考察

表 2 を見ると、FID の差は Stable Diffusion 2.1 と提案手法とでは、表 2 の日本語 FID ほど差がない。CLIP Score と日本語 CLIP Score も同様である。ここから英語の画像品質と指示追従は同等程度だと考えられる。逆に Stable Diffusion 2.1 の日本語は翻訳を挟まないと全く使い物にならないことは表 2 や図 1 を見れば明らかである。全体的に日本人らしきものがたくさん出る傾向が見られた。翻訳を挟むことでニュアンスが変わることは既存の調査(注7)でわかっており、ネイティブな言語モデルがプロンプトを処理する必要があることもわかっている。日本語ネイティブな画像生成は文化保護のためにも必要である。Japanese Stable Diffusion は透明性の観点から疑問があると言える。したがって、同等の性能であれば、AI ガバナンスの観点から提案手法の有用性があると考えられる。

しかし、提案手法は定量的な面では他のモデルと競い合う能力があると考えられるが、定性的に見ると違和感が残る。例えば、図 1 においてはシマウマの顔が不自然である。また、バイクも形状が歪である。指示追従の観点からも違和感が残る。例えば、白い花と赤い花は指示通り生成されているが、白い花瓶は生成されていない。このように統計的な定量評価では問題なくとも定性評価としては問題があることがわかる。今後は人手評価による定性評価が求められる。

図 2 の画像間に法的な意味での類似性は見られず、このため、メンバーシップ推論攻撃のような極端な利用法でも提案手法がユーザーに著作権を侵害させることはないと思われる。しかし、定量的に示せられていないことから更なる調査が必要とされる。今回のデータセットは公開されているため、他者による検証も大切だと考える。

以上より、日本語と英語を総合して勘案すれば、定量的に最も優れていると考えられる。しかし、人による評価がないために、定性的に言えるかはまだわからないと言える。

6. ま と め

本研究では、我が国のコンテンツ産業を支えながらも、負のリスクを抑えた、CommonArt β を提案した。テキストからの画像生成モデルを作るには LLaVA のようなモデルでキャプションづけをして Llama ベースの大規模言語モデルと拡散トランスフォーマーを使えば、インターネット上にある約 2500 万枚程度の画像で Stable Diffusion 2.1 のような画像生成モデルが NVIDIA L4 で安全に作れることがわかった。

今後の課題として、人手評価の方法の確立、動画生成への知識転移方法やテキストエンコーダーの改良等が挙げられる。

文　献

[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

(注6)：https://cocodataset.org/

(注7)：https://developers.cyberagent.co.jp/blog/archives/38532/

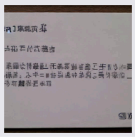
プロンプト→ モデル ↓	Pastry, water and a cup of coffee with spoon.	A gold and black motorcycle is parked by some grass.	白い花瓶に白や赤色の花が飾ってある	草原に顔を地面に付けたシマウマがいる
Stable Diffusion 2.1				
Common Canvas-S-C				
Japanese Stable Diffusion				
提案手法				

図 1　各種プロンプトに対する各モデルの生成結果（行: モデル , 列: プロンプト）

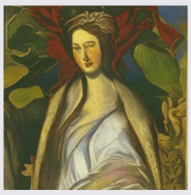
学習テキスト→	The image shows a painting of a woman with long hair and a crown on her head, surrounded by leaves and other objects. The painting is by Alphonse Mucha, a renowned French artist known for his work in the Art Nouveau style. The woman in the painting is wearing a white dress with a blue and gold patterned shawl draped over her shoulders, and her hair is pulled back in a bun. The leaves around her are a mix of green and yellow, and the background is a deep blue.	The image shows a man taking a selfie on a rocky mountain. He is wearing a black jacket and a purple beanie, and has a backpack on his back. The man is smiling and holding a camera in his hand. In the background, there are other hikers on the mountain, and the sky is blue with white clouds. The mountains are covered in green vegetation, and there is a layer of fog or mist in the air. The overall mood of the image is peaceful and serene.	この画像は、日本の伝統的な絵画スタイルの一つである「浮世絵」から取られているものです。画面中央に、一人の男性が手を上げて剣を握っている様子が描かれています。彼の髪は丸く束ねられており、表情は厳しく、目は鋭く見渡しています。背景には、網掛けや木々などが描かれ、その中には何かが見えません。この作品は、人物の表情と動きから物語を推測することができます。
学習画像			
提案手法の出力画像			

図 2　メンバーシップ推論攻撃の結果（列: プロンプト）

表 2　従来手法と比べた際の提案手法の定量評価結果。各指標の意味は本文参照。太字は比較した中で最も良かった値であり、イタリック体は２番目に良かった指標である。提案手法の値はおおよそ１番目か２番目に良かった値であることがわかる。

	FID-30k ↓	CLIP Score ↑	日本語 FID-30k ↓	日本語 CLIP Score ↑
Stable Diffusion 2.1 base	*15.6*	**31.3**	74.2	15.2
Japanese Stable Diffusion	56.2	18.6	19.5	**30.9**
CommonCanvas-S-C	**9.39**	**31.3**	177	11.7
提案手法 (CommonArt β)	22.8	*26.8*	**19.4**	*29.4*

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.

[2] James Betker, Gabriel Goh, Li Jing, Tim Brooks, Jianfeng Wang, Linjie Li, Long Ouyang, Juntang Zhuang, Joyce Lee, Yufe Guo, Wesam Manassra, Prafulla Dhariwal, Casey Chu, and Aditya Ramesh. Improving image generation with better captions. *preprint*, 2023.

[3] Ollin Boer Bohan. Megalith-10m. `https://huggingface.co/datasets/madebyollin/megalith-10m`, June 2024. Accessed: 2024-10-07.

[4] Junsong Chen, Chongjian Ge, Enze Xie, Yue Wu, Lewei Yao, Xiaozhe Ren, Zhongdao Wang, Ping Luo, Huchuan Lu, and Zhenguo Li. Pixart-σ: Weak-to-strong training of diffusion transformer for 4k text-to-image generation. In *The 18th European Conference on Computer Vision*, 2024.

[5] 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.

[6] Mehdi Cherti, Romain Beaumont, Ross Wightman, Mitchell Wortsman, Gabriel Ilharco, Cade Gordon, Christoph Schuhmann, Ludwig Schmidt, and Jenia Jitsev. Reproducible scaling laws for contrastive language-image learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2818–2829, 2023.

[7] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *North American Chapter of the Association for Computational Linguistics*, 2019.

[8] Aaron Gokaslan, A. Feder Cooper, Jasmine Collins, Landan Seguin, Austin Jacobson, Mihir Patel, Jonathan Frankle, Cory Stephenson, and Volodymyr Kuleshov. Commoncanvas: Open diffusion models trained on creative-commons images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8250–8260, June 2024.

[9] Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. CLIPScore: A reference-free evaluation metric for image captioning. In Marie-Francine Moens, Xuanjing Huang, Lucia Specia, and Scott Wen-tau Yih, editors, *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7514–7528, Online and Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics.

[10] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017.

[11] Imagen-Team-Google, :, Jason Baldridge, Jakob Bauer, Mukul Bhutani, Nicole Brichtova, Andrew Bunner, Kelvin Chan, Yichang Chen, Sander Dieleman, Yuqing Du, Zach Eaton-Rosen, Hongliang Fei, Nando de Freitas, Yilin Gao, Evgeny Gladchenko, Sergio Gómez Colmenarejo, Mandy Guo, Alex Haig, Will Hawkins, Hexiang Hu, Huilian Huang, Tobenna Peter Igwe, Christos Kaplanis, Siavash Khodadadeh, Yelin Kim, Ksenia Konyushkova, Karol Langner, Eric Lau, Shixin Luo, Soňa Mokrá, Henna Nandwani, Yasumasa Onoe, Aäron van den Oord, Zarana Parekh, Jordi Pont-Tuset, Hang Qi, Rui Qian, Deepak Ramachandran, Poorva Rane, Abdullah Rashwan, Ali Razavi, Robert Riachi, Hansa Srinivasan, Srivatsan Srinivasan, Robin Strudel, Benigno Uria, Oliver Wang, Su Wang, Austin Waters, Chris Wolff, Auriel Wright, Zhisheng Xiao, Hao Xiong, Keyang Xu, Marc van Zee, Junlin Zhang, Katie Zhang, Wenlei Zhou, Konrad Zolna, Ola Aboubakar, Canfer Akbulut, Oscar Akerlund, Isabela Albuquerque, Nina Anderson, Marco Andreetto, Lora Aroyo, Ben Bariach, David Barker, Sherry Ben, Dana Berman, Courtney Biles, Irina Blok, Pankil Botadra, Jenny Brennan, Karla Brown, John Buckley, Rudy Bunel, Elie Bursztein, Christina Butterfield, Ben Caine, Viral Carpenter, Norman Casagrande, Ming-Wei Chang, Solomon Chang, Shamik Chaudhuri, Tony Chen, John Choi, Dmitry Churbanau, Nathan Clement, Matan Cohen, Forrester Cole, Mikhail Dektiarev, Vincent Du, Praneet Dutta, Tom Eccles, Ndidi Elue, Ashley Feden, Shlomi Fruchter, Frankie Garcia, Roopal Garg, Weina Ge, Ahmed Ghazy, Bryant Gipson, Andrew Goodman, Dawid Górny, Sven Gowal, Khyatti Gupta, Yoni Halpern, Yena Han, Susan Hao, Jamie Hayes, Amir Hertz, Ed Hirst, Tingbo Hou, Heidi Howard, Mohamed Ibrahim, Dirichi Ike-Njoku, Joana Iljazi, Vlad Ionescu, William Isaac, Reena Jana, Gemma Jennings, Donovon Jenson, Xuhui Jia, Kerry Jones, Xiaoen Ju, Ivana Kajic, Christos Kaplanis, Burcu Karagol Ayan, Jacob Kelly, Suraj Kothawade, Christina Kouridi, Ira Ktena, Jolanda Kumakaw, Dana Kurniawan, Dmitry Lagun, Lily Lavitas, Jason Lee, Tao Li, Marco Liang, Maggie Li-Calis, Yuchi Liu, Javier Lopez Alberca, Peggy Lu, Kristian Lum, Yukun Ma, Chase Malik, John Mellor, Inbar Mosseri, Tom Murray, Aida Nematzadeh, Paul Nicholas, João Gabriel Oliveira, Guillermo Ortiz-Jimenez, Michela Paganini, Tom Le Paine, Roni Paiss, Alicia Parrish, Anne Peckham, Vikas Peswani, Igor Petrovski, Tobias Pfaff, Alex Pirozhenko, Ryan Poplin, Utsav Prabhu, Yuan Qi, Matthew Rahtz, Cyrus Rashtchian, Charvi Rastogi, Amit Raul, Ali Razavi, Sylvestre-Alvise Rebuffi, Susanna Ricco, Felix Riedel, Dirk Robinson, Pankaj Rohatgi, Bill Rosgen, Sarah Rumbley, Moonkyung Ryu, Anthony Salgado, Sahil Singla, Florian Schroff, Candice Schumann, Tanmay Shah, Brendan Shillingford, Kaushik Shivakumar, Dennis Shtatnov, Zach Singer, Evgeny Sluzhaev, Valerii Sokolov,

Thibault Sottiaux, Florian Stimberg, Brad Stone, David Stutz, Yu-Chuan Su, Eric Tabellion, Shuai Tang, David Tao, Kurt Thomas, Gregory Thornton, Andeep Toor, Cristian Udrescu, Aayush Upadhyay, Cristina Vasconcelos, Alex Vasiloff, Andrey Voynov, Amanda Walker, Luyu Wang, Miaosen Wang, Simon Wang, Stanley Wang, Qifei Wang, Yuxiao Wang, Ágoston Weisz, Olivia Wiles, Chenxia Wu, Xingyu Federico Xu, Andrew Xue, Jianbo Yang, Luo Yu, Mete Yurtoglu, Ali Zand, Han Zhang, Jiageng Zhang, Catherine Zhao, Adilet Zhaxybay, Miao Zhou, Shengqi Zhu, Zhenkai Zhu, Dawn Bloxwich, Mahyar Bordbar, Luis C. Cobo, Eli Collins, Shengyang Dai, Tulsee Doshi, Anca Dragan, Douglas Eck, Demis Hassabis, Sissie Hsiao, Tom Hume, Koray Kavukcuoglu, Helen King, Jack Krawczyk, Yeqing Li, Kathy Meier-Hellstern, Andras Orban, Yury Pinsky, Amar Subramanya, Oriol Vinyals, Ting Yu, and Yori Zwols. Imagen 3, 2024.

[12] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: bootstrapping language-image pre-training with frozen image encoders and large language models. In *Proceedings of the 40th International Conference on Machine Learning*, ICML'23. JMLR.org, 2023.

[13] Peiyuan Liao, Xiuyu Li, Xihui Liu, and Kurt Keutzer. The artbench dataset: Benchmarking generative models with artworks, 2022.

[14] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning, 2023.

[15] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2019.

[16] Cheng Lu, Yuhao Zhou, Fan Bao, Jianfei Chen, Chongxuan Li, and Jun Zhu. Dpm-solver: A fast ode solver for diffusion probabilistic model sampling in around 10 steps. *arXiv preprint arXiv:2206.00927*, 2022.

[17] Yang Luo, Xiaozhe Ren, Zangwei Zheng, Zhuo Jiang, Xin Jiang, and Yang You. Came: Confidence-guided adaptive memory efficient optimization. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4442–4453, 2023.

[18] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67, 2020.

[19] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10684–10695, June 2022.

[20] Kei Sawada, Tianyu Zhao, Makoto Shing, Kentaro Mitsui, Akio Kaga, Yukiya Hono, Toshiaki Wakatsuki, and Koh Mitsuda. Release of pre-trained models for the Japanese language. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 13898–13905, 5 2024. https://arxiv.org/abs/2404.01657.

[21] Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, Patrick Schramowski, Srivatsa Kundurthy, Katherine Crowson, Ludwig Schmidt, Robert Kaczmarczyk, and Jenia Jitsev. Laion-5b: An open large-scale dataset for training next generation image-text models. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, editors, *Advances in Neural Information Processing Systems*, volume 35, pages 25278–25294. Curran Associates, Inc., 2022.

[22] Makoto Shing and Kei Sawada. rinna/japanese-stable-diffusion.

[23] R. Shokri, M. Stronati, C. Song, and V. Shmatikov. Membership inference attacks against machine learning models. In *2017 IEEE Symposium on Security and Privacy (SP)*, pages 3–18, Los Alamitos, CA, USA, may 2017. IEEE Computer Society.

[24] Gowthami Somepalli, Vasu Singla, Micah Goldblum, Jonas Geiping, and Tom Goldstein. Diffusion art or digital forgery? investigating data replication in diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6048–6058, 2023.

[25] Gowthami Somepalli, Vasu Singla, Micah Goldblum, Jonas Geiping, and Tom Goldstein. Understanding and mitigating copying in diffusion models. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023.

[26] Chameleon Team. Chameleon: Mixed-modal early-fusion foundation models. *arXiv preprint arXiv:2405.09818*, 2024.

[27] Bart Thomee, David A. Shamma, Gerald Friedland, Benjamin Elizalde, Karl Ni, Douglas Poland, Damian Borth, and Li-Jia Li. Yfcc100m: the new data in multimedia research. *Commun. ACM*, 59(2):64–73, January 2016.

[28] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023.

[29] Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Yang Fan, Kai Dang, Mengfei Du, Xuancheng Ren, Rui Men, Dayiheng Liu, Chang Zhou, Jingren Zhou, and Junyang Lin. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*, 2024.

[30] B. Xiao, H. Wu, W. Xu, X. Dai, H. Hu, Y. Lu, M. Zeng, C. Liu, and L. Yuan. Florence-2: Advancing a unified representation for a variety of vision tasks. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4818–4829, Los Alamitos, CA, USA, jun 2024. IEEE Computer Society.

[31] Shuhei Yokoo, Shuntaro Okada, Peifei Zhu, Shuhei Nishimura, and Naoki Takayama. Clip japanese base.

[32] Yuya Yoshikawa, Yutaro Shigeto, and Akikazu Takeuchi. Stair captions: Constructing a large-scale japanese image caption dataset. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 417–421, Vancouver, Canada, July 2017. Association for Computational Linguistics.

[33] AI 戦略会議. Ai に関する暫定的な論点整理. May 2024.