CommonCanvas: An Open Diffusion Model Trained with Creative-Commons Images

Anonymous Authors[1]

Training high-quality text-to-image (T2I) models currently requires a lot of data. A lack of curated datasets that are large enough for the task has led people to turn to web-scraped solutions (Lee et al., 2023b;a), like LAION-2B (LAION-2Ben, 2022). The use of web-scraped data is a common practice, however, courts have yet to definitively rule that use of web-scraped data for training purposes is permissible under copyright law (Doe 1 v. GitHub, Inc., 2022; J.L. v. Alphabet Inc., 2023; Getty Images (US), Inc. v. Stability AI, Inc., 2023; Kadrey v. Meta Platforms, Inc., 2023; Tremblay v. OpenAI, Inc., 2023; Anderson v. Stability AI, Ltd., 2023). Some recent work has begun to investigate alternative methods of navigating copyright concerns (Min et al., 2023; Vyas et al., 2023; GitHub, 2023; Kumari et al., 2023), but has not addressed training T2I models. This raises a natural question: *Is it possible to efficiently produce a high-quality T2I model by training only on Creative-Commons-licensed data?*

We suggest a possible path forward, training a suite of T2I architectures using only open licensed, Creative-Commons (CC) images (Figures 1). This brings to light two significant challenges. 1) data incompleteness (CC images lack the captions necessary to train a high-quality T2I model) and 2) data scarcity (there are relatively few high-resolution CC images). We solve incompleteness with *telephoning*: an intuitive variant of transfer learning, which we use to synthesize captions for CC images. To investigate scarcity, we train multiple Stable Diffusion 2 (SD2-base) latent diffusion models (LDMs) on differently-sized subsets of LAION-2B, and find that models of this size saturate training on $< 3\%$ of LAION-2B. These results encourage us to train *CommonCanvas*, a suite of LDM architectures trained on our curated CC-image-synthetic-caption dataset, *CommonCatalog*. Our largest model achieves performance comparable to SD2-base on human evaluation of Parti Prompts (Yu et al., 2022), even though our CommonCatalog training dataset is $< 3\%$ the size of LAION and has synthetic captions.

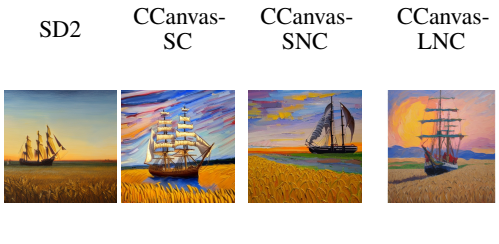

Figure 1: Using CC images and synthetic captions ($< 3\%$ the size of LAION-2B), we achieve comparable performance to SD2. We include results for two CommonCanvas architectures, small (S) and large (L), and two CC-image datasets, commercial (C) and non-commercial (SC) (Appendix **??** & **??**).

Preliminaries and motivation. T2I generative models (e.g., Stable Diffusion, or SD (Rombach et al., 2022)) refer to large neural networks trained on paired image-caption data. SD is a latent diffusion model (LDM) that converts images in latent representations and back again using Variational Autoencoders (VAEs) (Kingma and Welling, 2014); it uses an iterative sampling procedure (Sohl-Dickstein et al., 2015) and trains an underlying UNet (Ronneberger et al., 2015). The architecture also includes a text encoder, such as the Contrastive Language-Image Pre-training (CLIP) model (Podell et al., 2023). Stable Diffusion 2 (SD2)'s UNet has approximately 865 million trainable parameters; Stable Diffusion XL (SDXL) is larger, with 2.6 billion parameters, and has other advancements involving aspect ratio bucketing, micro-conditioning, and multiple text encoders and tokenizers. In terms of training data, the SD-family of models and OpenCLIP are both trained on subsets of the LAION-5B dataset (Beaumont, 2022; Schuhmann et al., 2022).

LAION-5B is a dataset derived from a snapshot of the Common Crawl, a massive corpus of data scraped from the web. From this snapshot, the LAION organization curated pairs of image URLs and their alt-text captions for the intended use of training T2I and I2T generative models (Beaumont, 2022; Schuhmann et al., 2022). Training models on this dataset requires visiting the URLs and downloading the associated images. It is often not known what the original image sources are (Lee et al., 2023b;a): they have unclear *provenance*. Courts have not yet decided if training on these datasets is "fair use" — an important exception in copyright (Leval, 1990; Sobel, 2017; Lee et al., 2023a; Samuelson, 2023). In the interim, there are several lawsuits for the alleged use of LAION-5B subsets to train generative models (Anderson v. Stability AI, Ltd., 2023; J.L. v. Alphabet Inc., 2023; Getty Images (US), Inc. v. Stability AI, Inc., 2023; Vincent, 2023). Further, since the datasets only contain the image URLs, they are plagued with *link rot* (Lakic et al., 2023).When accessing LAION-5B, there is no guarantee the images still exist at their URLs, making it impossible to fully reproduce the dataset and opening up the possibility of data poisoning attacks (Carlini et al., 2023).

Experiments. Equipped with commercial

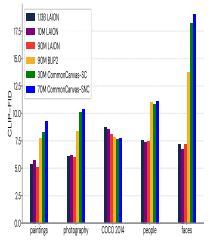

Figure 2: CLIP-FID for different models: Domain shift between MS-COCO and web-scraped conceptual captions. CLIP-FID likely favors SD2, as CLIP is trained on a similar style of text as LAION.

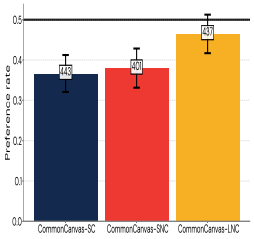

Figure 3: User preference study using Parti prompts. CommonCanvas-LNC model matches the performance of SD2 despite being trained with $< 3\%$ the amount of data.

(CommonCatalog-C) and non-commercial (CommonCatalog-NC) datasets, we train two different CommonCanvas models. We additionally train a larger variant of CommonCanvas-NC (CommonCanvas-LNC) that has a significantly larger U-Net. Figure 1 displays qualitative results from each of these model variants. For more details, see Appendix **??**. We measure performance with three automated image quality metrics on the MS COCO dataset (Lin et al., 2014): Frechet Inception Distance (FID) (Hessel et al., 2021), Kernal Inception Distance (KID) (Bińkowski et al., 2018), and CLIP-FID (Kynkäänniemi et al., 2022). Additionally, CLIP Score was evaluated to understand the alignment between captions and their respective images. Our model demonstrated comparable performance compared to the baseline of SD2 on the popular MS COCO benchmark. Like any model, ours has limitations. It underperformed in several categories, including faces, general photography, and paintings. These categories originated from the Conceptual Captions dataset (Sharma et al., 2018), which relies on web-scraped data. While abundant, web-sourced captions are often fully or semi-automatically generated or otherwise low quality (Nguyen et al., 2023), and may not always align with human-generated language nuances.

Human pairwise preference ratings for the three 512x512 resolution CommonCanvas models compared to SD2-base can be seen in Figure 3. In this experiment, human raters were shown a prompt (selected randomly from the PartiPrompts prompts set (Yu et al., 2022)) along with two generated images in randomized order, one from the reference model (SD2-base) and the other from a CommonCanvas model. We report the fraction of the time users selected the image generated by the CommonCanvas model over the corresponding generation from SD2 as the user preference rate for that model. We find that the two small CommonCanvas models are less perferred than SD2-base, with preference rates of 37% for CommonCanvas-SC and 38% for CommonCanvas-SNC, which we find surprisingly high considering the smaller and synthetic nature of the dataset. For the largest

Figure 4: We compare CommonCanvas-SNC (Ours) to SD2. Our model is less likely to generate iconic characters given suggestive prompts (drawn from Lee et al. (2023a)).

model, CommonCanvas-LNC, we do not measure a statistically significant difference in user preference between this model and SD2-base, indicating comparable model quality.

Although we train on Creative-Commons images, it is still possible for an adversarial prompt to produce content that, for example, includes iconic characters. In the Appendix (Figure 4), we subject our model to ambiguous prompts that are suggestive of such characters. Qualitatively, our model deviated more from these characters than SD2.

Discussion and related work. We note that several recent works study copyright. This work tends to concern text-to-text training data (Min et al., 2023), be primarily theoretical (Vyas et al., 2023; Scheffler et al., 2022), involve ablation studies (Kumari et al., 2023), or only handle verbatim memorization (Carlini et al., 2021) through the use of generation-time content filters (GitHub, 2023), which has been shown to be an incomplete solution (Ippolito et al., 2023). To the best of our knowledge, no prior open work attempts to train T2I models on only open licensed data. Our work on telephoning aligns with the trend of using advanced generative models to address data scarcity. This is evident in various modalities, such as producing audio captions from image-text pairs (Xiao et al., 2023) and text from audio (Radford et al., 2023). Similar approaches have also been used to generate instruction tuning datasets (Li et al., 2023). We coin this term to shorthand processes like these, which we believe will become more prevalent as generative models progress. Most prior work on text-caption-dataset creation has focused on extracting caption data from Common Crawl (Gadre et al., 2023; Desai et al., 2021; Laurençon et al., 2023) or re-captioning low-quality captions (Nguyen et al., 2023). We instead focus on synthesizing captions directly by using a pre-trained BLIP-2 model.

References

Anderson v. Stability AI, Ltd., 2023. No. 3:23-cv-00201 (N.D. Cal. Jan. 13, 2023).

Romain Beaumont. LAION-5B: A New Era of Large-Scale Multi-Modal Datasets. *LAION Blog*, March 2022. URL https://laion.ai/blog/laion-5b/.

Mikołaj Bińkowski, Danica J Sutherland, Michael Arbel, and Arthur Gretton. Demystifying mmd gans. *arXiv preprint arXiv:1801.01401*, 2018.

Nicholas Carlini, Florian Tramèr, Eric Wallace, Matthew Jagielski, Ariel Herbert-Voss, Katherine Lee, Adam Roberts, Tom Brown, Dawn Song, Úlfar Erlingsson, Alina Oprea, and Colin Raffel. Extracting Training Data from Large Language Models. In *30th USENIX Security Symposium (USENIX Security 21)*, pages 2633–2650. USENIX Association, August 2021. ISBN 978-1-939133-24-3.

Nicholas Carlini, Matthew Jagielski, Christopher A. Choquette-Choo, Daniel Paleka, Will Pearce, Hyrum Anderson, Andreas Terzis, Kurt Thomas, and Florian Tramèr. Poisoning Web-Scale Training Datasets is Practical, 2023.

Karan Desai, Gaurav Kaul, Zubin Aysola, and Justin Johnson. Redcaps: Web-curated image-text data created by the people, for the people. *arXiv preprint arXiv:2111.11431*, 2021.

Doe 1 v. GitHub, Inc., 2022. No. 4:22-cv-06823 (N.D. Cal. November 3, 2022).

Samir Yitzhak Gadre, Gabriel Ilharco, Alex Fang, Jonathan Hayase, Georgios Smyrnis, Thao Nguyen, Ryan Marten, Mitchell Wortsman, Dhruba Ghosh, Jieyu Zhang, et al. DataComp: In search of the next generation of multimodal datasets, 2023.

Getty Images (US), Inc. v. Stability AI, Inc., 2023. No. 1:23-cv-00135 (D. Del. February 3, 2023).

GitHub. Configuring github copilot in your environment, 2023. URL https://docs.github.com/en/copilot/configuring-github-copilot/configuring-github-copilot-in-your-environment.

Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. Clipscore: A reference-free evaluation metric for image captioning. *arXiv preprint arXiv:2104.08718*, 2021.

Daphne Ippolito, Florian Tramèr, Milad Nasr, Chiyuan Zhang, Matthew Jagielski, Katherine Lee, Christopher A. Choquette-Choo, and Nicholas Carlini. Preventing Verbatim Memorization in Language Models Gives a False Sense of Privacy, 2023.

J.L. v. Alphabet Inc., 2023. No. 3:23-cv-03440-LB (N.D. Cal July 11, 2023).

Kadrey v. Meta Platforms, Inc., 2023. No. 3:23-cv-03417 (N.D. Cal. July 7, 2023).

Dirk P. Kingma and Max Welling. Auto-Encoding Variational Bayes. In *International Conference on Learning Representations*, 2014.

Nupur Kumari, Bingliang Zhang, Sheng-Yu Wang, Eli Shechtman, Richard Zhang, and Jun-Yan Zhu. Ablating Concepts in Text-to-Image Diffusion Models, 2023.

Tuomas Kynkäänniemi, Tero Karras, Miika Aittala, Timo Aila, and Jaakko Lehtinen. The role of imagenet classes in fr\'echet inception distance. *arXiv preprint arXiv:2203.06026*, 2022.

LAION-2Ben, 2022. URL https://huggingface.co/datasets/laion/laion2B-en. Accessed September 23, 2023.

Viktor Lakic, Luca Rossetto, and Abraham Bernstein. Link-Rot In Web-Sourced Multimedia Datasets. In *MultiMedia Modeling: 29th International Conference, MMM 2023, Bergen, Norway, January 9–12, 2023, Proceedings, Part I*, page 476–488, Berlin, Heidelberg, 2023. Springer-Verlag. ISBN 978-3-031-27076-5. doi: 10.1007/978-3-031-27077-2_37. URL https://doi.org/10.1007/978-3-031-27077-2_37.

Hugo Laurençon, Lucile Saulnier, Léo Tronchon, Stas Bekman, Amanpreet Singh, Anton Lozhkov, Thomas Wang, Siddharth Karamcheti, Alexander M. Rush, Douwe Kiela, Matthieu Cord, and Victor Sanh. OBELICS: An Open Web-Scale Filtered Dataset of Interleaved Image-Text Documents, 2023.

Katherine Lee, A. Feder Cooper, and James Grimmelmann. Talkin' 'Bout AI Generation: Copyright and the Generative-AI Supply Chain, 2023a.

Katherine Lee, A. Feder Cooper, James Grimmelmann, and Daphne Ippolito. AI and Law: The Next Generation, 2023b.

Pierre N. Leval. Toward a Fair Use Standard. *Harvard Law Review*, 103(5):1105, 1990.

Xian Li, Ping Yu, Chunting Zhou, Timo Schick, Luke Zettlemoyer, Omer Levy, Jason Weston, and Mike Lewis. Self-alignment with instruction backtranslation. *arXiv preprint arXiv:2308.06259*, 2023.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, pages 740–755. Springer, 2014.

Sewon Min, Suchin Gururangan, Eric Wallace, Hannaneh Hajishirzi, Noah A. Smith, and Luke Zettlemoyer. SILO Language Models: Isolating Legal Risk In a Nonparametric Datastore, 2023.

Thao Nguyen, Samir Yitzhak Gadre, Gabriel Ilharco, Sewoong Oh, and Ludwig Schmidt. Improving multimodal datasets with image captioning. *arXiv preprint arXiv:2307.10350*, 2023.

Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. SDXL: Improving Latent Diffusion Models for High-Resolution Image Synthesis, 2023.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. Robust speech recognition via large-scale weak supervision. In *International Conference on Machine Learning*, pages 28492–28518. PMLR, 2023.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-Resolution Image Synthesis with Latent Diffusion Models. In *2022 IEEE Conference on Computer Vision and Pattern Recognition*, 2022.

Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-Net: Convolutional Networks for Biomedical Image Segmentation. *Medical Image Computing and Computer-Assisted Intervention*, pages 234–241, 2015.

Pamela Samuelson. Generative AI meets copyright. *Science*, 381(6654):158–161, 2023. doi: 10.1126/science.adi0656. URL https://www.science.org/doi/abs/10.1126/science.adi0656.

Sarah Scheffler, Eran Tromer, and Mayank Varia. Formalizing Human Ingenuity: A Quantitative Framework for Copyright Law's Substantial Similarity. In *Proceedings of the Symposium on Computer Science and Law*, pages 37–49, 2022.

Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, et al. LAION-5B: An open large-scale dataset for training next generation image-text models. *Advances in Neural Information Processing Systems*, 35: 25278–25294, 2022.

Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2556–2565, 2018.

Benjamin L.W. Sobel. Artificial Intelligence's Fair Use Crisis. *Columbia Journal of Law and The Arts*, 41:45, 2017.

Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathany, and Surya Ganguli. Deep Unsupervised Learning using Nonequilibrium Thermodynamics. In *Proceedings of the 32nd International Conference on Machine Learning*, 2015.

Tremblay v. OpenAI, Inc., 2023. No. 3:23-cv-03223 (N.D. Cal. June 28, 2023).

James Vincent. Getty Images is suing the creators of AI art tool Stable Diffusion for scraping its content. *The Verge*, January 2023. URL https://www.theverge.com/2023/1/17/23558516/ai-art-copyright-stable-diffusion-g

Nikhil Vyas, Sham Kakade, and Boaz Barak. On Provable Copyright Protection for Generative Models, 2023.

Feiyang Xiao, Qiaoxi Zhu, Jian Guan, Xubo Liu, Haohe Liu, Kejia Zhang, and Wenwu Wang. Synth-ac: Enhancing audio captioning with synthetic supervision. *arXiv preprint arXiv:2309.09705*, 2023.

Jiahui Yu, Yuanzhong Xu, Jing Yu Koh, Thang Luong, Gunjan Baid, Zirui Wang, Vijay Vasudevan, Alexander Ku, Yinfei Yang, Burcu Karagol Ayan, et al. Scaling autoregressive models for content-rich text-to-image generation. *arXiv preprint arXiv:2206.10789*, 2(3):5, 2022.