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

Aaron Gokaslan[1] **A. Feder Cooper**[1] **Jasmine Collins**[2] **Landan Seguin**[2] **Austin Jacobson**[2]

Mihir Patel[2] **Jonathan Frankle**[2] **Cory Stephenson**[2] **Volodymyr Kuleshov**[1]
[1]Cornell University ({akg87,afc78,vk379}@cornell.edu)
[2]Databricks ({firstname.lastname}@databricks.com)

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 [23, 24], like LAION-2B [20]. 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 [1, 9, 11, 15, 16, 51]. Some recent work has begun to investigate alternative methods of navigating copyright concerns [12, 18, 32, 53], 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 [57], even though our CommonCatalog training dataset is $<3\%$ the size of LAION and has synthetic captions.

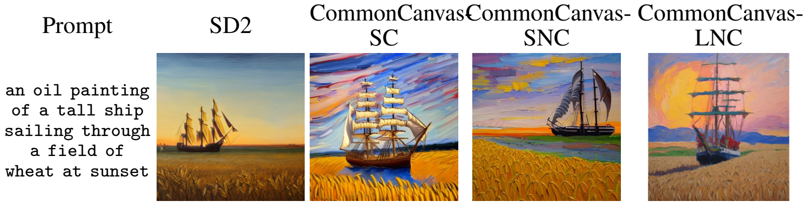

Prompt
SD2
CommonCanvas-SC
CommonCanvas-SNC
CommonCanvas-LNC
an oil painting of a tall ship sailing through a field of wheat at sunset

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 B & C.2).

Preliminaries and motivation. T2I generative models (e.g., Stable Diffusion, or SD [40]) 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) [17]; it uses an iterative sampling procedure [48] and trains an underlying UNet [41]. The architecture also includes a text encoder, such as the Contrastive Language-Image Pre-training (CLIP) model [36]. 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 [3, 45].

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 [3, 45]. Training

37th Conference on Neural Information Processing Systems (NeurIPS 2023).

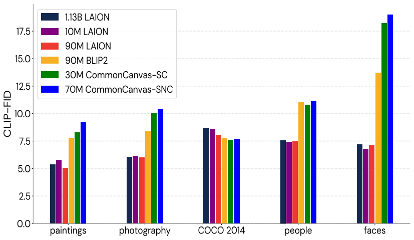


Figure 2: CLIP-FID for different models: Domain shift between MS-COCO and web-scraped conceptual captions. CLIP-FID likely favors SD2, as CLIP is trained on a similar style of text as LAION.

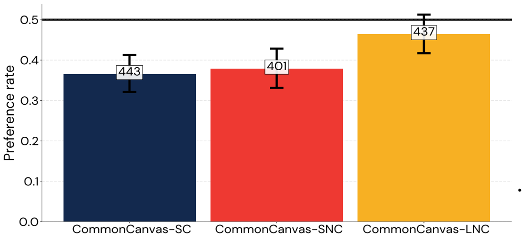


Figure 3: User preference study using Parti prompts. CommonCanvas-LNC model matches the performance of SD2 despite being trained with $< 3\%$ the amount of data.

models on this dataset requires visiting the URLs and downloading the associated images. It is often not known what the original image sources are [23, 24]: they have unclear *provenance*. Courts have not yet decided if training on these datasets is “fair use" — an important exception in copyright [23, 27, 42, 47]. In the interim, there are several lawsuits for the alleged use of LAION-5B subsets to train generative models [1, 11, 15, 52]. Further, since the datasets only contain the image URLs, they are plagued with *link rot* [21].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 [6].

Experiments. Equipped with commercial (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 C.2. We measure performance with three automated image quality metrics on the MS COCO dataset [30]: Frechet Inception Distance (FID) [13], Kernal Inception Distance (KID) [4], and CLIP-FID [19]. 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 [46], which relies on web-scraped data. While abundant, web-sourced captions are often fully or semi-automatically generated or otherwise low quality [35], 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 [57]) 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 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 8), 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 [32], be primarily theoretical [43, 53], involve ablation studies [18], or only handle verbatim memorization [5] through the use of generation-time content filters [12], which has been shown to be an incomplete solution [14]. 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 [55] and text from audio [39]. Similar approaches have also been used to generate instruction tuning datasets [29]. 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 [8, 10, 22] or re-captioning low-quality captions [35]. We instead focus on synthesizing captions directly by using a pre-trained BLIP-2 model.

References

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

[2] Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E Hinton. Layer normalization. *arXiv preprint arXiv:1607.06450*, 2016.

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

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

[5] 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.

[6] 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.

[7] Tri Dao, Daniel Y. Fu, Stefano Ermon, Atri Rudra, and Christopher Ré. FlashAttention: Fast and memory-efficient exact attention with IO-awareness. In *Advances in Neural Information Processing Systems*, 2022.

[8] 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.

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

[10] 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.

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

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

[13] 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.

[14] 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.

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

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

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

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

[19] 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.

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

[21] 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`.

[22] 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.

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

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

[25] Benjamin Lefaudeux, Francisco Massa, Diana Liskovich, Wenhan Xiong, Vittorio Caggiano, Sean Naren, Min Xu, Jieru Hu, Marta Tintore, Susan Zhang, Patrick Labatut, and Daniel Haziza. xFormers: A modular and hackable Transformer modelling library. `https://github.com/facebookresearch/xformers`, 2022.

[26] Mark A. Lemley. How Generative AI Turns Copyright Law on its Head, 2023. URL `https://ssrn.com/abstract=4517702orhttp://dx.doi.org/10.2139/ssrn.4517702`.

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

[28] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. *arXiv preprint arXiv:2301.12597*, 2023.

[29] 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.

[30] 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.

[31] Susan Box Mann. The Telephone Game, 2019. URL `https://icebreakerideas.com/telephone-game/`. Accessed September 27, 2023.

[32] 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.

[33] The Mosaic ML Team. composer. `https://github.com/mosaicml/composer/`, 2021.

[34] The Mosaic ML Team. streaming. `<https://github.com/mosaicml/streaming/>`, 2022.

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

[36] 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.

[37] Jacob Portes, Alexander R Trott, Sam Havens, Daniel King, Abhinav Venigalla, Moin Nadeem, Nikhil Sardana, Daya Khudia, and Jonathan Frankle. Mosaicbert: How to train bert with a lunch money budget. In *Workshop on Efficient Systems for Foundation Models@ ICML2023*, 2023.

[38] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning Transferable Visual Models From Natural Language Supervision. In *Proceedings of the 38th International Conference on Machine Learning*, 2021.

[39] 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.

[40] 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.

[41] 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.

[42] 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`.

[43] 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.

[44] Christoph Schuhmann. LAION-400-Million Open Dataset, 2021. URL `https://laion.ai/blog/laion-400-open-dataset/`.

[45] 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.

[46] 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.

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

[48] 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.

[49] Stability AI. Stable Diffusion v2-base Model Card, 2022. URL `https://huggingface.co/stabilityai/stable-diffusion-2-base`.

[50] 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. *Communications of the ACM*, 59(2):64–73, 2016.

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

[52] 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-getty-images-lawsuit`.

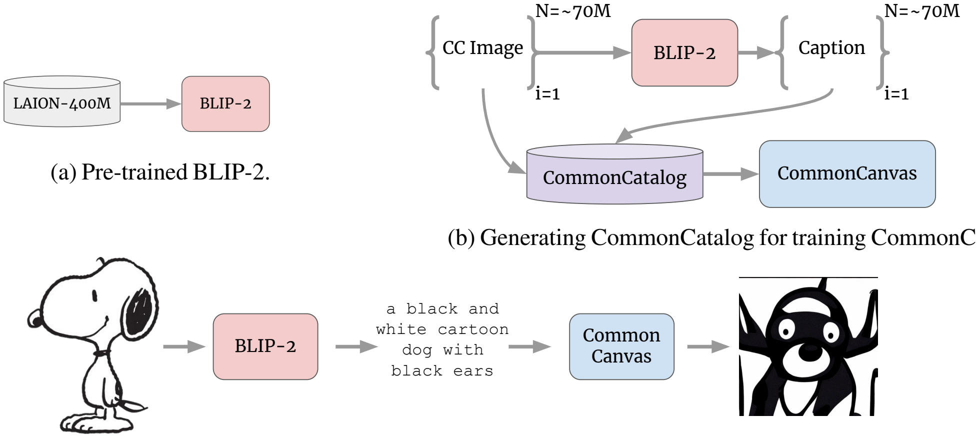
[53] Nikhil Vyas, Sham Kakade, and Boaz Barak. On Provable Copyright Protection for Generative Models, 2023.

[54] Yuxin Wu and Kaiming He. Group normalization. In *Proceedings of the European conference on computer vision (ECCV)*, pages 3–19, 2018.

[55] 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.

[56] Yuanzhong Xu, HyoukJoong Lee, Dehao Chen, Hongjun Choi, Blake Hechtman, and Shibo Wang. Automatic cross-replica sharding of weight update in data-parallel training. *arXiv preprint arXiv:2004.13336*, 2020.

[57] 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.



(a) Pre-trained BLIP-2.

(b) Generating CommonCatalog for training CommonCanvas.

(c) "Lossy compression" from image to BLIP-2 caption, back to image (via CommonCanvas generation).

Figure 4: *Telephoning*. (**a**) LAION's massive dataset of image-caption pairs is used to train BLIP-2, an image-to-text model. (**b**) We leverage BLIP-2 to produce synthetic captions for our caption-less CC images, and use the resulting synthetic image-caption pairs (the *CommonCatalog* dataset) to train our open diffusion model, *CommonCanvas*. (**c**) Although BLIP-2 was trained on LAION (e.g., a picture of Snoopy), the captions it produces behave like a "lossy compression." Like a game of telephone, the reconstructed output transforms to something that no longer resembles the original.

A Telephoning: A transfer-learning, image-captioning method

A natural alternative is to not use LAION datasets for training. One could instead independently curate a dataset of CC-licensed images with known provenance that expressly allow for copying, adaptation, and commercial use. As constituent images can be stored and distributed, this would also solve the link rot problem, thereby enabling greater reproducibility. While an attractive alternative to LAION-5B, CC images rarely contain the captions necessary to train T2I models.

Our solution for handling the lack of captions in CC images is called *telephoning*, an intuitive variant of transfer learning (Figure 4). Telephoning benefits from a simple analogy: it takes inputs from a high-dimensional modality (e.g., images), effectively performs a "lossy compression" to a low-dimensional modality (e.g., to a short-text caption), and then decompresses back to the high-dimensional modality. Because the intermediate compression step is "lossy", the ultimate output does not remotely resemble the original input, just like a game of telephone [31]. We derive the term telephoning from the above intuition, and employ it as useful shorthand to denote instances of transfer learning that solve data-scarcity problems in multimodal generative modeling.

In this work, CC images are the high-dimensional inputs, and we use a pre-trained BLIP-2 model (Figure 4a, Li et al. [28]) for "lossy compression" to short-text captions. Together, these CC-image-caption pairs comprise the CommonCatalog dataset (Section B), which we use to train our CommonCanvas T2I models (Figure 4b, Section **??**). Even though BLIP-2 was pre-trained on LAION-400M [44], CommonCatalog and CommonCanvas never have direct access to LAION-400M or, importantly, anything that is similar to the images that BLIP-2 was trained on. Instead, we only have access to the mapping in the model, which, given an image input, produces "lossy" output text that inherently does not literally resemble its image counterpart (Figure 4c).We defer to experts about fair use (Section **??**), with respect to LAION-5B's images and alt-text captions, and to models like BLIP-2 [23, 26, 42]. Generally, these experts seem to think many cases will fall under fair use, especially when model outputs do not resemble their inputs, which is the case with BLIP-2.

B CommonCatalog: A Dataset of CC Images & Synthetic Captions

In this section we introduce our open dataset, *CommonCatalog*. First, we describe the collection and curation process for open licensed, CC images. This process brings to light two challenges: caption-data incompleteness and image-data scarcity. To address the lack of CC captions, we show how telephoning can be instantiated to produce high-quality synthetic captions to accompany our set of curated images. We investigate the topic of data scarcity in the next section, where we also discuss necessary ML-systems optimizations that enable us to efficiently train several SD models.

Sourcing provenanced, licensed images for CommonCatalog. We focus on locating high-resolution Creative-Commons images that have open licenses. We began with the YFCC100M dataset, which consists of 100 million CC-licensed images and multimedia files, as well as Flickr IDs linking to the original data [50]. The images in the dataset associated with the original paper exhibit two issues that make it ill-suited for direct use to train Stable Diffusion: they are low-resolution, and many of them have licenses that do not expressly allow for the distribution of derivative works, which are an area of unsettled copyright law in the context of model training. We therefore re-scraped these images from Flickr, based on their IDs provided in the YFCC100M metadata. Our scraped images are very high resolution (exceeding 4k), which makes them suitable for training.

Some in the community believe that licenses prohibiting derivative works should not be used for training models. We don't have a view on that, but we also excluded non-derivative (ND) images from the dataset. The remaining images can be further divided into those that can be used for commercial (C) purposes and those that cannot (non-commercial/ NC). As shown in Table 5, we accordingly construct two datasets, CommonCatalog-C and CommonCatalog-NC. We defer additional details about licenses to Appendix D.1.1, but emphasize that all of the images we include have open licenses: we are free to use, adapt, and remix the images, so long as we attribute them. Ultimately, we are left with roughly 70 million NC CC-images, of which roughly 25 million can be used commercially.

Directly sourcing CommonCatalog avoids some concerns (Section **??**); however, it also comes with its own challenges. For one, CC images rarely have the alt-text captions necessary to train a text-to-image model like Stable Diffusion (Figure 5); those that do have associated text often just include the image title or a URL (Appendix). For another, we could *only* find roughly 70 million usable CC images, which pales in comparison to the 1+ billion LAION-2B images used to train SD2 (Section H). We take each of these challenges in turn. First, in the next subsection, we show how we instantiate telephoning (Section A) to produce high-quality, synthetic captions for CC images.

Synthesizing captions with telephoning. We compared a several captioning models and, based on qualitative analysis and its state-of-the-art performance on MS COCO, chose to use the pre-trained BLIP-2 OPT2.5B model for synthesizing captions [28]. BLIP-2 consists of three components: a pre-trained, fixed (i.e., frozen) visual encoder, a learned transformer network that converts the visual embeddings into a text prompt, and a frozen LLM that takes in the prompt. The LLM helps impart natural-language knowledge to the model, ensuring that the distribution of synthesized captions match patterns in English-natural language. The only trainable variables in the transformers are between the frozen visual encoder and frozen LLM layers.

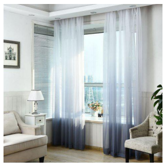
Given a LAION-2B image as input, we found that the resulting BLIP-2 caption is often qualitatively more descriptive than the corresponding LAION-2B ground-truth alt-text caption. LAION-2B captions often contain product names, irrelevant details, or poor grammar and syntax (Figure 6). This finding is corroborated by Nguyen et al. [35], which shows quantitatively (in terms of CLIP Score) that BLIP-2 captions are higher quality than ground-truth captions, at the cost of caption diversity.

Based on these preliminary results, we captioned all of the YFCC100M Creative-Commons images, which required about 1,120 GPU A100 hours. To do so, we center-cropped and resized all of the images to a maximum size of 512x512 pixels. We perform these transformations because captioning images at native resolution would be needlessly expensive. We release our commercial (CommonCatalog-C) and non-commercial (CommonCatalog-NC) CC-image and synthetic-caption datasets on HuggingFace at [REDACTED]. As an evaluation set, we also release the BLIP-2 captions that we produced for the non-derivative (ND) CC images that we did not use for training.

C Additional Details on Data Scarcity Analysis

C.1 Hypothesis: Diffusion Models are Too Small

A back-of-the-envelope calculation provides some insight on why this is the case. Consider a training dataset consisting of N images with resolution $H \times W$ and c channels. To completely memorize the training data, the model must be capable of storing $c \times H \times W \times N$ numbers. Given a number of trainable parameters N_p, it is natural to assume that on average each parameter is capable of storing roughly enough information to reconstruct a single number from the training dataset. Under this



Source	Caption
Alt-Text (LAION-2B)	Latest 1PC Transparent Gradient Color Voile Window Curtain
BLIP2-OPT-2.7B	A living room with a white couch and curtains

Figure 6: Original vs. BLIP-2-generated captions for an image from LAION-2B. BLIP-2 generates a caption that better aligns with what a human would write. See Figure 11 for more examples.

assumption, complete memorization is only possible if the size of the training dataset is at or below a critical size N_c ($N \leq N_c$) with N_c given by $N_c = \frac{N_p}{cHW}$. Note that this critical size assumes the data cannot be further compressed, which is obviously not the case for natural images. However, SD2 and SDXL are latent diffusion models, which first use a pretrained encoder to compress images by a factor of 8 in both H and W, and so when we train LDMS like SD2 and SDXL, we are training on data that has been significantly compressed already.

In our experiments, $c = 4$ and $H = W = 32$, corresponding to 256×256 resolution RGB images in the SD2 and SDXL latent space. The SD2 UNet has $N_p = 866 \times 10^6$ trainable parameters, and SDXL's UNet has $N_p = 2567 \times 10^6$. So we calculate $N_c \approx 0.2 \times 10^6$ for SD2 and $N_c \approx 0.6 \times 10^6$ for CommonCanvas-Large; both of these numbers are several orders of magnitude below the size of our YFCC derived datasets, and so even with significant additional data compression we expect that our CommonCatalog datasets should be sufficient to train both SD2 and SDXL. Additionally, this argument predicts that we should only begin to see significant overfitting in these models for datasets of size $N \sim 10^6$. These estimates are resolution dependent, and as image resolution increases we expect that N_c will decrease as more information is provided per image.

C.2 Increasing Model Capacity with CommonCanvas-LNC

We also train a variant of SD2 with more trainable parameters, taking the UNet from SDXL. We refer to this model as CommonCanvas-LNC. We adapt the SDXL UNet architecture to SD2 by changing the cross-attention dimensionality to match that of the SD2 text encoder hidden state dimensionality (1024 for SD2 vs. 2048 for SDXL). SDXL also retrains the VAE component in their model, and we use this improved performance VAE as well. Except for these changes, the architecture is identical to that of SD2.

D Training Dataset Details

D.1 LAION-2B

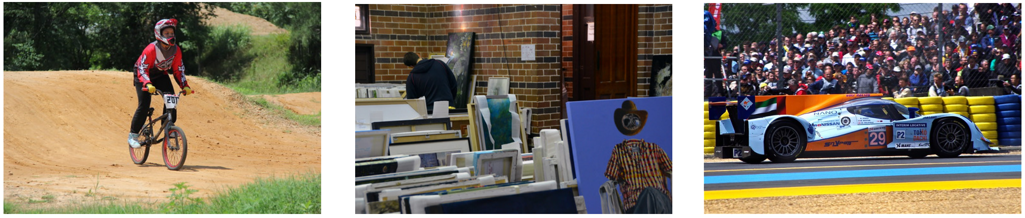
The fact that LAION is not a stable benchmark can lead to multiple reproducability and security issues. Data poisoning attacks would be difficult to detect at the scale of 2 billion parameters. While this could be mitigated by using hash values of the images, then any time the a site decide to re-encode the image, those images would now need to be excluded from the dataset. Furthermore, targeted data poisoning attacks for diffusion models are no longer just academic conjecture. Last year after the release of Stable Diffusion, a protest was launched on ArtStation that had uses upload images that said "NoAI" to taint future training data for generative models after artists felt as though their work had been unfairly used to train the models. With the high degree of link rot, targeted attacks are fairly easy. Furthermore, reproduction of the experiments becomes virtually impossible. This means any benchmarks that use copies of LAION as ground truth are are likely using differing subsets of the full dataset.

D.1.1 Sourcing Creative-Commons images

Table 1: CC licenses in YFCC100M. ND means derivative works are not licensed or the license doesn't allow the user to create derivative works. NC means images cannot be used in commercial contexts. CommonCatalog-C only contains data from the bottom two (yellow) rows, reflecting images licensed for commercial contexts (i.e., roughly 25 million images). CommonCatalog-NC contains CommonCatalog-C, and additionally includes the middle two (blue) rows, reflecting images licensed for non-commercial purposes. We do not include the roughly 30 million images in the top two (pink) rows in CommonCatalog, as they are non-derivative licenses. We do not train on these images. We do, however, produce BLIP-2 captions for them and release those captions as an evaluation set.

CC License	# Images	% Captioned
CC-BY-NC-ND-2.0	25,790,117	33.52%
CC-BY-ND-2.0	4,827,970	30.23%
CC-BY-NC-2.0	12,468,229	31.39%
CC-BY-NC-SA-2.0	28,314,685	31.57%
CC-BY-SA 2.0	9,270,079	34.05%
CC-BY 2.0	16,962,338	28.96%

Table 2: Randomly sampled images from the YFCC [50] training set. Our synthetic BLIP2 captions are also provided below.



a person riding a bike on a dirt road

a paintings on the wall

an orange and blue race car driving on a track

D.1.2 Release and documentation

We release CommonCatalog at [REDACTED], with an associated data sheet.

E YFCC Example Images

Table 3: This table shows the top 10 captions in the YFCC dataset. The most common captions are not user generated and would require quite a lot of cleaning be of any use. OpenAI released a subset of YFCC that may be usable for captioned data, but ultimately,

YFCC Original Caption	**Count**
OLYMPUS+DIGITAL+CAMERA	184889
SONY+DSC	123128
Exif_JPEG_PICTURE	104480
Barclays+Center+Arena%0AAtlantic+Yards%0A6th+and+Atlantic+A	68832
Olympus+digital+camera	54805
Effortlessly+uploaded+by Eye-Fi	48388
.	43227
-+Camera+phone+upload+powered+by ShoZu	38856
Sony+dsc	32709
Photo+by @Kmeron |Facebook page is this way|	23754

F Training Details

F.1 Model Architecture

We follow the model architecture and training recipe of Stable Diffusion 2 as closely as we can to best reproduce the model for CC-Small. The model has an identical number of params and structure as the original model. In fact, we can even load SD2's model weights into our framework due to the identical architecture and naming scheme. We are able to achieve virtually identical performance with SD2 in a much shorter training time with less data. We use the same VAE, tokenizers, and UNet archicture as SD2 except for reducing the precision of the normalization layers.

Table 4: Number of usable captions from OpenAI's YFCC14M dataset [38]. This table is actually a subset from 1 for which either the user description or image title were deemed usable. These figures provide an estimate on how many images in each category are actually potentially usable as captions.

License Name	count
CC-BY 2.0	2448002
CC-BY-ND 2.0	682273
CC-BY-NC 2.0	1925854
CC-BY-NC-ND 2.0	4058817
CC-BY-NC-SA 2.0	4146113
CC-BY-SA 2.0	1568336

Table 5: Performance (throughput) and approximate cost of training SD2 UNet with our optimizations. Depending on the number of GPUs used, the cost to train the same models without these optimizations range from $90,000-$140,000

Number of A100s	256x256 (img/s)	512x512 (img/s)	512x512 with EMA (img/s)	Days to Train	Cost ($)
8	1100	290	290	101.04	$38,800.00
16	2180	585	580	50.29	$38,630.00
32	4080	1195	1160	25.01	$38,420.00
64	8530	2340	2220	12.63	$38,800.00
128	11600	4590	3927	6.79	$41,710.00

Our CC-Large model takes SD2's model and replaces the UNet with the SDXL architecture [36]. Like CC-Small, we also replace the normalization layers with their low-precision version. The replacement of all the normalization layers is handled automatically by MosaicML's Composer library [33]. We perform all dataloading through MosaicML's streaming library [34]

F.2 Software and hardware speed-ups

Stability AI reports an estimated 200,000 A100 hours to train SD2 [49]. Depending on the available hardware, a single SD2 run could take anywhere from a few weeks to over a month to train. We sought out multiple avenues to reduce this training-time constraint. Ultimately we were able to achieve a speedup of 2.71X relative to the original SD2 implementation.

First, we applied Flash Attention [7] with the xFormers library [25]. We also pre-computed VAE and text encoder latents over the entire training dataset, cast all GroupNorm [54] and LayerNorm [2] to float16 precision, and applied fully-sharded data parallelism (FSDP) to our training run. Finally we opted to only keep an exponential moving average of the weights for the final 3.5% of training. More detail on each of these improvements can be found in Appendix G.

When applying all of the aforementioned strategies together, we are able to achieve a 2.71X speedup in A100 hours over our SD2-baseline implementation. We found that latent pre-computation helped the most at low resolutions, while FSDP also provided significant gains, especially at scale. The other optimizations helped reduce total memory usage, allowing us to increase the microbatch size for better hardware utilization. Figure 7 summarizes each of the proposed methods and the cumulative speedup that results from its application. Equipped with an optimized training setup, we are able to more easily study effect of varying training dataset size.

G Additional Details on efficiency optimizations

In this section we provide additional details on the optimizations we implemented to achieve SD2 training speedups. We also report the approximate cost of training our implementation of SD2 on various hardware configurations in Table 5.

Flash Attention. Cross attention operations are a very expensive part of training that occurs in dozens of layers in diffusion model UNets [40]. Flash Attention is an efficient implementation that is optimized to work well with reduced precision and GPU hardware [7], which was implemented using the XFormers library [25], allowing us to save compute and memory usage.

Precomputing Latents. Each forward pass of SD2 requires computing a latent representation of the input image, as well as transforming the caption into a text embedding. Instead of computing the latents for each example during training, we can precompute latents for the entire dataset, amortizing

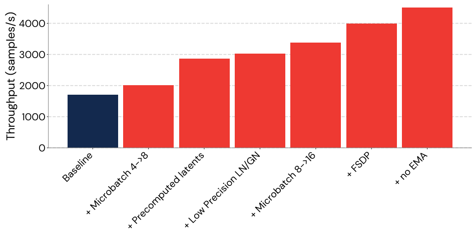
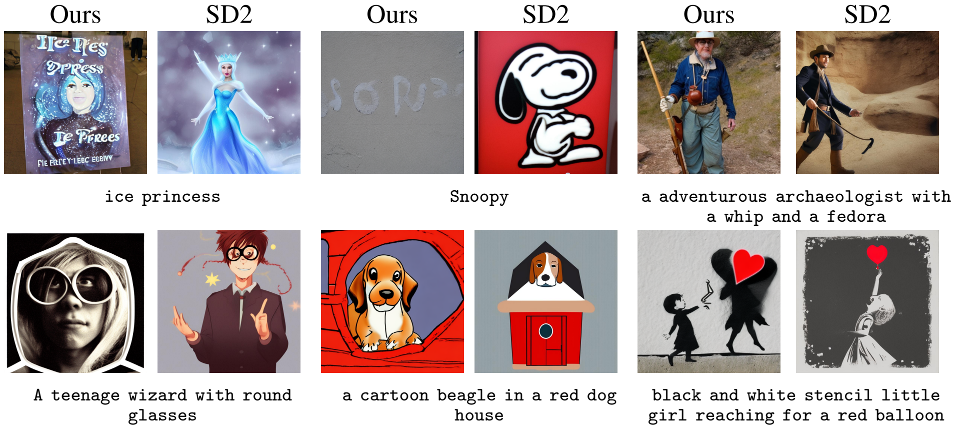


Figure 7: Cumulative effect of various speed-ups in our SD2 training pipeline.



ice princess

Snoopy

a adventurous archaeologist with a whip and a fedora

A teenage wizard with round glasses

a cartoon beagle in a red dog house

black and white stencil little girl reaching for a red balloon

Figure 8: We compare CommonCanvas-SNC (Ours) to SD2. Our model is less likely to generate iconic characters given suggestive prompts (drawn from Lee et al. [23]).

the cost. Doing so speeds up training of the model, especially at lower resolutions, in exchange for a one-time fixed cost of precomputing all the latents over 1 epoch.

Reduced-Precision GroupNorm and LayerNorm. Most layers in SD2 are implemented in float16 precision, but GroupNorm and LayerNorm are implemented in float32, in part because it was assumed to be necessary for training stability. The resulting, frequent upcasting causes a major bottleneck in training speed. Recent work shows that it is safe to implement LayerNorm using float16 precision [37], and we found the same to be true of GroupNorm. We thus cast all GroupNorm and LayerNorm operators to float16 and are able to further reduce total memory consumption and accelerate training.

Fully-Sharded Data Parallelism (FSDP). FSDP is a variant of data-parallel training that shards the models parameters, gradients and optimizer state across multiple devices. When training data batches do not fit into memory, we do several forward and backward passes on smaller microbatches, followed by a single gradient update. At GPU scale, there may only be a single microbatch, so the time for the gradient update can become a significant bottleneck. In standard data distributed training, each GPU communicates all its gradients to every other GPU, and then each GPU updates its local copy of the model. Instead, we use a different paradigm inspired by [56] where each GPU only gets the gradients and updates the weights for a small part of the model before sending the updated weights for that part of the model to all of the other GPUs. By dividing the update step across all the GPUs, we can ensure that the amount of work per GPU decreases as we increase the number of GPUs, helping us achieve linear scaling. To tackle this problem, we use PyTorch's experimental support for Fully Sharded Data Parallelism (FSDP), specifically, FSDP's SHARD_GRAD_OP mode.

Scheduled Exponential Moving Average (EMA). SD2 uses EMA, which maintains an exponential moving average of the weights at every gradient update for the entire training period. This can be slow due to the memory operations required to read and write all the weights at every step. Since the old weights are decayed by a factor of 0.9999 at every batch, the early iterations of training only contribute minimally to the final average. We decide to only apply EMA for the final 50K steps (about 3.5% of the training period), and are able to avoid adding overhead and still achieve a nearly equivalent EMA model.

H Training Efficiency Optimizations and Data Scarcity Analysis

High-resolution CC images are indeed much less abundant than arbitrary web-scraped ones, but the amount of data necessary to train high-quality SD2 models has not been well-studied. We set out to quantify this amount by training multiple SD2 models on differently-sized subsets of LAION-2B. However, training a single SD2 model, even with hundreds of GPUs, can take several days. To make our data scarcity analysis more tractable, we first implement several efficiency optimizations. For the sake of space, we detail these optimizations in the Appendix, and note that we were able to achieve a 2.71X training speedup relative to the original Stability AI SD2 implementation [49].

YFCC100M contains 1 million images, about 10% the size of the 1.1B LAION examples we could access, thus about 5% of the original LAION-2B dataset. We ask whether or not is necessary to train on 1+ billion images to get results that are as good as the original LAION-trained SD2. Our results show, surprisingly, that this is not the case, and with a slightly larger model (CommonCanvas-L), which replaces SD2's U-Net with SDXL's larger one [36]). Further, our larger model can achieve the same results on human evaluation as we do on a model using 33X more training data. We train

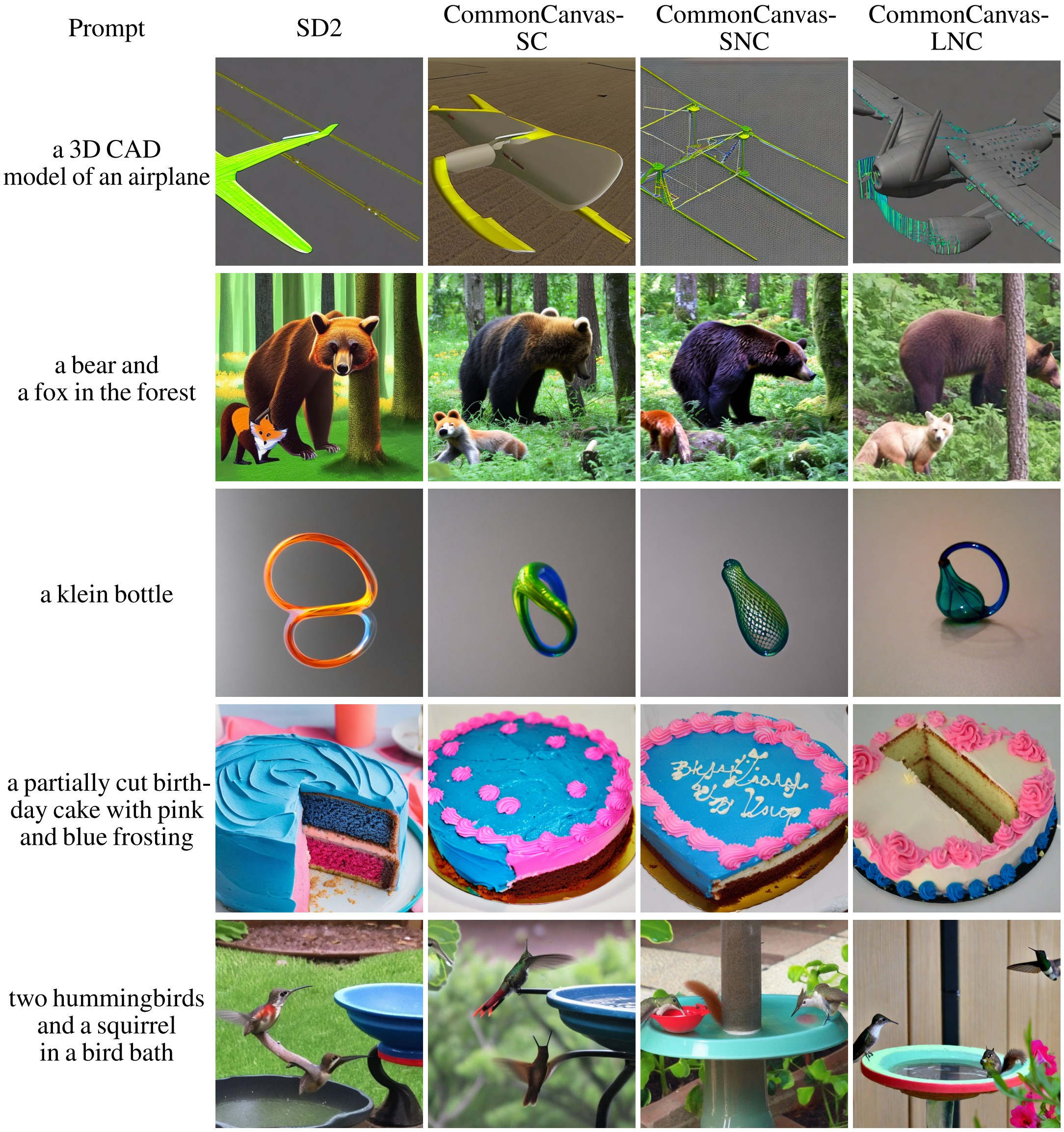

Prompt
SD2
CommonCanvas-SC
CommonCanvas-SNC
CommonCanvas-LNC
a 3D CAD model of an airplane
a bear and a fox in the forest
a klein bottle
a partially cut birthday cake with pink and blue frosting
two hummingbirds and a squirrel in a bird bath

Figure 9: Additional qualitative examples comparing SD2 to our model trained on the commerical split (CommonCanvas-SC), non-commerical split (CommonCanvas-SNC), and the larger UNet model trained on the non-commercial (CommonCanvas-LNC).

Figure 10: More qualitative examples of our models' outputs

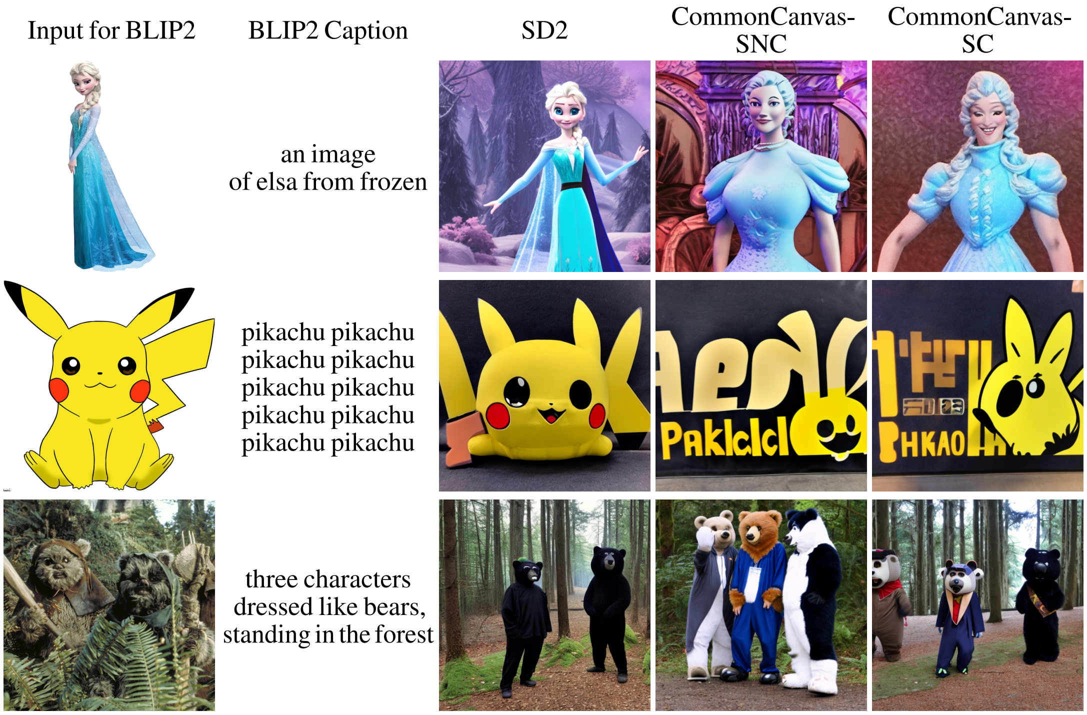

Input for BLIP2
BLIP2 Caption
SD2
CommonCanvas-SNC
CommonCanvas-SC
an image of elsa from frozen
pikachu pikachu pikachu pikachu pikachu pikachu pikachu pikachu pikachu pikachu
three characters dressed like bears, standing in the forest

Figure 11: Additional qualitative examples comparing our CommonCanvas models to SD2, given synthetic BLIP2 captions as prompts. While not perfect, our models are better at avoiding generating potentially problematic data.

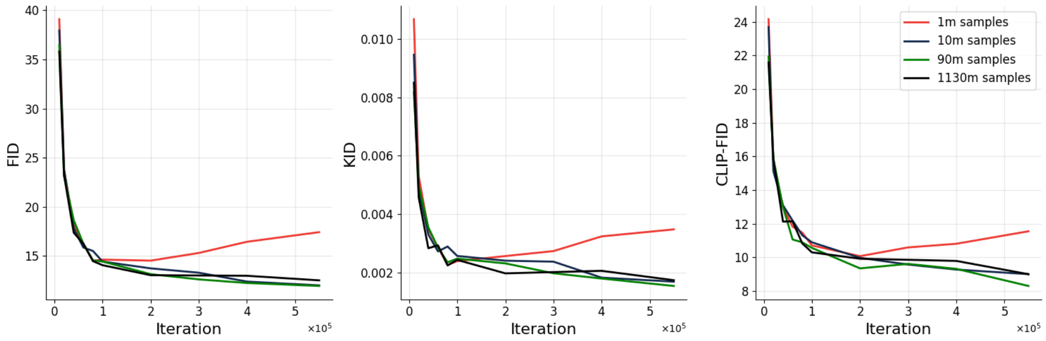


Figure 12: How does reducing the amount of training data affect the training dynamics? We find a noticeable improvement drop when training with less than 10 million samples.

on increasingly smaller, random subsets of data from our LAION-1.1B model and find that we can achieve a similar result on the commonly reported MS COCO numbers, but with <3% the amount of SD2's training data (Figure 13). In fact, we run experiments down to 1-million LAION-1.1B images, and find that only 10 million images are required for stable training behavior (Appendix, Figure 12).

This findings suggest that SD2 models may be underparameterized. In fact, when we use CommonCanvas-LNC, we achieve competitive performance with SD2 on user preferences, despite training on significantly less data (Section 3). Further, in spite of the drastic reduction in dataset size, we observe that the larger model (CommonCanvas-LNC) outperforms the smaller one (CommonCanvas-SNC), consistent with the notion that these models are still underparameterized. We hypothesize about why this might be the case and how much data is actually necessary to saturate the model in Appendix C.1.

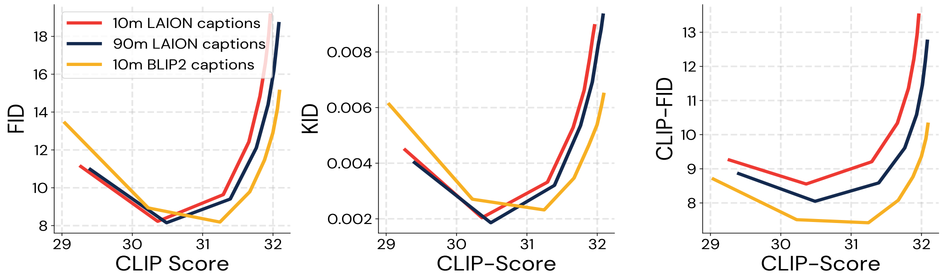


Figure 13: FID, KID, and CLIP-FID vs. CLIP-Score computed on 30k samples from COCO2014 for different SD2 models trained on smaller subsets of LAION (10M, 90M, using either original captions or synthetic BLIP2 captions. Interestingly, increasing the amount of training data from 10M to 90M samples does not lead to improved quantitative metrics across guidance scales 1 to 8.

Figure 5: CommonCatalog-C contains images licensed only for commercial use; -NC contains -C as well as images licensed for non-commercial use.

Dataset	**# Images**	**% Alt Text**
CommonCatalog-C	26,232,417	30.76%
CommonCatalog-NC	67,015,331	31.22%

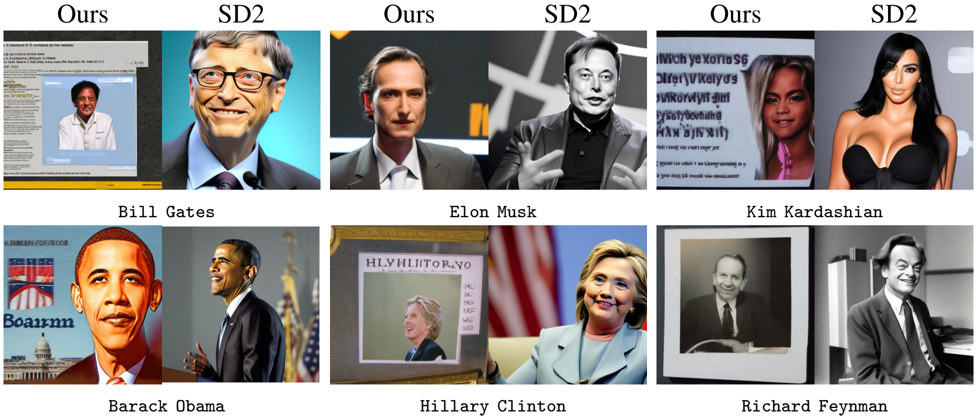


Figure 14: Using CommonCanvas-SNC (Ours) to generate celebrities. Our model is worse at synthesizing individual people than SD2, but is capable of generating some noteworthy public figures.