

# CommonCanvas: Open Diffusion Models Trained on Creative-Commons Images

Aaron Gokaslan<sup>1\*</sup>, A. Feder Cooper<sup>1,2</sup>, Jasmine Collins<sup>3</sup>, Landan Seguin<sup>3</sup>,  
Austin Jacobson<sup>3</sup>, Mihir Patel<sup>3</sup>, Jonathan Frankle<sup>3</sup>, Cory Stephenson<sup>3</sup>, Volodymyr Kuleshov<sup>1</sup>  
\*akg87@cornell.edu

<sup>1</sup>Cornell University <sup>2</sup>The GenLaw Center <sup>3</sup>Mosaic Research Databricks



Figure 1. We achieve comparable performance to public Stable Diffusion 2 (SD2), using entirely Creative-Commons images and a synthetic captioning approach that requires only  $<3\%$  of the amount of the data used to train previous models. We include results for two CommonCanvas architectures, small (S) and large (L), and two CC-image datasets, commercial (C) and non-commercial (NC).

## Abstract

We train a set of open, text-to-image (T2I) diffusion models on a dataset of curated Creative-Commons-licensed (CC) images, which yields models that are competitive with Stable Diffusion 2 (SD2). This task presents two challenges: (1) high-resolution CC images lack the captions necessary to train T2I models; (2) CC images are relatively scarce. To address these challenges, we use an intuitive transfer learning technique to produce a set of high-quality synthetic captions paired with our assembled CC images. We then develop a data- and compute-efficient training recipe that requires as little as 3% of the LAION data (i.e., roughly 70 million examples) needed to train existing SD2 models, but obtains the same quality. These results indicate that we have a sufficient number of CC images (also roughly 70 million) for training high-quality models. Our recipe also implements a variety of optimizations that achieve  $2.71\times$  training speed-ups, enabling rapid model iteration. We leverage this recipe to train several high-quality T2I models, which we dub the CommonCanvas family. Our largest model achieves comparable performance to SD2 on human evaluation, even though we use a synthetically captioned CC-image dataset that is only  $<3\%$  the size of LAION for training. We release our models, data, and code on [GitHub](#).

## 1. Introduction

Most high-quality text-to-image (T2I) models are trained using large-scale, web-scraped datasets, like LAION-

2B [34]. Even though this is a very common practice, U.S. courts have yet to definitively rule if this is permissible under copyright law [15, 17, 24, 25, 69]. In response, recent work in ML has begun to investigate alternative methods of navigating copyright concerns in text generation [44], code completion [18, 57], and image generation [28]. Nevertheless, matching the performance of state-of-the-art models remains a challenge. In this work, we study the following natural question: *is it possible to efficiently produce a high-quality T2I model by training only on Creative-Commons-licensed data?*

We suggest a path forward, training a suite of T2I architectures using *only* open-licensed, Creative-Commons (CC) images (Figures 1 & 2). This task brings to light two significant challenges. The first problem is data incompleteness: almost all CC images lack the captions necessary to train a high-quality T2I model. The second is data scarcity: there are relatively few high-resolution CC images — roughly 70 million, compared to LAION-2B’s roughly 2 billion [30].

We address the data incompleteness problem by using a pre-trained BLIP-2 model [39] to produce high-quality, synthetic captions for a set of curated, open-licensed CC images. This is an intuitive transfer-learning solution: we leverage a powerful pre-trained generative model to produce synthetic labels for an unlabeled dataset, which we can then use to train a different multimodal generative model. To deal with data scarcity, we propose a data- and compute-efficient training recipe that obtains the same quality as Stable Diffusion 2 (SD2) [64], but, perhaps surprisingly,

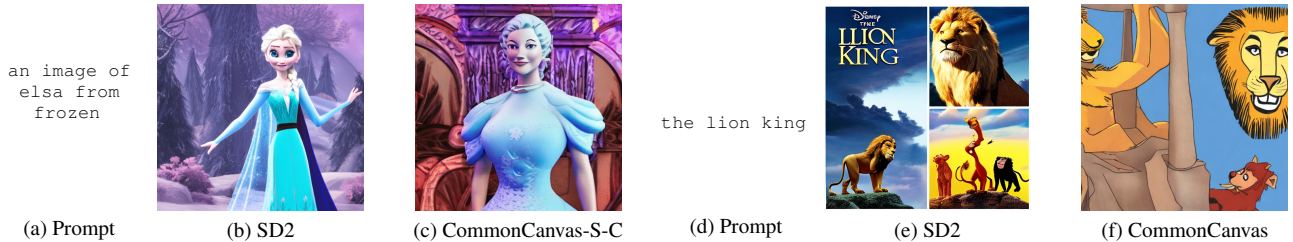


Figure 2. Prompting with Disney concepts (a, d). SD2 generates a recognizable image of Elsa from *Frozen* (b) and an image with a misshapen Disney logo and characters resembling those from *The Lion King* (e); CommonCanvas-S-C (small, commercial) does not (c, f).

requires as little as 3% of the LAION-2B data (i.e., roughly 70 million examples) originally used to train SD2. We call this model *SD2-90M*. These results indicate that we have a sufficient number of CC images (also roughly 70 million) for training high-quality models. Our training recipe also implements a variety of optimizations that achieve  $2.71\times$  training speed-ups, enabling rapid model iteration.

The above methods enable us to create *CommonCanvas*, a suite of latent diffusion model (LDM) architectures trained on our curated dataset of CC images and synthetic captions, which we denote *CommonCatalog*. For one of our architectures, we swap SD2’s UNet for SDXL’s larger network to demonstrate how, even with less data, larger models do not overfit to this smaller dataset. Our largest model (CommonCanvas-L-NC) achieves performance comparable to SD2-90M on human evaluation of Parti Prompts [75], even though our CommonCatalog training dataset is 3% the size of LAION and has synthetically generated captions. Although this is a larger and more capable model architecture than SD2, we find it surprising and important that it is possible to train an SD2-quality model *at all* based on such a limited dataset with synthetic captions. This reveals a promising path forward for future research on highly capable, open T2I models. In summary, we:

- Curate *CommonCatalog*, a multimodal training dataset of roughly 70 million open-licensed CC images (Section 4) for which we synthesize a set of high-quality captions. We note that synthesizing training data using generative models is an increasingly common transfer-learning technique, and we give it the shorthand name *telephoning* (Sections 3).
- Train *CommonCanvas*, a suite of LDM architectures trained on CommonCatalog. The largest of these models, CommonCanvas-L-NC, produces qualitative results that are competitive with public SD2 (Section 6). To make this analysis tractable, we implement training optimizations that achieve  $2.71\times$  speed-ups in training SD2-90M (Section 5).
- We will release our CommonCatalog dataset along with our trained CommonCanvas models at <https://github.com/mosaicml/diffusion/blob/main/assets/common-canvas.md>.

## 2. Preliminaries and Motivation

In this section, we present background on training the T2I Stable Diffusion model, which was originally trained on the web-scraped LAION-2B dataset. We then discuss copyright and reproducibility with respect to LAION datasets. This discussion motivates the creation of an alternative dataset composed of open-licensed CC images with synthetic captions, which we introduce in Section 4.

### 2.1. Text-to-image generative models

Text-to-image (T2I) generative models are neural networks trained on image-caption pairs. One family of T2I models is Stable Diffusion (SD) [53]: a latent diffusion model (LDM) that converts images to latent representations and back again using Variational Autoencoders (VAEs) [27], and which uses an iterative sampling procedure [63] to train an underlying UNet [54]. The architecture also includes a text encoder, such as the Contrastive Language-Image Pre-training (CLIP) model [49] — the original OpenAI CLIP [51] or its open-source counterpart, OpenCLIP [11, 22].

Stable Diffusion 2 (SD2)’s UNet has approximately 865 million trainable parameters; Stable Diffusion XL (SDXL) has 2.6 billion parameters and other advancements involving aspect ratio bucketing, micro-conditioning, and multiple text encoders and tokenizers. In terms of training data, SD models and OpenCLIP are both trained on subsets of the LAION-5B dataset [3, 59]. The exact training dataset for CLIP is unknown, but it is likely web-scraped data [51].

### 2.2. Copyright, reproducibility, & LAION datasets

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 corresponding alt-text captions for the intended use of training T2I and image-to-text (I2T) generative models [3, 59]. In practice, T2I models are typically trained on filtered subsets of the full LAION-5B dataset (e.g. LAION-2B [30]). Training T2I models on this dataset requires visiting the URLs and downloading the associated images. There are two elements of LAION datasets that are relevant to our work:

**Copyright.** The images associated with LAION datasets have unclear *provenance*: it is often not known what the original image sources are [34]. Although LAION datasets are released under the open MIT license, some experts note that it is unclear if this is sufficient to allow for training on the underlying images and captions, which often have their own copyrights [12, 19, 33–35]. Courts have not yet decided if training on these datasets is “fair use” — an important exception in copyright [33, 35, 38, 56, 62]. There are several copyright lawsuits for the alleged use of LAION-5B subsets to train generative models [1, 17, 24, 70, e.g.].

**Reproducibility.** Since LAION datasets only contain the image URLs, and not the images themselves, they are plagued with *link rot* [31].<sup>1</sup> 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 [9]. A natural alternative is to not use LAION datasets for training. Instead, one could independently curate a dataset of CC-licensed images with known provenance that explicitly allow for copying, adaptation, and commercial use. As constituent images can be stored and distributed, this would also solve the link-rot problem, enabling greater reproducibility. (Further, LAION datasets are no longer public because they contain CSAM [6, 67].) We defer our discussion of sourcing CC-licensed images to Section 4, where we detail CommonCatalog: our new, open dataset. While CC images are an attractive alternative to LAION-5B, we note that CC images rarely contain the captions necessary to train T2I models. Therefore, we first need a method for captioning CC images.

### 3. Transfer Learning for Image Captioning

Our solution for handling the lack of captions in CC images is an intuitive type of transfer learning for producing high-quality synthetic labels. We describe this method, and note that there are various similar methods in prior literature on generative modeling. Altogether, these methods indicate that this type of transfer learning has become an increasingly common pattern: producing synthetic labels that later serve as inputs to training other generative models. We therefore give this method a shorthand name: *telephoning*.

#### 3.1. Telephoning

Telephoning (Figure 3) proceeds in two steps. First, shown in Figure 3b, it takes inputs from a high-dimensional modality (e.g., images) and effectively performs a “lossy compression” to a (scarce) low-dimensional modality (e.g., short-text captions). Second, shown in Figure 3d, it takes the “lossy compression” and decompresses back to the high-dimensional modality. Because the intermediate compression step is “lossy,” the ultimate output often does not re-

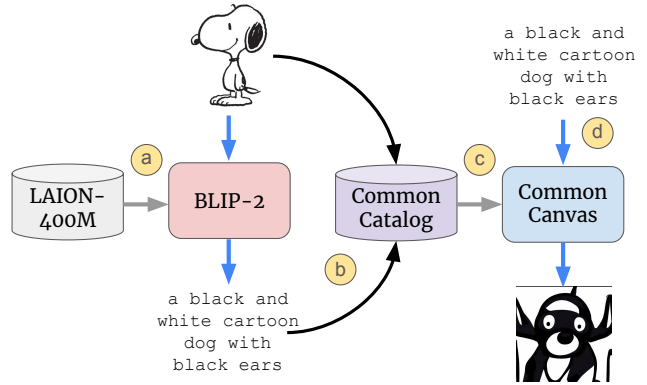


Figure 3. (a) We use the LAION-400M-pre-trained, I2T BLIP-2 model to produce synthetic captions for our uncaptioned CC images (e.g., the Wikipedia CC-licensed image of Snoopy). The synthetic captions are “lossy compressions” of the input images (e.g., a black and white cartoon dog with black ears has no mention of Snoopy). (b) We compile the resulting synthetic image-caption pairs into *CommonCatalog*, which (c) we use to train our open, T2I *CommonCanvas* models. (d) When we supply “lossy” captions to a T2I model, like a game of telephone, **it produces outputs that no longer resemble the original images** (e.g., *CommonCanvas* produces an image that matches the caption, but does not look like Snoopy).

motely resemble the original input, just like a game of telephone [43]. We derive the term telephoning from the above intuition and use it as 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 [39] for “lossy compression” to short-text captions (Figure 3a). Together, these CC-image-caption pairs comprise the *CommonCatalog* dataset (Section 4), which we use to train our *CommonCanvas* T2I models (Figure 3b). While BLIP-2 was pre-trained on LAION-400M [58], we emphasize that, for training *CommonCanvas*, we only ever have access to the captions — to the “lossy compressions” it produces. We 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.

**Telephoning & Copyright** We defer to experts about fair use (Section 2.2) — namely, regarding models like BLIP-2, and LAION-5B’s images and alt-text captions. Generally, these experts seem to think that many cases will fall under fair use [33, 37, 56], especially when model outputs do not resemble their inputs (i.e., the use is “non-expressive” or “non-consumptive” [12]). This is the case with our use of BLIP-2 to produce “lossy” captions.

Nevertheless, it is possible that BLIP-2 could produce captions that resemble those in its LAION training data. This might seem to present a copyright concern similar to

<sup>1</sup>This also applies to other web-scrapes, e.g., DataComp [16].



those that others have expressed about T2I generations that resemble LAION images. However, according to the U.S. Copyright Office, short phrases (like captions) may often not be copyrightable: “short phrases” often contain “an insufficient amount of authorship” to meet the threshold for copyright protection [66]. So, even if hypothetically BLIP-2 were to regurgitate captions from LAION verbatim, according to legal experts [33], the copyright considerations are likely to be different than they are for generated images or generated long-form text. We defer to experts for more precise legal arguments, but note that this is another reason why we believe it is reasonable for us to rely on BLIP-2 for captioning our CC images.

### 3.2. Related work on telephoning

Our work 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 [73] and text from audio [52]. Similar approaches have also been used to generate instruction-tuning datasets for both text and images [40, 42]. Concurrent work, e.g. LLaVA [42], has used visual question-answer models to augment existing caption datasets, such as the ones used in training DALLE-3 [4] and Chen et al. [10]. Our model is one of the first works to train on a dataset without any ground-truth captions, and one of the first to release our dataset along with a fully trained diffusion model. The caption upsampling approaches described in these other works could be used to further improve the captions of CommonCatalog in future work.

Captioning models have also been used to create descriptive captions to guide a diffusion model to create an image visually similar to a specific image. In concurrent work, SynthCap [7] generates a synthetic captioning dataset using a diffusion model to generate images from captions — the inverse of our problem statement. We coin the term telephoning to short-hand processes like these, which include our work and prior work, and which we believe will become more prevalent as generative-model capabilities advance.

## 4. A CC-Image, Synthetic-Caption Dataset

We now introduce our open dataset, *CommonCatalog*. First, we describe the collection and curation process for the 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 concretely how we use telephoning 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 systems-level training optimizations that enable efficient model iteration.

### 4.1. Sourcing 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 [68]. 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 — a use that is in unsettled copyright law in the context of model training [33].

We therefore re-scraped these images from Flickr, based on the IDs provided in the YFCC100M metadata. Our scraped images are of very high resolution (exceeding 4K), which makes them more suitable for T2I training. We exclude images with non-derivative (ND) licenses. The remaining images can be further divided into those that can be used for commercial (C) purposes and those that cannot (NC). As shown in Table 4, we accordingly construct two datasets, CommonCatalog-C and CommonCatalog-NC. We defer additional details about licenses to Appendix B.1.1, but emphasize that all of the included images have open licenses: individuals are free to use, adapt, and remix the images, so long as they attribute them. In total, CommonCatalog contains roughly 70 million images that can be used non-commercially, of which a approximately 25 million images can also be used commercially.

Directly sourcing CommonCatalog avoids some concerns (Section 2.2); however, it also comes with its own challenges. For one, CC images rarely have the alt-text captions necessary to train a T2I model like Stable Diffusion (Figure 4); those that do have associated text often just include the image title or a URL. For another, we could *only* find roughly 70 million usable CC images, which pales in comparison to the billions of images in LAION used to train SD2 (Section 5). We take each of these challenges in turn. First, in the next subsection, we show how we instantiate telephoning (Section 3) to produce high-quality, synthetic captions for CC images.

### 4.2. Synthesizing captions with telephoning

We compared several captioning models and chose the pre-trained BLIP-2 OPT2.5B model for synthesizing Common-

Figure 4. 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%




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 5. Original vs. BLIP-2-generated captions for an image from LAION-2B. In this example, BLIP-2’s caption better aligns with what a human would write. See appendix for more examples.

Catalog’s captions [39], based on qualitative analysis and state-of-the-art performance on MS COCO. BLIP-2 consists of three components: a pre-trained, frozen (i.e., fixed) visual encoder, a learned transformer network that converts the visual embeddings into a text prompt, and a frozen large language model (LLM) that takes in the prompt. The only trainable variables in the transformers are between the frozen visual encoder and the 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 5). This finding is corroborated by Nguyen et al. [48], which quantitatively shows that (in terms of CLIP Score) 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. We center-cropped and resized all of the images to a maximum size of 512x512 pixels, since captioning images at native resolution would be very expensive. At training time for CommonCanvas models, we use the high-resolution images.

We release our commercial (CommonCatalog-C) and non-commercial (CommonCatalog-NC) CC-image and synthetic-caption datasets with associated data cards. 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.

## 5. Optimizations and Data-Scarcity Analysis

High-resolution CC images are indeed much less abundant than web-scraped images; however, it is unclear if this scarcity presents a problem for training. Prior work has not studied in depth how much data is actually necessary to train high-quality SD2 models. 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. So, to make our data scarcity analysis more tractable, we first implemented several efficiency optimizations.

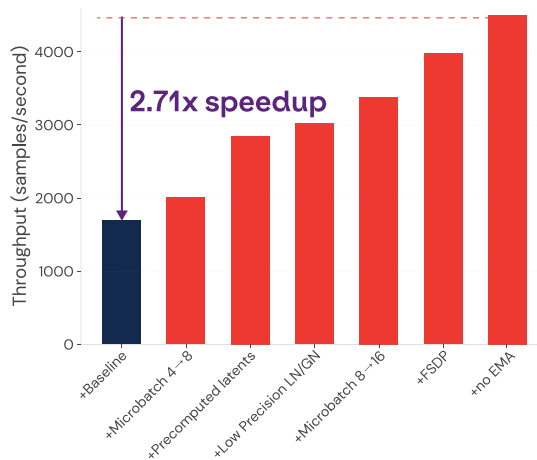


Figure 6. Cumulative effect of various speed-ups (totalling 2.71 $\times$ ) in our SD2 training pipeline evaluated on 128 A100s.

### 5.1. Software and hardware speed-ups

Stability AI reports an estimated 200,000 A100 hours to train SD2 [65]. Depending on hardware, a single SD2 training run could take anywhere from a few weeks to over a month. We sought out multiple avenues to reduce this training-time constraint. We applied Flash Attention [13] with the xFormers library [36], pre-computed VAE and text encoder latents over the entire training dataset, cast all GroupNorm [72] 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. Altogether, 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 6 summarizes each of the proposed methods and the cumulative speedup that results from their application. Equipped with an optimized training setup, it is more feasible for us to study the effect of varying training-dataset size. More details can be found in Appendix D.

### 5.2. Investigating data scarcity

YFCC100M contains 100 million images, about 10% the size of the 1.1B LAION examples we could access (due to link rot) — about 5% of the original LAION-2B dataset. An interesting question remains: *how much data is actually needed to train these diffusion models effectively; do we really need billions of images to get high-quality results?*

To answer this question, we train multiple SD2 architectures on increasingly smaller, random subsets of data from our LAION-1.1B dataset: 1.1B, 90M, 10M, and 1M sam-

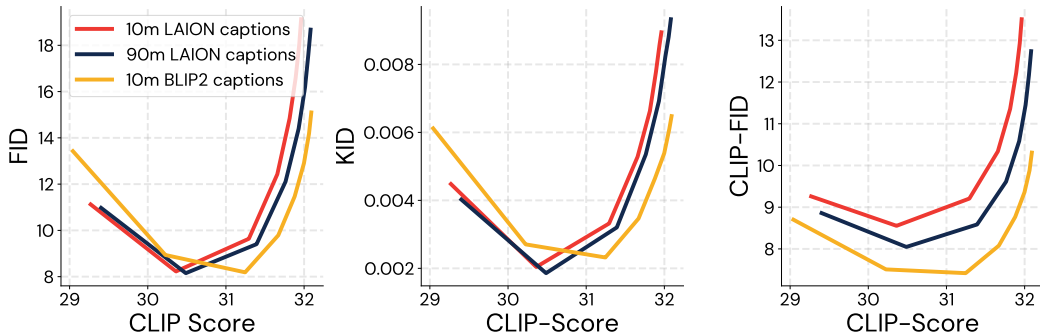


Figure 7. For different SD2 models trained on subsets of LAION (90M, 10M using either original captions or synthetic BLIP-2 captions), we compute FID [21], KID [5], CLIP-FID [29], and CLIP-Score [20] on 30K samples from MS COCO. We compute these metrics across a text-guidance scale of 1-8, with higher values indicating the model should respect the text prompt more. Lower FID, KID, and CLIP-FID indicate higher quality; higher CLIP-Score indicates higher quality. Together, these plots show that increasing the amount of training data from 10M to 90M samples does not lead to quantitative improvements. BLIP-2 re-captions provide nearly identical performance to LAION in terms of FID and KID; the re-captions indicate slightly better performance when using CLIP-FID as the quality metric.

ple subsets. While human evaluation remains the gold standard for evaluating generative models, we use proposed automated metrics like Fréchet-Inception Distance [21], Kernel Inception Distance [5] and caption-alignment metrics such as CLIP Score [20] (Section 6). We find that performance (FID and KID on MS COCO) does not degrade until training with as few as 1 million images; our models trained on 10M and 90M subsets perform comparably to the entire 1.1B dataset (Appendix Figure 16). Figure 7 further compares our SD2 variants trained on 10M and 90M LAION subsets across different guidance scales. We also plot the effect of using the original LAION captions vs. BLIP-2 synthetic captions at these size regimes (discussed further in Section 6.1). These findings suggest that SD2 models may be underparameterized. We hypothesize about why this might be the case and how much data is actually necessary to saturate the model in the appendix.

## 6. Experiments

In this section, our model evaluations use automated, quantitative image-quality metrics from the literature. We measure performance with three metrics on the commonly used MS COCO dataset [41]: Fréchet Inception Distance (FID) [21], Kernel Inception Distance (KID) [5], and CLIP-FID [29]. Each captures a slightly different measures of generated-image quality and diversity, in relation to statistics in the training data, with lower values corresponding to higher quality. Additionally, we evaluated CLIP-Score [20], which can help us understand the alignment between captions and their respective images, with higher values signaling better alignment. While these automated metrics are intended to be efficient proxies for human preferences in image quality, they often fall short; the gold standard for T2I model evaluation still remains human evaluation. Since synthetic captions differ so much from human-designed ones [48], we also set up a pairwise preference rating task

to measure the relative quality of our trained models.

### 6.1. Training with Synthetic Captions

First, we look at the effect of training with synthetic captions instead of ground-truth captions from LAION. Interestingly, we observe that synthetic captions can enhance the alignment of our model. For instance, the CLIP-Score for synthetic captions exceeded that of ground-truth captions as seen in Figure 7 (for CLIP-FID).

To get a more nuanced perspective on the effect of our synthetic captions, we assess CLIP-FID for image generations from different models on human- and computer-

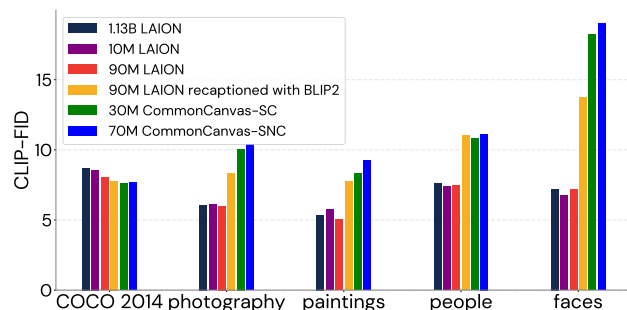


Figure 8. Evaluating models at 256 resolution on different subsets of the Conceptual Captions dataset and MS COCO. LAION models are trained on 1.1 billion, 90 million (SD2-90M), and 10 million subsets. We also train a model with a 90 million subset re-captioned with BLIP-2 to evaluate distribution shift. The last two models are trained on on the CommonCatalog-C, and CommonCatalog-NC. We observe a domain shift between MS COCO and web-scraped Conceptual Captions. CLIP-FID may exhibit a preference for SD2 models, given that CLIP has been trained on a text style akin to that found in LAION. Subsampling the LAION dataset from 1.13B to 10M images does not seem to affect quantitative performance. Using synthetic captions causes a significant performance drop on the LAION dataset when evaluated on Conceptual Caption test datasets, but not MS COCO.

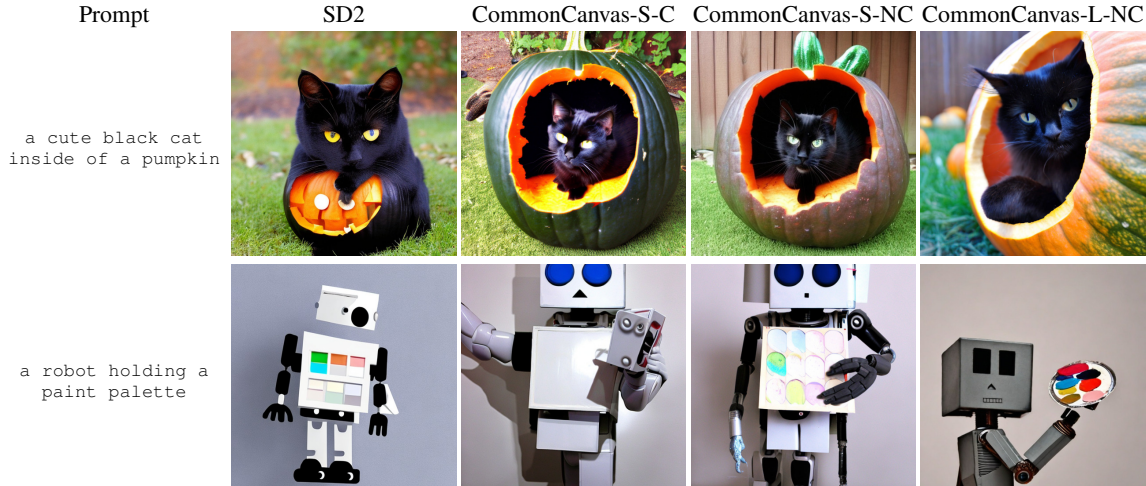


Figure 9. Using entirely Creative-Commons images and our synthetic captioning approach, we achieve comparable qualitative performance to public SD2, as seen in CommonCanvas generations, while only requiring a small fraction ( $< 3\%$ ) of the amount of training data. We include results for two CommonCanvas architectures, small (S) and large (L) (Section 6), and two CC-image datasets, commercial (C) and non-commercial (NC) (Section 4). We label our results accordingly as CommonCanvas-<architecture>-<dataset>.

generated captions (Fig. 8). In Figure 8, we compute CLIP-FID for various models trained using LAION, CommonCatalog, or LAION images re-captioned with BLIP-2; CLIP-FID is computed based on generating for prompts from MS COCO and the Conceptual Captions dataset. Unlike other caption datasets, MS COCO captions are human written. Most captions from web-based datasets (like LAION) are computer-generated [48]. BLIP-2 captions are also generated, but the BLIP-2 model is then fine-tuned to align with human-written captions. Given the higher quality of our synthetic captions, it is unsurprising that CommonCanvas’s CLIP-FID is better (i.e., lower) for MS COCO (i.e., aligns better with human-written captions).

However, like any model, ours has limitations. CommonCanvas under-performed in several categories, including faces, general photography, and paintings. These datasets all originated from the Conceptual Captions dataset [61], which relies on web-scraped data. These web-sourced captions, while abundant, may not always align with human-generated language nuances [4, 7, 48]. Although transitioning to synthetic captions introduces certain performance challenges, the drop in performance is not as dramatic as one might assume. Moreover, we speculate that the model will perform better if users provide their more specialized datasets to the model, such as FFHQ [26].

## 6.2. CommonCanvas vs. LAION-trained SD2

Given that our data-scarcity analysis suggests that CommonCatalog is large enough to train a high-quality SD2 model and that synthetic captions can perform well (Section 6.1), we train two different CommonCanvas models: one trained on commercial (CommonCatalog-C) images, another on non-commercial (CommonCatalog-NC). For

a fair comparison with SD2, we use the OpenCLIP text encoder. Like BLIP-2, OpenCLIP is trained on LAION captions (Section 2.2). For example generations, see Figure 9.

We also note that, although we train on Creative-Commons images, it is still possible for an adversarial prompt to produce content that includes iconic characters. In Figure 10, we subject our model to ambiguous prompts that are suggestive of such characters. Examples include visuals closely resembling Elsa from Frozen, Indiana Jones resembling Harrison Ford, and even a likeness to Harry Potter (Figure 10). Qualitatively, our model deviated more from these characters than SD2.

## 6.3. Reaching SD2 quality with CommonCanvas-L

We also did a human study measuring pairwise preference ratings for the 512x512 resolution CommonCanvas models compared to SD2 (Figure 12). In this experiment, human raters were shown a prompt (selected randomly from the PartiPrompts prompts set [75]) along with two generated images in randomized order, one from the reference model (public SD2) 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 our CommonCanvas models are slightly less preferred than SD2-90M, with preference rates of 37% for CommonCanvas-S-C and 38% for CommonCanvas-S-NC, which we find surprisingly high considering the smaller and synthetic nature of the dataset. Figure 9 displays the results from our human study.

Our previous results suggest that SD2 may be underparameterized. We additionally train a larger variant of CommonCanvas-N-C (CommonCanvas-L-NC) that





Figure 10. We compare CommonCanvas-S-NC (Ours) to SD2. Our model is less likely to generate iconic characters given suggestive prompts (drawn from Lee et al. [33]).

has a significantly larger U-Net (the U-Net architecture from SDXL ([49], see the appendix). When we use CommonCanvas-L-NC, we achieve competitive performance with SD2 on user preferences (Figure 9). For the largest model, CommonCanvas-L-NC, we do not measure a statistically significant difference in user preference between this model and SD2.

## 7. Discussion and Related Work

In this paper, we train the CommonCanvas family of text-to-image, latent diffusion models using only Creative-Commons images and synthetic captions. We discuss and address data incompleteness and scarcity issues associated with CC images. For data incompleteness, we propose telephoning, an intuitive type of transfer learning (Section 3), which we instantiate with BLIP-2 to produce synthetic captions for CC images (together, the CommonCatalog dataset; Section 4). Regarding data scarcity, we hypothesize that only a small fraction of the data contained in LAION-2B is actually necessary to saturate SD2, and that the examples in CommonCatalog should be sufficient for training. To make testing this hypothesis more efficient, we implement a variety of ML-systems optimizations, which achieve a  $2.71 \times$  speed-up over our SD2 baseline.

Ultimately, we find that we can train the SD2 model on  $<3\%$  of LAION-2B (i.e., roughly 70 million images; Section 5), yielding a model we call SD2-90M. This encourages us to train on CommonCatalog’s commercially usable (also roughly 70 million) and non-commercially usable (roughly 25 million) examples. Compared to SD2, our CommonCanvas models under-perform in some categories, like faces, but CommonCanvas-L-NC demonstrates statistically equivalent performance with SD2 on human evaluation (Section 6).

While several recent works similarly address ML topics

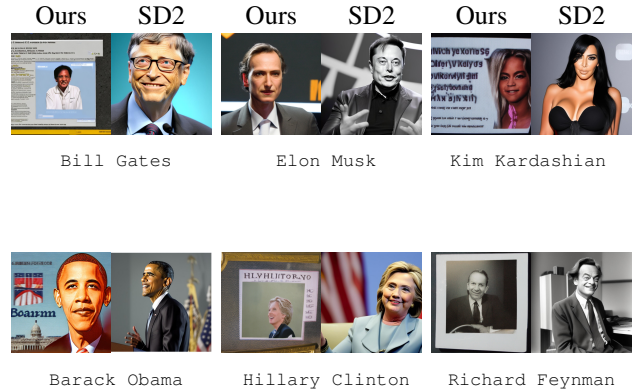


Figure 11. Using CommonCanvas-SNC (Ours) to generate celebrities. Our model is worse at synthesizing individual people than SD2, but is capable of generating some noteworthy public figures. This result demonstrates how our model struggles to generate specific celebrities, which may be desirable from a privacy perspective.

relating to copyright, the literature tends to concern text-to-text training data [44], be primarily theoretical [57, 71], involve ablation studies [28], or only handle verbatim memorization [8, 47] through the use of generation-time content filters [18], which has been shown to be an incomplete solution [23]. To the best of our knowledge, no prior open work attempts to train T2I models on only open-licensed data. Most prior work on image-caption-dataset creation has extracted caption data from Common Crawl [14, 16, 32]. We instead focus on synthesizing captions directly by using a pre-trained BLIP-2 model. Nguyen et al. [48] demonstrates that existing caption datasets can be improved by using BLIP-2 to replace low-quality image captions (e.g., in Datacomp), but does not focus on creating a new dataset of synthetic captions.

Another limitation is that the YFCC100M data is about a decade old; its CC images are not as current as those in LAION-2B. In the future, we plan to augment CommonCatalog with Creative-Commons images from other sources, as well as test larger model architectures and more advanced captioning models, like LLaVA [42].

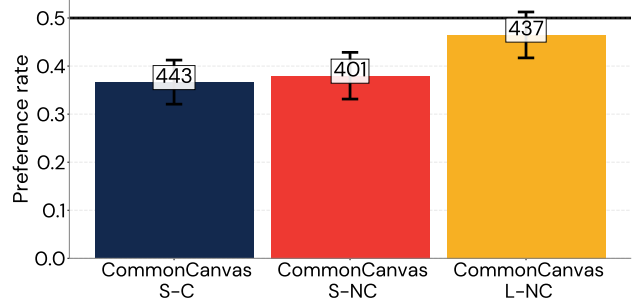


Figure 12. User preference study using Parti prompts. Preference rate (compared to SD2, the thick black horizontal line). CommonCanvas-L-NC matches the performance of SD2.

## References

- [1] Anderson v. Stability AI, Ltd., 2023. No. 3:23-cv-00201 (N.D. Cal. Jan. 13, 2023). [3](#)
- [2] Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E Hinton. Layer normalization. *arXiv preprint arXiv:1607.06450*, 2016. [5](#)
- [3] Romain Beaumont. LAION-5B: A New Era of Large-Scale Multi-Modal Datasets. *LAION Blog*, 2022. [2](#)
- [4] James Betker, Gabriel Goh, Li Jing, Tim Brooks, Jianfeng Wang, Linjie Li, Long Ouyang, Juntang Zhuang, Joyce Lee, Yufei Guo, Wesam Manassra, Prafulla Dhariwal, Casey Chu, Yunxin Jiao, and Aditya Ramesh. Improving image generation with better captions. 2023. [4](#), [7](#)
- [5] Mikolaj Bińkowski, Danica J Sutherland, Michael Arbel, and Arthur Gretton. Demystifying mmd gans. *arXiv preprint arXiv:1801.01401*, 2018. [6](#)
- [6] Abeba Birhane, Vinay Uday Prabhu, and Emmanuel Kahembwe. Multimodal datasets: misogyny, pornography, and malignant stereotypes, 2021. [3](#)
- [7] Davide Caffagni, Manuele Barraco, Marcella Cornia, Lorenzo Baraldi, and Rita Cucchiara. Synthcap: Augmenting transformers with synthetic data for image captioning. In *International Conference on Image Analysis and Processing*, pages 112–123. Springer, 2023. [4](#), [7](#)
- [8] 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, 2021. [8](#)
- [9] 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. [3](#)
- [10] Junsong Chen, Jincheng Yu, Chongjian Ge, Lewei Yao, Enze Xie, Yue Wu, Zhongdao Wang, James Kwok, Ping Luo, Huchuan Lu, and Zhenguo Li. Pixart- $\alpha$ : Fast training of diffusion transformer for photorealistic text-to-image synthesis, 2023. [4](#)
- [11] 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, 2022. [2](#)
- [12] A. Feder Cooper, Katherine Lee, James Grimmelmann, Daphne Ippolito, Christopher Callison-Burch, Christopher A. Choquette-Choo, Niloofar Mireshghallah, Miles Brundage, David Mimno, Madiha Zahrah Choksi, Jack M. Balkin, Nicholas Carlini, Christopher De Sa, Jonathan Frankle, Deep Ganguli, Bryant Gipson, Andres Guadamuz, Swee Leng Harris, Abigail Z. Jacobs, Elizabeth Joh, Gautam Kamath, Mark Lemley, Cass Matthews, Christine McLeavey, Corynne McSherry, Milad Nasr, Paul Ohm, Adam Roberts, Tom Rubin, Pamela Samuelson, Ludwig Schubert, Kristen Vaccaro, Luis Villa, Felix Wu, and Elana Zeide. Report of the 1st Workshop on Generative AI and Law. *arXiv preprint arXiv:2311.06477*, 2023. [3](#)
- [13] 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. [5](#), [2](#)
- [14] 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. [8](#)
- [15] Doe I v. GitHub, Inc., 2022. No. 4:22-cv-06823 (N.D. Cal. November 3, 2022). [1](#)
- [16] Samir Yitzhak Gadre, Gabriel Ilharco, Alex Fang, Jonathan Hayase, Georgios Smyrnis, Thao Nguyen, Ryan Marten, Mitchell Wortsman, Dhruva Ghosh, Jieyu Zhang, et al. DataComp: In search of the next generation of multimodal datasets, 2023. [3](#), [8](#)
- [17] Getty Images (US), Inc. v. Stability AI, Inc., 2023. No. 1:23-cv-00135 (D. Del. February 3, 2023). [1](#), [3](#)
- [18] GitHub. Configuring github copilot in your environment, 2023. [1](#), [8](#)
- [19] Peter Henderson, Xuechen Li, Dan Jurafsky, Tatsunori Hashimoto, Mark A. Lemley, and Percy Liang. Foundation Models and Fair Use, 2023. [3](#)
- [20] 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. [6](#)
- [21] 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. *Advances in neural information processing systems*, 30, 2017. [6](#)
- [22] Gabriel Ilharco, Mitchell Wortsman, Ross Wightman, Cade Gordon, Nicholas Carlini, Rohan Taori, Achal Dave, Vaishaal Shankar, Hongseok Namkoong, John Miller, Hananeh Hajishirzi, Ali Farhadi, and Ludwig Schmidt. OpenCLIP, 2021. [2](#)
- [23] 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. [8](#)
- [24] J.L. v. Alphabet Inc., 2023. No. 3:23-cv-03440-LB (N.D. Cal July 11, 2023). [1](#), [3](#)
- [25] Kadrey v. Meta Platforms, Inc., 2023. No. 3:23-cv-03417 (N.D. Cal. July 7, 2023). [1](#)
- [26] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4401–4410, 2019. [7](#)
- [27] Dirk P. Kingma and Max Welling. Auto-Encoding Variational Bayes. In *International Conference on Learning Representations*, 2014. [2](#)
- [28] Nupur Kumari, Bingliang Zhang, Sheng-Yu Wang, Eli Shechtman, Richard Zhang, and Jun-Yan Zhu. Ablating Concepts in Text-to-Image Diffusion Models, 2023. [1](#), [8](#)
- [29] 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. 6
- [30] LAION-2Ben, 2022. Accessed September 23, 2023. 1, 2
- [31] 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. 3
- [32] Hugo Launçon, 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. 8
- [33] Katherine Lee, A. Feder Cooper, and James Grimmelmann. Talkin’ ’Bout AI Generation: Copyright and the Generative-AI Supply Chain. *arXiv preprint arXiv:2309.08133*, 2023. 3, 4, 8
- [34] Katherine Lee, A. Feder Cooper, James Grimmelmann, and Daphne Ippolito. AI and Law: The Next Generation, 2023. 1, 3
- [35] Katherine Lee, A. Feder Cooper, and James Grimmelmann. Talkin’ ’Bout AI Generation: Copyright and the Generative-AI Supply Chain (The Short Version). In *Proceedings of the Symposium on Computer Science and Law*, page 48–63, New York, NY, USA, 2024. Association for Computing Machinery. 3
- [36] 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. 5, 2
- [37] Mark A. Lemley. How Generative AI Turns Copyright Law on its Head, 2023. 3
- [38] Pierre N. Leval. Toward a Fair Use Standard. *Harvard Law Review*, 103(5):1105, 1990. 3
- [39] 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. 1, 3, 5
- [40] 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. 4
- [41] 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. 6
- [42] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual Instruction Tuning. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. 4, 8
- [43] Susan Box Mann. The Telephone Game, 2019. Accessed September 27, 2023. 3
- [44] 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. 1, 8
- [45] The Mosaic ML Team. composer. <https://github.com/mosaicml/composer/>, 2021. 2
- [46] The Mosaic ML Team. streaming. [<https://github.com/mosaicml/streaming/>](https://github.com/mosaicml/streaming/), 2022. 2
- [47] Milad Nasr, Nicholas Carlini, Jonathan Hayase, Matthew Jagielski, A. Feder Cooper, Daphne Ippolito, Christopher A. Choquette-Choo, Eric Wallace, Florian Tramèr, and Katherine Lee. Scalable Extraction of Training Data from (Production) Language Models. *arXiv preprint arXiv:2311.17035*, 2023. 8
- [48] Thao Nguyen, Samir Yitzhak Gadre, Gabriel Ilharco, Sewoong Oh, and Ludwig Schmidt. Improving multimodal datasets with image captioning. *arXiv preprint arXiv:2307.10350*, 2023. 5, 6, 7, 8
- [49] 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. 2, 8
- [50] 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. 2
- [51] 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. 2, 3
- [52] 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. 4
- [53] 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. 2
- [54] 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. 2
- [55] Matthew Sag. Copyright Safety for Generative AI. *Houston Law Review*, 2023. Forthcoming. 3
- [56] Pamela Samuelson. Generative AI meets copyright. *Science*, 381(6654):158–161, 2023. 3
- [57] 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. 1, 8



- [58] Christoph Schuhmann. LAION-400-Million Open Dataset, 2021. [3](#)
- [59] 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. [2](#)
- [60] Charles M. Schultz. Snoopy\_Peanuts, 2020. Accessed September 26, 2023. [3](#)
- [61] 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. [7](#)
- [62] Benjamin L.W. Sobel. Artificial Intelligence’s Fair Use Crisis. *Columbia Journal of Law and The Arts*, 41:45, 2017. [3](#)
- [63] 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. [2](#)
- [64] Stability AI. Stable Diffusion 2.0 Release, 2022. [1](#)
- [65] Stability AI. Stable Diffusion v2-base Model Card, 2022. [5](#)
- [66] The US Copyright Office. Works Not Protected by Copyright (Circular 33), 2021. [4](#)
- [67] David Thiel. Identifying and eliminating csam in generative ml training data and models. Technical report, Technical report, Stanford University, Palo Alto, CA, 2023. URL [https://purl ...](https://purl...), 2023. [3](#)
- [68] 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. [4](#), [2](#)
- [69] Tremblay v. OpenAI, Inc., 2023. No. 3:23-cv-03223 (N.D. Cal. June 28, 2023). [1](#)
- [70] James Vincent. Getty Images is suing the creators of AI art tool Stable Diffusion for scraping its content. *The Verge*, 2023. [3](#)
- [71] Nikhil Vyas, Sham Kakade, and Boaz Barak. On Provable Copyright Protection for Generative Models, 2023. [8](#)
- [72] Yuxin Wu and Kaiming He. Group normalization. In *Proceedings of the European conference on computer vision (ECCV)*, pages 3–19, 2018. [5](#)
- [73] 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. [4](#)
- [74] 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. [2](#)
- [75] 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. [2](#), [7](#)

# CommonCanvas: Open Diffusion Models Trained on Creative-Commons Images

## Supplementary Material

### A. Details on Data Scarcity Analysis

#### A.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 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  $256 \times 256$  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.

#### A.2. Increasing model capacity

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.

### B. Training Dataset Details

#### B.1. LAION-2B

The fact that LAION is not a stable benchmark can lead to multiple reproducibility 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 likely using differing subsets of the full dataset.

#### B.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%

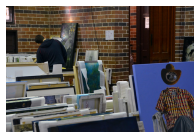
### B.1.2 Release and documentation

## C. YFCC Example Images

Table 2. Randomly sampled images from the YFCC [68] 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

### 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 architecture as SD2 except for reducing the precision of the normalization layers.

Our CC-Large model takes SD2’s model and replaces the UNet with the SDXL architecture [49]. 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 [45]. We perform all dataloading through MosaicML’s streaming library [46].

## D. 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 [53]. Flash Attention is an efficient implementation that is optimized to work well with reduced precision and GPU hardware [13], which was implemented using the XFormers library [36], 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

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 [50], 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 micro-batches, 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 [74] 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.

## E. Telephoning

We dub our solution for handling the lack of captions in CC images as *telephoning*, a type of transfer learning (Figure 3). Telephoning assumes the existence of a large labeled dataset  $\mathcal{D}_1 = \{(x^{(i)}, y^{(i)})\}_{i=1}^n$ , consisting of pairs



Table 3. Top 10 highest frequency captions in the YFCC dataset. The most common captions are not user generated and are not very descriptive of the corresponding image.

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

Table 4. Number of usable captions from OpenAI’s YFCC14M dataset [51]. 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

of high-dimensional  $x^{(i)}$  (e.g., images, audio) that map to a compact, structured label  $y^{(i)}$  (e.g., caption, audio transcript). Telephoning trains a forward model  $q(y|x)$  on  $\mathcal{D}_1$  to learn the mapping of  $y$  given  $x$  via maximum likelihood learning  $\max_{q \in \mathcal{Q}} \sum_{i=1}^n \log q(y^{(i)}|x^{(i)})$ . It then uses  $q$  as training signal for a reverse model  $p(x|y)$  trained on a separate dataset  $\mathcal{D}_2 = \{x^{(i)}\}_{i=1}^m$  by maximizing  $\sum_{i=1}^m \mathbb{E}_{y \sim q(y|x^{(i)})} [\log p(x^{(i)}|y^{(i)})]$ , the likelihood of the data  $\mathcal{D}_2$  and the predicted label  $y$  under  $q$ . This forms a type of knowledge transfer from the forward labeling task defined by  $\mathcal{D}_1$  to the reverse task of inverting  $x$  from  $y$  on a separate  $\mathcal{D}_2$ .

While telephoning can be viewed as a type of synthetic labeling, it becomes particularly interesting when  $x$  is a type of protected modality (e.g., a copyrighted image), while  $y$  is a compact representation of  $x$  that does not encode sensitive aspects of  $y$  (e.g., a generic caption). Effectively, telephoning performs a type of “lossy compression” or “distillation” from a high-dimensional or information-rich  $x$  (e.g., an image of Snoopy) to a low-dimensional or information-poor  $y$  that loses the sensitive content in  $x$  (e.g., the visual characteristics of Snoopy). Because this compression step is “lossy”, a reconstruction  $x'$  of  $x$  from  $p(x|y)$  via  $y$  of-

ten does not remotely resemble the original input, just like in a game of telephone [43]. 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.

**Telephoning for text-to-image modeling.** In this work, we apply telephoning to the image and text domains, where CC images are the high-dimensional inputs  $x$ , and we use a pre-trained BLIP-2 model [39] for “lossy compression” to short-text captions  $y$  (Figure 3a). Together, these CC-image-caption pairs comprise the CommonCatalog dataset, which we use to train our CommonCanvas T2I models (Figure 3b). Even though BLIP-2 was pre-trained on LAION-400M [58], 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 3c).<sup>2</sup>

<sup>2</sup>We draw on the example of Snoopy from [55]. Figure 3’s Snoopy is CC-licensed [60].

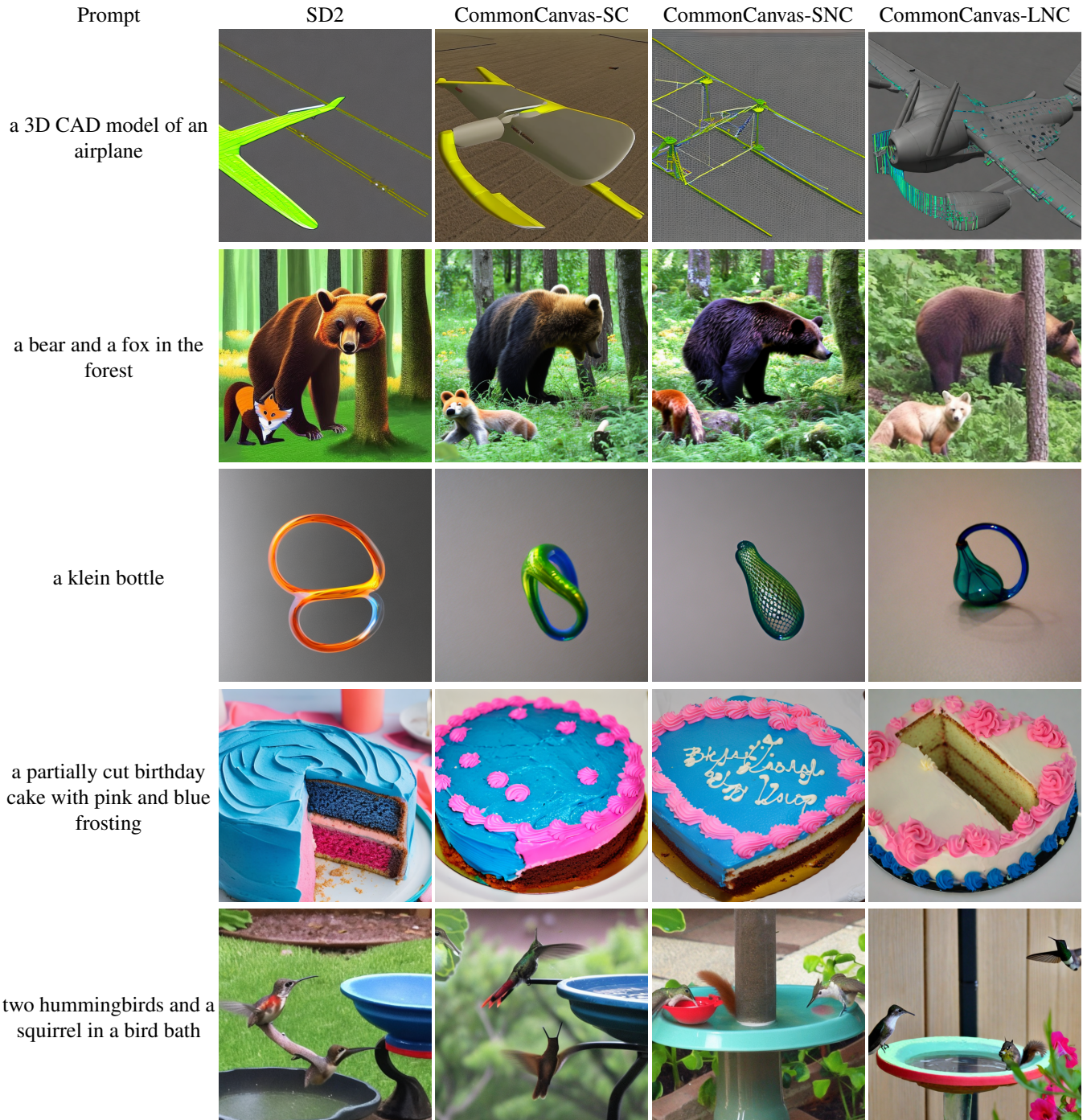


Figure 13. Additional qualitative examples comparing SD2 to our model trained on the commercial split (CommonCanvas-SC), non-commercial split (CommonCanvas-SNC), and the larger UNet model trained on the non-commercial (CommonCanvas-LNC).



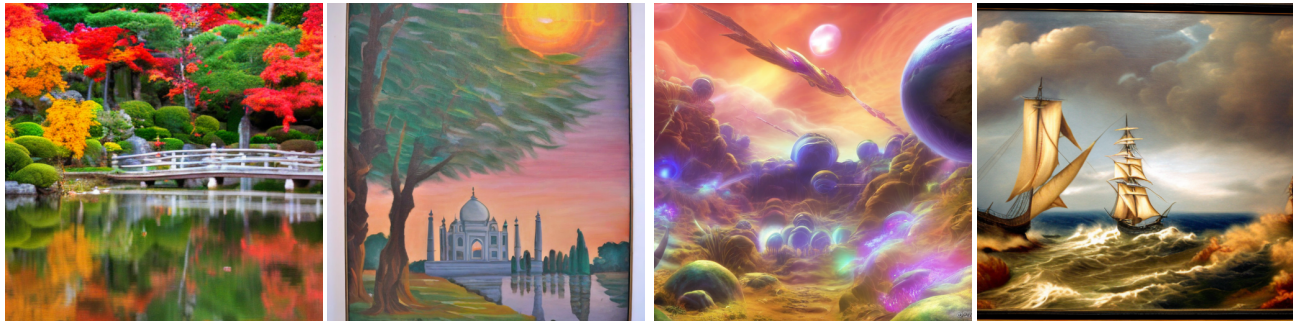


Figure 14. Additional qualitative examples of our CommonCanvas models.



Figure 15. 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.

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

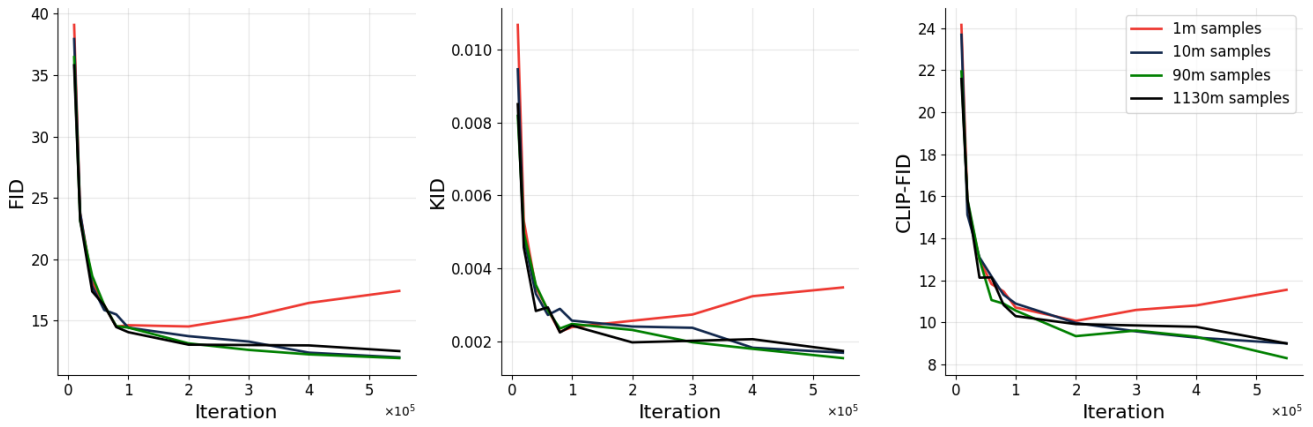


Figure 16. MS COCO metrics over training duration for various dataset sizes. We investigate how reducing the size of the training dataset affects training dynamics, and find that performance is largely unchanged until dropping below 10 million samples. We show that the FID of the eval set remains stable as training progresses. However, reducing the number of samples in our training dataset to 1 million leads to divergence. This finding suggests that only 10 million to 1 million synthetic image caption pairs are needed for good performance on MS COCO.