How Do Data Owners Say No? A Case Study of Data Consent Mechanisms in Web-Scraped Vision-Language AI Training Datasets

Chung Peng Lee[1], Rachel Hong[2], Harry H. Jiang[3], Aster Plotnik[4], William Agnew[3], Jamie Heather Morgenstern[2]

[1]Princeton University
[2]University of Washington
[3]Carnegie Mellon University
[4]University of Toronto
cl6486@princeton.edu, hongrach@cs.washington.edu, hhj@andrew.cmu.edu, amanda.plotnik@mail.utoronto.ca, wagnew@andrew.cmu.edu, jamiemmt@cs.washington.edu

Abstract

The internet has become the main source of data to train modern text-to-image or vision-language models, yet it is increasingly unclear whether web-scale data collection practices for training AI systems adequately respect data owners' wishes. Ignoring the owner's indication of consent around data usage not only raises ethical concerns but also has recently been elevated into lawsuits around copyright infringement cases. In this work, we aim to reveal information about data owners' consent to AI scraping and training, and study how it's expressed in DataComp, a popular dataset of 12.8 billion text-image pairs. We examine both the *sample-level* information, including the copyright notice, watermarking, and metadata, and the *web-domain-level* information, such as a site's Terms of Service (ToS) and Robots Exclusion Protocol. We estimate at least 122M of samples exhibit some indication of copyright notice in CommonPool, and find that 60% of the samples in the top 50 domains come from websites with ToS that prohibit scraping. Furthermore, we estimate 9-13% with 95% confidence interval of samples from CommonPool to contain watermarks, where existing watermark detection methods fail to capture them in high fidelity. Our holistic methods and findings show that data owners rely on various channels to convey data consent, of which current AI data collection pipelines do not entirely respect. These findings highlight the limitations of the current dataset curation/release practice and the need for a unified data consent framework taking AI purposes into consideration.

Code — https://github.com/Anderson-Lee-Git/tracing-data-consent-in-datacomp

Extended Version — https://arxiv.org/abs/2511.08637

Introduction

Web-scraped vision-language datasets (VLD) comprising billions of samples have enabled the success of CLIP (Radford et al. 2021) as well as text-to-image models like Stable Diffusion v1 (Rombach et al. 2022), DALL-E (Ramesh et al. 2021), and MidJourney (Midjourney 2025). However, the reliance on copyrighted material from the web to train foundation text-to-image or vision language models remains the subject of much recent debate, especially in recent lawsuits against OpenAI, Stability AI, and Meta[1]. While efforts toward transparent use of copyrighted training data have been explored in text-based pre-training datasets (Longpre et al. 2024; Elazar et al. 2024), the data consent landscape of web-scraped VLDs remains relatively underexplored, especially as multimodal image-text models become increasingly common.

The shift from the text modality to the image-text modality results in several changes in data consent mechanisms: (1) The signals of data consent in image-text samples are heterogeneous, and (2) image content is often delivered via third-party cloud providers, making the practice of tracking data provenance more challenging. Despite these changes, the impact of violating data consent in the vision-language landscape is no less concerning than that in the text-based counterpart, especially as visual artist communities have spoken out about potential economic loss and reputational harm as a result of generative AI systems (Jiang et al. 2023).

Furthermore, in recent cases involving Anthropic and Meta[2], although the training on copyrighted material was deemed "fair use," the alleged collection of content from pirated sources remains contentious and has precluded the dismissal of the case. This decision raises questions around how dataset curation methods gather data in the first place, and whether such sourcing is allowed. In light of the lack of transparency in web-scraped VLD's data consent (Hardinges, Simperl, and Shadbolt 2024), we aim to *demystify the data consent mechanisms throughout the life cycle of curating, releasing, and using a web-scraped VLD.*

Specifically, we use DataComp's CommonPool (Gadre et al. 2023) as a case study of the web-scraped VLDs. They sourced image-text pairs from CommonCrawl (CommonCrawl 2025), an archive of web pages crawled from the internet, and performed deduplication and minimal filter-

[1]*Andersen v. Stability AI*, No. 3:23-cv-00201 (N.D. Cal.), *Getty v. Stability AI* [2025] EWHC 38 (Ch), *Kadrey v. Meta*, Nos. 3:23-cv-03417, 3:24-cv-06893 (N.D. Cal.), *NYT v. Microsoft*, No. 1:23-cv-11195 (S.D.N.Y.)

[2]*Kadrey v. Meta* (see *supra.*), Doc. 598 (Partial Summary Judgment), and *Bartz v. Anthropic PBC*, 3:24-cv-05417, (N.D. Cal.), Doc. 231 (Partial Summary Judgment)

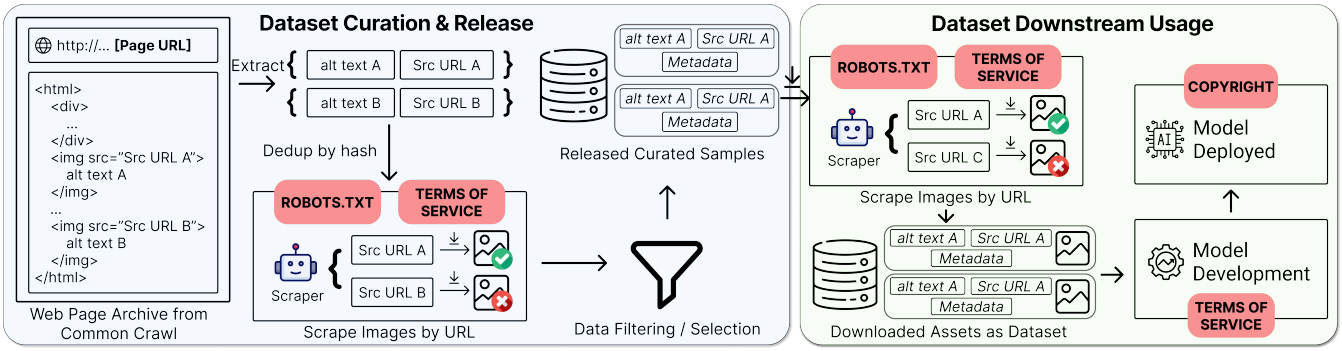

Figure 1: The life cycle of curating, releasing, and using the web-scraped VLD. Even though the *Dataset Curator* initially downloads the image assets in their curation process, the released samples only contain the caption, *src url* pointing to the image asset, and image metadata. To access the dataset, the *Dataset User* must download the images following the released URLs. The red tags on each step indicate the data consent mechanism we consider involved.

ing to produce a set of 12.8B *url-text* pairs, where the *url* points to the image content. As of July 2025, CommonPool has over 2M downloads (Huggingface 2025). Pulling from the same web archive, CommonPool has substantial overlap with its precursor, LAION-5B (Schuhmann et al. 2022), which enabled the early version of Stable Diffusion v1, MidJourney, and Google's Imagen (Rombach et al. 2022; Midjourney 2025; Saharia et al. 2022). Even though the data used to train OpenAI's CLIP or DALL-E were not disclosed, the corresponding papers claim to have sourced the training datasets from the internet (Radford et al. 2021; Ramesh et al. 2021), similar to CommonPool. Therefore, we believe CommonPool as a case study not only informs the open-source vision-language model development community but also provides a lens into commercially protected datasets.

We recognize and take advantage of various signals provided by the image, text, metadata, and their associated data host. We use both sample-level characteristics, such as copyright notice, the exchangeable image file format (EXIF) [3] metadata, and watermark detection, and web-domain-level characteristics, such as Terms of Service (ToS) and Robots Exclusion Protocols (REP), also known as robots.txt. We make the following contributions:

1. Investigate data consent mechanisms in a web-scraped VLD provided by the information in the released artifact
2. Estimate approximately 122M of samples in CommonPool have included copyright information, and over 60% of samples from the top 50 domains, in the `small-en` scale of CommonPool, are sourced from sites restricting scraping in their ToS.
3. Demonstrate that data owners often rely on inconsistent channels to convey data consent, of which AI data collection pipelines do not fully respect, surfacing issues of a lack of a uniform consent mechanism.
4. Use our findings to outline various limitations and recommendations for future web-scraped VLD curation.

[3]https://en.wikipedia.org/wiki/Exif

Background

Terminology

In this section, we outline the scope of each term and the role they play in the explicit permission granted to use the data. We limit our focus to examining data consent and copyright implications within the United States.

Copyright As defined by the U.S. Copyright Office (U.S. Copyright Office 2025), copyright protects the expression of original work. As long as the work is *fixed*, *expressed in tangible forms*, and not an idea, concept, fact, or other exception, it automatically becomes copyright-protected. Notably, the role of the *copyright notice*, like "© John Doe 2025", is to publicly claim that the work is protected by copyright. As such, it becomes more difficult for defendants in infringement cases to argue they were not aware of the work being copyrighted (U.S. Copyright Office 2021).

License A license, or agreement, grants specified rights to someone to use the work for purposes protected by copyright, such as reproduction, display, or making derivatives. A license could be useful for the creator to limit the use of the work in certain scenarios without placing it in the *public domain*, which is outside the scope of copyright protection.

Data Consent We refer to data consent as the "permission" granted for the user to use the data for model training purposes. This is not limited to any form of written consent, such as ToS, copyright notice, claims, or license. In other words, data consent is obtained when the user follows the acceptable pipeline to retrieve data proposed by the data host or data owner. As an example, even if the data is not copyright-registered through the U.S. Copyright Office, a written ToS to restrict the use of such data for model training purposes would be considered a "restriction to use" in the scope of data consent we consider.

Involved Parties

The pipeline to curate, release, and download a web-scraped dataset involves multiple entities. To study the data consent

landscape, we first define how the stakeholders are involved in the life cycle of such datasets.

- *Dataset Curator* – The curator of the dataset releases a set of *url-text* pairs for downstream use. In the case of DataComp (Gadre et al. 2023), it would be their authors.
- *Dataset User* – The user of the dataset downloads the pairs of URLs and texts released by the *Dataset Curator*.
- *Data Owner* – The owner of the image data itself. Since tracing data ownership on the internet is extremely difficult, we relax the ownership to be the action of embedding the image on their web page. This relaxation builds on the assumption that the actor of embedding the image respects the copyright of the image and shares it per the level of consent they obtain.
- *Data Host* – The data host is the entity that owns the image URL referred to by the sample. Since the delivery of image content is often optimized through content delivery network (CDN) and cloud providers, this entity may exhibit little information about the *Data Owner*.

Life Cycle of Web-Scraped VLD

Curation & Release The top-level raw source of data originates from CommonCrawl (CommonCrawl 2025). The collection of *url-text* pairs comes from extracting the `<img src=URL>alt text</img>` from the internet. This extraction *does not* consider the *page url* where the image appears. Figure 1 illustrates the distinction between *page url* and *src url*. With the extracted *url-text* pairs, the *Dataset Curator* uses tools like img2dataset (Beaumont 2021) to automatically download all the images from these URLs, referred to as *scraping*. Since the URLs are extracted from archives of the internet, *not all download attempts are successful or align with the original image*. For instance, the owner of the URL could replace the image with another image or take down the image completely. With the downloaded assets, the *Dataset Curator* experiment with data filtering, cleaning, augmentation, and model training/evaluation to curate the best set for release. Finally, the release of the curated dataset comprises *url-text* pairs along with metadata they obtain from either their experiments or downloading, *without the actual image assets.*

Downstream Usage The *Dataset User* first obtains the index of *url-text* pairs released by the *Dataset Curator*. Since the released dataset artifact comes without the image assets, the *Dataset User* has to utilize similar tools to *scrape* through the provided URLs. In the case of DataComp (Gadre et al. 2023), the scraping functionality is provided as part of the release. This mechanism inherits the same drawback of potentially inconsistent or failed downloads. Not only does it potentially diverge from the *Dataset User*'s expectation of the released dataset, but it might also expose the *Dataset User* to the risk of data poisoning (Carlini et al. 2024). Furthermore, since the *Dataset User* is scraping the web with the index of the URLs, the *Dataset User* is responsible for abiding by any ToS or other data consent mechanism specified by the website hosting the content. With the image assets downloaded, the *Dataset User* then experiments with the downloaded samples in their storage.

Scale	Released	Accessible	"Top 50"
`small`	12.8M	9.8M	–
`small-en`	6.3M	4.8M	2.1M
`medium`	128.0M	98.3M	–
`medium-en`	63.0M	47.7M	21.5M

Table 1: Sample counts of CommonPool's configurations considered in our work. `scale-en` refers to the English-filtered version of the original scale. Accessible counts refer to images downloadable through the released link. "Top 50" refers to the subset in the top 50 *base domains*.

Methods

We first outline the concrete experiment setup for our audit, including data filtering, sizes, and scales that we audit. Then, we present the methods in two categories, one at the sample level and the other at the web domain level. These two angles allow us to audit how image owners and website owners disclose consent for scraping and AI training.

Setup

CommonPool was released at four scales: `xlarge` (12.8B), `large` (1.28B), `medium` (128M), and `small` (12.8M), where the largest contains 12.8B samples and the lower scale is a subset of the larger ones. Due to limited storage space and compute resources, we study both `small` and `medium` such that we can verify whether results found in `small` are also observed in `medium`.

Moreover, since legal mechanisms of data consent are dependent on specific jurisdictions, we restrict our target data to be English-based. Particularly, we follow the same measure in Gadre et al. (2023) to use `fasttext` (Joulin et al. 2016) to filter the original dataset by English-only captions. Table 1 summarizes the audited dataset.

Sample-level Characteristics

At the sample level, we use text, visual, and metadata information to source characteristics of data consent. Particularly, we search for samples with the presence of *copyright notice*, *copyright field in metadata*, and *image watermark*. With the presence of this information, it is difficult for a defendant on copyright infringement to argue ignorance of the fact that the material was copyright-protected (U.S. Copyright Office 2021).

Copyright Notice We crafted a set of regular expressions to capture common copyright notices such as "©" and "copr." These rules are applied to both caption and OCR-extracted text, where we use open-source PaddleOCR (PaddlePaddle Authors 2020) for extraction. The full list of search patterns is included in Appendix.

Copyright Field in Metadata *Exchangeable image file format* (EXIF) is a standard of image metadata to specify information about the image as well as the digital device that produced the image. For instance, some tags include original height, width, and focal length. We search for samples

of which the metadata contains a non-empty copyright tag field keyed by "Copyright" or "0x8298," following the EXIF standard version 2.3. (Standardization Committee 2012).

Image Watermark A watermark detection classifier aims to output whether or not a given image contains a watermark. We (1) use off-the-shelf watermark-finetuned YoloV8 (ultralytics community 2025; mnemic 2024), (2) build a watermark-finetuned MobileViTv2 (Mehta and Rastegari 2022), (3) use two open-source VLMs, Rolm OCR (Reducto AI 2025) and Gemma-3-12b-it (Gemma Team 2025) as our detection methods. To validate the faithfulness of these methods, we evaluate them on (1) *watermark-eval*: Felice Pollano (2019)'s validation set, with a balance of ∼3200 images for both watermarked and non-watermarked images, and (2) *datacomp-watermark-eval*: a random 955-image subset of CommonPool we annotate, to validate the robustness of our detection methods on web-scraped images. Last but not least, we question the faithfulness of LAION-5B's release of *watermark score* by annotating a subset of LAION-5B and analyzing the utility of those scores[4]. The full training and evaluation details can be found in Appendix.

Web-Domain-Level Characteristics

At the web-domain level, the administrator who hosts the content typically specifies rules on permitted usage of their content. Particularly, we examine the top 50 web domains' ToS and their REP, which specifies the restriction of scraping/crawling bots. The top 50 domains are defined by the counts of samples sourced from these domains. In both `small-en` and `medium-en` scales, the top 50 domains cover ∼45% of all samples, namely 2.1M and 21.5M samples respectively.

The web domains are extracted from *src url* as provided by CommonPool, which points to the image asset, rather than the original website where the content is embedded, which we call *page url*. Furthermore, since most content is delivered through domains designed for static content or a content delivery network (CDN), we extract the *base domain* by trimming off the prefix to aggregate the sharded domain URLs. For instance, Pinterest uses bucketed web domains like `i.pinimg.com` and `i-h1.pinimg.com` to deliver content. Through extracting only the *base domain*, which would be `pinimg.com` in the example, we have a more accurate estimate of sample counts for each web domain.

Terms of Service (ToS) Following Longpre et al. (2024), we annotate each web domain with the following attributes: (1) Category: the core function of the *Data Host*, (2) License Type: the permission granted to the end user, and (3) Scraping Policy: the restriction on web-scraping. In this work, we focus on the act of *scraping*, the action of automatically downloading/copying a vast majority of data through an index of links, because both the *Dataset User* and *Dataset Curator* directly engage in this act.[5]

Similar to Fiesler, Lampe, and Bruckman (2016)'s qualitative analysis process, we have two coders to annotate each web domain's attributes, but we start with the codebook for (2) and (3) from Longpre et al. (2024). For the Category, the primary coder first builds the codebook when iteratively going through the web domains. After creating the initial codebook and first pass, the second coder annotates the web domains. The two coders resolve any conflict through adjusting either the annotations or the codebook. The types in each attribute and the full codebook are included in the Appendix.

Robots Exclusion Protocol (REP) REP, implemented via robots.txt, allows website administrators to specify which automated clients (user agents) can access their sites. Administrators can allow or disallow access for specific agents, such as "CCBot" (CommonCrawl), "GPTBot" (OpenAI), or any agent using the wildcard "*". They can also restrict access to certain website paths. In Germany, robots.txt is legally enforceable, with exceptions for scientific research (Hamburg 2024; Official Journal of the European Union 2019).

For each of the top 50 *base domains*, we map the *base domain* to a list of *full domains*, which are the web domains with the original prefix. For instance, the *base domain*, `pinimg.com`, maps to a list of *full domains*, `[i.pinimg.com, i-h1.pinimg.com, ...]`. We retrieve robots.txt by appending "robots.txt" at the end of the *full domains*. In the `small-en` scale, there are 96,436 unique URLs requested, and 81,273 of them successfully return with a non-empty robots.txt[6].

We parse each robots.txt following Longpre et al. (2024) to three categories: *All Disallowed*, *Some Disallowed*, and *None Disallowed* for agents listed in the robots.txt file. *All Disallowed* is when a particular agent is mentioned and disallowed from all parts of the site. *None Disallowed* is when "the particular agent is mentioned and allowed for all parts of the site," or "has no disallowed parts." *Some Disallowed* is when "a particular agent is mentioned and disallowed from some parts of the website." *Some Disallowed* is when a particular agent is mentioned and disallowed from some parts of the website. An agent must be listed in robots.txt to determine the category.

Results

In this section, we present our findings according to the sample-level and web-domain-level methods of determining data consent.

Sample-Level Statistics

Approximately 122M English samples contain characteristics of copyright notice or claims in CommonPool.

[4]LAION-5B releases watermark scores per sample to estimate the probability of the presence of watermark in the image.

[5]In contrast, the term *crawling* refers to the act of developing a spider to recursively follow links from web pages to store content.

[6]In the `medium-en` scale, there are 434,498 URLs requested, and 392,286 of them successfully return with a non-empty robots.txt.

Model	*wm-eval*				*datacomp-wm-eval*			
	Accuracy	Precision	Recall	F1	Accuracy	Precision	Recall	F1
Finetuned YoloV8	96.69	97.44	95.90	96.66	86.91	42.63	51.88	46.80
Finetuned MobileViTv2	89.25	90.43	86.63	88.49	30.37	11.02	74.53	19.20
Rolm-OCR	74.62	99.15	49.74	66.25	89.10	50.80	59.43	54.78
Gemma-3-12b-it	90.66	99.22	81.87	89.71	85.34	41.05	73.58	52.70

Table 2: Evaluation of watermark detection methods on both standard watermark detection dataset, *wm-eval* with 3289 clean and 3299 watermark images, and an annotated set of web-scraped images from CommonPool, *datacomp-wm-eval* with 849 clean and 106 watermark images.

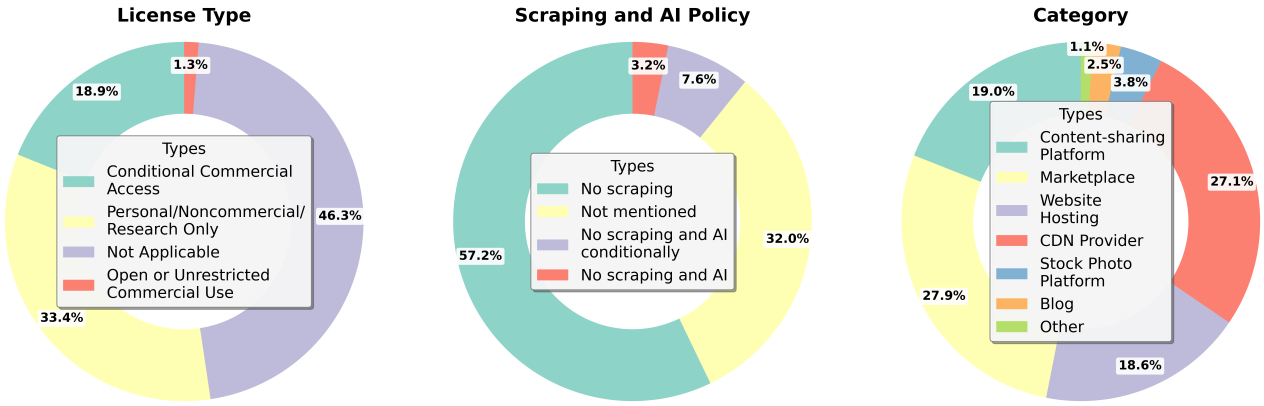

Figure 2: Terms of Service annotations. The full population in each chart is all samples in the top 50 base domains of `small-en`. The portion is determined by the exact number of samples in each type. For License Type, "Not Applicable" indicates that the ToS from the base domain does not specify or provide any license type information. For Category, "Other" indicates that the base domain is for a very domain-specific service. For instance, *4sqi.net* is delivered by Foursquare, a location-intelligence service provider.

Measure	`small-en`	`medium-en`
Caption	10,585 (0.22%)	98,555 (0.21%)
OCR	4,307 (0.09%)	38,697 (0.08%)
EXIF Metadata	108,951 (2.27%)	1.09M (2.28%)
Caption ∪ OCR ∪ EXIF	123,096 (2.56%)	1.22M (2.55%)

Table 3: Number of samples found through each measurement method, where Caption and OCR refer to searching the copyright notice through samples' captions and OCR-extracted texts.

We find 1.22M samples exhibiting characteristics of copyright notice or claims in the `medium-en` scale. We further validate the faithfulness as the portions of the found samples through each method scale similarly from `small-en` to `medium-en`, as shown in Table 3. This extends our results to implications on the full dataset of 12.8B samples, where approximately 122M of English samples may contain copyright notices or claims. We observe very little overlap between the keyword search methods across image, text, and EXIF metadata. This signifies that copyright claims are heterogeneously disclosed for images on the internet, which emphasizes the need to examine each modality to adequately determine copyright information from web-scraped samples.

Watermarks are present in web-scraped images, but detecting them remains a major challenge — even for conventionally advanced methods.

In our evaluation suites, we use (1) *watermark-eval*, comprising a balance of 3289 clean and 3299 watermarked images, and (2) *datacomp-watermark-eval*, a random sample of 955 images from CommonPool we annotate. We find that 106 of those images, or 11.09%, are watermarked, resulting in a 9% to 13% of the distribution with 95% confidence interval. From Table 2, we observe that across all models, the F1-score significantly drops on *datacomp-wm-eval*. This indicates a distribution shift between the traditional watermark detection dataset and the web-scraped images "in the wild." Upon investigation, we determine that traditional methods tend to have lower precision on *datacomp-watermark-eval* because of the text appearing in the image, where the models tend to output *True* for images with texts in them.

Is LAION-5B's released watermark score faithful or informative for understanding and respecting data consent?

In light of our watermark detection experiments, we question the fidelity of the watermark score released in LAION-5B (Schuhmann et al. 2022). We annotate 1308 random samples from LAION-5B and find that 176 have a watermark, or 13.45%. Furthermore, using the standard threshold of 0.5 on the watermark scores released, the precision and recall are only at 34.09% and 51.13%. The area under the receiver operating characteristic (ROC) curve is 0.74. These statistics further demonstrate the difficulty of watermark detection for web-scraped images "in the wild" observed in our experiments. Moreover, the low performance of LAION-5B's watermark score reveals the low utility of this watermark probability score if a dataset user wishes to avoid training AI systems on watermarked images.

Web-Domain-Level Statistics

Since the top 50 base domains in `small-en` and `medium-en` only differ by 1 base domain, we present the results for `small-en` for conciseness. The distribution of the top 50 base domains can be found in the Appendix. For robots.txt, we primarily present our results with the top six user agents in terms of the number of "observations," or samples that come from sites with robots.txt files that mention the top six agents. The total number of observed agents, weighted by sample counts, is 1.1M. Full results are included in Appendix.

60% of samples in the top 50 base domains prohibit scraping, and 33% of them are restricted to Personal/Research/Non-commercial Only Use.

Through our analysis of the ToS in Figure 2, 57.1% of the top 50 base domains prohibit general scraping without mentioning AI, and 3.2% prohibit scraping and AI unconditionally. This not only emphasizes the responsibility of the *Dataset Curator* but also that of the *Dataset User*, who scrapes these sites as well while downloading CommonPool. Furthermore, 33.4% of samples in the top 50 base domains come from websites with ToS limiting usage of content for Personal/Research/Non-commercial purposes.

The practice of releasing only url-text pairs restricts the ability to examine data consent through ToS.

Web-scraped VLDs, such as CommonPool, LAION-400M, and LAION-5B, all use the practice of releasing only the *src url* and caption as described in Section . We find that 27.1% and 18.6% of the samples in the top 50 base domains are under CDN Provider and Website Hosting Service categories, respectively. Yet, the ToS of `amazonaws.com` cannot fully reflect the actual ToS used by the website offering the content stored at those *src urls*. The core reason is that image content delivered via *src url* is often through a CDN or static content host, and only those *src urls* are released instead of the original *page url*. Without the context of *page url*, the website URL where the *url-text* pair is extracted, a thorough examination of data consent is infeasible. This characteristic also primarily accounts for the reason why 46.9% of samples' License Type in Figure 2 are categorized as "Not Applicable," meaning that the provided *src urls*' base domain's ToS may not have the right to specify the License Type.

robots.txt *is mostly adopted to convey restrictions for AI-purpose scrapers/crawlers.*

In the top 6 agents by number of samples covered by observations, we see that traditional web-indexing (googlebot-image) or wildcard ($*$) agents don't have very high *All Disallowed* rate compared to agents related to AI-purposes such as GPTBot, Bytespider, and claudebot. This phenomenon implies that the website administrator disallowing these AI-purpose agents wishes to prevent the use of their content for model development. However, a dataset user downloading CommonPool to train a model does not specify the user agent by default and therefore can bypass REP to scrape many of these same samples from sites that ban GPTBot, Bytespider, and claudebot. Only 3.9% of samples come from sites that disallow any agent, so many sites that specifically block AI-purpose bots may miss dataset users scraping open-source VLDs to train models.

Moreover, even though CommonPool is sourced from CommonCrawl, which respects robots.txt when sourcing the web pages, we still observe CCBot in 353K robots.txt. The most likely reason is that the user adopts robots.txt to revoke their consent after CommonCrawl archives their pages. Despite this adoption, the collection of CommonPool as an index of url-text pairs continues to direct scraping traffic to those websites that chose to revoke consent when the *Dataset User* downloads CommonPool using a non-CCBot user agent name.

Discussion

Limitation of Current Release Practice

Problem Our results reveal several drawbacks in the current release practice of web-scraped VLDs. Firstly, the lack of *page url* greatly restricts the ability to probe whether an image is prohibited from use by the associated ToS. This issue originates from a combination of how image content is usually delivered through CDN, how each sample is collected by only an HTML tag, and how the website itself (*page url*) is not always related to the extracted HTML tag. Secondly, releasing an index of the web through *url-text* pairs allows the *Dataset Curator* to avoid hosting any image asset, and thus any copyright infringement claim or responsibility of providing a convenient channel for the *Dataset User* to access the copyrighted/restricted-to-use data. This shift of accountability may not be made aware to the *Dataset User*, creating an illusion that the curation of an open-sourced web-scraped VLD has already dealt with data consent, so usage of that dataset is in the clear.

Recommendation For better data provenance and transparency, we recommend that future releases include the website page where the samples are collected. Moreover, the *Dataset Curator* should either *clearly inform or warn* the *Dataset User* about the potential responsibility of scraping when using their dataset, or take the responsibility to construct the dataset with standalone image assets respecting the *Data Owner*'s consent, through the various mechanisms we used in our audit.

Agent	Observed	*All Disallowed*		*Some Disallowed*		*None Disallowed*	
		Count	% of observed	Count	% of observed	Count	% of observed
"All Agents"	1,126,876	6,442	0.6%	1,014,576	90.0%	105,858	9.4%
GPTBot 🤖	578,498	538,431	93.1%	40,028	6.9%	39	0.0%
*	475,139	18,595	3.9%	391,799	82.5%	64,745	13.6%
CCBot 🤖	353,324	313,920	88.8%	39,365	11.1%	39	0.0%
Bytespider 🤖	301,344	262,029	87.0%	39,274	13.0%	41	0.0%
googlebot-image	224,268	0	0.0%	224,166	100.0%	102	0.0%
claudebot 🤖	224,200	224,199	100.0%	1	0.0%	0	0.0%

Table 4: Top results from robots.txt analysis for `small-en` scale's top 50 *base domains*, accounting for 96,436 attempted *full domains*, 81,273 successful robots.txt, and 1,126,876 samples observed. For each agent, the number of observed cases is broken down by the number and percentage (relative to observed) of cases where all, some, or none were disallowed. The dark gray background highlights rows that have over 80% *All Disallowed* rate, and the 🤖 icon indicates that the agent is AI-purposed. "All Agents" row refers to an aggregation of all agents found in all the examined robots.txt. The aggregation rule is as follows: If for all agents, a robots.txt has *All Disallowed*, then the decision is *All Disallowed*. If for any agent in all agents, a robots.txt has *All Disallowed* or *Some Disallowed*, then a robots.txt has *Some Disallowed*. Otherwise, it has *None Disallowed*.

Call for a Unified Data Consent Framework

Problem In our case study of DataComp CommonPool, we find that each audit approach surfaced a distinct set of samples restricting data usage with very few overlaps. This observation indicates that the data consent is conveyed through multiple channels, such as image metadata, copyright notice, or image watermark. Even though this highlights the importance of auditing through our comprehensive techniques, it presents a problem of lacking a universally recognized framework to convey data consent, particularly in the life cycle of AI data collection. For instance, robots.txt was constructed for web scraping, but web scraping is only a part of the life cycle. As another example, the copyright notice goes beyond the consent for model development, but also for display, re-distribution, and so on. In addition to the divergent channels to convey data consent, Longpre et al. (2024) reveals a contradiction between these channels where ToS have different restrictions from REP.

Recommendation All the involved parties highlighted in this work need a common protocol such that data owners can communicate data consent, specifically for the use of model development. The Robots Exclusion Protocol is not sufficient because we showed that website maintainers often are not the owners of the data. We believe that a unified channel not only helps the *Data Owner* to protect their works from misuse, but also guides the *Dataset Curator* and *Dataset User* to respect their data consent. Such a framework should not only be adopted but also treated as the source of truth to represent data consent. In addition, we encourage the adoption of an opt-in understanding of consent, as supported by many data owner stakeholders (Kyi et al. 2025; Cultural Intellectual Property Rights Initiative 2017). Existing solutions, e.g. an opt-out model Spawning (2025), do not address the obscurity of scraping and training to many data owners, and implicitly obfuscate consent. Recently proposed Human Commons (Kosmyna and Hauptmann 2025) can be viewed as a specialized consent mechanism acknowledging the complexity and uniqueness of the problem, but the community adoption is still in progress. In short, although a variety of frameworks have been proposed with their merits, there is still a lack of consensus on which to adopt and which to respect.

Related Work

Prior work on auditing web-scale pre-training datasets ranges from data governance, privacy to social biases. In the text modality, Dodge et al. (2021) highlighted the importance of documenting datasets with the excluded data's characteristic, web domain distribution, and other aspects of Colossal Clean Crawled Corpus (C4) (Raffel et al. 2020). Elazar et al. (2024) extended the goal to understand these datasets to several pre-training datasets, such as C4, LAION-2B-en, and The Pile (Raffel et al. 2020; Schuhmann et al. 2022; Gao et al. 2020), by documenting their domain statistics, contamination with evaluation sets, and PII inclusion. More specific to data consent, Longpre et al. (2024) investigated the consent mechanism of text-based pre-training datasets including C4, dolma, and RefinedWeb (Raffel et al. 2020; Soldaini et al. 2024; Penedo et al. 2023). They focus on the temporal changes in data consent in both ToS and robots.txt and highlight the increasing restrictions on the web to train AI models with web-scraped data.

In the vision-language datasets landscape, Hong et al. (2024) studied the impact of data filtering on the exclusion/inclusion statistics concerning minority groups across gender, religion, and race. Hong et al. (2025) presented a legally-grounded study on private information existing in CommonPool and its implications from a legal perspective. Our work studies the data consent mechanism in the landscape of web-scraped VLDs.

Acknowledgements

We would like to thank Christina Yeung for the thoughtful feedback on licensing, policy, and the writing revisions. This research is supported by the NSF Graduate Research Fellowship Program.

References

Beaumont, R. 2021. img2dataset: Easily turn large sets of image urls to an image dataset. https://github.com/rom1504/img2dataset.

Carlini, N.; Jagielski, M.; Choquette-Choo, C. A.; Paleka, D.; Pearce, W.; Anderson, H.; Terzis, A.; Thomas, K.; and Tramèr, F. 2024. Poisoning web-scale training datasets is practical. In *2024 IEEE Symposium on Security and Privacy (SP)*, 407–425. IEEE.

CommonCrawl. 2025. CommonCrawl. https://commoncrawl.org. Accessed: 2025-07-04.

Cultural Intellectual Property Rights Initiative. 2017. Consent Credit Compensation: The Legal Literacy Campaign.

Dodge, J.; Sap, M.; Marasović, A.; Agnew, W.; Ilharco, G.; Groeneveld, D.; Mitchell, M.; and Gardner, M. 2021. Documenting large webtext corpora: A case study on the colossal clean crawled corpus. *arXiv preprint arXiv:2104.08758*.

Elazar, Y.; Bhagia, A.; Magnusson, I. H.; Ravichander, A.; Schwenk, D.; Suhr, A.; Walsh, E. P.; Groeneveld, D.; Soldaini, L.; Singh, S.; Hajishirzi, H.; Smith, N. A.; and Dodge, J. 2024. What's In My Big Data? In *The Twelfth International Conference on Learning Representations*.

Felice Pollano. 2019. Watermarked / Not watermarked images. https://www.kaggle.com/datasets/felicepollano/watermarked-not-watermarked-images/data. A suite of images with and without a random watermark, divided into training and validation sets. Dataset licensed under CC BY-NC-SA 4.0. Accessed: 2025-07-08.

Fiesler, C.; Lampe, C.; and Bruckman, A. S. 2016. Reality and perception of copyright terms of service for online content creation. In *Proceedings of the 19th ACM conference on computer-supported cooperative work & social computing*, 1450–1461.

Gadre, S. Y.; Ilharco, G.; Fang, A.; Hayase, J.; Smyrnis, G.; Nguyen, T.; Marten, R.; Wortsman, M.; Ghosh, D.; Zhang, J.; et al. 2023. Datacomp: In search of the next generation of multimodal datasets. *Advances in Neural Information Processing Systems*, 36: 27092–27112.

Gao, L.; Biderman, S.; Black, S.; Golding, L.; Hoppe, T.; Foster, C.; Phang, J.; He, H.; Thite, A.; Nabeshima, N.; Presser, S.; and Leahy, C. 2020. The Pile: An 800GB Dataset of Diverse Text for Language Modeling. *arXiv preprint arXiv:2101.00027*.

Gemma Team. 2025. Gemma 3. Accessed: 2025-07-04.

Hamburg, L. G. 2024. Urteil vom 27.09.2024 - 310 O 227/23. https://openjur.de/u/2495651.html.

Hardinges, J.; Simperl, E.; and Shadbolt, N. 2024. We must fix the lack of transparency around the data used to train foundation models. *Harvard Data Science Review (Special Issue 5). https://doi. org/10.1162/99608f92. a50ec6e6*.

Hong, R.; Agnew, W.; Kohno, T.; and Morgenstern, J. 2024. Who's in and who's out? A case study of multimodal CLIP-filtering in DataComp. In *EAAMO*.

Hong, R.; Hutson, J.; Agnew, W.; Huda, I.; Kohno, T.; and Morgenstern, J. 2025. A Common Pool of Privacy Problems: Legal and Technical Lessons from a Large-Scale Web-Scraped Machine Learning Dataset. *arXiv preprint arXiv:2506.17185*.

Huggingface. 2025. Hugginface API. https://huggingface.co/api/datasets/mlfoundations/datacomp_pools?expand%5B%5D=downloads&expand%5B%5D=downloadsAllTime. Accessed: 2025-07-04.

Jiang, H. H.; Brown, L.; Cheng, J.; Khan, M.; Gupta, A.; Workman, D.; Hanna, A.; Flowers, J.; and Gebru, T. 2023. AI Art and its Impact on Artists. In *Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society*, 363–374.

Joulin, A.; Grave, E.; Bojanowski, P.; and Mikolov, T. 2016. Bag of Tricks for Efficient Text Classification. *arXiv preprint arXiv:1607.01759*.

Kosmyna, N.; and Hauptmann, E. 2025. Humans Commons. https://www.humanscommons.org. Content licensed under Humans Commons AI0-BY-NC-ND-1.0.

Kyi, L.; Mahuli, A.; Silberman, M. S.; Binns, R.; Zhao, J.; and Biega, A. J. 2025. Governance of Generative AI in Creative Work: Consent, Credit, Compensation, and Beyond. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*, CHI '25. New York, NY, USA: Association for Computing Machinery. ISBN 9798400713941.

Longpre, S.; Mahari, R.; Lee, A.; Lund, C.; Oderinwale, H.; Brannon, W.; Saxena, N.; Obeng-Marnu, N.; South, T.; Hunter, C.; et al. 2024. Consent in crisis: The rapid decline of the ai data commons. *Advances in Neural Information Processing Systems*, 37: 108042–108087.

Mehta, S.; and Rastegari, M. 2022. Separable self-attention for mobile vision transformers. *arXiv preprint arXiv:2206.02680*.

Midjourney. 2025. Midjourney. https://www.midjourney.com/home. Accessed: 2025-07-04.

mnemic. 2024. mnemic/watermarks_yolov8. https://huggingface.co/mnemic/watermarks_yolov8. Accessed: 2025-07-04.

Official Journal of the European Union. 2019. Directive (EU) 2019/790 of the European Parliament and of the Council of 17 April 2019 on copyright and related rights in the Digital Single Market and amending Directives 96/9/EC and 2001/29/EC. https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:32019L0790. Accessed 2025-07-28.

PaddlePaddle Authors. 2020. PaddleOCR, Awesome multilingual OCR toolkits based on PaddlePaddle. https://github.com/PaddlePaddle/PaddleOCR.

Penedo, G.; Malartic, Q.; Hesslow, D.; Cojocaru, R.; Alobeidli, H.; Cappelli, A.; Pannier, B.; Almazrouei, E.; and Launay, J. 2023. The RefinedWeb Dataset for Falcon LLM: Outperforming Curated Corpora with Web Data Only. In *Thirty-seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; Krueger, G.; and Sutskever, I. 2021. Learning Transferable Visual Models From Natural Language Supervision. In Meila, M.; and Zhang, T., eds., *Proceedings of the 38th International Conference on Machine Learning*, volume 139

of *Proceedings of Machine Learning Research*, 8748–8763. PMLR.

Raffel, C.; Shazeer, N.; Roberts, A.; Lee, K.; Narang, S.; Matena, M.; Zhou, Y.; Li, W.; and Liu, P. J. 2020. Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer. *Journal of Machine Learning Research*, 21(140): 1–67.

Ramesh, A.; Pavlov, M.; Goh, G.; Gray, S.; Voss, C.; Radford, A.; Chen, M.; and Sutskever, I. 2021. Zero-shot text-to-image generation. In *International conference on machine learning*, 8821–8831. Pmlr.

Reducto AI. 2025. RolmOCR: A Faster, Lighter Open Source OCR Model.

Rombach, R.; Blattmann, A.; Lorenz, D.; Esser, P.; and Ommer, B. 2022. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 10684–10695.

Saharia, C.; Chan, W.; Saxena, S.; Li, L.; Whang, J.; Denton, E. L.; Ghasemipour, K.; Gontijo Lopes, R.; Karagol Ayan, B.; Salimans, T.; et al. 2022. Photorealistic text-to-image diffusion models with deep language understanding. *Advances in neural information processing systems*, 35: 36479–36494.

Schuhmann, C.; Beaumont, R.; Vencu, R.; Gordon, C.; Wightman, R.; Cherti, M.; Coombes, T.; Katta, A.; Mullis, C.; Wortsman, M.; et al. 2022. Laion-5b: An open large-scale dataset for training next generation image-text models. *Advances in neural information processing systems*, 35: 25278–25294.

Soldaini, L.; Kinney, R.; Bhagia, A.; Schwenk, D.; Atkinson, D.; Authur, R.; Bogin, B.; Chandu, K.; Dumas, J.; Elazar, Y.; et al. 2024. Dolma: an Open Corpus of Three Trillion Tokens for Language Model Pretraining Research. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 15725–15788.

Spawning. 2025. Spawning. https://spawning.ai. Accessed: 2025-07-31.

Standardization Committee. 2012. Exchangeable image file format for digital still camera: Exif Version 2.3. https://www.cipa.jp/std/documents/e/DC-008-2012_E.pdf. Accessed: 2025-07-04.

ultralytics community. 2025. ultralytics. https://github.com/ultralytics/ultralytics. Accessed: 2025-07-04.

U.S. Copyright Office. 2021. Copyright Notice. https://www.copyright.gov/circs/circ03.pdf. Accessed 2025-07-27.

U.S. Copyright Office. 2025. What is Copyright. https://www.copyright.gov/what-is-copyright/. Accessed: 2025-07-07.