

ROM BAR-NISSIM (SBN: 293356)
Rom@HeahBarNissim.com
HEAH BAR-NISSIM LLP
1801 Century Park East, Suite 2400
Los Angeles, CA 90067
Telephone: 310.432.2836

JARRETT LEE ELLZEY (*pro hac vice application forthcoming*)
Jellzey@EKSM.com
TOM KHERKHER (*pro hac vice application forthcoming*)
TKherkher@EKSM.com
LEIGH S. MONTGOMERY (*pro hac vice application forthcoming*)
LMontgomery@EKSM.com
NATISCHA VOLPE (*pro hac vice application forthcoming*)
NVolpe@EKSM.com
ELLZEY KHERKHER SANFORD MONTGOMERY LLP
4200 Montrose Street, Suite 200
Houston, TX 77006

Attorneys for Plaintiffs
TED ENTERTAINMENT, INC.
MATT FISHER
GOLFHOLICS, INC.

UNITED STATES DISTRICT COURT
NORTHERN DISTRICT OF CALIFORNIA

TED ENTERTAINMENT, INC., MATT
FISHER, and GOLFHOLICS, INC., each
individually and on behalf of all others
similarly situated,

Plaintiff,

v.

NVIDIA CORPORATION,

Defendant

Case No.: 3:25-cv-10287

CLASS ACTION COMPLAINT

DEMAND FOR JURY TRIAL

1 Plaintiffs Ted Entertainment, Inc. (“TEI”), Matt Fisher, and Golfholics, Inc. (collectively,
2 where appropriate “Plaintiffs”), each individually and on behalf of all others similarly situated, by
3 and through their undersigned counsel, file this Complaint against Nvidia Corporation
4 (“Defendant”), and alleges as follows:

5 **NATURE OF THE ACTION**

6 1. This lawsuit arises from Defendant unlawfully circumventing technological
7 measures to access and scrape millions of copyrighted videos from the online video viewing
8 platform, YouTube, in order to feed, train, improve, and commercialize Defendant’s large-scale
9 generative artificial intelligence (“AI”) foundational model named “Cosmos.”

10 2. Cosmos is a video world model—conceived and marketed as a foundational AI
11 model designed to be layered between Defendant’s full suite of services and products, including
12 GeForce, Avatar, GR00T, DGX Cloud, GPU Arch, DL Frameworks and autonomous vehicle
13 systems.

14 3. YouTube allows the public to view audiovisual works only through controlled
15 streaming and never provides access to the underlying video files. Defendant intentionally bypassed
16 those restrictions by deploying scraping tools designed to evade YouTube’s access protections.
17 Defendant used a video-downloading program combined with virtual machines that rotated IP
18 addresses to avoid detection and blocking—enabling the mass extraction of, at times, eighty years
19 of video content per day. Defendant used Plaintiff’s and Class Members’ intellectual property for
20 its own commercial gain. In doing so, Defendant has violated YouTube’s Terms of Service, which
21 were intended to protect Plaintiffs and others similarly situated.

22 4. By scraping and downloading those files to build Cosmos, Defendant deliberately
23 circumvented YouTube’s access controls to create a core AI infrastructure that would power
24 Defendant’s entire ecosystem.

25 5. Plaintiffs and the Class Members whom Plaintiffs seek to represent are content
26 creators who upload their audiovisual content to YouTube. In doing so, the content creators are
27 authorizing and instructing YouTube to provide protection to the video content through YouTube’s
28 anti-circumvention software and Terms of Service. In fact, YouTube’s anti-circumvention software

1 and protective Terms of Service are a driving factor behind content creators' decision to upload their
2 video content to YouTube.

3 6. Rather than seek permission or pay a fair price for the audiovisual content hosted on
4 YouTube, Defendant harvested content creators' protected and copyrighted videos for commercial
5 use and at scale without consent or compensation.

6 7. Defendant's actions were not only unlawful, but an unconscionable attack on the
7 community of content creators whose content is used to fuel the multi-trillion-dollar generative AI
8 industry without any compensation.

9 8. Content creators such as Plaintiffs and the Class Members will never be able to claw
10 back the intellectual property unlawfully copied and used by Defendant to train its generative AI.
11 Once AI ingests content, that content is stored in its neural network and not capable of deletion or
12 retraction. Defendant's actions constitute abuse and exploitation of content creators' work for
13 Defendant's profit.

14 9. Most YouTube videos are not registered with the U.S. Copyright Office. That lack
15 of registration, however, does not render them valueless or leave them unprotected. Content creators
16 invest time, skill, and resources into producing their works, and they rely on YouTube's
17 technological protection measures to safeguard their files from unauthorized access. Because
18 copyright registration is not a prerequisite for protection against unlawful circumvention of access
19 controls, this claim is especially critical where Defendant's misconduct lies in breaking through
20 access barriers that prevent anyone from obtaining the underlying files in the first place.

21 10. Plaintiff Ted Entertainment, Inc. is an independent content creator with over 5,800
22 original videos on YouTube, with a combined total of over 4,000,000,000 YouTube views. Plaintiff
23 has additionally amassed a substantial following on YouTube of over 2,600,000 subscribers.
24 Plaintiff has invested substantial time and money into bringing awareness around his content.

25 11. Plaintiff Matt Fisher is an individual and resident of the State of California. He is a
26 golf content creator who posts videos on YouTube, many of which are instructional. His channel is
27 @Mr.ShortGame on the YouTube platform. He has over 500,000 subscribers and hundreds of
28 millions of views. Plaintiff has invested substantial time and money into bringing awareness around

1 his content.

2 12. Plaintiff Golfholics is a corporate entity organized pursuant to the laws of the State
3 of California. It is a golf content channel who posted videos on YouTube. The channel is
4 @Golfholics on the YouTube platform. It has over 130,000 subscribers and millions of views.
5 Plaintiff invested substantial time and money into bringing awareness around its content.

6 13. Like Plaintiffs, the Class Members in this case are independent content creators who
7 uploaded their video content to YouTube's video platform. The video content of Class Members
8 was also part of the dataset utilized by Defendant to train its foundational AI model.

9 14. An essential component of Defendant's business model—powering AI features and
10 services with large-scale training data—includes the mass acquisition and ingestion of creators'
11 videos scraped from YouTube.

12 15. Defendant has profited substantially from its infringement of Plaintiffs' and Class
13 Members' video content through Defendant's training of its generative AI products. Defendant's
14 massive financial success would not have been possible without the video content created by
15 Plaintiffs and Class Members, which was intended for streaming on YouTube.

16 16. Plaintiffs bring this class action on behalf of themselves and of a nationwide class
17 of YouTube creators whose works were scraped, ingested, and trained on without authorization,
18 seeking statutory damages, injunctive relief, restitution, and all other remedies allowed by law
19 pursuant to the Digital Millennium Copyright Act, 17 U.S.C.A. § 1201(a) seeking an injunction and
20 damages commensurate with the scope of Defendant's massive and ongoing infringement. More
21 particularly, Defendant's conduct violates the provisions of the DMCA regarding anti-
22 circumvention (§1201) by bypassing technological protection measures that control access to and
23 copying of YouTube videos

24 **THE PARTIES**

25 17. Plaintiff TEI is a California based media company that owns and operates the
26 YouTube channels "h3h3 Productions" and "H3 Podcast Highlights." Both channels appear
27 extensively in the datasets Defendant used to train its Cosmos model. H3H3 Productions appears
28 with 146 videos in HDVILA-100M, 83 videos in HD-VG-130M, and 24 videos in HowTo100M.

1 H3 Podcast Highlights appears with 285 videos in HDVILA-100M, 209 videos in HD-VG-130M,
2 and 1 video in HowTo100M. TEI invested significant resources into producing and publishing this
3 video content, and every one of these videos was downloaded, copied, and ingested by Defendant
4 without authorization.

5 18. TEI is the creator of the videos identified in Exhibit A, all of which were uploaded
6 exclusively to YouTube by Plaintiff and all of which were included in the datasets used by Defendant
7 to train its generative AI models. Plaintiff derives value from the works identified in Exhibit A
8 through viewership, advertising, sponsorships, licensing, and related monetization.

9 19. TEI and its owners – Ethan and Hila Klein – are longtime champions of the rights
10 of YouTube creators. The Kleins helped define what constitutes fair use reaction videos and the
11 good faith belief standard for DMCA counternotifications. TEI has championed online free speech
12 and is currently helping define what reaction content does not constitute fair use to ensure YouTube
13 content creators can enjoy the fruits of their labor.

14 20. Plaintiff Matt Fisher is an individual and resident of the State of California. He is a
15 golf content creator who posts videos on YouTube, many of which are instructional. His channel is
16 “Mr.ShortGame Golf” on the YouTube platform. He has over 500,000 subscribers and hundreds of
17 millions of views. Plaintiff has invested substantial time and money into bringing awareness around
18 his content. MrShort Game appears with 2 videos in HDVILA-100M, 1 video in HD-VG-130M,
19 and 4 videos in HowTo100M.

20 21. Plaintiff Golfholics is a corporate entity organized pursuant to the laws of the State
21 of California. It is a golf content channel who posted videos on YouTube. The channel is
22 “Golfholics” on the YouTube platform. It has over 130,000 subscribers and millions of views.
23 Plaintiff invested substantial time and money into bringing awareness around its content. Golfholics
24 appears with 62 videos in HDVILA-100M and 37 video in HD-VG-130M.

25 22. Defendant NVIDIA Corporation is a Delaware corporation with its principal place
26 of business at 2788 San Tomas Expressway, Santa Clara, California 95051.

27 **JURISDICTION AND VENUE**

28 23. This is a civil action seeking damages and injunctive relief for infringement under

1 the Digital Millennium Copyright Act, 17 U.S.C. §§ 1201–1205. As such, this Court has subject
2 matter jurisdiction over this action pursuant to 28 U.S.C. §§ 1331 and 1338(a), based on federal
3 question jurisdiction.

4 24. The Court has supplemental jurisdiction over related state-law claims, if any, under
5 28 U.S.C. § 1367.

6 25. Personal jurisdiction is proper because Defendant’s principal place of business is in
7 Santa Clara, California, transacts business nationwide, purposefully avails itself of this forum, and
8 a substantial part of the events or omissions giving rise to these claims occurred or were directed
9 here.

10 26. Venue lies in this judicial district pursuant to 28 U.S.C. § 1391(b)(1) and/or 1400(a)
11 because Nvidia resides in, is found in, or may be sued in this District, and a substantial part of the
12 events giving rise to the claims occurred here.

13 **FACTUAL BACKGROUND**

14 **A. Content Creators, YouTube and Technological Protection Measures**

15 27. As noted above, Plaintiffs and the Class Members are independent content creators
16 of audiovisual content.

17 28. Plaintiff TEI is the creator of the videos listed in Exhibit A.

18 29. Plaintiff Matt Fisher is the creator of the videos listed in Exhibit B.

19 30. Plaintiff Golfaholics, Inc. is the creator of the videos listed in Exhibit C.

20 31. Plaintiffs and the Class Members create video content and upload their video content
21 onto YouTube’s video sharing platform.

22 32. YouTube’s “users”—the individuals who view the digital content available on
23 YouTube—can watch and listen to music videos for free on YouTube’s ad-supported service, but
24 YouTube does not give users access to or allow downloading of the digital files for the content
25 viewed by the user.

26 33. YouTube deploys technological protection measures (“TPMs”) designed to control
27 access to the underlying video files and prevent direct downloading outside permitted channels (*e.g.*,
28 streaming-only delivery, application programming interface (“API”) usage limits, and access

controls).

34. YouTube's Terms of Service expressly prohibit scraping, unauthorized downloading, bulk extraction, or other forms of data mining of audiovisual content except through expressly permitted features or licensed APIs.¹ These contractual restrictions operate together with TPMs to prevent unlicensed access to creators' videos.

35. According to YouTube's Terms of Service, Content creators such as Plaintiffs who upload content onto YouTube grant license to YouTube for certain uses as well as to other users of YouTube to access content through YouTube's services; however, the license makes clear that it "does not grant any rights or permissions for a user to make use of [the] Content independent of the Service."²

36. This language confirms that end users are not given access to the file itself, only the ability to view (*i.e.*, stream) through YouTube's controlled environment. YouTube's Terms of Service reflect YouTube's intent to restrict access to the digital files underlying the videos that YouTube's users are allowed to stream through YouTube's platform.

37. Streaming through YouTube and downloading permanent copies provide the user with different value propositions—watching and listening for free but seeing ads, versus possessing a permanent digital copy.

38. Accordingly, scraping or bulk downloading is not merely copying material already provided; it is an act of unauthorized access to data files that YouTube affirmatively withholds from public download.

39. In order to enforce its prohibitions on downloading content, YouTube does not provide a downloading option that is readily available to users.

40. Although YouTube does offer downloading options to customers who subscribe and pay for YouTube's "Premium" plan, YouTube also employs TPMs to prevent access for Premium customers. Specifically, YouTube prohibits all downloading audiovisual content except to the

¹ "Terms of Service," YouTube, <https://www.youtube.com/t/terms> (last viewed October 16, 2025) (prohibiting "access[ing] the Service using any automated means (such as robots, botnets or scrapers) except (a) in the case of public search engines, in accordance with YouTube's robots.txt file; or (b) with YouTube's prior written permission").

² *Id.*

1 YouTube app. Further, the “download” option only makes audiovisual content available for offline
2 streaming—it does not provide the Premium customer with access to the audiovisual files.
3 Accordingly, the audiovisual files cannot be transferred to any other device, but remain only for
4 streaming on the app. Finally, the files are only available for offline streaming for a limited amount
5 of time, at which point they will become available only for online streaming once again.

6 41. For users who do not have a “Premium” plan, the “download” option on YouTube’s
7 player page is non-functional.

8 42. In order to further enforce its prohibitions on downloading content, YouTube uses
9 technological processes and tools to detect and block unauthorized downloading. For example,
10 YouTube occasionally updates APIs, which operate to interfere with downloaders utilizing the non-
11 updated API to extract files. As another example, YouTube monitors downloading activity and may
12 block IP addresses that make too many download attempts in a specified period.

13 43. YouTube’s refusal to provide users with a readily available downloading option,
14 along with the processes and tools YouTube employs operate to restrict users’ ability to access
15 audiovisual material, are part of the TPMs that YouTube has employed to prohibit access to
16 audiovisual files.

17 44. Content creators, including Plaintiffs and Class Members, rely on the TPMs
18 contained in YouTube’s Terms of Service in deciding to upload their audiovisual content to
19 YouTube. Content creators, including Plaintiffs and Class Members, expect that their works will not
20 be copied at scale without consent.

21 **B. Defendant Improperly Obtained Millions of YouTube Videos to Train Its**
22 **Foundational AI Model**

23 45. As noted above, Defendant requires significant amounts of data in order to feed,
24 train, improve, and commercialize Cosmos, the video world model created, owned and controlled
25 by Defendant.
26
27
28

46. Cosmos was not a research tool or isolated project; it was Defendant’s foundational video world model. Defendant’s own internal materials show that Cosmos was designed to sit at the center of Defendant’s architecture, providing a common layer of training data and video understanding that other services—such as GR00T, Avatar, OV, and GeForce—would rely upon, as reflected in *Figure 1*,³ below.

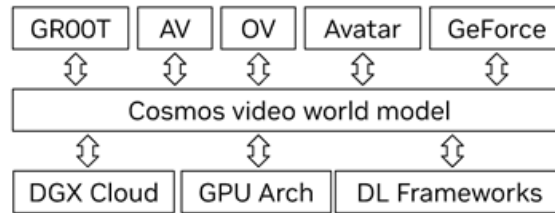


Figure 1

47. Because Cosmos was foundational, Defendant had an overwhelming incentive to acquire training data on an unprecedented scale. Rather than negotiate for lawful licenses, Defendant broke through YouTube’s access protections to obtain the massive datasets necessary to fuel Cosmos and, by extension, Defendant’s broader product lines.

48. Defendant has invested in its foundational AI model and has also fundamentally collaborated on improving its generative capabilities. Particular to the current case, Defendant has sought to develop Cosmos so it is capable, among other things, of turning language prompts into audiovisual content. In other words, Defendant has sought to, and has had some success with, creating an AI product that accepts language instruction and produces a video based on that instruction.

49. In order to adequately train its generative AI models, Defendant requires vast amounts of data—the more data, the better the AI product. At the direction of Defendant, Defendant’s employees collaborated to obtain those datasets, at times communicating via chats on the “Slack” platform, including a particular Slack channel aptly titled “#cosmos-dataset-creation,” which was intended as a channel to “generate huge curated video dataset for video generative

³ All Figures contained in this Complaint are images taken from the article, “Leaked Documents Show Nvidia Scraping ‘A Human Lifetime’ of Videos Per Day to Train AI,” Samantha Cole, 404 Media (August 5, 2024).

modeling,” as reflected in *Figure 2*,⁴ below.

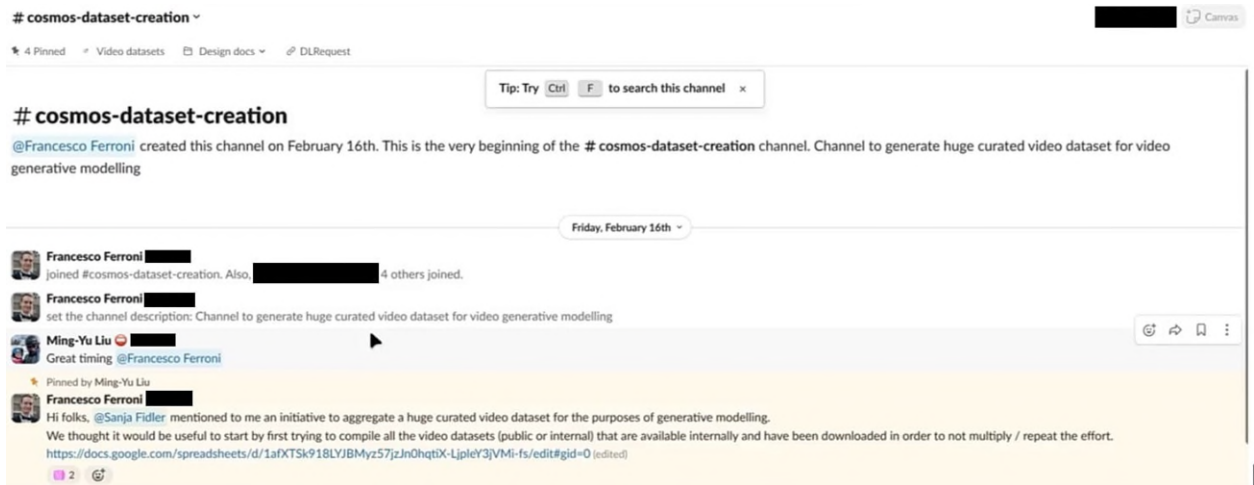


Figure 2

50. Defendant obtained datasets from a variety of sources, including academic repositories, research compilations, and other large-scale video collections created by universities, corporations, and independent researchers. These datasets were treated by Defendant as raw material for Cosmos, even when the datasets were expressly licensed for academic or non-commercial use and prohibited commercial exploitation, redistribution, or any use that would involve downloading the underlying copyrighted works. The link identified in *Figure 2*, above, is to a spreadsheet containing links to datasets, as referenced in *Figure 3*, below.

⁴ The redactions in *Figure 2* were made in the context of the 404 Media article from which the images were taken, which redacted the names of lower-level employees from the original images, but retained the names of “several high-level engineers and executives who were involved in the project, because of their public profiles as leaders in the AI industry.” See “Leaked Documents Show Nvidia Scraping ‘A Human Lifetime’ of Videos Per Day to Train AI,” Samantha Cole, 404 Media (August 5, 2024).

	A	B	C	D	E	F
1	Dataset name	Link	# of videos	Resolution	Domain	Total length (hrs)
2	AudioSet	http://research.google.com/audioset/	2M clips, 10s each		open	5,500
3	ACAV100M	https://acav100m.github.io	100M clips		open	238,000
4	HowTo100M	https://www.di.ens.fr/willow/research/howto100m/	137M clips	240p	instructional	134,000
5	MSR-VTT	https://www.microsoft.com/en-us/research/publication	10K clips, ~15s		open	42
6	Kinetics-400	https://paperswithcode.com/dataset/kinetics-400-1	300K clips, 10s each		open	830
7	VGG Sound		200K clips, 10s each		open	
8	WebVid-10M	https://maxbain.com/webvid-dataset/		360p	open	52,000
9	Ego4D	https://ego4d-data.org			egocentric	3670
10	Ego-Exo4D	https://ego-exo4d-data.org/			egocentric	1422
11	MERLOT-Reserve	https://rowanzellers.com/merlotreserve/	20M clips		instructional	9300
12	HD-VILA-100M	https://github.com/microsoft/XPretrain/blob/main/hd-v	100M clips	720p	open	371,500
13	YouCook2	http://youcook2.eecs.umich.edu/	11682 clips		instructional	
14	VATEX	https://eric-xw.github.io/vatex-website/about.html	22725 clips		open	
15	NextQA	https://github.com/doc-doc/NEXT-QA/tree/main	34132 clips		open	
16	iVQA	https://opendatalab.com/OpenDataLab/iVQA	5394 clips		instructional	
17	ActivityNet QA	https://arxiv.org/abs/1906.02467	2835 clips		open	
18	HDVG-130M	https://github.com/daoshee/HD-VG-130M	130M clips	720p	open	
19	InternVid-18M	https://github.com/OpenGVLab/InternVideo/tree/main	234M clips			760,000
20	InternVid-10M-FLT	https://huggingface.co/datasets/OpenGVLab/InternVi	10M clips	720p	open	32,000
21	HowToInterlink-7M		7M clips		open	
22	VideoMatting240K	https://paperswithcode.com/dataset/videomatte240k	484 clips	4K/HD		
24	DLSS dataset	-	n.a.	HD		1
25	Maxine GreenScreen Youtube	https://www.youtube.com/	n.a.	Various	5500	
26	Maxine GreenScreen RTX	https://crowdsourcing.nvidia.com/en-us/maxine/		360p	5000	
27	Maxine GreenScreen internal	-		Various		183
28	Isaac Sim	-	n.a.	4k		100
29	AV uncured	-	n.a.	4k		1,500,000
30	AV on ORD - V0	-	345K clips, 10s each	360p		100
31	AV on ORD - V1 (eta 2/27)	-	7M clips, 10s each			20,000
32	AV on ORD - V2 (eta 3/25)	-	95M clips, 10s each			270,000
33	MPEG & AOM seqs (gaming)	https://confluence.nvidia.com/display/VHT/Super-re	~70 clips, 15s each	~3480x2160		
33	MPEG & AOM seqs (gaming)	https://confluence.nvidia.com/display/VHT/Super-re	~70 clips, 15s each	~3480x2160		
34	Gaming Captures (gaming)	https://confluence.nvidia.com/display/VHT/Super-re	~50s clips, 15s each	~3480x2160		
35	Camera Captures (gaming)	https://confluence.nvidia.com/display/VHT/Super-re	~420 clips, 15s each	~3480x2160		
36	Gaming Captures (RL)	https://drive.google.com/drive/folders/1jVoYC-6r	200 clips, 1-2hrs long	~3480x2160		
37	Human and Camera Motion	https://nvlabs.github.io/PACE/#:~:text=Example%20F	Synthetic data generator			
38	MovieNet	https://arxiv.org/abs/2007.10937	1.1K videos	240p		
39	Video-Instruct 100k	https://github.com/mbzuai-oryx/Video-ChatGPT	same as activitynet			
40	UCF-101	https://www.crcv.ucf.edu/data/UCF101.php	13320	240p		
41	OpenXEmbodiment: ~2M episodes (CC-BY-4 license)					
42	DROID: 350 hours (CC-BY-4)					

Figure 3

51. This case is about Defendant's use of the HD-VG-130M, HDVILA100M, and HowTo100M datasets. Each comprised of YouTube videos ingested to create the Cosmos foundational AI model. Each of these datasets is made up entirely of YouTube source material that must be downloaded from YouTube to be used for AI training. These datasets do not contain the

1 videos themselves. A video dataset functions only as an index file that lists pointers such as URLs,
2 YouTube IDs, or similar location identifiers, sometimes with limited metadata attached. A dataset
3 contains no audiovisual files. To use any such dataset, a company must retrieve the actual video files
4 by downloading them directly from YouTube. Doing so requires new copying of each copyrighted
5 work and requires bypassing YouTube's technological restrictions, terms of service, and licensing
6 limitations. The dataset merely identifies where the videos are located. All copying, downloading,
7 and circumvention is performed by the user.

8 52. These datasets do not contain complete videos. Rather, they list thousands of
9 specific timestamps within those videos that are designated as clips. A single YouTube video can
10 be divided into many such clips, and each clip is treated as a separate training sample. As a result,
11 the scraping process requires repeated retrieval of the same underlying video for each clip. This
12 means the same copyrighted work is copied multiple times, from different starting and ending points,
13 for the purpose of extracting each designated clip.

14 53. To obtain any clip listed in any of the datasets, the user must newly copy the
15 underlying YouTube video at the specific timestamp identified for that clip. Each clip requires a
16 separate retrieval of the source video, which results in a distinct act of copying. Because YouTube
17 provides no lawful mechanism for downloading these works, every retrieval requires the user to
18 circumvent YouTube's technological restrictions, terms of service, and licensing limits. The dataset
19 provides only the video location and the start and end times for each clip. All downloading, copying,
20 and circumvention occurs on a clip-by-clip basis, and each instance constitutes its own violation.

21 54. The first dataset is HD-VG-130M. It contains location pointers for more than 130
22 million video clips created by slicing longer YouTube videos into short segments for machine
23 learning research. Each clip requires its own download to obtain the underlying audiovisual file.
24 HD-VG-130M is a compilation of 1,549,408 YouTube videos that were broken into approximately
25 130 million high-definition clips.

26 55. The compilers sought out videos with high aesthetic scores as measured by an AI
27 model and excluded videos that displayed creator names or visible watermarks.⁵ The dataset was

28 ⁵ "Swap Attention in Spatiotemporal Diffusions for Text-to-Video Generation," Wenjing Wang, et al., arXiv:2305.10874v4 [cs.CV] (April 24, 2024).

1 created for academic research in 2023 and was never licensed for commercial use of any kind.

2 56. Defendant used HD-VG-130M to retrieve every clip and incorporate those files into
3 its commercial training pipeline. Because each clip corresponds to a separate act of copying, the
4 ingestion of HD-VG-130M required roughly 130 million unauthorized downloads, each constituting
5 a separate circumvention event.

6 57. Microsoft created a document identifying the content of the HD-VG-130M dataset.
7 Microsoft then released this list into the broader AI ecosystem. Not only did Microsoft assemble the
8 HD-VG-130M dataset internally, but it also posted the dataset publicly on GitHub—a code-sharing
9 platform owned by Microsoft—at <https://github.com/daoshee/HD-VG-130M>.

10 58. The GitHub repository includes a license agreement that restricts the dataset to
11 “academic use only,” acknowledging that any use for commercial purposes would be unauthorized.

12 59. In addition, the GitHub license agreement prohibits distribution, reproduction,
13 modification, or exploitation of the dataset content without the permission of the copyright owner,
14 thus recognizing that copyright ownership resides with the original creators of the source material,
15 including YouTube content creators such as Plaintiff and the Class Members.

16 60. By including these limitations on its GitHub repository, Microsoft implicitly
17 acknowledged that the HD-VG-130M dataset is comprised of protected, copyrighted works
18 obtained without license, permission, or authorization.

19 61. Defendant obtained the dataset that was earmarked for academic use only, including
20 the HD-VG-130M dataset. Defendant was aware that the datasets were for research purposes only
21 but used the datasets for commercial development (i.e., “training”) of its generative AI model.

22 62. Further, Defendant used the HD-VG-130M dataset to scrape audiovisual content
23 from YouTube without the consent of YouTube or the creators of the audiovisual content.

24 63. The second dataset is HDVILA-100M. This dataset is comprised entirely of
25 YouTube videos and contains 3,098,462 source videos that were divided into roughly 100 million
26 clips. Each clip represents a distinct segment of the original work. The dataset was compiled by
27 Microsoft Research Asia and published in 2021 for research involving video-based AI models. It
28 lists pointers to YouTube videos and includes subtitle information extracted from YouTube’s

1 closed captioning system.

2 64. The usage license on the dataset states it was created “solely for Computational Use
3 for non-commercial research. This restriction means that you may engage in non-commercial
4 research activities (including non-commercial research undertaken by or funded via a commercial
5 entity), but you may not use the Data or any Results in any commercial offering, including as part
6 of a product or service (or to improve any product or service) you use or provide to others.”

7 65. HDVILA-100M does not provide the video files themselves. Any user of the dataset
8 must download each referenced clip directly from YouTube. Defendant did so at scale, which
9 triggered roughly 100 million separate acts of unauthorized access and copying.

10 66. By including these limitations on its usage license, Microsoft implicitly
11 acknowledged that the HDVILA100M dataset is comprised of protected, copyrighted works
12 obtained without license, permission, or authorization.

13 67. Defendant obtained the dataset that was earmarked for non-commercial use only,
14 including the HDVILA100M dataset. Defendant was aware that the datasets were for research
15 purposes only but used the datasets for commercial development (i.e., “training”) of its generative
16 AI model. Further, Defendant used the HDVILA100M dataset to scrape audiovisual content from
17 YouTube without the consent of YouTube or the creators of the audiovisual content.

18 68. The third dataset is HowTo100M. This dataset consists of 1,238,911 YouTube
19 videos that were split into more than 100 million short clips. The clips depict 23,611 different tasks
20 related to cooking, handcrafting, personal care, gardening, and other activities. The task labels and
21 descriptions were generated using the taxonomy developed by WikiHow. The dataset is hosted by
22 the European research institutions that created it. Like the others, it contains only identifiers and
23 metadata, not any audiovisual material. To use HowTo100M in training, a company must download
24 each clip directly from YouTube. Defendant carried out these downloads as part of its Cosmos
25 training pipeline, resulting in more than 100 million additional unauthorized copying events.

26 69. Although the HowTo100M dataset is posted online with what its creators label as a
27 “commercial license.” That license has no effect on the legal rights associated with the underlying
28 YouTube videos. The dataset creators do not own the videos and therefore cannot grant any license

1 that authorizes the reproduction, downloading, or commercial use of those works. The dataset
2 license applies only to the dataset itself, meaning the text files, identifiers, and metadata created by
3 the researchers. It does not and cannot confer any right to copy, download, or redistribute the
4 millions of YouTube videos identified in the dataset. Each of those videos remains protected by
5 YouTube's terms of service and by technological access restrictions. As a result, every act of
6 downloading a HowTo100M clip from YouTube is a separate unauthorized copying event and a
7 separate circumvention violation, regardless of any license attached to the dataset file.

8 70. The HDVILA-100M, HD-VG-130M, and HowTo100M datasets consists of
9 location identifiers that point to millions of YouTube videos or clips. None of them contains the
10 underlying audiovisual files. To use them in training, a company must retrieve and download every
11 referenced video directly from YouTube. By targeting these datasets, Defendant initiated millions
12 of individual downloads of protected YouTube content, all without authorization, all in violation
13 of YouTube's access restrictions, and all for the commercial purpose of building its foundational
14 video model.

15 71. Upon information and belief, in order to acquire "high-quality text-to-video
16 generation," Defendant, directly and through agents, contractors, and affiliates, intentionally
17 accessed large volumes of YouTube videos by scraping and/or using tools and workflows that
18 bypass or evade YouTube's TPMs and usage restrictions, and then reproduced those videos to
19 assemble training corpora for Defendant's AI models and services.

20 72. "Scraping" content involves downloading audio and video files from a website, in
21 this case YouTube, using an automation of some type.

22 73. When Defendant scraped audio and video files from YouTube, Defendant did not
23 simply download those files onto the YouTube app for offline streaming, as envisioned by
24 YouTube's Premium plan. Instead, Defendant improperly accessed the actual audio and video files
25 and downloaded those files into Defendant's own system, where Defendant had control over the
26 files and where Defendant could store them indefinitely. These actions are inconsistent with
27 Defendant's Premium plan.
28

74. Defendant's employees discussed downloading video content from links provided in Microsoft's HDVILA-100M video dataset, as reflected in *Figure 4*, below.

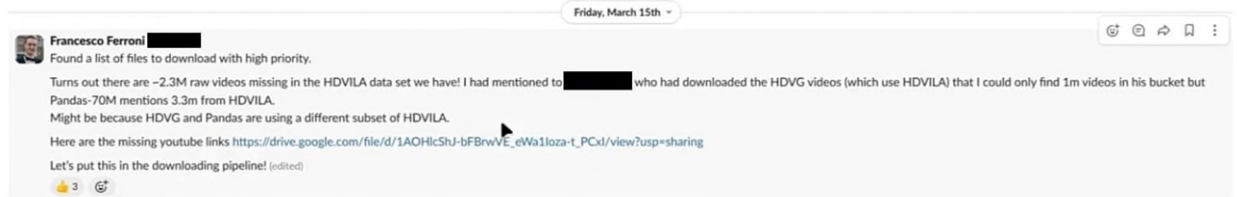


Figure 4

75. Defendant also obtained its own separate audiovisual datasets directly through its own independent scraping of YouTube's video sharing platform.

76. The scraping activities Defendant conducted on YouTube's video sharing platform required Defendant to use processes to defeat protections that YouTube put in place to prevent unauthorized access to the digital files underlying the audiovisual content. As detailed below, the processes used by Defendant to scrape audiovisual files from YouTube are not used in the ordinary course of YouTube's operations from the point of view of an ordinary consumer.

77. For example, with the explicit knowledge and at the direction of Defendant, Defendant's employees used tools and processes such as the open-source YouTube video downloader "yt-dlp" combined with virtual machines that refresh IP addresses. To access audiovisual content from YouTube's platform. Such tools and processes are necessary for Defendant to avoid being blocked by YouTube.⁶

78. The use of virtual machines to refresh IP addresses is necessary because YouTube uses programs to monitor the activity from IP addresses for downloading activity and block those IP addresses from accessing audiovisual data on the YouTube platform. Refreshing the IP addresses operates to defeat YouTube's monitoring programs.

79. The yt-dlp downloader is a tool used for downloading videos and audio from online platforms such as YouTube. The yt-dlp downloader can be used to automatically merge separate audio and video streams and download entire playlists, among other tasks, and Defendant used the yt-dlp downloader for this very purpose in downloading files from YouTube.

⁶ "Leaked Documents Show Nvidia Scraping 'A Human Lifetime' of Videos Per Day to Train AI," Samantha Cole, 404 Media (August 5, 2024).

1 80. To use yt-dlp, Defendant first had to install the program, which can be downloaded
2 from the official yt-dlp page on GitHub, along with other programs necessary to effectively
3 download files, such as programs for merging video and audio files. Once installed, yt-dlp can be
4 used to download individual files or entire playlists by adding the Uniform Resource Locator
5 (“URL”), which is essentially a web address and mechanism for retrieving a specific file, for each
6 desired video to the yt-dlp program.

7 81. Defendant used programs such as yt-dlp to improperly access and download files
8 and merge data in order to feed complete audiovisual packages to its generative AI model for training
9 purposes.

10 82. Processes such as these allowed Defendant to bypass YouTube’s player page and
11 avoid YouTube’s monitoring systems in order to scrape audio and video content from YouTube,
12 and feed content directly to its generative AI model.

13 83. Defendant did not obtain the consent of YouTube, Plaintiff or other creators of the
14 audiovisual content to conduct its scraping activities.

15 84. Defendant’s project managers responsible for acquiring audiovisual data for training
16 purposes discussed efforts to “build a data factory that can yield a human lifetime visual experience
17 worth of training data per day.”⁷

18 85. On information and belief, and based on the sheer volume of video data included,
19 the datasets of videos that included YouTube videos were scraped without permission from
20 YouTube and in violation of YouTube’s Terms and Services.

21 86. The scraping and acquisition processes used by Defendant were inconsistent with,
22 and in violation of, YouTube’s Terms of Service, which forbid scraping and mass downloading of
23 videos.⁸

24 87. The automated system used by Defendant to access video content from YouTube
25 was intentionally designed and utilized to circumvent YouTube’s TPMs.

26 ⁷ *Id.*

27 ⁸ “Terms of Service,” YouTube, <https://www.youtube.com/t/terms> (last viewed November 26,
28 2025) (prohibiting “access, reproduce, download, distribute, transmit, broadcast, display, sell,
license, alter, modify or otherwise use any part of the Service or any Content except: (a) as expressly
authorized by the Service; or (b) with prior written permission from YouTube and, if applicable,
the respective rights holders;”).

88. To be clear, Defendant's generative AI models are not "watching" YouTube in order to be trained. Rather, data improperly accessed and downloaded from YouTube is being uploaded into Defendant's generative AI models in order to develop and improve Defendant's products.

89. Defendant knew or should have known its conduct contravened YouTube's rules and the TPMs that enforce them.

90. Defendant's knowledge that its activities violated YouTube's rules and the TPMs that enforce them is evidenced by conversations among Defendant's employees regarding efforts to avoid detection of the use of scrapers by YouTube and the associated risk of having their IP addresses blocked from access to YouTube, as reflected in *Figure 5*, below.

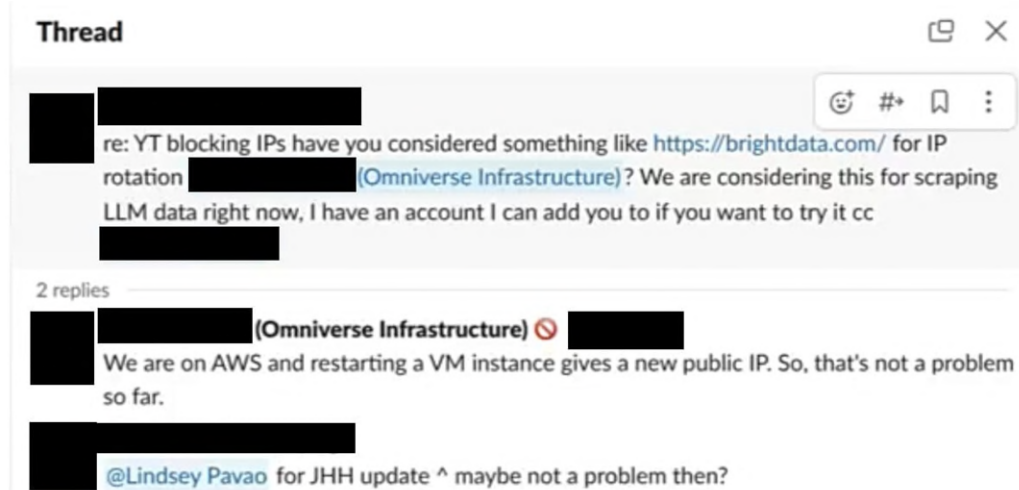


Figure 5

91. Defendant's knowledge regarding the illegality of its actions is evidenced by questions raised by Defendant's employees. For example, one of Defendant's employees specifically inquired as to whether a dataset referred to as "YouTube-8M," which had been designated for research purposes, could be used for non-research purposes, as reflected in *Figure 6*, below.



Figure 6

92. In another example, an employee explicitly expressed concern over the legality of downloading YouTube videos, but instead of receiving assurance that the process was legal, the employee received assurance that the decision to do so was an “executive decision” for which the

team had “umbrella approval,” as reflected in *Figure 7*, below.

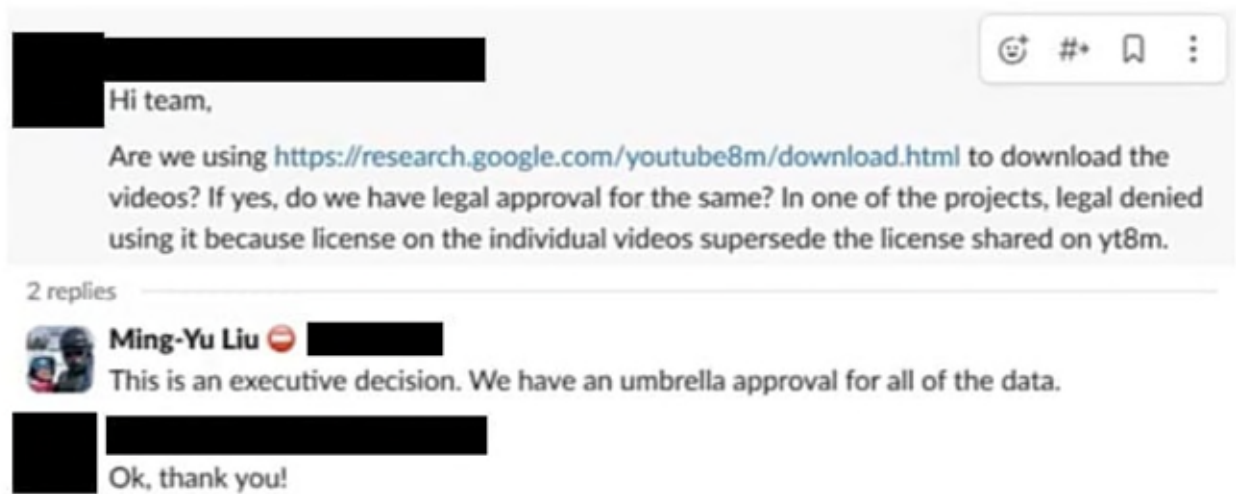


Figure 7

93. In another example, Defendant’s employees explicitly discussed YouTube’s Terms of Service that prohibit downloading of video content. Defendant conceded that “whether using copyrighted material for training is fair use is currently an open legal issue.”⁹

94. In other words, Defendant was at least aware that there were questions regarding the legality of its use of YouTube video datasets to train its generative AI models.

95. Despite these clear concerns, Defendant continued its program of improperly accessing YouTube data and using the audiovisual content it improperly accessed to train its generative AI models.

96. By May of 2024, Defendant announced that they had downloaded 38.5 million video URLs and included charts reflecting the types of content they had downloaded and cumulative downloading data, as reflected in *Figure 8*, below.

⁹ *Id.*

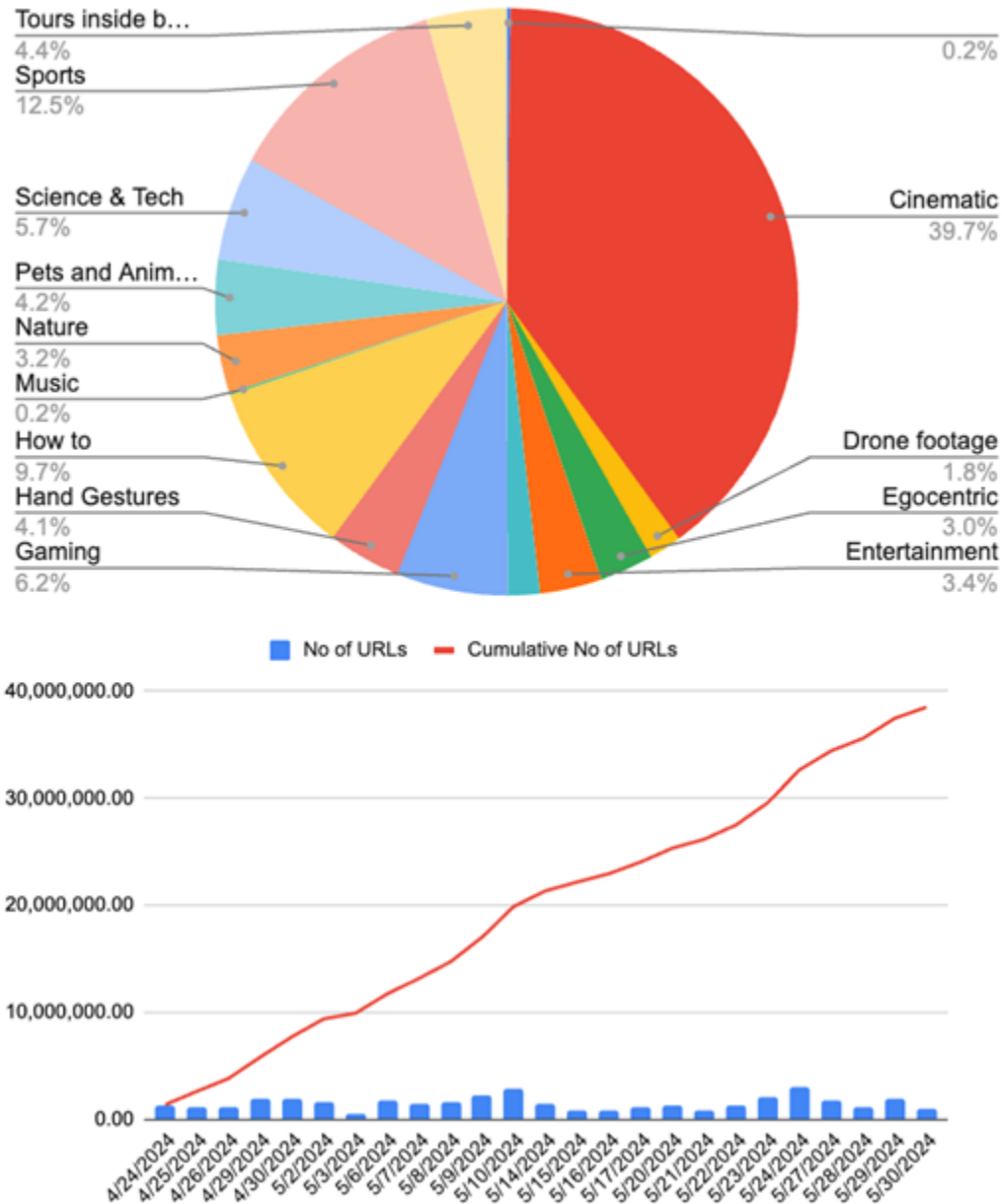


Figure 8

C. Defendant's Actions Caused Harm to Plaintiffs and Class Members

97. At no time did Defendant seek or obtain Plaintiffs' or Class Members' authorization to access and use the videos Defendant used to train its generative AI models.

98. At no time did Defendant seek or obtain Plaintiffs' or Class Members' authorization, nor did Defendant compensate Plaintiffs or the Class Members for the access, copying, ingestion, and use of their YouTube videos in AI training and related workflows.

99. Plaintiffs and Class Members used YouTube as their platform of choice to upload their video content in substantial part due to YouTube's terms and service that prohibit the type of scraping, unauthorized accessing and downloading, bulk extraction, and other forms of data mining of audiovisual content utilized by Defendant to obtain Plaintiffs' video content.

CLASS ACTION ALLEGATIONS

100. Plaintiffs bring this action on behalf of themselves and on behalf of all other persons similarly situated.

101. Plaintiffs propose the following Class definition, subject to amendment as appropriate:

All individuals and entities in the United States who uploaded original videos to YouTube and whose videos (or material portions thereof) were included in the HDVILA-100M, HD-VG-130M, and HowTo100M datasets used by Defendant to train, refine or otherwise improve its generative AI models.

102. Excluded from the Class are Defendant's officers and directors, and any entity in which Defendant has a controlling interest; and the affiliates, legal representatives, attorneys, successors, heirs, and assigns of Defendant. Excluded also from the Class are Members of the judiciary to whom this case is assigned, their families and Members of their staff.

103. Plaintiffs reserve the right to amend or modify the class definition with greater specificity or division after having an opportunity to conduct discovery. The proposed Class meets the criteria for certification under Rule 23 of the Federal Rules of Civil Procedure.

104. Numerosity. The Members of the Class are so numerous that joinder of all of them is impracticable. The exact number of Class Members is unknown to Plaintiffs now, but Plaintiffs estimates that there are thousands of Class Members.

105. Commonality. There are questions of law and fact common to the Class, which predominate over any questions affecting only individual Class Members. These common questions of law and fact include, without limitation:

a. Whether Defendant used an automated tool to obtain unauthorized access to the

1 Plaintiffs and Class Members copyrighted content on YouTube download or otherwise obtain
2 video content created and uploaded by Plaintiffs and Class Members onto YouTube;

3 b. Whether Defendant's downloading of the video content was intended to bypass the
4 TPMs put in place by YouTube at the authorization of Plaintiffs and Class Members;

5 c. Whether Defendant used Plaintiffs' and Class Members' content for purposes of
6 training its generative AI models;

7 d. Whether Plaintiffs and Class Members suffered damages from Defendant's
8 misconduct;

9 e. Whether Plaintiffs and Class Members are entitled to actual damages, statutory
10 damages and/or injunctive relief.

11 106. These issues are common to the Class, and their resolution would advance matter
12 and the parties' interests therein.

13 107. Typicality. Plaintiffs' claims are typical of those of other Class Members because
14 Plaintiffs' video content, like that of every other Class Member, was made available on YouTube
15 and improperly altered and used by Defendant to train Defendant's generative AI models for
16 commercial purposes. Plaintiffs' claims are typical of those of the other Class Members because,
17 among other things, all Class Members were injured through the common misconduct of Defendant.
18 Plaintiffs are advancing the same claims and legal theories on behalf of themselves and all other
19 Class Members, and no defenses are unique to Plaintiffs. Plaintiffs' claims and those of Class
20 Members arise from the same operative facts and are based on the same legal theories.

21 108. Adequacy of Representation. Plaintiffs will fairly and adequately represent and
22 protect the interests of the Members of the Class. Plaintiffs' Counsel is competent and experienced
23 in litigating class actions.

24 109. Predominance. Defendant has engaged in a common course of conduct toward
25 Plaintiffs and Class Members, in that Plaintiffs' and Class Members' video content was unlawfully
26 downloaded, altered and used by Defendant in the same way. The fact that Cosmos was a
27 foundational model underscores that Defendant's conduct was uniform across the Class: every video
28 unlawfully accessed from YouTube was ingested into the same core model that underpins

1 Defendant's product ecosystem. This commonality further demonstrates the predominance of shared
2 legal and factual issues among Class Members. The common issues arising from Defendant's
3 conduct affecting Class Members set out above predominate over any individualized issues.
4 Adjudication of these common issues in a single action has important and desirable advantages of
5 judicial economy.

6 110. Superiority. A Class action is superior to other available methods for the fair and
7 efficient adjudication of the controversy. Class treatment of common questions of law and fact is
8 superior to multiple individual actions or piecemeal litigation. Absent a class action, most Class
9 Members would likely find that the cost of litigating their individual claims is prohibitively high and
10 would therefore have no effective remedy. The prosecution of separate actions by individual Class
11 Members would create a risk of inconsistent or varying adjudications with respect to individual
12 Class Members, which would establish incompatible standards of conduct for Defendant. In
13 contrast, the conduct of this action as a class action presents far fewer management difficulties,
14 conserves judicial resources and the parties' resources, and protects the rights of each Class member.

15 111. Defendant has acted on grounds that apply generally to the Class as a whole, so that
16 class certification, injunctive relief, and corresponding declaratory relief are appropriate on a Class-
17 wide basis.

18 112. Ascertainability. Finally, all members of the proposed Class are readily
19 ascertainable. Defendant used datasets that contain the full lists of URLs and video identifiers for
20 every YouTube video incorporated into the training pipeline. Those identifiers allow the parties to
21 determine exactly which videos were used and to match each video to its creator through YouTube's
22 public channel and authorship information. Because the datasets provide a complete map of the
23 videos Defendant downloaded, the identities of the affected creators are identifiable through
24 straightforward reference to the URLs and corresponding channel data.

25 **CLAIM FOR RELIEF**

26 **FIRST CAUSE OF ACTION**

27 **Violation of the DMCA (Anti-Circumvention) 17 U.S.C. § 1201(a)**

28 113. Plaintiffs repeat, reallege, and incorporates the allegations contained in the previous

1 paragraphs as if fully set forth herein.

2 114. YouTube's Terms of Service prohibit scraping, unauthorized downloading, bulk
3 extraction, or other forms of data mining of audiovisual content. These restrictions operate together
4 with TPMs to prevent unlicensed access to the underlying video files, not simply reproduction of
5 content.

6 115. Plaintiffs and Class Members uploaded their original video content to YouTube's
7 video sharing platform due in substantial part to YouTube's Terms of Service and TPMs prohibiting
8 bulk downloading of creators' videos.

9 116. Content creators, including Plaintiffs and Class Members, expect that their works
10 will not be accessed at the file level and copied at scale without consent.

11 117. YouTube's Terms of Service and associated access controls constitute "effective
12 technological measures" within the meaning of 17 U.S.C. §1201.

13 118. Defendant used automated tools for the sole purpose of circumventing YouTube's
14 access barriers and extracting files never made available to the public. In doing so, Defendant
15 improperly obtained millions of videos from YouTube's platform.

16 119. This distinction is critical: viewing a YouTube video through YouTube's platform
17 does not provide access to the underlying file. Defendant's circumvention tools broke through that
18 access barrier, triggering liability under §1201.

19 120. Defendant accessed, downloaded, stored and utilized those videos to assemble
20 training corpora for Defendant's AI models and services.

21 121. Each act of circumvention constitutes a separate violation of §1201. Plaintiffs and
22 the Class Members are entitled to statutory damages, injunctive relief, impoundment, and attorneys'
23 fees and costs under 17 U.S.C. §1203.

24 122. Each of Defendant's acts of infringement is a willful violation as Defendant
25 specifically utilized automated tools designed to evade YouTube's TPMs.

26 **PRAYER FOR RELIEF**

27 WHEREFORE, Plaintiffs respectfully request a judgment in their favor and against
28 Defendant as follows:

1 A. For a declaration that Defendant has willfully circumvented the copyright
2 protection systems of YouTube intended to protect Plaintiffs' and the Class Members's
3 audiovisual content.

4 B. For statutory damages (up to the maximum allowed by law per violation),
5 injunctive relief, and attorneys' fees and costs under 17 U.S.C. §1203;

6 C. For such equitable relief under Title 17, Title 28, and/or the Court's inherent
7 authority as is necessary to prevent or restrain infringement of Plaintiffs' and the Class
8 Members' copyright-protected content, including a preliminary and permanent injunction
9 requiring that Defendant and its officers, agents, servants, employees, attorneys, directors,
10 successors, assigns, licensees, and all others in active concert or participation with any of them,
11 cease infringing, or causing, aiding, enabling, facilitating, encouraging, promoting, inducing,
12 or materially contributing to or participating in the infringement of any of Plaintiff's or the
13 Class Members' exclusive rights under federal law, including without limitation in the content
14 identified in Exhibit A;

15 D. For an award of pre-judgment and post-judgment interest, to the fullest extent
16 available, on any monetary award made part of the judgment against Defendant; and

17 E. For such other and further relief as the Court may deem just and proper.

18 **JURY DEMAND**

19 Plaintiff demands a trial by jury on all claims for which trial by jury is proper.

20 Dated: November 26, 2025

HEAH BAR-NISSIM LLP

21 By /s/ ROM BAR-NISSIM
22 ROM BAR-NISSIM

23 **ELLZEY KHERKHER SANFORD**
24 **MONTGOMERY LLP**

25 JARRET LEE ELLZEY
26 TOM KHERKHER
27 LEIGH S. MONTGOMERY
28 NATISCHA VOLPE

Attorneys for Plaintiffs