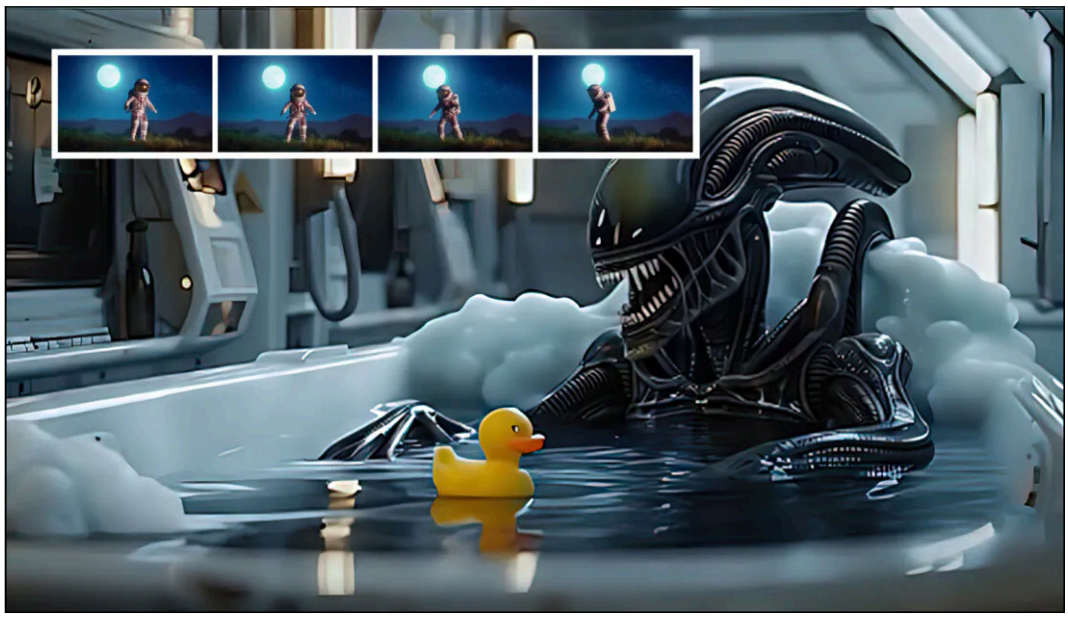
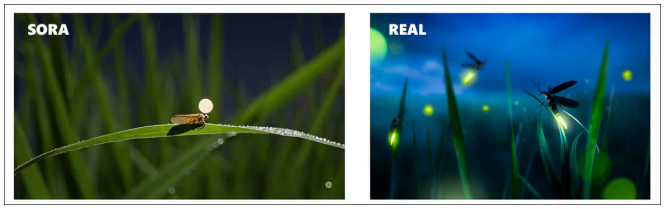
Rethinking Video AI Training with User-Focused Data

Originally published at https://www.unite.ai/rethinking-video-ai-training-with-user-focused-data/



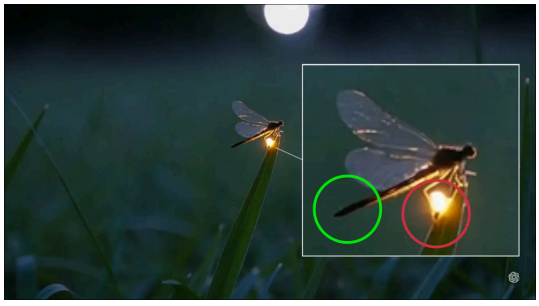
The kind of content that users might want to create using a generative model such as Flux or Hunyuan Video may not be always be easily available, even if the content request is fairly generic, and one might guess that the generator could handle it.

One example, illustrated in a new paper that we'll take a look at in this article, notes that the increasingly-eclipsed OpenAI Sora model has some difficulty rendering an anatomically correct firefly, using the prompt *'A firefly is glowing on a grass's leaf on a serene summer night'*:



OpenAI's Sora has a slightly wonky understanding of firefly anatomy. Source: https://arxiv.org/pdf/2503.01739

Since I rarely take research claims at face value, I tested the same prompt on Sora today and got a slightly better result. However, Sora still failed to render the glow correctly – rather than illuminating the tip of the firefly's tail, where bioluminescence occurs, it misplaced the glow near the insect's feet:



My own test of the researchers' prompt in Sora produces a result that shows Sora does not understand where a Firefly's light actually comes from.

Ironically, the Adobe Firefly generative diffusion engine, trained on the company's copyright-secured stock photos and videos, only managed a 1-in-3 success rate in this regard, when I tried the same prompt in Photoshop's generative AI feature:

Only the final of three proposed generations of the researchers' prompt produces a glow at all in Adobe Firefly (March 2025), though at least the glow is situated in the correct part of the insect's anatomy.

This example was highlighted by the researchers of the new paper to illustrate that the distribution, emphasis and coverage in training sets used to inform popular foundation models may not align with the user's needs, even if the user is not asking for anything particularly challenging – a topic that brings up the challenges involved in adapting hyperscale training datasets to their most efficient and performative outcomes as generative models.

The authors state:

> *'[Sora] fails to capture the concept of a glowing firefly while successfully generating grass and a summer [night]. From the data perspective, we infer this is mainly because [Sora] has not been trained on firefly-related topics, while it has been trained on grass and night. Furthermore, if [Sora had] seen the video shown in [above image], it will understand what a glowing firefly should look like.'*

They introduce a newly curated dataset and suggest that their methodology could be refined in future work to create data collections that better align with user expectations than many existing models.

Data for the People

Essentially their proposal posits a data curation approach that falls somewhere between the custom data for a model-type such as a LoRA (and this approach is far too specific for general use); and the broad and relatively indiscriminate high-volume collections (such as the LAION dataset powering Stable Diffusion) which are not specifically aligned with any end-use scenario.

The new approach, both as methodology and a novel dataset, is (rather tortuously) named *Users' FOcus in text-to-video*, or *VideoUFO*. The VideoUFO dataset comprises 1.9 million video clips spanning 1291 user-focused topics. The topics themselves were elaborately developed from an existing video dataset, and parsed through diverse language models and Natural Language Processing (NLP) techniques:

time	police	spider	bloom
crow	conversion	wealth	doll
nightclub	phone	botany	apple
night sky	sport	trump	balloon
pizza	earth	cryptocurrency	crystal
parody	banana	rotation	environment
black hole	heart	dolphin	planet
death	haunted mansion	superman	knight
dragonfly	giant	russia	krishna
cake	bunny	countryside	statue
politics	cowboy	origami	gather
childhood	penguin	drift	act
new york	reflection	perseverance	star
dystopia	cricket	horseback riding	jellyfish
routine	classroom	sheep	rise
money	climb	witchcraft	deity
chaos	laughter	fractal	evolution
snowfall	clothe	treasure	flying car
cosmology	skeleton	king	pyramid
street	glass	fear	candy
dream	ancient rome	color	agriculture
prayer	book	meet	disaster
island	virtual reality	sea creature	tokyo
parent	teddy bear	investigation	food

hulk	anger	taylor swift	hybridization
hunt	communication	happiness	meadow
background	field	skull	dinner
lightning	tennis	candle	intelligence
swimming pool	energy	disco	relaxation
gift	business	bus	illness
seaside	program	alcohol	disapproval
imprisonment	kitchen	gesture	cry
punk	hand	nutrition	social media
record	floral	life	soviet
truck	aerial	draw	shiva
compassion	digital	ski	bubble
italy	retro	computer	cinema
wrestle	fog	tribalism	hallway
crocodile	restaurant	lego	cybersecurity
giraffe	cristiano ronaldo	speech	camp
sand	introspection	character design	peacock
textile	raccoon	joy	cuteness
victoriana	spaghetti	church	manufacture
blonde	living room	skydive	fatherhood
heroism	moonlight	satire	helicopter
picnic	bikini	safety	extraterrestrial
leaf	motivation	telephone	hotel
bee	spongebob	occult	nuclear

Samples of the distilled topics presented in the new paper.

The VideoUFO dataset features a high volume of novel videos trawled from YouTube – 'novel' in the sense that the videos in question do not feature in video datasets that are currently popular in the literature, and therefore in the many subsets that have been curated from them (and many of the videos were in fact uploaded subsequent to the creation of the older datasets thar the paper mentions).

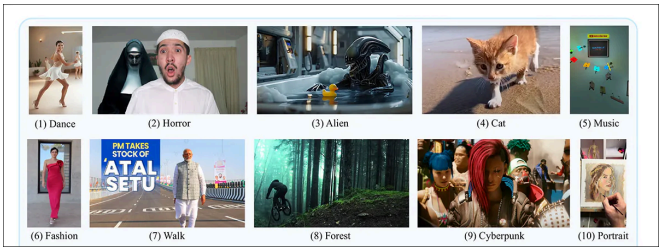
In fact, the authors claim that there is *only 0.29% overlap with existing video datasets* – an impressive demonstration of novelty.

One reason for this might be that the authors would only accept YouTube videos with a Creative Commons license that would be less likely to hamstring users further down the line: it's possible that this category of videos has been less prioritized in prior sweeps of YouTube and other high-volume platforms.

Secondly, the videos were requested on the basis of pre-estimated user-need (see image above), and not indiscriminately trawled. These two factors in combination could lead to such a novel collection. Additionally, the researchers checked the YouTube IDs of any contributing videos (i.e., videos that may later have been split up and re-imagined for the VideoUFO collection) against those featured in existing collections, lending credence to the claim.

Though not everything in the new paper is quite as convincing, it's an interesting read that emphasizes the extent to which we're still rather at the mercy of uneven distributions in datasets, in terms of the obstacles the research scene is often confronted with in dataset curation.

The new work is titled *VideoUFO: A Million-Scale User-Focused Dataset for Text-to-Video Generation*, and comes from two researchers, respectively from the University of Technology Sydney in Australia, and Zhejiang University in China.



Select examples from the final obtained dataset.

A 'Personal Shopper' for AI Data

The subject matter and concepts featured in the total sum of internet images and videos do not necessarily reflect what the average end user may end up asking for from a generative system; even where content and demand *do* tend to collide (as with porn, which is plentifully available on the internet and of great interest to many gen AI users), this may not align with the developers' intent and standards for a new generative system.

Besides the high volume of NSFW material uploaded daily, a disproportionate amount of net-available material is likely to be from advertisers and those attempting to manipulate SEO. Commercial self-interest of this kind makes the distribution of subject matter far from impartial; worse, it is difficult to develop AI-based filtering systems that can cope with the problem, since algorithms and models developed from meaningful hyperscale data may in themselves reflect the source data's tendencies and priorities.

Therefore the authors of the new work have approached the problem by reversing the proposition, through determining what users are likely to want, and obtaining videos that align with these needs.

On the surface, this approach seems just as likely to trigger a semantic race to the bottom as to achieve a balanced, Wikipedia-style neutrality. Calibrating data curation around user demand risks amplifying the preferences of the lowest-common-denominator while marginalizing niche users, since majority interests will inevitably carry greater weight.

Nonetheless, let's take a look at how the paper tackles the challenge.

Distilling Concepts with Discretion

The researchers used the 2024 VidProM dataset as the source for topic analysis that would later inform the project's web-scraping.

This dataset was chosen, the authors state, because it is the only publicly-available 1m+ dataset 'written by real users' – and it should be stated that this dataset was itself curated by the two authors of the new paper.

The paper explains*:

> *'First, we embed all 1.67 million prompts from VidProM into 384-dimensional vectors using SentenceTransformers Next, we cluster these vectors with K-means. Note that here we preset the number of clusters to a relatively large value, i.e., 2, 000, and merge similar clusters in the next step.*
>
> *'Finally, for each cluster, we ask GPT-4o to conclude a topic [one or two words].'*

The authors point out that certain concepts are distinct but notably adjacent, such as *church* and *cathedral*. Too granular a criteria for cases of this kind would lead to concept embeddings (for instance) for each type of dog breed, instead of the term *dog*; whereas too broad a criteria could corral an excessive number of sub-concepts into a single over-crowded concept; therefore the paper notes the balancing act necessary to evaluate such cases.

Singular and plural forms were merged, and verbs restored to their base (infinitive) forms. Excessively broad terms – such as *animation*, *scene*, *film* and *movement* – were removed.

Thus 1,291 topics were obtained (with the full list available in the source paper's supplementary section).

Select Web-Scraping

Next, the researchers used the official YouTube API to seek videos based on the criteria distilled from the 2024 dataset, seeking to obtain 500 videos for each topic. Besides the requisite creative commons license, each video had to have a resolution of 720p or higher, and had to be shorter than four minutes.

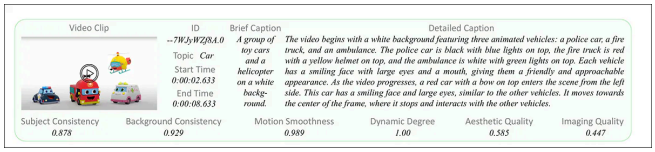
In this way 586,490 videos were scraped from YouTube.

The authors compared the YouTube ID of the downloaded videos to a number of popular datasets: OpenVid-1M; HD-VILA-100M; Intern-Vid; Koala-36M; LVD-2M; MiraData; Panda-70M; VidGen-1M; and WebVid-10M.

They found that only 1,675 IDs (the aforementioned 0.29%) of the VideoUFO clips featured in these older collections, and it has to be conceded that while the dataset comparison list is not exhaustive, it does include all the biggest and most influential players in the generative video scene.

Splits and Assessment

The obtained videos were subsequently segmented into multiple clips, according to the methodology outlined in the Panda-70M paper cited above. Shot boundaries were estimated, assemblies stitched, and the concatenated videos divided into single clips, with brief and detailed captions provided.



Each data entry in the VideoUFO dataset features a clip, an ID, start and end times, and a brief and a detailed caption.

The brief captions were handled by the Panda-70M method, and the detailed video captions by Qwen2-VL-7B, along the guidelines established by Open-Sora-Plan. In cases where clips did not successfully embody the intended target concept, the detailed captions for each such clip were fed into GPT-4o mini, in order to ascertain whether it was truly a fit for the topic. Though the authors would have preferred evaluation via GPT-4o, this would have been too expensive for millions of video clips.

Video quality assessment was handled with six methods from the VBench project.

Comparisons

The authors repeated the topic extraction process on the aforementioned prior datasets. For this, it was necessary to semantically-match the derived categories of VideoUFO to the inevitably different categories in the other collections; it has to be conceded that such processes supply only approximated equivalent categories, and therefore this may be too subjective a process to vouchsafe empirical comparisons.

Nonetheless, in the image below we see the results the researchers obtained by this method:

Dataset	Venue	Caption	#Vid.	Len.	Words	Resolution	Domain	#Topic	License
WebVid-10M [3]	ICCV21	Download	10M	18.0s	14.2	<360p	Open	1,000	Retracted
HD-VILA-100M [41]	CVPR22	Recognition	103M	13.4s	32.5	720p	Open	648	R-UDA
InternVid [5]	ICLR24	Generated	234M	11.7s	17.6	720p	Open	1,051	Apache 2.0
Panda-70M [5]	CVPR24	Generated	70M	8.5s	13.2	720p	Open	719	R-UDA
LVD-2M [40]	NeurIPS24	Generated	2M	20.2s	88.7	Diverse	Open	814	R-UDA
MiraData [9]	NeurIPS24	Generated	0.33M	72.1s	318.0	720p	Open	639	GPL 3.0
Koala-36M [33]	Arxiv24	Generated	36M	17.2s	202.1	720p	Open	724	R-UDA
VidGen-1M [29]	Arxiv24	Generated	1M	10.6s	89.3	720p	Open	835	R-UDA
OpenVid-1M [20]	ICLR25	Generated	1M	7.2s	127.3	Diverse	Open	671	R-UDA
VideoUFO	Ours	Generated	1M	12.6s	155.5	720p	**Users**	**1,291**	**CC BY**

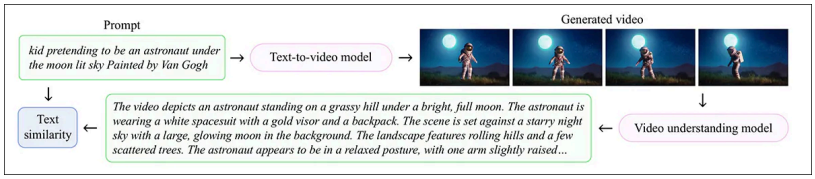
Comparison of the fundamental attributes derived across VideoUFO and the prior datasets.

The researchers acknowledge that their analysis relied on the existing captions and descriptions provided in each dataset. They admit that re-captioning older datasets using the same method as VideoUFO could have offered a more direct comparison. However, given the sheer volume of data points, their conclusion that this approach would be prohibitively expensive seems justified.

Generation

The authors developed a benchmark to evaluate text-to-video models' performance on user-focused concepts, titled *BenchUFO*. This entailed selecting 791 nouns from the 1,291 distilled user topics in VideoUFO. For each selected topic, ten text prompts from VidProM were then randomly chosen.

Each prompt was passed through to a text-to-video model, with the aforementioned Qwen2-VL-7B captioner used to evaluate the generated results. With all generated videos thus captioned, SentenceTransformers was used to calculate cosine similarity for both the input prompt and output (inferred) description in each case.



Schema for the BenchUFO process.

The evaluated generative models were: Mira; Show-1; LTX-Video; Open-Sora-Plan; Open Sora; TF-T2V; Mochi-1; HiGen; Pika; RepVideo; T2V-Zero; CogVideoX; Latte-1; Hunyuan Video; LaVie; and Pyramidal.

Besides VideoUFO, MVDiT-VidGen and MVDit-OpenVid were the alternative training datasets.

The results consider the 10th-50th worst-performing and best-performing topics across the architectures and datasets.

Models	Low 10	Low 50	Top 50	Top 10
Mira [10]	0.236	0.282	0.508	0.550
Show-1 [44]	0.266	0.303	0.524	0.564
LTX-Video [15]	0.268	0.310	0.532	0.574
Open-Sora-Plan [16]	0.314	0.361	0.559	0.598
TF-T2V [36]	0.316	0.359	0.560	0.595
Mochi-1 [30]	0.323	0.367	0.580	0.616
HiGen [24]	0.352	0.394	0.589	0.625
Open-Sora [45]	0.363	0.409	0.601	0.639
Pika [22]	0.365	0.404	0.583	0.619
RepVideo [28]	0.368	0.402	0.589	0.619
T2V-Zero [11]	0.375	0.410	0.586	0.616
CogVideoX [42]	0.383	0.419	0.601	0.629
Latte-1 [18]	0.384	0.421	0.592	0.636
HunyuanVideo [12]	0.388	0.427	0.612	0.645
LaVie [37]	0.399	0.426	0.595	0.632
Pyramidal [8]	0.400	0.433	0.606	0.647
MVDiT-VidGen [29]	0.382	0.426	0.594	0.626
MVDiT-OpenVid [20]	0.401	0.437	0.609	0.645
MVDiT-VideoUFO	**0.442**	**0.465**	**0.619**	**0.651**

Results for the performance of public T2V models vs. the authors' trained models, on BenchUFO.

Here the authors comment:

> *'Current text-to-video models do not consistently perform well across all user-focused topics. Specifically, there is a score difference ranging from 0.233 to 0.314 between the top-10 and low-10 topics. These models may not effectively understand topics such as "giant squid", "animal cell", "Van Gogh", and "ancient Egyptian" due to insufficient training on such videos.*
>
> *'Current text-to-video models show a certain degree of consistency in their best-performing topics. We discover that most text-to-video models excel at generating videos on animal-related topics, such as 'seagull', 'panda', 'dolphin', 'camel', and 'owl'. We infer that this is partly due to a bias towards animals in current video datasets.'*

Conclusion

VideoUFO is an outstanding offering if only from the standpoint of fresh data. If there has been no error in evaluating and eliminating YouTube IDs, and if the dataset contains so much material that is new to the research scene, it is a rare and potentially valuable proposition.

The downside is that one needs to give credence to the core methodology; if you don't believe that user demand should inform web-scraping formulas, you'd be buying into a dataset that comes with its own sets of troubling biases.

Further, the utility of the distilled topics depends on both the reliability of the distilling method used (which is generally hampered by budget constraints), and also the formulation methods for the 2024 dataset that provides the source material.

That said, VideoUFO certainly merits further investigation – and it is available at Hugging Face.

* *My substitution of the authors' citations for hyperlinks.*

First published Wednesday, March 5, 2025

Writer on machine learning, domain specialist in human image synthesis. Former head of research content at Metaphysic.ai, until its dissolution into DNEG's Brahma.ai.
Portfolio site: martinanderson.ai
Contact: martin@martinanderson.ai

Discover More