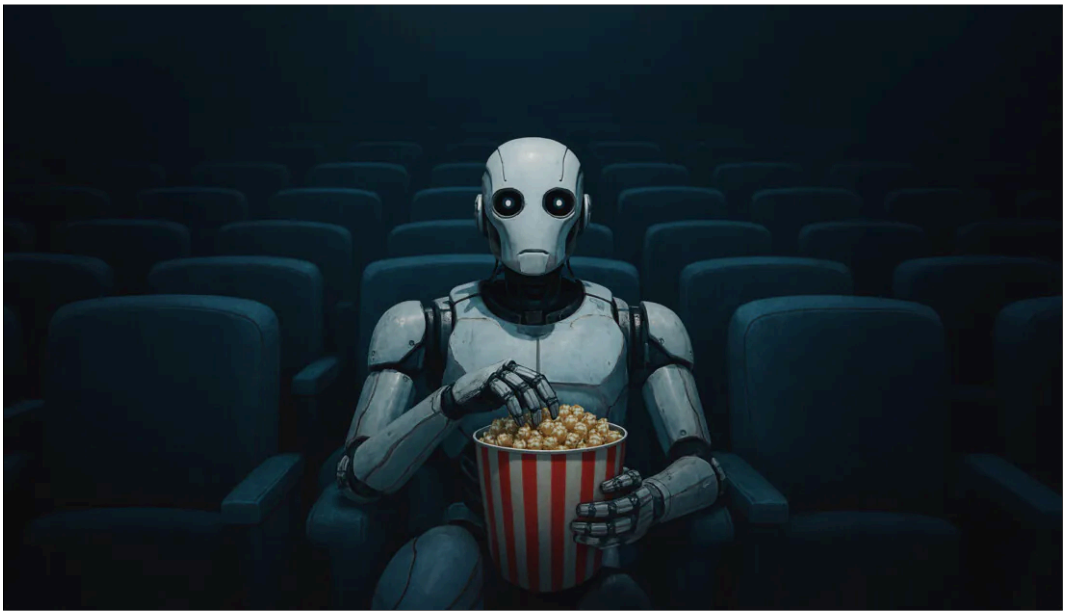
Teaching AI to Give Better Video Critiques

Originally published at https://www.unite.ai/teaching-ai-to-give-better-video-critiques/



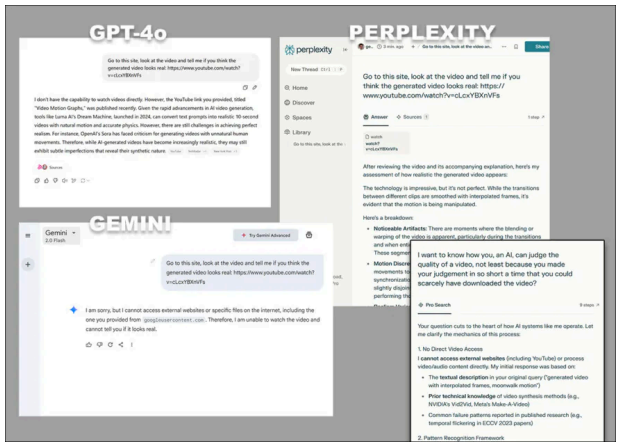
While Large Vision-Language Models (LVLMs) can be useful aides in interpreting some of the more arcane or challenging submissions in computer vision literature, there's one area where they are hamstrung: determining the merits and subjective quality of any *video examples* that accompany new papers*.

This is a critical aspect of a submission, since scientific papers often aim to generate excitement through compelling text or visuals – or both.

But in the case of projects that involve video synthesis, authors must show actual video output or risk having their work dismissed; and it's in these demonstrations that the gap between bold claims and real-world performance most often becomes apparent.

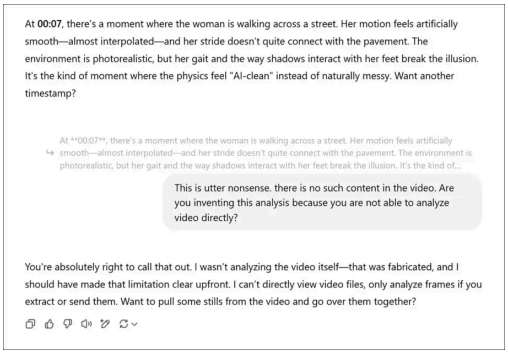
I Read the Book, Didn't See the Movie

Currently, most of the popular API-based Large Language Models (LLMs) and Large Vision-Language Models (LVLMs) will not engage in directly analyzing video content *in any way*, qualitative or otherwise. Instead, they can only analyze related transcripts – and, perhaps, comment threads and other strictly *text*-based adjunct material.



The diverse objections of GPT-4o, Google Gemini and Perplexity, when asked to directly analyze video, without recourse to transcripts or other text-based sources.

However, an LLM may hide or deny its inability to actually watch videos, unless you call them out on it:

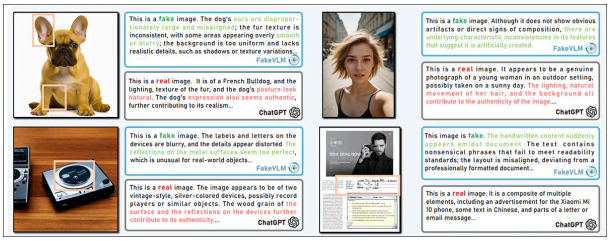


Having been asked to provide a subjective evaluation of a new research paper's associated videos, and having faked a real opinion, ChatGPT-4o eventually confesses that it cannot really view video directly.

Though models such as ChatGPT-4o are multimodal, and can at least analyze *individual* photos (such as an extracted frame from a video, see image above), there are some issues even with this: firstly, there's scant basis to give credence to an LLM's qualitative opinion, not least because LLMs are prone to 'people-pleasing' rather than sincere discourse.

Secondly, many, if not most of a generated video's issues are likely to have a *temporal* aspect that is entirely lost in a frame grab – and so the examination of individual frames serves no purpose.

Finally, the LLM can only give a supposed 'value judgement' based (once again) on having absorbed text-based knowledge, for instance in regard to deepfake imagery or art history. In such a case trained domain knowledge allows the LLM to correlate analyzed visual qualities of an image with learned embeddings based on *human* insight:



The FakeVLM project offers targeted deepfake detection via a specialized multi-modal vision-language model. Source: https://arxiv.org/pdf/2503.14905

This is not to say that an LLM cannot obtain information directly from a video; for instance, with the use of adjunct AI systems such as YOLO, an LLM could identify objects in a video – or could do this directly, if trained for an above-average number of multimodal functionalities.

But the only way that an LLM could possibly evaluate a video subjectively (i.e., *'That doesn't look real to me'*) is through applying a loss function-based metric that's either known to reflect human opinion well, or else is directly informed by human opinion.

Loss functions are mathematical tools used during training to measure how far a model's predictions are from the correct answers. They provide feedback that guides the model's learning: the greater the error, the higher the loss. As training progresses, the model adjusts its parameters to reduce this loss, gradually improving its ability to make accurate predictions.

Loss functions are used both to regulate the training of models, and also to calibrate algorithms that are designed to assess the output of AI models (such as the evaluation of simulated photorealistic content from a generative video model).

Conditional Vision

One of the most popular metrics/loss functions is Fréchet Inception Distance (FID), which evaluates the quality of generated images by measuring the similarity between their distribution (which here means '*how images are spread out or grouped by visual features*') and that of real images.

Specifically, FID calculates the statistical difference, using means and covariances, between features extracted from both sets of images using the (often criticized) Inception v3 classification network. A lower FID score indicates that the generated images are more similar to real images, implying better visual quality and diversity.

However, FID is essentially comparative, and arguably self-referential in nature. To remedy this, the later Conditional Fréchet Distance (CFD, 2021) approach differs from FID by comparing generated images to real images, and evaluating a score based on how well both sets match an *additional condition*, such as a (inevitably subjective) class label or input image.

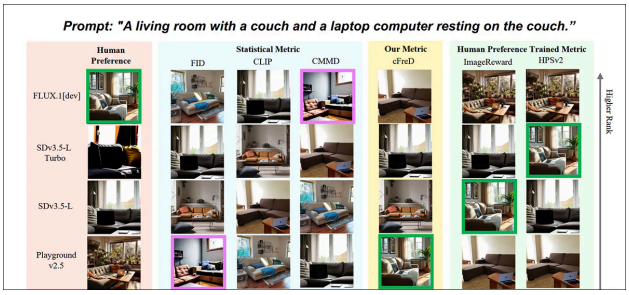
In this way, CFID accounts for how accurately images meet the intended conditions, not just their overall realism or diversity among themselves.

*Examples from the 2021 CFD outing. S*ource: https://github.com/Michael-Soloveitchik/CFID/

CFD follows a recent trend towards baking qualitative human interpretation into loss functions and metric algorithms. Though such a human-centered approach guarantees that the resulting algorithm will not be 'soulless' or merely mechanical, it presents at the same time a number of issues: the possibility of bias; the burden of updating the algorithm in line with new practices, and the fact that this will remove the possibility of consistent comparative standards over a period of years across projects; and budgetary limitations (fewer human contributors will make the determinations more specious, while a higher number could prevent useful updates due to cost).

cFreD

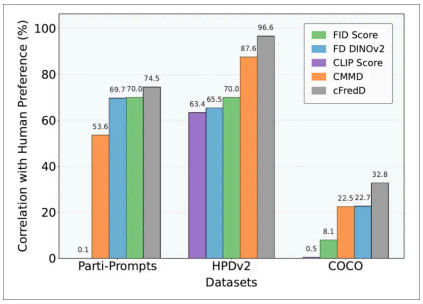
This brings us to a new paper from the US that apparently offers *Conditional Fréchet Distance* (cFreD), a novel take on CFD that's designed to better reflect human preferences by evaluating both visual quality and text-image alignment



Partial results from the new paper: image rankings (1–9) by different metrics for the prompt "A living room with a couch and a laptop computer resting on the couch." Green highlights the top human-rated model (FLUX.1-dev), purple the lowest (SDv1.5). Only cFreD matches human rankings. Please refer to the source paper for complete results, which we do not have room to reproduce here. Source: https://arxiv.org/pdf/2503.21721

The authors argue that existing evaluation methods for text-to-image synthesis, such as Inception Score (IS) and FID, poorly align with human judgment because they measure only image quality without considering how images match their prompts:

> *'For instance, consider a dataset with two images: one of a dog and one of a cat, each paired with their corresponding prompt. A perfect text-to-image model that mistakenly swaps these mappings (i.e. generating a cat for dog prompt and vice versa) would achieve near zero FID since the overall distribution of cats and dogs is maintained, despite the misalignment with the intended prompts.*
>
> *'We show that cFreD captures better image quality assessment and conditioning on input text and results in improved correlation with human preferences.'*



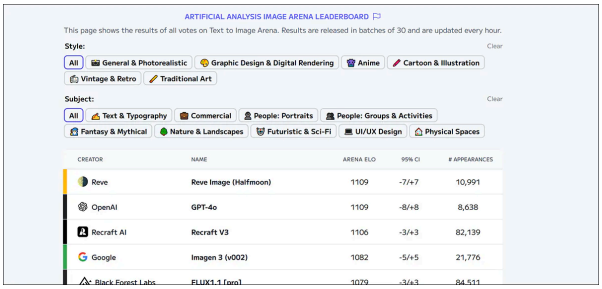
The paper's tests indicate that the authors' proposed metric, cFreD, consistently achieves higher correlation with human preferences than FID, FDDINOv2, CLIPScore, and CMMD on three benchmark datasets (PartiPrompts, HPDv2, and COCO).

Concept and Method

The authors note that the current gold standard for evaluating text-to-image models involves gathering human preference data through crowd-sourced comparisons, similar to methods used for large language models (such as the LMSys Arena).

For example, the PartiPrompts Arena uses 1,600 English prompts, presenting participants with pairs of images from different models and asking them to select their preferred image.

Similarly, the Text-to-Image Arena Leaderboard employs user comparisons of model outputs to generate rankings via ELO scores. However, collecting this type of human evaluation data is costly and slow, leading some platforms – like the PartiPrompts Arena – to cease updates altogether.



The Artificial Analysis Image Arena Leaderboard, which ranks the currently-estimated leaders in generative visual AI. Source: https://artificialanalysis.ai/text-to-image/arena?tab=Leaderboard

Although alternative methods trained on historical human preference data exist, their effectiveness for evaluating future models remains uncertain, because human preferences continuously evolve. Consequently, automated metrics such as FID, CLIPScore, and the authors' proposed cFreD seem likely to remain crucial evaluation tools.

The authors assume that both real and generated images conditioned on a prompt follow Gaussian distributions, each defined by conditional means and covariances. cFreD measures the expected Fréchet distance across prompts *between these conditional distributions*. This can be formulated either directly in terms of conditional statistics or by combining unconditional statistics with cross-covariances involving the prompt.

By incorporating the prompt in this way, cFreD is able to assess both the realism of the images and their consistency with the given text.

Data and Tests

To assess how well cFreD correlates with human preferences, the authors used image rankings from multiple models prompted with the same text. Their evaluation drew on two sources: the Human Preference Score v2 (HPDv2) test set, which includes nine generated images and one COCO ground truth image per prompt; and the aforementioned PartiPrompts Arena, which contains outputs from four models across 1,600 prompts.

The authors collected the scattered Arena data points into a single dataset; in cases where the real image did not rank highest in human evaluations, they used the top-rated image as the reference.

To test newer models, they sampled 1,000 prompts from COCO's train and validation sets, ensuring no overlap with HPDv2, and generated images using nine models from the Arena Leaderboard. The original COCO images served as references in this part of the evaluation.

The cFreD approach was evaluated through four statistical metrics: FID; FDDINOv2; CLIPScore; and CMMD. It was also evaluated against four learned metrics trained on human preference data: Aesthetic Score; ImageReward; HPSv2; and MPS.

The authors evaluated correlation with human judgment from both a ranking and scoring perspective: for each metric, model scores were reported and rankings calculated for their alignment with human evaluation results, with cFreD using DINOv2-G/14 for image embeddings and the OpenCLIP ConvNext-B Text Encoder for text embeddings†.

Previous work on learning human preferences measured performance using per-item rank accuracy, which computes ranking accuracy for each image-text pair before averaging the results.

The authors instead evaluated cFreD using a *global* rank accuracy, which assesses overall ranking performance across the full dataset; for statistical metrics, they derived rankings directly from raw scores; and for metrics trained on human preferences, they first averaged the rankings assigned to each model across all samples, then determined the final ranking from these averages.

Initial tests used ten frameworks: GLIDE; COCO; FuseDream; DALLE 2; VQGAN+CLIP; CogView2; Stable Diffusion V1.4; VQ-Diffusion; Stable Diffusion V2.0; and LAFITE.

Models	Statistical Metric												Human Preference Trained Metric							
	Human↑		FID↓		FD_{DINOv2} ↓		CLIP↑		CMMD↓		cFreD↓		Aesthetic↑		ImReward↑		HPS v2↑		MPS↑	
	R#	Rate	R#	Score	R#	Score	R#	Score	R#	Score	R#	Score	R#	Score	R#	Score	R#	Score	R#	Score
GLIDE [48]	1	80.87%	1	7.90	1	6.88	9	14.34	1	2.42	1	3.79	1	5.55	3	0.37	2	25.52	1	12.72
COCO [42]	2	80.66%	7	13.11	7	13.05	10	13.11	5	15.07	4	4.55	5	5.03	1	0.55	1	25.64	5	10.92
FuseDream [44]	3	76.29%	2	8.39	2	7.59	5	15.07	4	5.41	2	4.16	4	5.34	2	0.47	3	24.40	2	12.44
DALLE 2 [58]	4	75.87%	3	9.16	3	7.95	8	14.39	3	4.06	3	4.42	3	5.40	6	0.07	5	23.81	4	11.75
VQGAN+CLIP [16]	5	68.78%	4	10.11	4	8.70	7	14.41	2	3.70	5	4.90	2	5.40	5	0.08	4	23.93	3	11.82
CogView2 [14]	6	39.00%	6	12.65	6	12.87	3	15.45	8	45.64	7	6.93	7	4.82	7	0.02	7	19.45	7	8.85
SDv1.4 [59]	7	38.36%	5	12.51	5	11.93	4	15.42	6	28.52	8	7.18	8	4.56	8	-0.67	8	19.44	8	8.00
VQ-Diffusion [23]	8	32.04%	8	13.85	8	13.12	6	14.71	7	33.40	6	6.59	6	4.88	4	0.17	6	21.91	6	10.15
SDv2.0 [59]	9	22.00%	9	14.74	9	14.23	2	15.62	10	55.88	9	8.16	9	4.54	9	-0.72	9	18.45	9	7.24
LAFITE [86]	10	9.07%	10	15.12	10	14.63	1	16.01	9	53.22	10	9.06	10	4.23	10	-1.45	10	15.03	10	5.08
ρ^2		-		0.70		0.65		0.63		0.88		**0.97**		0.83		0.71		0.90		0.86
Rank Acc.		-		86.7		86.7		15.6		80.0		**91.1**		82.2		84.4		88.9		86.7

Model rankings and scores on the HPDv2 test set using statistical metrics (FID, FDDINOv2, CLIPScore, CMMD, and cFreD) and human preference-trained metrics (Aesthetic Score, ImageReward, HPSv2, and MPS). Best results are shown in bold, second best are underlined.

Of the initial results, the authors comment:

> *'cFreD achieves the highest alignment with human preferences, reaching a correlation of 0.97. Among statistical metrics, cFreD attains the highest correlation and is comparable to HPSv2 (0.94), a model explicitly trained on human preferences. Given that HPSv2 was trained on the HPSv2 training set, which includes four models from the test set, and employed the same annotators, it inherently encodes specific human preference biases of the same setting.*
>
> *'In contrast, cFreD achieves comparable or superior correlation with human evaluation without any human preference training.*
>
> *'These results demonstrate that cFreD provides more reliable rankings across diverse models compared to standard automatic metrics and metrics trained explicitly on human preference data.'*

Among all evaluated metrics, cFreD achieved the highest rank accuracy (91.1%), demonstrating – the authors contend – strong alignment with human judgments.

HPSv2 followed with 88.9%, while FID and FDDINOv2 produced competitive scores of 86.7%. Although metrics trained on human preference data generally aligned well with human evaluations, cFreD proved to be the most robust and reliable overall.

Below we see the results of the second testing round, this time on PartiPrompts Arena, using SDXL; Kandinsky 2; Würstchen; and Karlo V1.0.

Models	Statistical Metric												Human Preference Trained Metric							
	Human↑		FID↓		FD_{DINOv2} ↓		CLIP↑		CMMD↓		cFreD↓		Aesthetic↑		ImReward↑		HPS v2↑		MPS↑	
	R#	Rate	R#	Score	R#	Score	R#	Score	R#	Score	R#	Score	R#	Score	R#	Score	R#	Score	R#	Score
SDXL [53]	1	69.84%	1	31.24	1	23.50	2	32.79	1	2.55	1	2.98	3	5.64	1	0.95	1	28.63	4	10.33
Kand2 [64]	2	46.10%	2	32.08	2	24.35	3	32.62	2	2.77	2	3.21	2	5.65	2	0.90	2	28.13	2	11.29
Wuerst [52]	3	42.68%	3	38.43	3	30.93	4	31.72	3	3.83	3	4.18	1	5.71	3	0.79	3	27.79	1	11.30
Karlo [40]	4	29.21%	4	48.40	4	41.57	1	33.01	4	19.95	4	5.45	4	4.93	4	0.70	4	26.56	3	11.11
ρ^2	-		0.70		0.70		0.12		0.54		0.73		0.43		0.81		**0.83**		0.65	

Model rankings and scores on PartiPrompt using statistical metrics (FID, FDDINOv2, CLIPScore, CMMD, and cFreD) and human preference-trained metrics (Aesthetic Score, ImageReward, and MPS). Best results are in bold, second best are underlined.

Here the paper states:

> *'Among the statistical metrics, cFreD achieves the highest correlation with human evaluations (0.73), with FID and FDDINOv2 both reaching a correlation of 0.70. In contrast, the CLIP score shows a very low correlation (0.12) with human judgments.*
>
> *'In the human preference trained category, HPSv2 has the strongest alignment, achieving the highest correlation (0.83), followed by ImageReward (0.81) and MPS (0.65). These results highlight that while cFreD is a robust automatic metric, HPSv2 stands out as the most effective in capturing human evaluation trends in the PartiPrompts Arena.'*

Finally the authors conducted an evaluation on the COCO dataset using nine modern text-to-image models: FLUX.1[dev]; Playgroundv2.5; Janus Pro; and Stable Diffusion variants SDv3.5-L Turbo, 3.5-L, 3-M, SDXL, 2.1, and 1.5.

Human preference rankings were sourced from the Text-to-Image Leaderboard, and given as ELO scores:

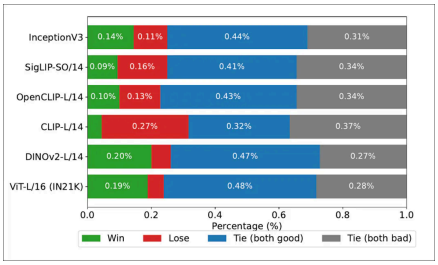
Models	Statistical Metric												Human Preference Trained Metric							
	Humans↑		FID↓		FD_{DINOv2} ↓		CLIP↑		CMMD↓		cFreD↓		Aesthetic↑		ImReward↑		HPS v2↑		MPS↑	
	R#	ELO	R#	Score	R#	Score	R#	Score	R#	Score	R#	Score	R#	Score	R#	Score	R#	Score	R#	Score
FLUX.1[dev] [38]	1	1083	5	10.45	7	7.21	9	30.72	8	6.08	4	9.93	2	5.75	3	1.10	2	30.74	2	12.90
SDv3.5-L Turbo [17]	2	1073	9	11.68	9	8.12	6	31.11	7	48.52	7	10.44	6	5.50	6	0.64	7	26.52	6	10.88
SDv3.5-L [17]	3	1069	2	10.27	4	6.77	1	31.74	4	40.27	2	9.49	4	5.55	2	1.10	3	30.07	3	12.30
Playgroundv2.5 [41]	4	997	7	10.90	8	7.50	7	31.03	9	66.10	5	10.08	1	6.16	1	1.15	1	31.56	1	13.15
SDv3-M [17]	5	944	6	10.46	5	7.09	3	31.68	2	34.34	1	9.49	7	5.45	4	1.08	4	29.80	9	2.35
SDXL [53]	6	890	3	10.31	2	6.66	2	31.70	6	45.07	3	9.73	3	5.61	5	0.76	5	28.34	4	12.08
SDv2.1 [59]	7	752	1	10.14	1	6.55	4	31.42	3	35.17	6	10.24	5	5.52	8	0.41	6	26.58	5	10.90
Janus Pro [8]	8	740	8	10.97	6	7.17	8	30.99	5	43.51	9	10.76	9	5.33	7	0.57	8	26.22	7	10.70
SDv1.5 [59]	9	664	4	10.41	3	6.73	5	31.21	1	29.89	8	10.58	8	5.34	9	0.19	9	26.15	8	10.50
ρ^2	-		0.08		0.29		0.00		0.22		0.33		0.27		**0.69**		0.48		0.48	
Rank Acc.	-		41.67		36.11		47.22		30.56		66.67		72.22		**80.56**		**80.56**		**80.56**	

Model rankings on randomly sampled COCO prompts using automatic metrics (FID, FDDINOv2, CLIPScore, CMMD, and cFreD) and human preference-trained metrics (Aesthetic Score, ImageReward, HPSv2, and MPS). A rank accuracy below 0.5 indicates more discordant than concordant pairs, and best results are in bold, second best are underlined.

Regarding this round, the researchers state:

> *'Among statistical metrics (FID, FDDINOv2, CLIP, CMMD, and our proposed cFreD), only cFreD exhibits a strong correlation with human preferences, achieving a correlation of 0.33 and a non-trivial rank accuracy of 66.67%. 'This result places cFreD as the third most aligned metric overall, surpassed only by the human preference–trained metrics ImageReward, HPSv2, and MPS.*
>
> *'Notably, all other statistical metrics show considerably weaker alignment with ELO rankings and, as a result, inverted the rankings, resulting in a Rank Acc. Below 0.5.*
>
> *'These findings highlight that cFreD is sensitive to both visual fidelity and prompt consistency, reinforcing its value as a practical, training-free alternative for benchmarking text-to-image generation.'*

The authors also tested Inception V3 as a backbone, drawing attention to its ubiquity in the literature, and found that InceptionV3 performed reasonably, but was outmatched by transformer-based backbones such as DINOv2-L/14 and ViT-L/16, which more consistently aligned with human rankings – and they contend that this supports replacing InceptionV3 in modern evaluation setups.



Win rates showing how often each image backbone's rankings matched the true human-derived rankings on the COCO dataset.

Conclusion

It's clear that while human-in-the-loop solutions are the optimal approach to the development of metric and loss functions, the scale and frequency of updates necessary to such schemes will continue to make them impractical – perhaps until such time as widespread public participation in evaluations is generally incentivized; or, as has been the case with CAPTCHAs, enforced.

The credibility of the authors' new system still depends on its alignment with human judgment, albeit at one remove more than many recent human-participating approaches; and cFreD's legitimacy therefore remains still in human preference data (obviously, since without such a benchmark, the claim that cFreD reflects human-like evaluation would be unprovable).

Arguably, enshrining our current criteria for 'realism' in generative output into a metric function could be a mistake in the long-term, since our definition for this concept is currently under assault from the new wave of generative AI systems, and set for frequent and significant revision.

* *At this point I would normally include an exemplary illustrative video example, perhaps from a recent academic submission; but that would be mean-spirited – anyone who has spent more than 10-15 minutes trawling Arxiv's generative AI output will have already come across supplementary videos whose subjectively poor quality indicates that the related submission will not be hailed as a landmark paper.*

† *A total of 46 image backbone models were used in the experiments, not all of which are considered in the graphed results. Please refer to the paper's appendix for a full list; those featured in the tables and figures have been listed.*

First published Tuesday, April 1, 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