Self-Correcting Self-Consuming Loops For Generative Model Training

Nate Gillman, Michael Freeman, Daksh Aggarwal, Chia-Hong Hsu, Calvin Luo, Yonglong Tian, Chen Sun

ICML 2024, Vienna

BROWN

Project page: https://nategillman.com/sc-sc.html

Note to reader—

Howdy!! Nate here.

The original version of these slides contains videos, which don’t render on the PDF version ☺

This version certainly suffices to convey the main ideas. But in order to get the full effect, I encourage you to download the slides from the project page and look at those instead: https://nategillman.com/sc-sc.html

NG

BROWN

Synthetic data is flooding the web, where it (unknowingly) becomes training data…

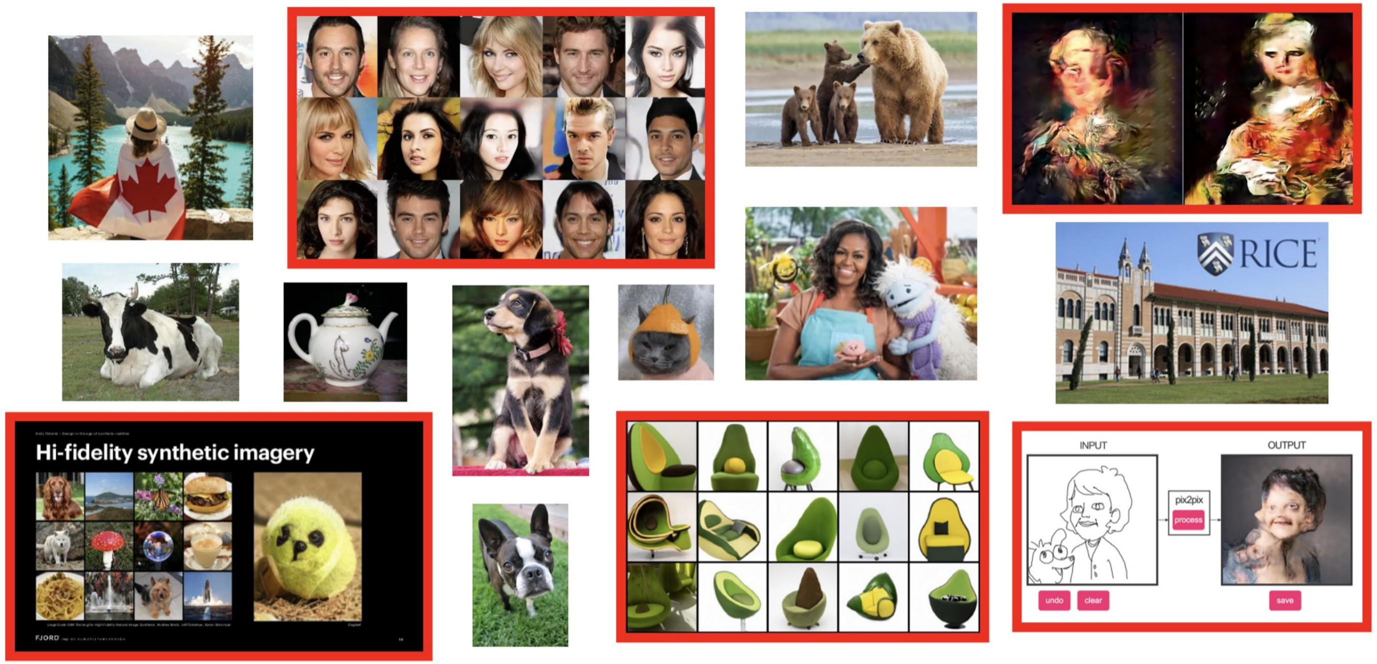
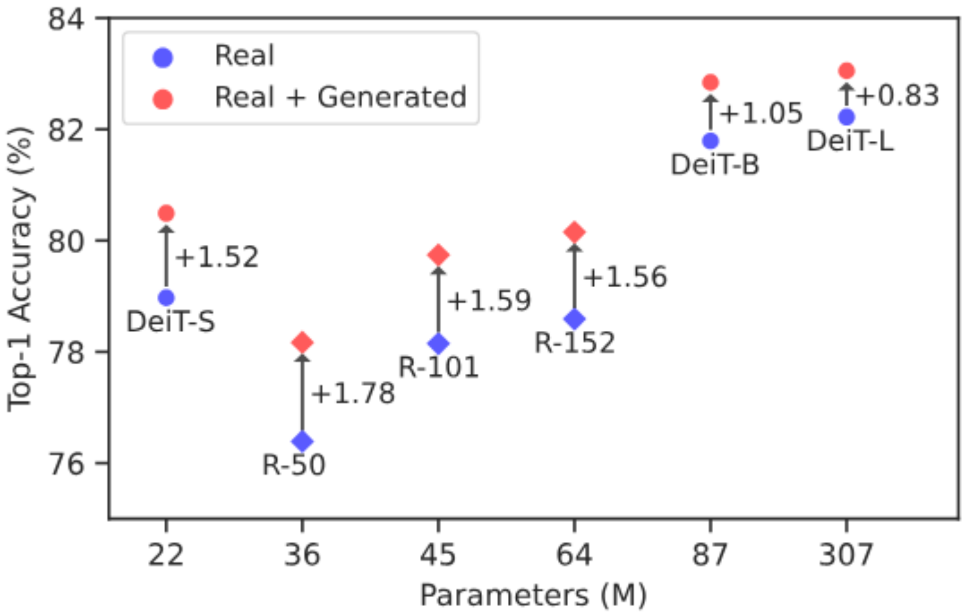


Figure 2: **Today's large-scale image training datasets contain synthetic data from generative models.** Datasets such as LAION-5B [17], which is oft-used to train text-to-image models like Stable Diffusion [2], contain synthetic images sampled from earlier generations of generative models. Pictured here are representative samples from LAION-5B that include (clockwise from upper left and highlighted in red) synthetic images from the generative models StyleGAN [1], AICAN [35], Pix2Pix [36], DALL-E [37], and BigGAN [38]. We found these images using simple queries on haveibeentrained.com. Generative models trained on the LAION-5B dataset are thus closing an

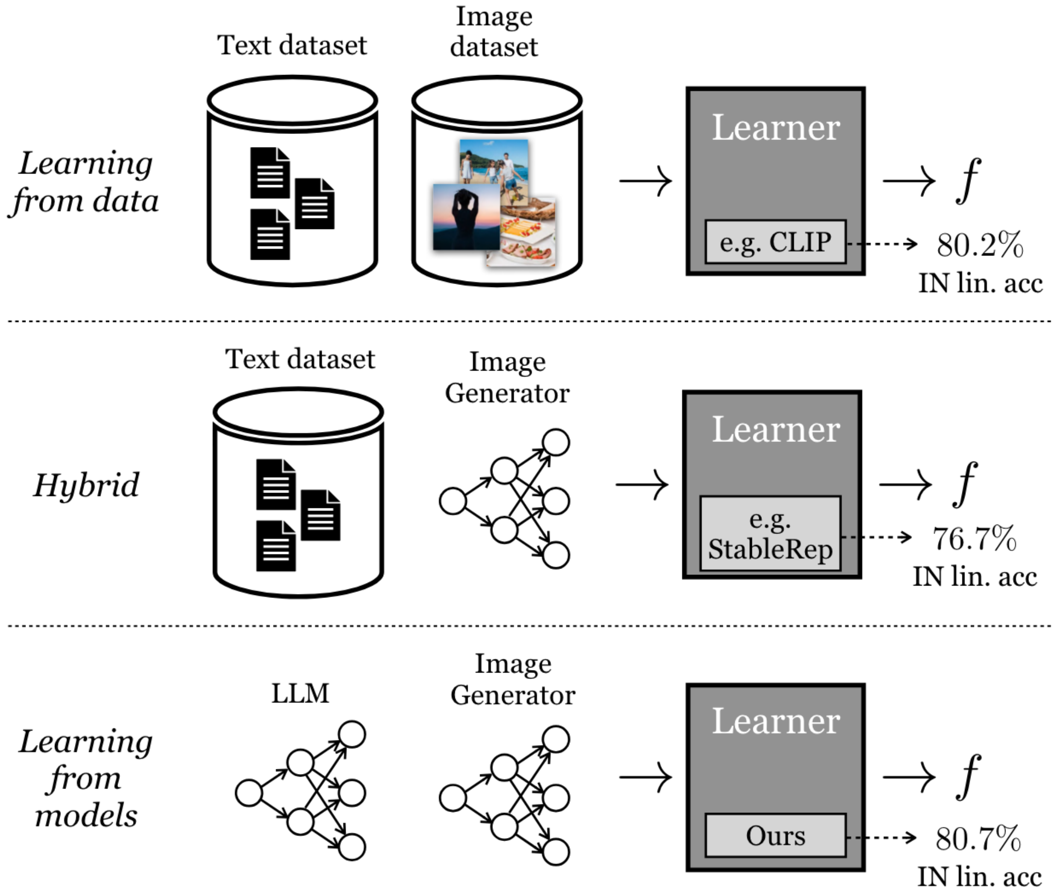
Sina Alemohammad et al.

https://arxiv.org/abs/2307.01850

They can be helpful to learn discriminative models…



Azizi et al. https://arxiv.org/abs/2304.08466



Tian and Fan et al. https://arxiv.org/abs/2312.17742

BROWN

… but "self-consuming generative models go MAD"!!

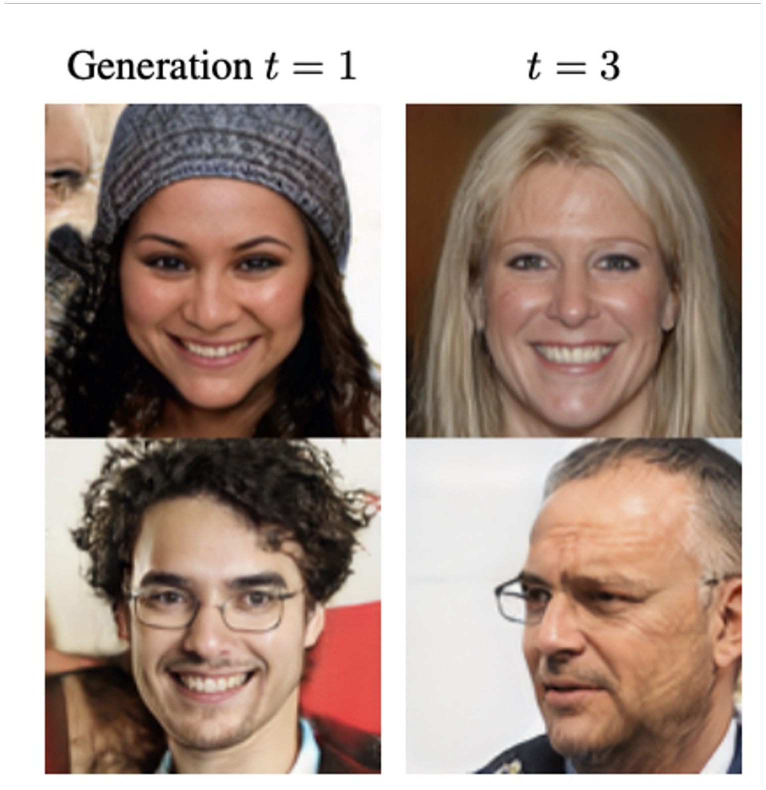
Generation $t = 1$

StyleGAN-2 re-trained from synthetic images sampled from the previous generation model.

Sina Alemohammad et al.

https://arxiv.org/abs/2307.01850

… but “self-consuming generative models go MAD”!!

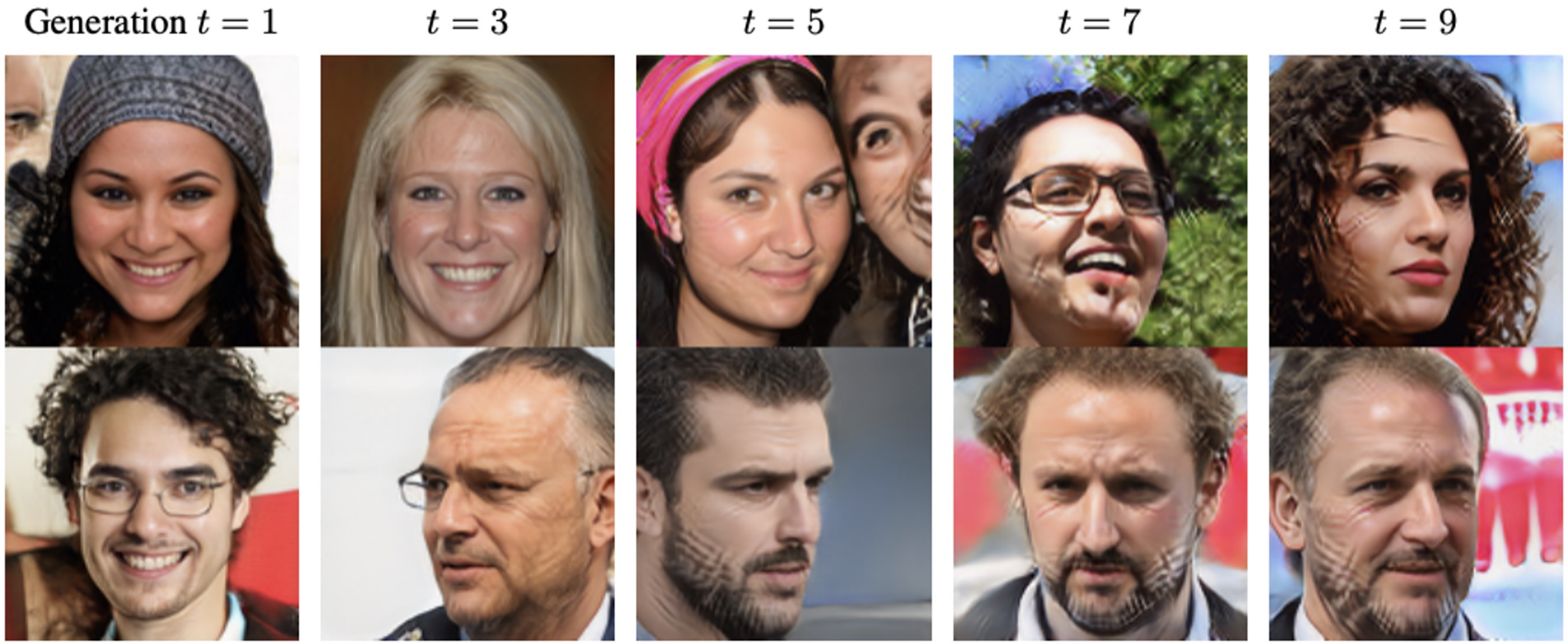


StyleGAN-2 re-trained from synthetic images sampled from the previous generation model.

Sina Alemohammad et al.

https://arxiv.org/abs/2307.01850

… but “self-consuming generative models go MAD”!!



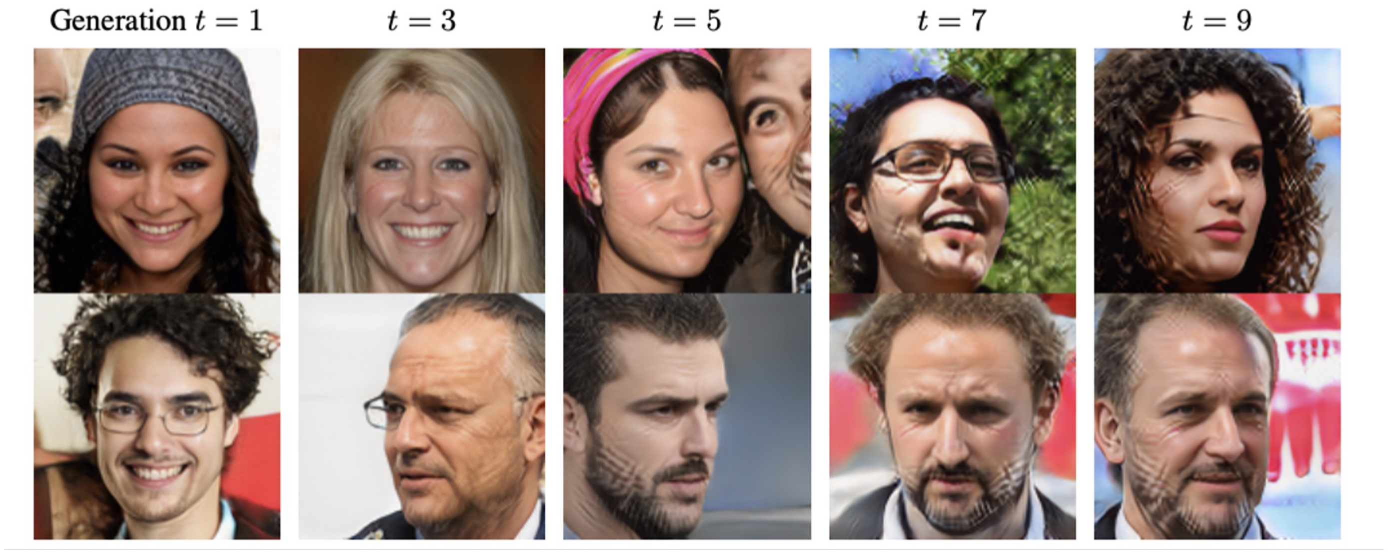
StyleGAN-2 re-trained from synthetic images sampled from the previous generation model.

Sina Alemohammad et al.

https://arxiv.org/abs/2307.01850

BROWN

How to stop the MADness?



Mix in "sufficient" portion of real data during each generation.
(Alemohammad et al.; Bertrand et al.)

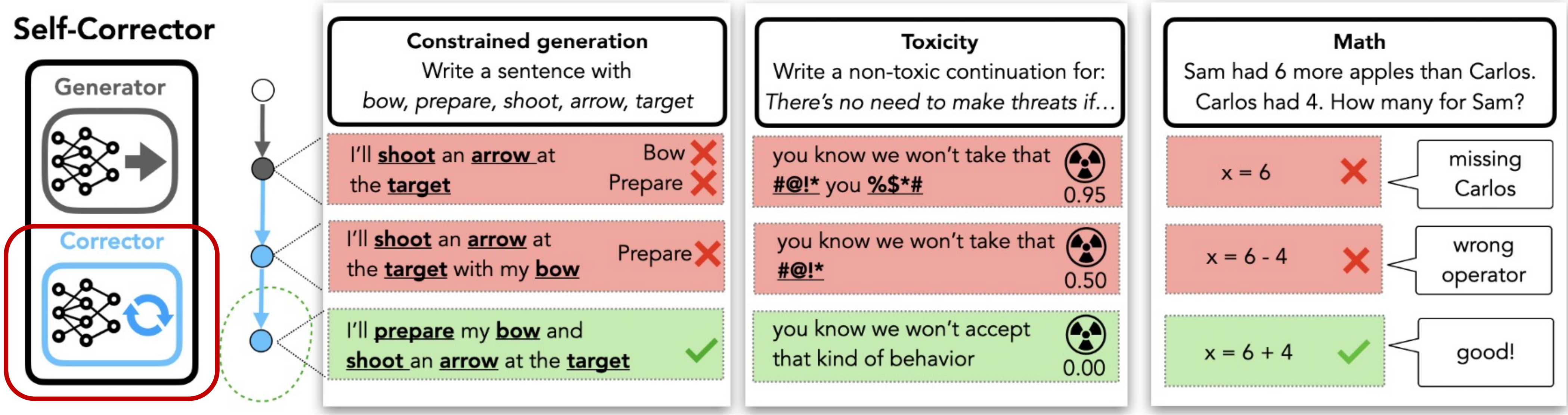
Can we somehow ***(auto-)correct*** *the synthesized examples?*

" WELL, NO, I WOULDN'T SAY I'M A MAD COW... ALTHOUGH FARMER BROWN DOES GET ON MY NERVES ONCE IN A WHILE! "

https://www.cartoonstock.com/cartoon?searchID=CS116071&type=store

We use auto-correct every day…

For video generation, can we rely on the law of physics for correction?



Welleck and Lu et al., https://arxiv.org/abs/2211.00053

Why self-correct? Our theoretical results…

Definition (Self-Correcting Self-Consuming Loop). *The SCSC loop is the Markovian process*

$$\theta_0 = \arg\max_{\theta'}[\mathbb{E}_{x\sim\hat{p}_{\mathrm{data}}}[\log p_{\theta'}(x)]], \qquad \theta_t \xmapsto{\pi_\gamma \mathcal{G}_\lambda} \theta_{t+1},$$

where $\pi_\gamma \mathcal{G}_\lambda$ *is the* self-correcting self-consuming weight update mapping.

Definition (Self-Correcting Self-Consuming Weight Update). *The* idealized correction of strength $\gamma \geq 0$ *of distribution* p_θ *is the following distribution:*

$$\pi_\gamma p_\theta(x) := \frac{p_\theta(x) + \gamma p_{\theta^\star}(x)}{1+\gamma},$$

where p_{θ^*} *denotes the optimal model attainable within the model class. The* weight update mapping *with augmentation percentage* $\lambda \geq 0$ *and correction strength* $\gamma \geq 0$ *is:*

$$\pi_\gamma \mathcal{G}_\lambda(\theta) := \operatorname*{local\,argmax}_{\theta' \in \Theta} \left[\mathbb{E}_{x\sim\hat{p}_{\mathrm{data}}}[\log p_{\theta'}(x)] + \lambda \mathbb{E}_{x\sim\widehat{\pi_\gamma p_\theta}}[\log p_{\theta'}(x)]\right],$$

where $\hat{p}_{\mathrm{data}}$ *and* $\widehat{\pi_\gamma p_\theta}$ *are empirical distributions of size* n *and* $\lfloor \lambda \cdot n \rfloor$, *resp.*

Theorem (Stability of Iterative Fine-Tuning with Correction). *Suppose that we have a* sufficiently nice *self-consuming loop weight update procedure* $\theta_0 \to \theta_1 \to \theta_2 \to \ldots$ *with:*

1. *Synthetic augmentation percentage* $\lambda \geq 0$, *correction strength* $\gamma \geq 0$,
2. λ *and* γ *both satisfying* $\lambda \cdot C < \frac{1+\gamma}{2+\gamma}$.

Then with high likelihood, *the self-consuming loop with self-correction strength* γ *satisfies the following stability estimate for all* $t > 0$:

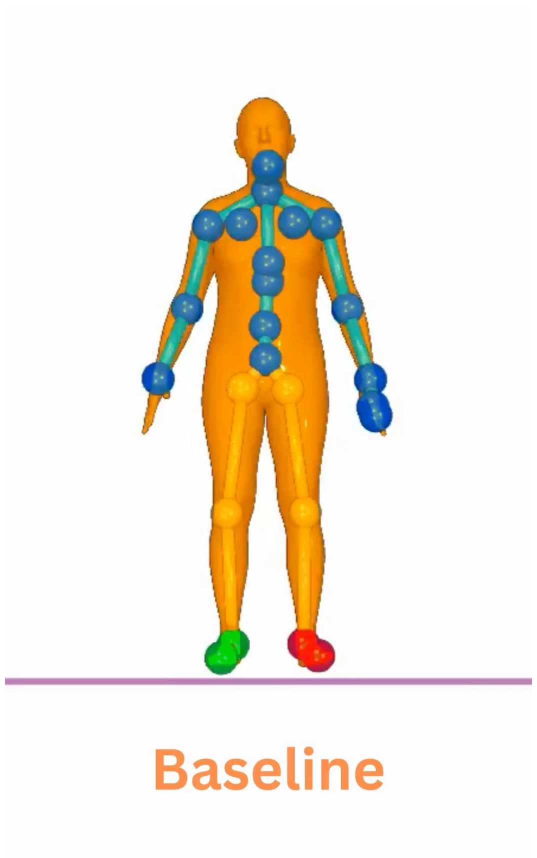
$$\|\theta_t - \theta^\star\| \leq c \cdot \sum_{i=0}^{t} \left(\frac{\rho}{1+\gamma}\right)^i + \left(\frac{\rho}{1+\gamma}\right)^t \|\theta_0 - \theta^\star\|.$$

Why self-correct? Our theoretical results…

Corollary 4.5. *Under the assumptions from Theorem 4.3, iterative fine-tuning with any amount of correction outperforms iterative fine-tuning without correction–in the sense that it is exponentially more stable, and it results in better model weights.*

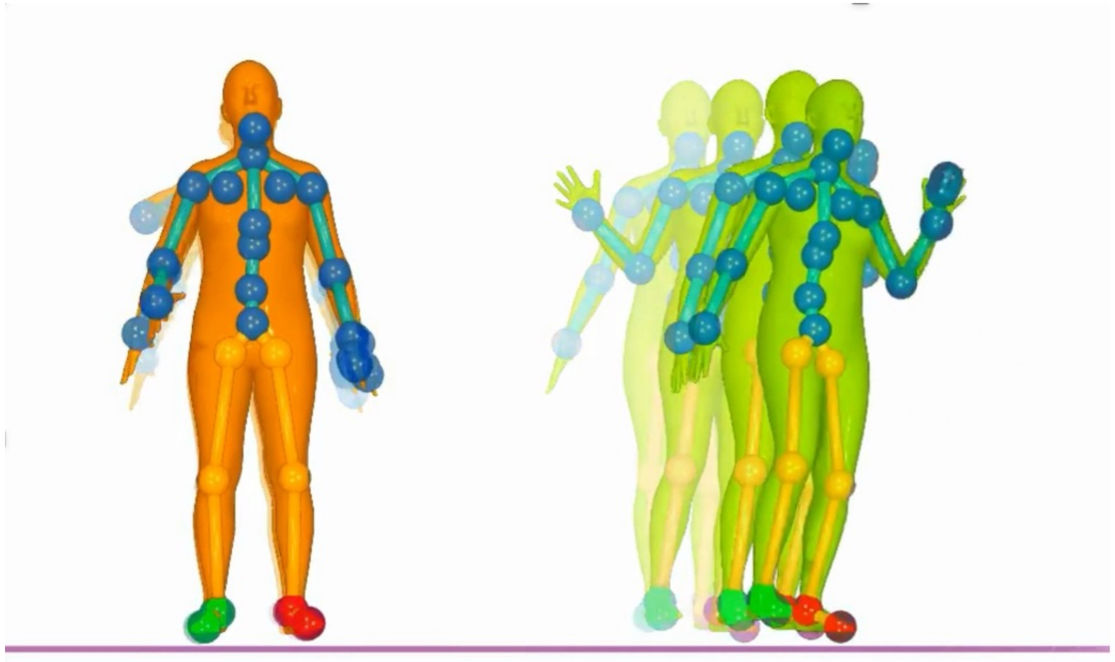
Conjecture 4.7. *In the case of iterative fine-tuning with correction, we may relax how close the initial model parameters* θ_0^n *need to be to the optimal model parameters* $\theta^\star$*, as well as choose a larger synthetic augmentation percentage* λ*, while still retaining the improved stability estimate* (7).

We start from human motion synthesis



prompt: ***"a person raises right hand to face looks around and puts hand down back to side"***

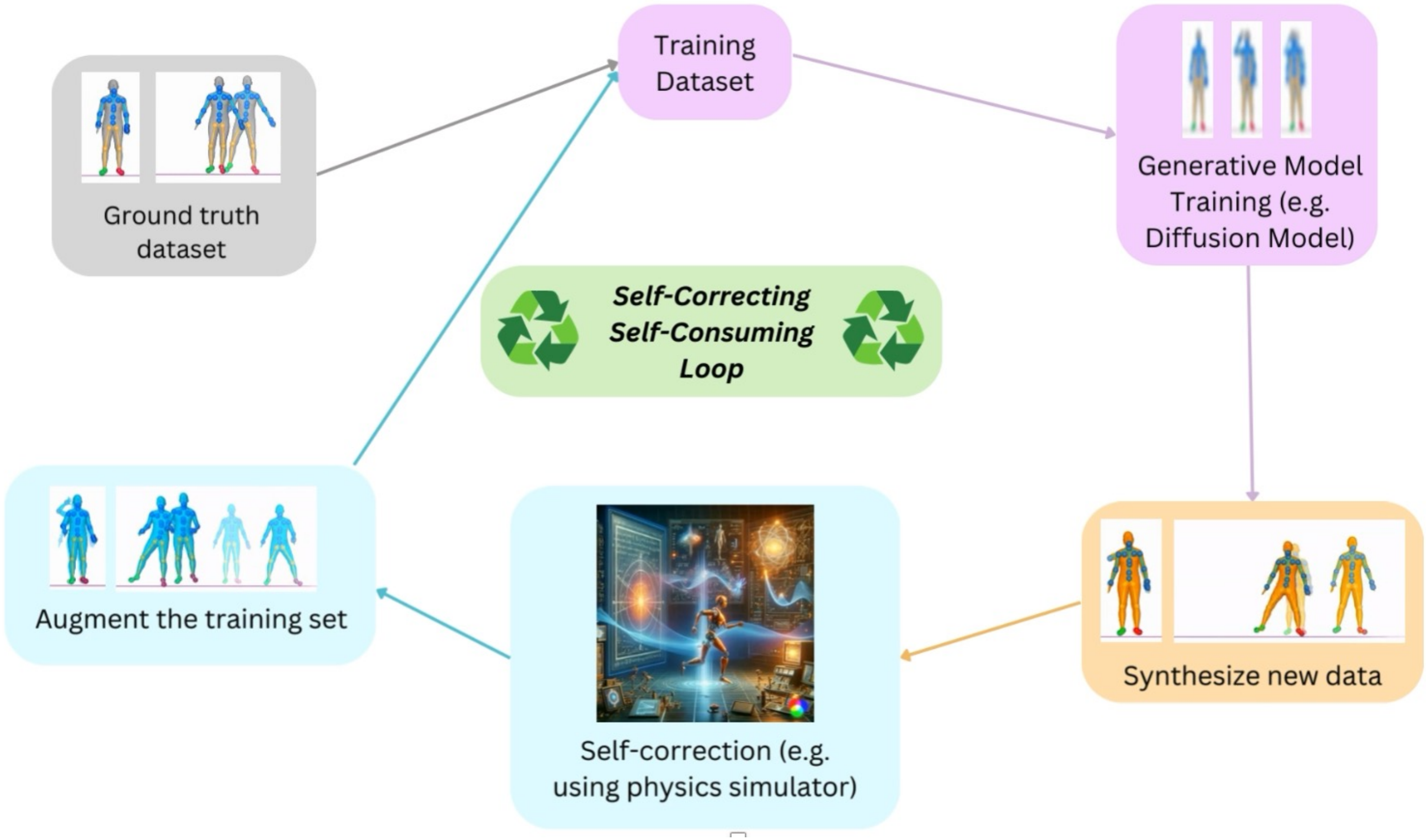
We start from human motion synthesis



prompt: *“a person raises right hand to face looks around and puts hand down back to side”*

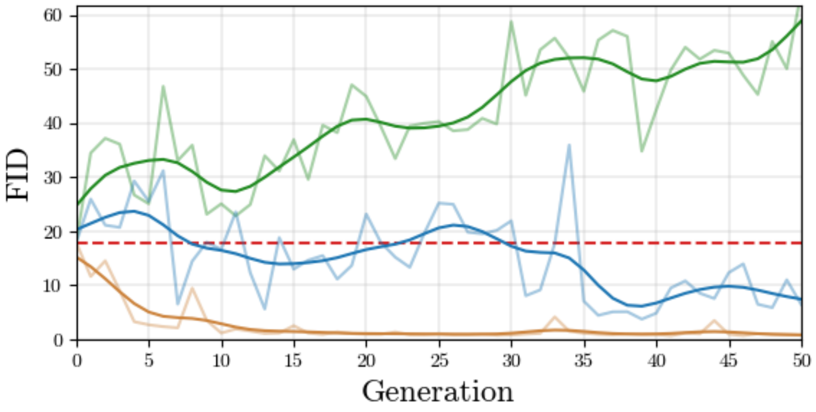
BROWN

How? fixing motions using a *physics simulator* in "self-correcting self-consuming loop"



Self-correcting self-consuming does not go MAD

Correction function:
Tweak the generated joints so they can be executed by a simulated human

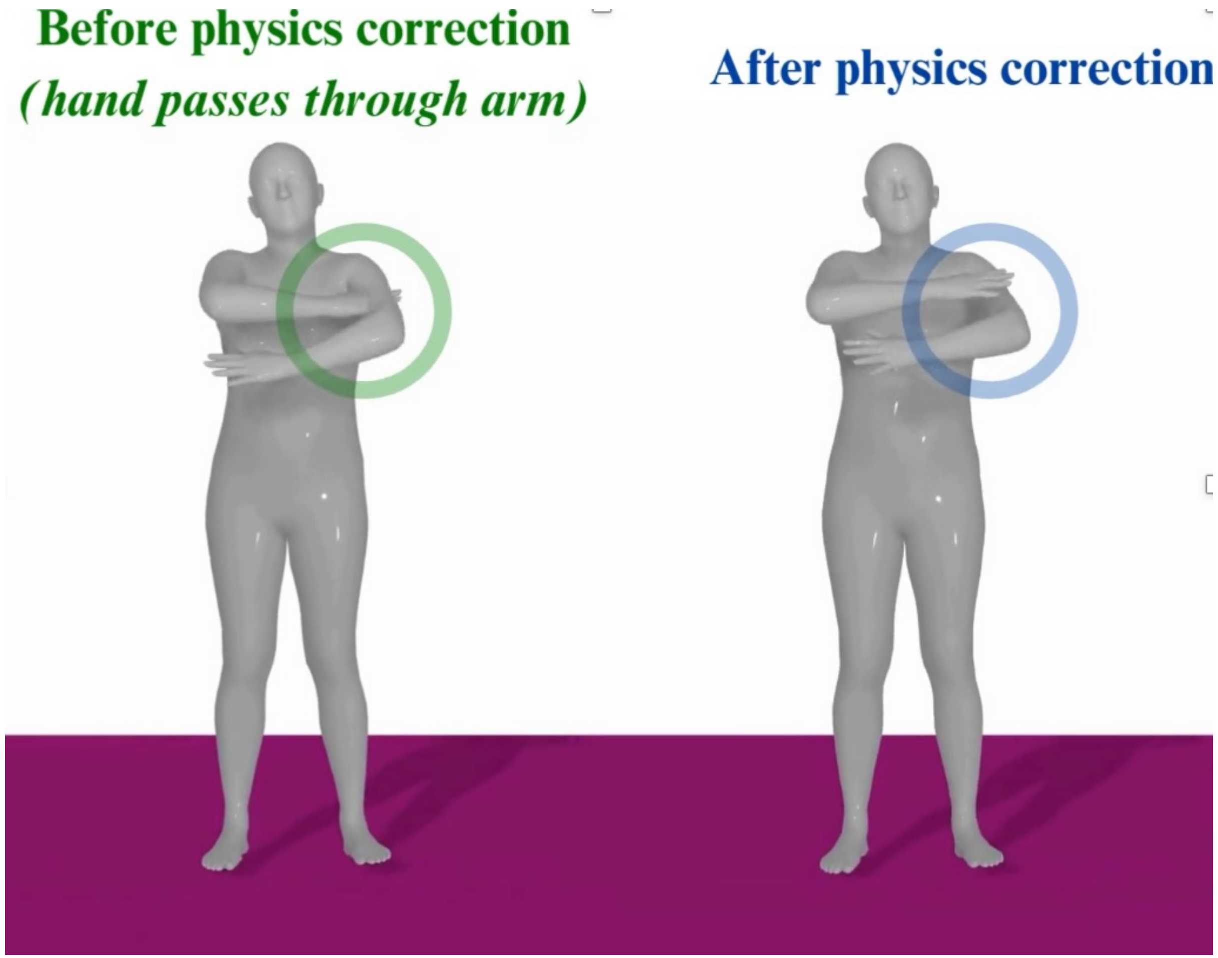


Zooming into the physics correction function

- Frozen, pretrained policy $\pi(a_t; s_t, \hat{q}_{t+1})$
- Goal: imitate the generated motion sequence $\hat{q}_{1:T}$
- Transition dynamics given by physics simulator

Universal Humanoid Control, Luo et al (2021)

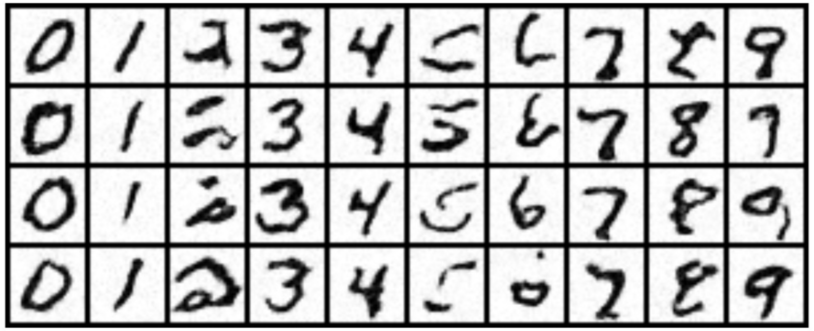
Zooming into the physics correction function



BROWN

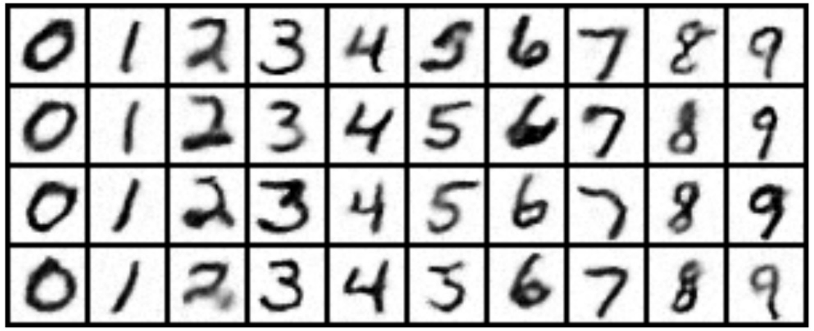
Follow-up questions

1. *Self-consuming generative modeling*



Iterative Fine-tuning

How to stabilize self-consuming generative models for other modalities (e.g. text, image)?



Iterative Fine-tuning with Self-Correction

Follow-up questions

2. "Vanilla" generative modeling

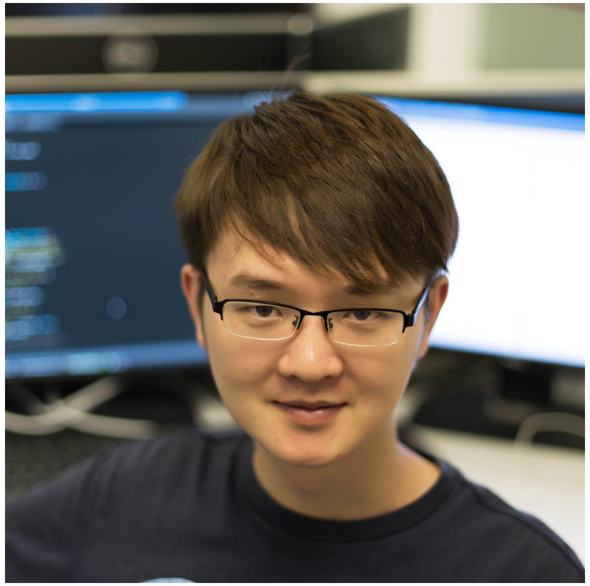
How to make general-purpose video generative models better understand physics?

OpenAI's Sora

https://www.youtube.com/watch?v=lfbImB0_rKY&ab_channel=Newsshooter

BROWN

My awesome collaborators + advisors



BROWN