MITIGATING MODEL COLLAPSE IN LATENT DIFFUSION MODELS USING A LEWIS SIGNALING REGULARISER

Anonymous authors
Paper under double-blind review

ABSTRACT

Model Collapse, a degenerative process in which models forget the underlying data distribution when trained on sequences of synthetic data, is a looming problem as generated content proliferates on the internet. This stands to not only undermine the general utility of the models but also introduces mechanisms that lead to systemic bias. In this work, we aim to mitigate the collapse by introducing a novel regularisation method that frames latent diffusion as a Lewis Signaling Game. In this framing, the diffusion model is a "sender" that transforms a text embedding, the "state", into an image which acts as the "signal". This signal must then be processed by a "receiver", a pretrained CLIP vision encoder whose goal is to retrieve the original text embedding. This provides a communication pressure which forces the diffusion model to remain sufficiently expressive. Our results show a clear improvement and mitigation of model collapse over multiple generations. Further, we find that the specificity of text prompts significantly impacts our method's utility. Characteristics of the image described in the text tend toward a minimal set of stereotypical motifs. As such, precise, detailed prompts help encourage generations in the tail of the distribution, helping to prevent collapse along those axes. Finally, we show how this method takes a step towards avoiding bias in latent diffusion, such as the lightening of skin tones. Overall, this method supports the notion that model collapse results from ungrounded content generation and demonstrates that explicitly modeling a receiver or verification step in the training pipeline helps prevent collapse.

1 INTRODUCTION

The advent of modern large language models has revolutionized many industries, however the ability of generative models to generate the high quality and diverse output that they do hinges on access to huge quantities of data, with the capabilities of a model scaling with the amount of data it is trained on (Kaplan et al., 2020). Until now this data has been human generated, however the massive rise in popularity of these models has led to large portions of the internet being overtaken by generated text and images. Dolezal et al. (2026) shows that by the mid 2025 over 35% of new internet uploads were AI generated. This leads to semantic contraction, with information on the internet shifting towards the means of the distributions that the models are trained off of, particularly presenting as a collapse towards western culture, ideas and images (Agarwal et al., 2025).

The shift from training data being *human generated* to *model generated* leads to a kind of model "autophagy" (Xing et al., 2024), where the new generations of models will train off of the data generated by the older generations. This will cause newer models to forget information that previous generations deemed unimportant, i.e. the tails of the distribution, due to a finite ability to sample off of their data (Shumailov et al., 2023). This leads to a phenomenon known as model collapse which manifests itself as a progressive loss of both diversity and quality of the generated data (Alemohammad et al., 2023; Schaeffer et al., 2025) and propagates systemic bias in machine learning (Wang et al., 2025; Wilson et al., 2025; Jarvis et al., 2026a). This ailment is not unique to language models and has been shown to occur in a range of generative models (Alemohammad et al., 2023; Shumailov et al., 2024; Dohmatob et al., 2024; Wilson et al., 2025). This degradation of generated data is caused by a loss of diversity in the training data but more importantly an entanglement of the

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

modes of the distribution (Shumailov et al., 2023) meaning the models lose their ability to express diversity and concepts are confused within their latent spaces.

This process is reminiscent of the development of human language, with successive generations learning from language "generated" by their ancestors (Smith, 2009; Kirby et al., 2014; Jarvis et al., 2026b). Despite this, human language has maintained a rich ability to generate novel ideas and thoughts that are fully understandable (Kirby et al., 2015). One of the underlying mechanisms thought to drive this ability is cultural transmission (Saldana et al., 2019), which is the pressure to communicate with other members of the group while remaining grounded by the needs of the environment. This pressure causes language to maintain compositional structures that allow for diverse and novel concepts to be communicated (Saldana et al., 2019). If placed in a setting without sufficient need to remain expressive about a diverse set of stimuli in the environment, human language also collapses to become overly regular and repetitive (Kirby et al., 2008).

The role of language to convey sufficient information between members of a group such that a particular outcome is achieved by the group is often modeled computationally through a process known as a Lewis Signaling game (Lewis, 2008). In this theoretical framework two people communicate with each other to solve a problem that can be objectively measured. The game generally has a "sender" who observes a stimulus or "state" in the environment that must be conveyed as a "signal" to a "receiver". Upon receiving the signal the receiver must take an action or make a decision where it is mutually desirable that the correct decision is made. Both the sender and receiver are penalized if the message is incorrectly interpreted. When combined with language evolutions in a setting with sufficiently many states and actions, a form of linguistic collapse is once again observed where multiple states map to a shared signal (Huttegger, 2007; Pawlowitsch, 2008). This is known as a "pooling equilibrium" and, crucially, the degree of pooling is dependent on the difficulty of the task and the precision needed by the receiver.

Motivated by the observation that the need for a receiver to decode language grounds cultural transmission to avoid collapse across multiple generations we model the training of a latent diffusion model (Rombach et al., 2022) as a Lewis Signaling game. We do this by introducing a frozen pretrained CLIP (Radford et al., 2021) model into the training pipeline of the diffusion model, the diffusion model is then tasked with minimizing the distance between the CLIP embeddings of the seed text used to generate the image and the image generated, i.e. the model must relay the meaning of the text to the CLIP vision encoder. Structurally, this forms a similar role to cycle consistency which has long been used to regularise to learning (Dwibedi et al., 2019) and in image generation models (Xu et al., 2023; Kwon & Park, 2019). Specifically, we make three primary contributions:

- We show that a CLIP model playing the role of a receiver to a generated image is able to act as a regulariser that mitigates model collapse.
- We show that we are able to control the collapse by emphasizing different traits of an image in the prompt, recovering the insight that pooled equilibria are determined by the difficulty of the Lewis Signaling Game in a machine learning setting.
- By leveraging the directed collapse we can control the diffusion away from bias and fairness concerns such lightening of skin tones (Wilson et al., 2025).

2 Background

2.1 Model Collapse

Model collapse can be defined as a degenerative process where models forget the true underlying data distribution as a result of the inclusion of synthetic data in the training dataset (Shumailov et al., 2023; Schaeffer et al., 2025). The mechanism of generated data being introduced into the training pipeline can take multiple forms and is typically referred to as "poisoning". The most severe mechanism is the *fully synthetic* case, in which each generation receives an entirely new dataset generated using the model from the previous generation. This is the technique used by Alemohammad et al. (2023); Shumailov et al. (2023) when studying model collapse. Poisoning can also occur where generated samples replace real samples (*augmentation*) or are added (*accumulation*) as a percentage λ of the dataset size (Gerstgrasser et al., 2024). The remaining data is a partial copy of the dataset used by the previous generation (Alemohammad et al., 2023; Bohacek & Farid, 2023; Bertrand et al.,

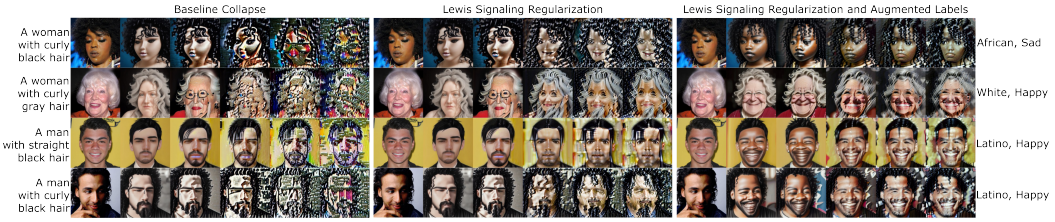

Figure 1: **Effectiveness of Lewis Signaling regularisation** The first image shows the baseline collapse with each column being a new generation. The texts on the left of the image are the labels used as prompts to the diffusion. In the middle image we see the result of our regularisation on the same prompts. In the final image we see the results of our regularisation combined with more detailed prompts, the prompts now include the attributes listed to the right. Note how in the final image we are able to better retain race and emotion compared to the normal regularisation.

2023) and the new sampled data. In some cases, a proportion of unseen real data is reintroduced each generation (Alemohammad et al., 2023). In this work we use the augmentation strategy where 90% of the dataset is replaced with new synthetic samples at each generation. We do not reintroduce real data samples. The effect is demonstrated in Figure 1 for the CelebA-HQ dataset.

Model collapse of image generation models has been measured using two primary metrics. Firstly, the distance between the real data and generated data can be measured using the Fréchet inception distance (FID) (Heusel et al., 2017) which will grow as the model collapses (Alemohammad et al., 2023). Secondly, Bohacek & Farid (2023) use the CLIP Score to measure the distance between the text and the generated image in the CLIP embedding space (Radford et al., 2021). As the models collapse this distance grows, reflecting the loss of semantic content in the generated images. Thus, the goal of mitigating model collapse aims to decrease the FID Score and increase the CLIP Score.

Shumailov et al. (2023) characterizes model collapse into two phases, *early collapse* - the loss of the tails of the distribution causing a decrease in diversity, and *late collapse* - the entanglement of the different modes causing an inability to differentiate across classes. This is attributed to two error sources that compound over time. The first being a *statistical approximation error*, since a finite number of samples are used for each generation and there is a chance that information is omitted in the samples for a generation. This is typically termed the transmission bottleneck from the linguistic evolution perspective (Kirby et al., 2008; 2014). Secondly there is a *functional approximation error* where the neural network itself fails to completely learn the data distribution. This is known as the learning bottleneck (Kirby et al., 2008; Jarvis et al., 2026b). Mitigation of the model collapse requires solving for both sources of error. Gerstgrasser et al. (2024) proposes accumulating data, rather than augmenting or replacing, and show that this stops model collapse from occurring. This however assumes infinite compute as you would need to train on all data and not just some subset (Kazdan et al., 2024). This stands to worsen the learning bottleneck and can become computationally infeasible. Data pruning and curation has been proposed as a solution (Shumailov et al., 2024; Schaeffer et al., 2025) and recent progress has been made on automating this process (Liu et al., 2025; Thudi et al., 2025) to ensure the data remains diverse and anomalies are filtered out. However, detecting synthetic samples is becoming increasingly difficult (Boutadjine et al., 2025) and curation remains a source of potential bias injection by the curator (Jo & Gebru, 2020; Birhane & Prabhu, 2021; Bender et al., 2021).

2.2 CLIP and Latent Diffusion Models

Contrastive Language-Image Pretraining (CLIP) (Radford et al., 2021) is a technique for learning a joint embedding space for images and texts. This is accomplished by jointly training a text encoder, f_{txt}, and an image encoder, f_{img}, on large scale (image, caption) datasets. During training, images and their captions are embedded by their respective encoders with the objective of embedding matched image-caption pairs close to each other and unmatched pairs far from each other, as measured by the cosine similarity in Equation 1:

$$\text{sim}(I, T) = \frac{f_{\text{img}}(I) \cdot f_{\text{txt}}(T)}{\|f_{\text{img}}(I)\| \, \|f_{\text{txt}}(T)\|} \tag{1}$$

Once training has been completed, the encoder weights are frozen and the encoders can be used to measure alignment between new unseen image-text pairs using the cosine similarity between their embeddings. This measuring of this alignment can be seen in Figure 2 in the CLIP section. This allows for effective zero shot performance on bimodal tasks that can be rephrased as the similarity between text and an image.

Diffusion Model (DMs) (Sohl-Dickstein et al., 2015) learn to reverse a fixed noising process on images, this allows them to start with noise and generate high quality images. Latent Diffusion Models (LDMs) (Rombach et al., 2022), a variant of DMs, operate on a compressed latent space rather than directly on pixels. In this setup a pretrained Variational Autoencoder (VAE), consisting of an encoder that maps an image into a low dimensional latent z and a decoder that maps z back to the image, is used. Diffusion then occurs on z rather than on the pixels of the image itself which is significantly cheaper. During training Gaussian noise ϵ is added to z_0, t times to produce z_t. A U-Net is then trained to reverse the noising process by predicting the noise added at each timestep (ϵ_θ) such that the previous latent embedding ($\hat{z}_{t-1}$) could be produced by subtracting the predicted noise (ϵ_θ) from the current latent embedding (z_t). The diffusion loss is then the expected squared error between the true and predicted noise shown in Equation 2:

$$\mathcal{L}_{\text{diff}} = \mathbb{E}_{z_0,\epsilon,t}\left[\|\epsilon - \epsilon_\theta(z_t, t)\|_2^2\right]. \tag{2}$$

Repeated denoising will then allow the network to approximate a real image embedding (z_0) even if the original embedding is sampled directly from a Gaussian distribution rather than obtained from a noising process. This allows for new samples to be generated at inference time where a latent embedding ($z_t \sim \mathcal{N}(0, I)$) is drawn and iteratively denoised. The approximation of a real image embedding ($\hat{z}_0$) that this produces is mapped using the decoder from the VAE to an image. During the denoising process the U-Net is additionally conditioned on a text caption by letting the network attend to a CLIP embedding of the caption (Ho & Salimans, 2022). This conditioning serves to guide the model towards images which are consistent with the caption. See Figure 2 for an overview of the process.

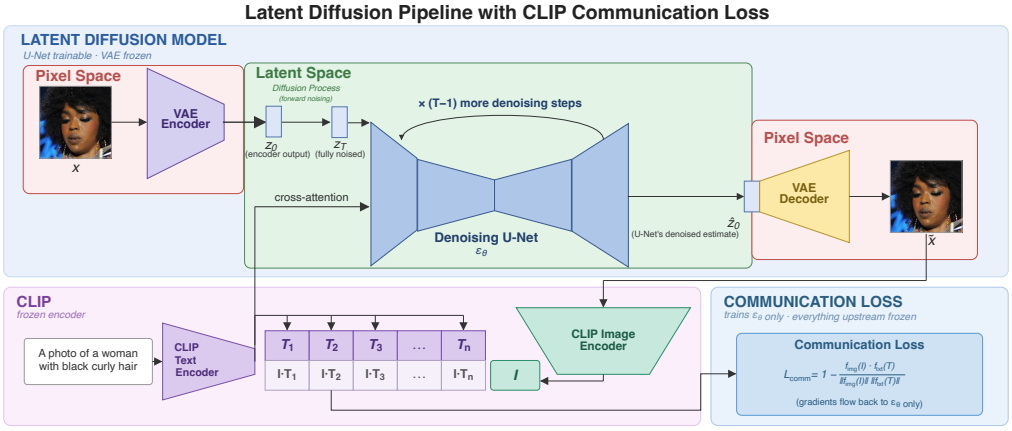

Figure 2: **Architecture of Communication Loss**: During training the VAE Encoder encodes an image into a latent z_0, a forward pass then occurs that adds noise to the latent, the U-Net is then trained with the objective of denoising the latent to reproduce z_0, this process is conditioned on a CLIP embedding of the image label. The latent produced by the denoising process, $\hat{z}_0$ is then decoded by the VAE decoder. This decoded image is then encoded by a CLIP Image Encoder and the communication loss calculated against the original text embedding.

3 Methodology

3.1 Datasets and Recursive Finetuning

We follow the methodology outlined by Bohacek & Farid (2023) when studying model collapse on two datasets, CelebA-HQ (Karras et al., 2018) and COCO general scenes (Lin et al., 2015). Similar

to Bohacek & Farid (2023) we subsample the datasets to create training sets that are near balanced across different categories of the data. We used the COCO 2017 train split of 118k captioned images and sourced 900 and 2000 random images from it to use as training data with a separate 200 images being set aside for evaluation. The labels used for these images were their human-written captions from the original dataset. We outline below how we sampled images from the CelebA-HQ dataset and generated three corresponding image captions.

Once the original training images were selected, they were used as input to the image-to-image pipeline of the stable diffusion (SD) model[1] to generate new, similar images while also being guided by the image caption. This corresponds to the Generation 0 model being used to sample data which seeds the first generation of recursive training as outlined by Bohacek & Farid (2023). From the original dataset 10% of the images are then set aside to serve as the "real" component of the next generation dataset and the remaining 90% are passed through the image-to-image pipeline of our generation 0 SD model to produce synthetic data for the next generation dataset. This data is then used to finetune the denoising U-Net component of the SD model which becomes our generation 1 model and is sampled to obtain synthetic data for generation 2. This process is repeated for 5 generations where 90% of the dataset is replaced with new synthetic samples each generation.

3.1.1 CelebA-HQ Dataset Curation

Taking the CelebA-HQ dataset of 30,000 1024x1024 images as well as their corresponding attributes from the original CelebA dataset (Liu et al., 2015), we combine the *gender* (male and female), *hair type* (straight and wavy) and *hair colour* (black, blond, brown, gray) attributes to create 16 categories. We aim to form a balanced training dataset across these categories and sample 50 images from each. If a category does not have 50 images then we just keep all of them. This produced a total of 712 images for training and Table 1 shows the number of samples from the categories that did not have 50 images. A further 187 images that nearly evenly distributed among the categories were set aside for evaluation of the model.

Table 1: The four categories from CelebA-HQ that did not reach the 50-image target

Gender	Hair type	Hair colour	No. training
Woman	Straight	Grey	0
Man	Wavy	Blond	21
Man	Wavy	Grey	45
Man	Straight	Blond	46

Three variants of labels for this subset of images were then created. The first we refer to as *Short Structured Labels* which are of the form "a photo of a *[gender]* with *[hair colour] [hair type]* hair". This gave us labels such as *"A photo of a woman with brown curly hair"*. The second type is *Unstructured Labels* where we utilized Claude Sonnet 4.6 (Anthropic, 2026) to create more natural variations of these labels using the prompt outlined in Appendix A. This produced labels such as *"A female with undulating tawny locks"*. These natural variants of the labels were then randomly assigned to images from each category such that each image had a unique label. Finally, we augmented the label space with more specific attributes using pretrained classifiers[2]. Specifically we add *age* (young, middle aged, senior), *race* (East Asian, Indian, Black, White, Middle Eastern, Latino, Southeast Asian) and *emotion* (angry, fearful, smiling, neutral, sad, surprised) labels. The distribution of these augmented labels on our image dataset is outlined in Table 2.

3.2 Lewis Signaling regularisation

In order to introduce a communication pressure on the stable diffusion model we introduce a Communication loss to the diffusion pipeline. This loss rewards image to text alignment. During training for each sample the model estimates the clean latent from the model's prediction using the DDPM inversion (Ho et al., 2020) outlined in Equation 3:

$$\hat{x}_0 = \frac{x_t - \sqrt{1-\bar{\alpha}_t}\,\hat{\epsilon}_\theta(x_t, t)}{\sqrt{\bar{\alpha}_t}} \tag{3}$$

[1]Experiments are run using the PyTorch Library (Paszke et al., 2019) with the *segmind/tiny-sd* implementation of stable diffusion (Kim et al., 2024) from the Hugging Face Library (Wolf et al., 2020).

[2]All classifiers were ViT models (Dosovitskiy, 2020) obtained from Hugging Face (Wolf et al., 2020) For age we use *nateraw/vit-age-classifier*, for race we use *NikhilJaddu/fairface-race-vit* which is pretrained on the FairFace Dataset (Karkkainen & Joo, 2021) and for expressions we use *trpakov/vit-face-expression*.

Table 2: Distribution of augmented labels over training dataset

Race		Emotion		Age	
Category	%	Category	%	Category	%
White	78.23%	Happy	55.62%	Young	61.10%
Latino	11.38%	Neutral	34.13%	Adult	14.61%
East Asian	4.49%	Angry	6.46%	Middle-aged	13.90%
Middle Eastern	2.67%	Sad	2.11%	Senior	10.39%
Indian	1.83%	Fear	0.98%		
Black	1.26%	Surprise	0.70%		
Southeast Asian	0.14%				

As outlined in Figure 2, this denoised latent is then decoded to an image using the frozen VAE decoder which is then passed through the CLIP image encoder. At the same time the sample's label is passed through the text encoder component of the CLIP model. The communication loss is then calculated as $1 - \mathrm{sim}(I, T)$ using the cosine similarity of Equation 1:

$$\mathcal{L}_{\text{comm}} = \mathbb{E}\left[1 - \mathrm{sim}(I, T)\right]. \tag{4}$$

The diffusion loss from Equation 2 is then updated with this Lewis Signaling regularisation term for a weighting $w \in [0, 1]$. In our experiments this weighting is varied between 10%, 50% and 85%:

$$\mathcal{L}_{\text{LS Reg}} = (1 - w)\mathcal{L}_{\text{diff}} + w\mathcal{L}_{\text{comm}} \tag{5}$$

We investigated applying this Lewis Signaling regularised diffusion loss throughout training as well as waiting for generation 3 once the collapse has visibly and measurably started. Additionally we varied the denoising timesteps that this loss was applied to, testing timesteps in ranges $[0, 50]$, $[0, 350]$ and $[0, \infty]$.

3.2.1 Evaluation

The model is evaluated using the Frechet Inception Distance (FID) (Heusel et al., 2017) which quantifies the realism of the generated images. It does this by comparing them to a set of real images, which in our case are sampled once from the real images dataset at the beginning of training. A smaller FID means the generated images are more similar to the real images. As the model collapses it is expected that this number increases, which is confirmed in our quantitative results shown in Figure 3. We also evaluate the model on CLIP Score (Hessel et al., 2022), which uses a CLIP model to evaluate whether the generated image is semantically similar to the text prompt by calculating the cosine similarity between the image and the text. The higher the CLIP Score the more similar the image and text are. Finally, we also perform qualitative evaluations by comparing different interventions and hyper-parameters on consistent images.

4 Results

4.1 Performance on CelebA-HQ

The process described in Section 3.1 on poisoning demonstrably shows collapse as the generations proceed as can be seen in Figure 1 and in Figure 6. This collapse presents as images that are progressively noisier and features that are less realistic. The addition of Lewis Signaling regularisation with small structured labels of the form "a photo of a *[gender]* with *[hair colour] [hair type]* hair" is enough to pull the model away from the collapse and direct the model to create less noisy images with more realistic features as shown in Figure 4. Quantitatively, Figure 3 shows the benefit of the Lewis Signaling regulariser as its inclusion yields improved FID and CLIP Scores over the baseline (no regulariser). Crucially, increasing the weighting of the regulariser in the loss results in improved performance as well. We also report the CLIP Score on CelebA-HQ in Figure 3 for completeness. However, we note that the regulariser is explicitly optimizing for this metric and so it is unsurprising that our method improves on this metric. Thus, the stronger the influence of the regulariser the better the model resists collapse and our results show that the regulariser is able to keep the generated images closer to the real data distribution.

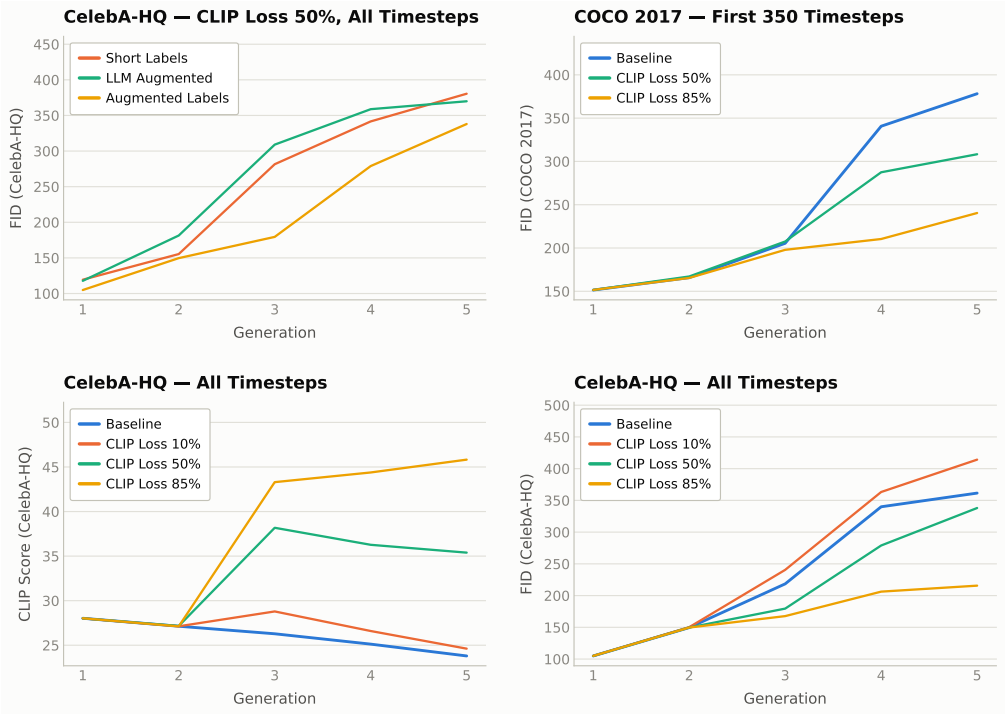

Figure 3: **Top Left**: Changing of FID Score across generations compared across label generation scheme showing Augmented labels to best stabilize FID Score. **Top Right**: Changing of FID Score across generations on the COCO 2017 dataset with the communication loss applied to all timesteps with $t \in [0, 350]$. We compare this across different weightings of communication loss showing that the more influential the loss, the more stable the FID remains. This shows that the behavior is consistent on both faces and natural images. **Bottom**: Changing of both CLIP and FID Score across generations. We compare this across different weightings of communication loss showing that the behavior on CelebA-HQ with generated labels is consistent with the COCO 2017 dataset.

4.2 The Role of Prompt Detail and Bias Mitigation

The strongest effect we observe in our results is that the quality of the prompt plays a significant role in the benefit of Lewis Signaling regularisation. While predictable, the way this appears is instructive. Firstly, augmenting the labels to include more details mitigates model collapses more. This is reflected in the quantitative and qualitative results of Figures 3 and 4 respectively. However, simply making the label more varied moderately mitigates the collapse but can become under-specified. In these cases the model collapses towards caricature like images and uses a "play-on-words" in some cases to minimize the loss. This is demonstrated in Figure 4 where the LLM Augmented labels resulted in highly stylized samples. An extreme example of the satirical emphasis of characteristics described in the prompt is shown in Figure 10 where a prompt which described the subjects hair as "hazelnut" resulted in actual hazelnut strands appearing over the image.

Most strikingly, augmentation of the structured prompts with race, age and emotion was highly effective at allowing the model to better retain these characteristics through the collapse. This is despite the bias in the data as Table 2 clearly shows that the training data is heavily skewed towards people of white (78.23%) descent who are happy (55.62%). When analyzing Figure 4 we can clearly see how this bias influences the short labels during the collapse, with the skin lightening and the smile increasing as the generations progress. The addition of a more detailed prompt that highlights desired characteristics allows the model to more accurately preserve these characteristics. Additionally the stronger the influence of our regulariser the better the model preserves the characteristics, see Figure 7. However the detailed labels can lead to the model overcompensating for common attributes. For example, in Figure 1 there is an over-emphasis of the smile on happy faces. Thus, while Lewis Signaling regularisation has a clear benefit and can mitigate model col-

lapse, some properties of the data are still lost if they are not included in the captions. For example, in many samples we see a rotation of the face towards the camera as this is the most typical angle of photographs. Thus, the property of subject angle is lost. Similarly in the bottom row of Figure 1, the man's hand is placed on his chin and this property is removed within a single generation. This effect is consistent both with the literature on language evolutions (Kirby et al., 2008; 2014) and Lewis Signaling Games with the notion of pooling equilibria (Huttegger, 2007; Pawlowitsch, 2008) and we discuss this more in Section 5. However, it is important to note that collapse is mitigated only as much as the Lewis Signaling Game needs.

4.3 Performance on Natural Scenes in the COCO Dataset

The benefit obtained by increasing the influence of the regulariser is consistent across datasets and we see a clear improvement on the FID Score on the COCO 2017 dataset with the inclusion of Lewis Signaling regularisation. We also find that it is most important to apply the regulariser early during sampling. In the case of COCO 2017 it was sufficient to apply the regulariser for the first 350 inference steps out of 1000 to obtain clear improvements on FID. Thus, the added time complexity introduce by the regulariser can be mitigated significantly by only using it for the early phase of inference. We demonstrate this in Appendix B.2 in Figure 9 which shows a diminishing improvement in FID Score when exposing more timesteps to the regulariser. However, this can result in some loss of detail which is only introduced later during sampling, as shown in Figure 4. In Appendix B.2 we show a similar result for the CelebA-HQ dataset. Thus, it is important to apply the regularisation for long enough to guide the diffusion and with a sufficiently large regularisation rate, however for later stages in sampling the regularisation can be removed to allow for quicker inference. This is also consistent with recent results showing that diffusion goes through distinct stages during sampling (Biroli et al., 2024). Note that for the COCO 2017 dataset we did not augment the labels at all and were able to show a positive influence from the Lewis Signaling regulariser using the human-provided captions. The utility of the method does seem to rely on it being introduced at the appropriate generation, once the early signs of collapse are observable. The application of the regulariser in the early generations still shows a clear benefit over not applying the regulariser. However, it can lead to an over-reliance on the prompt and distortions in the final sample. In our experimental setup the collapse appeared to start accelerating after generation 3 and so we introduced the regulariser at this point in the training run. This provided the best balance of mitigating model collapse, while not leading to an over-emphasis on the prompt (for example by adding more "boys in suits" than is necessary). Thus, our method mitigates model collapse for natural scenes and natural language captions, as shown qualitatively in Figure 5.

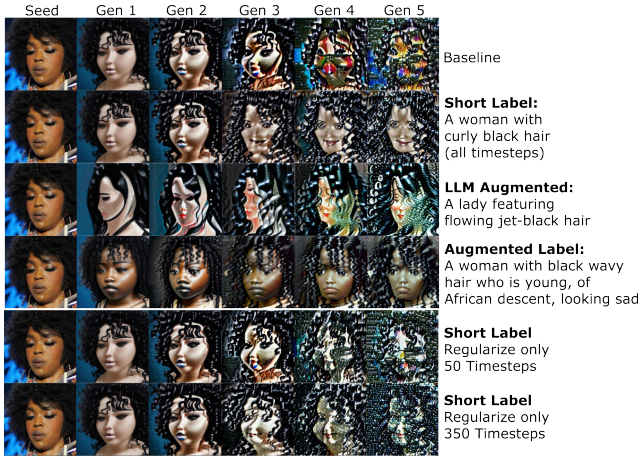

Figure 4: **The role of prompt detail and regularisation duration on qualitative samples:** Comparing various prompt strategies and regularisation durations for the image with base label *"A photo of a woman with curly black hair"* demonstrates that increasing exposure to the Lewis Signaling regularisation during the generational collapse shows a marked decrease in noise and feature distortion. However, a sufficient number of steps is needed since the duration of regularisation influences which features are forgotten. For example, regularising to timestep 350 maintains the structure of the image and some details, like the lips and cheeks, up to the last generation but does lose some details. Regularising for the full duration maintains the structure and specified details. Similarly, more detailed and clear prompts yield significantly better samples. This can be used to combat the typical injection of bias from latent diffusion, such as the lightening of skin tone (Wilson et al., 2025). We observe this lightening with the Short Label where race is not mentioned. By including it in the Augmented Label the race is maintained.

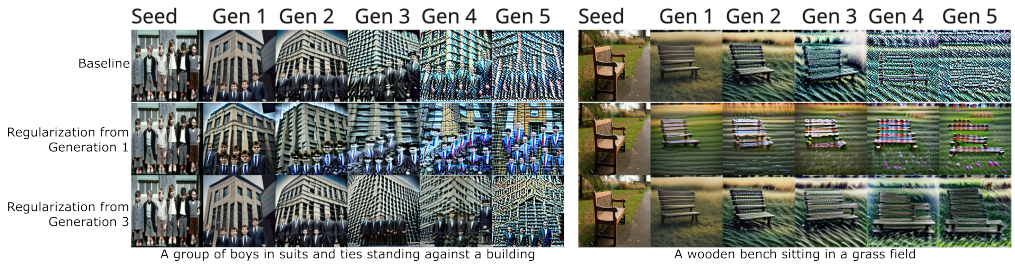

Figure 5: Method applied to natural images and labels from the COCO dataset. Sensitivity to the time of introduction of the regulariser

5 Discussion & Societal Impact

In our experimental setup we poisoned 90% of the dataset at each generation to speed up the collapse and make the experimentation tractable. This is admittedly a worst-case scenario with collapse likely happening far slower in reality (Schaeffer et al., 2025). In addition our experiments are run using the *segmind/tiny-sd* version of stable diffusion which is an extremely lightweight model and produces less realistic images than are usually produced from stable diffusion. This is necessary to obtain a severe degree of collapse within a reasonable time frame. Both design decisions for the experiments exacerbate model collapse and, as we aim to mitigate its effects, are justified to test our approach under the most difficult conditions. However, it is important to not overstate the impact of model collapse or its likely time frame on the internet, since precisely defining model collapse is difficult as noted by Schaeffer et al. (2025). Under this setup we were able to show a clear progressive collapse through the generations. The collapse was visible across all categories and displayed itself visually as an increase in noise, unnatural features and unnatural colours in the generated images. Additionally it was clear in the increasing FID Score and decreasing CLIP Score that the images were diverging from the real distribution. Thus, our experiments reproduce the key properties of model collapse provided by prior work (Shumailov et al., 2023; Dohmatob et al., 2024; Schaeffer et al., 2025) but it remains necessary to note that characterizing model collapse outside of controlled experiments remains open.

In this work, we reframe diffusion model training as a Lewis Signaling Game by introducing communication pressure through a pretrained CLIP model. The diffusion model acts as the "sender", conveying its text-conditioned state through generated images, while the CLIP image encoder acts as the "receiver". This pressure mitigated model collapse and reduced the emergence of systemic biases during retraining on generated content. We note that prompt specificity plays a critical role in determining which information is retained and that verification pressure can preserve model expressivity by requiring generated content to remain interpretable to a "receiver". Thus, our work contributes to the growing evidence that synthetic data is useful in settings where a secondary pressure exists to enforce expressivity. Cases include automated theorem solvers where synthetic problems are generated and yields reliable data to train a subsequent model on, as long as the problem generating model is grounded in the student's real world performance (Sundaram et al., 2026).

Future work using the Lewis Signaling regulariser can further ground the model studies of language evolution, similar to other recent approaches aiming to utilize iterated learning in a machine learning context (Vani et al., 2021; Zheng et al., 2024). For example, in simulated communities consisting of multiple agents compositionality emerges far quicker (Brighton et al., 2001) and in some case within a single generation (Raviv et al., 2019). Thus, moving beyond a single interlocutor (the one CLIP vision embedding model) towards a community of receiving agents will likely yield further improvements. It is also important to note that the instantiation of the Lewis Signaling Game in our regulariser forms a type of cycle consistency, where a CLIP text embedding guides the diffusion process of the sender and must subsequently be retrieved by the receiver. Yet, cycle consistency in various forms has been used successfully for a number of generative modeling frameworks (Dwibedi et al., 2019; Xu et al., 2023; Kwon & Park, 2019). Further, exploring cycle consistency as a means of grounding could also yield future improvements on our regulariser.

6 AI DECLARATION

In this work we made use of generative AI tools for coding assistance, to polish writing, retrieval of papers and generation of figures. We have not used AI for any other tasks. All AI assisted work was reviewed and manually tested. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

7 ETHICS STATEMENT

There are a number of important ethics concerns which should be noted. Model collapse stands to have a destructive effect on culture and representation on the internet and in human communities (Bohacek & Farid, 2023; Shumailov et al., 2024; Wilson et al., 2025; Schaeffer et al., 2025; Wang et al., 2025; Jarvis et al., 2026b). While nuance is needed when discussing the severity of these concerns (Schaeffer et al., 2025), it is important to note this effect on any work which aims to improve the performance of generative models and how to train them recursively. While we aim to mitigate the negative impacts of collapsing diffusion models (Bohacek & Farid, 2023; Wilson et al., 2025) and promote fairness in AI by providing a mechanism to enforce diversity in generated data samples, the under-specification results in Figures 4 and 10 show that our method can have unintended consequences if insufficient care is given to the prompts. Further, the clear effect of the regulariser and the fact that it can even over-rule the image prompt leaves room for misuse. Two perspectives mitigate this risk. Firstly, it is plausible that the under-specification effect will weaken with the introduction of multiple receivers. These are predominantly examples of the diffusion model leveraging the biases in the CLIP vision embeddings, such as a "happy person" having a very large smile. Thus, with more receivers and diverse prompts it should become more difficult to manipulate any one receiver. Secondly, model collapse as it currently stands appears arbitrary and not only linked to data imbalance and representation (Wang et al., 2025). Thus, a form of guided collapse, even if misused, introduces room for accountability and justice.

REFERENCES

Dhruv Agarwal, Mor Naaman, and Aditya Vashistha. Ai suggestions homogenize writing toward western styles and diminish cultural nuances. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*, CHI '25, New York, NY, USA, 2025. Association for Computing Machinery. ISBN 9798400713941. doi: 10.1145/3706598.3713564. URL https://doi.org/10.1145/3706598.3713564.

Sina Alemohammad, Josue Casco-Rodriguez, Lorenzo Luzi, Ahmed Imtiaz Humayun, Hossein Babaei, Daniel LeJeune, Ali Siahkoohi, and Richard Baraniuk. Self-consuming generative models go mad. In *The Twelfth International Conference on Learning Representations*, 2023.

Anthropic. Claude sonnet 4.6 system card, 2026. URL https://www-cdn.anthropic.com/bbd8ef16d70b7a1665f14f306ee88b53f686aa75/Claude%20Sonnet%204.6%20System%20Card.pdf.

Emily M Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency*, pp. 610–623, 2021.

Quentin Bertrand, Avishek Joey Bose, Alexandre Duplessis, Marco Jiralerspong, and Gauthier Gidel. On the stability of iterative retraining of generative models on their own data. *arXiv preprint arXiv:2310.00429*, 2023.

Abeba Birhane and Vinay Uday Prabhu. Large image datasets: A pyrrhic win for computer vision? In *2021 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pp. 1536–1546. IEEE, 2021.

Giulio Biroli, Tony Bonnaire, Valentin De Bortoli, and Marc Mézard. Dynamical regimes of diffusion models. *Nature Communications*, 15(1):9957, 2024.

Matyas Bohacek and Hany Farid. Nepotistically trained generative-ai models collapse. *arXiv e-prints*, pp. arXiv–2311, 2023.

Amal Boutadjine, Fouzi Harrag, and Khaled Shaalan. Human vs. machine: A comparative study on the detection of ai-generated content. *ACM Transactions on Asian and Low-Resource Language Information Processing*, 24(2):1–26, 2025.

Henry Brighton, Simon Kirby, and Kenny Smith. Situated cognition and the role of multi-agent models in explaining language structure. In *Symposium on Adaptive Agents and Multi-Agent Systems*, pp. 88–109. Springer, 2001.

Elvis Dohmatob, Yunzhen Feng, and Julia Kempe. Model collapse demystified: The case of regression. *Advances in Neural Information Processing Systems*, 37:46979–47013, 2024.

Jonas Dolezal, Sawood Alam, Mark Graham, and Maty Bohacek. The impact of ai-generated text on the internet. *arXiv preprint arXiv:2604.26965*, 2026.

Alexey Dosovitskiy. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

Debidatta Dwibedi, Yusuf Aytar, Jonathan Tompson, Pierre Sermanet, and Andrew Zisserman. Temporal cycle-consistency learning. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1801–1810. IEEE, 2019.

Matthias Gerstgrasser, Rylan Schaeffer, Apratim Dey, Rafael Rafailov, Henry Sleight, John Hughes, Tomasz Korbak, Rajashree Agrawal, Dhruv Pai, Andrey Gromov, et al. Is model collapse inevitable? breaking the curse of recursion by accumulating real and synthetic data. *arXiv preprint arXiv:2404.01413*, 2024.

Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. Clipscore: A reference-free evaluation metric for image captioning, 2022. URL https://arxiv.org/abs/2104.08718.

Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30, 2017.

Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*, 2022.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020.

Simon M Huttegger. Evolution and the explanation of meaning. *Philosophy of science*, 74(1):1–27, 2007.

Devon Jarvis, Richard Klein, Benjamin Rosman, Steven James, and Stefano Sarao Mannelli. Position: the stochastic parrot in the coal mine. model collapse is a threat to low-resource communities. In *Forty-third International Conference on Machine Learning Position Paper Track*, 2026a. URL https://openreview.net/forum?id=WlI0OsnkN2.

Devon Jarvis, Richard Klein, Benjamin Rosman, and Andrew M Saxe. Compositionality and systematicity emerge from iterated learning in deep linear networks. *Proceedings of the National Academy of Sciences*, 123(19):e2509739123, 2026b.

Eun Seo Jo and Timnit Gebru. Lessons from archives: Strategies for collecting sociocultural data in machine learning. In *Proceedings of the 2020 conference on fairness, accountability, and transparency*, pp. 306–316, 2020.

Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*, 2020.

Kimmo Karkkainen and Jungseock Joo. Fairface: Face attribute dataset for balanced race, gender, and age for bias measurement and mitigation. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pp. 1548–1558, 2021.

Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. Progressive growing of gans for improved quality, stability, and variation, 2018. URL https://arxiv.org/abs/1710.10196.

Joshua Kazdan, Rylan Schaeffer, Apratim Dey, Matthias Gerstgrasser, Rafael Rafailov, David L Donoho, and Sanmi Koyejo. Collapse or thrive? perils and promises of synthetic data in a self-generating world. *arXiv preprint arXiv:2410.16713*, 2024.

Bo-Kyeong Kim, Hyoung-Kyu Song, Thibault Castells, and Shinkook Choi. Bk-sdm: A lightweight, fast, and cheap version of stable diffusion, 2024. URL https://arxiv.org/abs/2305.15798.

Simon Kirby, Hannah Cornish, and Kenny Smith. Cumulative cultural evolution in the laboratory: An experimental approach to the origins of structure in human language. *Proceedings of the national academy of sciences*, 105(31):10681–10686, 2008.

Simon Kirby, Tom Griffiths, and Kenny Smith. Iterated learning and the evolution of language. *Current opinion in neurobiology*, 28:108–114, 2014.

Simon Kirby, Monica Tamariz, Hannah Cornish, and Kenny Smith. Compression and communication in the cultural evolution of linguistic structure. *Cognition*, 141:87–102, 2015. ISSN 0010-0277. doi: https://doi.org/10.1016/j.cognition.2015.03.016. URL https://www.sciencedirect.com/science/article/pii/S0010027715000815.

Yong-Hoon Kwon and Min-Gyu Park. Predicting future frames using retrospective cycle gan. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1811–1820. IEEE, 2019.

David Lewis. *Convention: A philosophical study*. John Wiley & Sons, 2008.

Tsung-Yi Lin, Michael Maire, Serge Belongie, Lubomir Bourdev, Ross Girshick, James Hays, Pietro Perona, Deva Ramanan, C. Lawrence Zitnick, and Piotr Dollár. Microsoft coco: Common objects in context, 2015. URL https://arxiv.org/abs/1405.0312.

Qian Liu, Xiaosen Zheng, Niklas Muennighoff, Guangtao Zeng, Longxu Dou, Tianyu Pang, Jing Jiang, and Min Lin. Regmix: Data mixture as regression for language model pre-training. In *International Conference on Learning Representations*, volume 2025, pp. 38305–38339, 2025.

Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild, 2015. URL https://arxiv.org/abs/1411.7766.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32, 2019.

Christina Pawlowitsch. Why evolution does not always lead to an optimal signaling system. *Games and Economic Behavior*, 63(1):203–226, 2008.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pp. 8748–8763. PmLR, 2021.

Limor Raviv, Antje Meyer, and Shiri Lev-Ari. Compositional structure can emerge without generational transmission. *Cognition*, 182:151–164, 2019.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10684–10695, 2022.

Carmen Saldana, Simon Kirby, Robert Truswell, and Kenny Smith. Compositional hierarchical structure evolves through cultural transmission: an experimental study. *Journal of Language Evolution*, 4(2):83–107, 2019.

Rylan Schaeffer, Joshua Kazdan, Alvan Caleb Arulandu, and Sanmi Koyejo. Position: Model collapse does not mean what you think. *arXiv preprint arXiv:2503.03150*, 2025.

Ilia Shumailov, Zakhar Shumaylov, Yiren Zhao, Yarin Gal, Nicolas Papernot, and Ross Anderson. The curse of recursion: Training on generated data makes models forget. *arXiv preprint arXiv:2305.17493*, 2023.

Ilia Shumailov, Zakhar Shumaylov, Yiren Zhao, Nicolas Papernot, Ross Anderson, and Yarin Gal. Ai models collapse when trained on recursively generated data. *Nature*, 631(8022):755–759, 2024.

Kenny Smith. Iterated learning in populations of bayesian agents. In *Proceedings of the annual meeting of the cognitive science society*, volume 31, 2009.

Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *International conference on machine learning*, pp. 2256–2265. pmlr, 2015.

Shobhita Sundaram, John Quan, Ariel Kwiatkowski, Kartik Ahuja, Yann Ollivier, and Julia Kempe. Teaching models to teach themselves: Reasoning at the edge of learnability. In *Forty-third International Conference on Machine Learning*, 2026. URL https://openreview.net/forum?id=GnqHK8Ww98.

Anvith Thudi, Evianne Rovers, Yangjun Ruan, Tristan Thrush, and Chris J Maddison. Mixmin: Finding data mixtures via convex minimization. *arXiv preprint arXiv:2502.10510*, 2025.

Ankit Vani, Max Schwarzer, Yuchen Lu, Eeshan Dhekane, and Aaron Courville. Iterated learning for emergent systematicity in vqa. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=Pd_oMxH8IlF.

Ze Wang, Zekun Wu, Yichi Zhang, Xin Guan, Navya Jain, Qinyang Lu, Saloni Gupta, and Adriano Koshiyama. Bias amplification: Large language models as increasingly biased media. In *Proceedings of the 14th International Joint Conference on Natural Language Processing and the 4th Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics*, pp. 115–132, 2025.

Kyra Wilson, Sourojit Ghosh, and Aylin Caliskan. Bias amplification in stable diffusion's representation of stigma through skin tones and their homogeneity. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, volume 8, pp. 2705–2717, 2025.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, et al. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 conference on empirical methods in natural language processing: system demonstrations*, pp. 38–45, 2020.

Xiaodan Xing, Fadong Shi, Jiahao Huang, Yinzhe Wu, Yang Nan, Sheng Zhang, Yingying Fang, Mike Roberts, Carola-Bibiane Schönlieb, Javier Del Ser, et al. When ai eats itself: on the caveats of ai autophagy. *arXiv preprint arXiv:2405.09597*, 2024.

Sihan Xu, Ziqiao Ma, Yidong Huang, Honglak Lee, and Joyce Chai. Cyclenet: Rethinking cycle consistency in text-guided diffusion for image manipulation. *Advances in Neural Information Processing Systems*, 36:10359–10384, 2023.

Chenhao Zheng, Jieyu Zhang, Aniruddha Kembhavi, and Ranjay Krishna. Iterated learning improves compositionality in large vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13785–13795, 2024.

A LLM PROMPT FOR CAPTION-VARIANT GENERATION

The following system prompt was used for Claude Sonnet 4.6 (Anthropic, 2026) via the Open Router API.

You rewrite Stable Diffusion image prompts. Given the gender, hair colour and hair type of a face photo, produce N short, natural-language caption variants that all describe that exact combination. Maximise vocabulary diversity: use synonyms and varied descriptors for every attribute (e.g. hair colour: blond, golden, flaxen, light-haired; hair type: wavy, curly, undulating, flowing; gender: woman, lady, female, girl). Also vary sentence structure, tone and phrasing. Keep each caption under 20 words and end with a period. Do not mention attributes that were not given. Return ONLY a JSON array of strings, no commentary.

With labels for each category being created using the prompt

Gender: [gender]. Hair colour: [hair colour]. Hair type: [hair type]. Produce 64 short caption variants for a photo of this person.

B ADDITIONAL RESULTS

B.1 AUGMENTED LABEL RUNS

When using the augmented labeling scheme the collapse still clearly occurs without the regulariser as shown by Figure 6, in addition we can clearly see that the stronger the regulariser influence the better the retention is of skin colour as reflected in other results, see Figure 7.

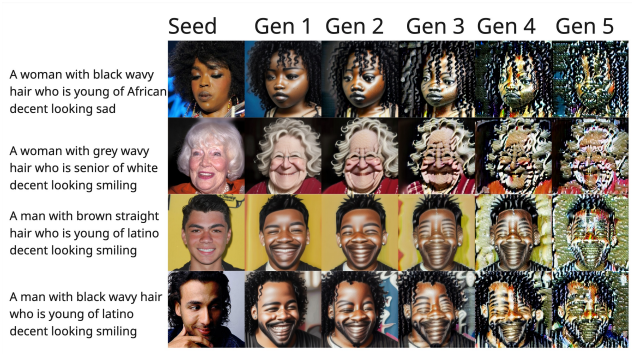

Figure 6: A baseline run showing collapse across a number of experimental augmented prompts when using the augmented label regime.

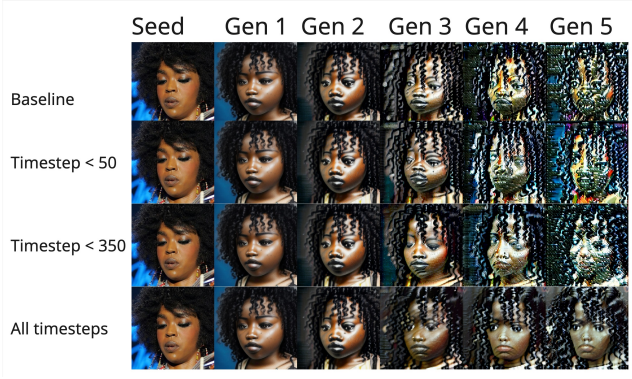

Figure 7: The effect of increasing the number of denoising timesteps exposed to the regulariser under the augmented label regime clearly shows better retention of skin colour during collapse with increased influence of the regulariser

B.2 Effect of regularising Time on FID in CelebA-HQ

The improvement to FID Score afforded by our regulariser can be seen across datasets and hyper-parameter choices. Figure 3 shows that increasing the weighting of the regulariser results in a reliable improvement in FID Score on both CelebA-HQ and COCO 2017. In the case of COCO 2017 the benefit could be obtained even if the regulariser was only applied for the first 350 timesteps. We report a similar analysis here for the CelebA-HQ dataset in Figure 8. We observe that applying the regulariser for longer during inference results in a consistent improvement, with the first 350 timesteps once again being the most important. However as Figure 9 shows although these timesteps offer improvement there are diminishing returns from increasing the number of timesteps with the later timesteps being less useful.

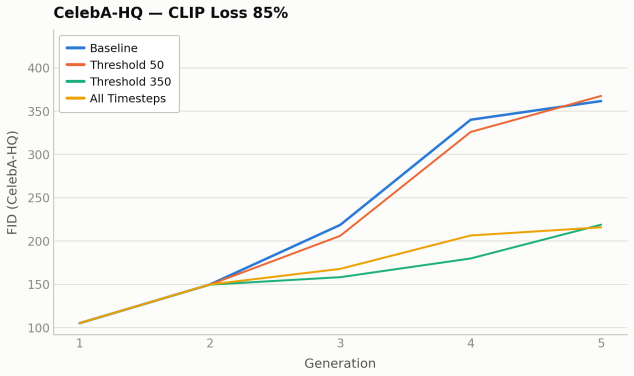

Figure 8: FID Score improves with increasing duration that the Lewis Signaling regulariser is applied on the CelebA-HQ dataset.

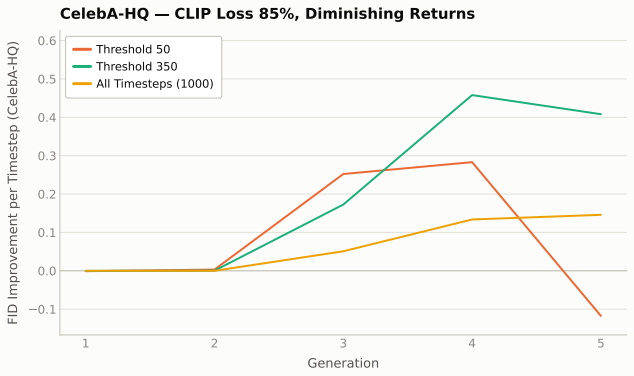

Figure 9: Per timestep improvement in FID Score compared across different number of timesteps exposed during training. This supports the notion that there are diminishing returns when increasing the number of timesteps

B.3 LLM Generated Labels and Under-specification

On occasion the elaborate language of the LLM generated labels seems to direct the collapse towards more satirical versions of the labels while ignoring the seed image. In the worst case this results in nonsensical images. For example a prompt for “hazelnut hair” results in the girl having hair made of literal hazelnuts. Figure 10 shows examples of this effect and demonstrates the risks of the prompt being under-specified and the propensity of the model to “game” the loss function. This is important to note as it demonstrates that the prompt itself has a very significant effect on guiding the collapse, and also the need to take care in prompting the model. This is necessary not only to obtain good performance, but also to avoid the injection of bias.

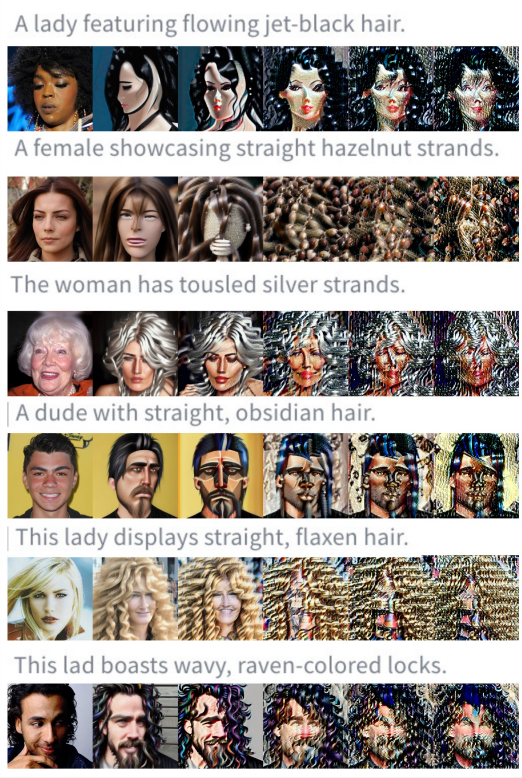

Figure 10: LLM generated labels lead to satirical and nonsensical images if the resulting prompt is under-specified or includes double-meanings to the chosen words.