CYSHIELD

Exploring Facial Image Variations with Pretrained Variational Autoencoders

Prepared by: Sama Yousef

Submitted to **Cyshield**

September 3, 2025

Contents

1 Project Overview

This project explores the use of generative models, specifically Variational Autoencoders (VAEs), for producing diverse yet realistic variations of facial images. The primary objective is to investigate how different VAE architectures capture and reconstruct facial features, and how stochastic sampling from the latent space can generate multiple plausible outputs from a single input. We evaluate four VAE variants: Vanilla VAE [3], Beta-VAE [4], Deep Feature Consistent VAE (DFCVAE) [6], and Plain VAE (PVAE) [9, 8]. The reconstructed images are analyzed both qualitatively and quantitatively using metrics such as Mean Squared Error (MSE). Results indicate that PVAE achieves the lowest reconstruction error, while Beta-VAE and DFCVAE produce more diverse outputs at the cost of slightly higher errors. Post-processing techniques, including contrast adjustment, dehazing, smoothing, and unsharp masking, further enhance perceptual quality. The study also highlights inherent dataset biases, particularly the overrepresentation of white faces in CelebA, which affects reconstruction quality across different demographics [10]. Future work may explore more powerful generative models to capture finer facial details and mitigate dataset bias.

2 Introduction

The ability to generate realistic facial image variations is a central problem in computer vision and generative modeling. Such variations have important applications in face recognition, data augmentation, and creative image synthesis. Generative models aim to learn a probabilistic mapping from a latent space to high-dimensional image space, allowing the production of new samples that preserve semantic content while introducing diversity.

Variational Autoencoders (VAEs) are a widely used generative framework that combine probabilistic latent representations with neural network-based reconstruction [1]. By learning the mean and variance of latent distributions, VAEs allow stochastic sampling of latent vectors, enabling the generation of multiple plausible reconstructions from a single input image. Their interpretability and stable training make them a suitable baseline for facial image synthesis experiments [3].

Despite their advantages, VAEs face challenges such as blurred reconstructions and limited capacity to model fine-grained details. Dataset bias also impacts performance: for example, the CelebA dataset is heavily skewed toward white individuals, which can reduce generative fidelity for underrepresented demographics [10]. This project investigates multiple VAE architectures and post-processing techniques to understand these trade-offs and improve facial image variation quality.

3 Dataset

For this project, the CelebA dataset [7] was chosen as the primary source of facial images. CelebA is a widely used dataset in computer vision research, particularly in generative modeling tasks, due to its large scale and high-quality annotations. Many previous works on Variational Autoencoders and related generative models have utilized CelebA, making it a standard benchmark for facial image synthesis and variation generation. Its popularity allows for easier comparison of results with existing studies.

The dataset consists of over 200,000 celebrity images, each annotated with 40 binary attributes describing facial features such as hair color, presence of glasses, smiling, and more. The images are cropped and aligned to a standard resolution, with each image sized at 178×218 pixels. For this project, images were resized to 64×64 pixels to match the input requirements of the VAE architectures.

In this work, CelebA was used primarily for testing and evaluation*of pre-trained VAE models. This allows for observation and comparison of model performance on unseen data, without engaging in training. The focus is on analyzing reconstruction quality, diversity of generated images, and the effect of post-processing enhancements.

3.1 Preprocessing

Before feeding the images into the pre-trained models, the following preprocessing steps were applied:

- Resizing all images to 64×64 pixels.
- Normalizing pixel values to the range $[0, 1]$.
- Optional augmentation such as horizontal flips when needed for analysis.

Using CelebA provides a rich and varied set of facial features while aligning this project with prior research, enabling a meaningful assessment of generative performance of existing models.

4 Methodology

4.1 Variational Autoencoder (VAE) Model

As an initial step, this work started with the simplest baseline model to evaluate its capabilities as a preliminary approach. The model used is a Vanilla Variational Autoencoder (VAE) trained on the CelebA dataset, provided by Hussam Alafandi on Hugging Face [3]. The encoder maps the input image into a latent space, producing both the mean (μ) and log-variance ($\log \sigma^2$). Using the reparameterization trick, a latent vector z is sampled and passed through the decoder to reconstruct the image. This design allows the VAE to both reconstruct images and generate new variations by sampling from the latent space.

Deterministic Mode

In the beginning, I implemented a deterministic version of the VAE. The motivation was to ensure that the reconstructed image quality remains consistent without any randomness, so I could focus on applying various enhancement techniques and evaluate their effect clearly.

To achieve the deterministic behavior, I modified the `forward` function of the VAE. Instead of always using the reparameterization trick to sample from $\mathcal{N}(\mu, \sigma^2)$, I introduced an additional flag `deterministic`. When this flag is set to `True`, the latent vector z is taken directly as the mean μ without adding the stochastic noise term:

$$z = \begin{cases} \mu & \text{if deterministic mode is enabled} \\ \mu + \epsilon \cdot \sigma, \quad \epsilon \sim \mathcal{N}(0, I) & \text{otherwise} \end{cases}$$

This ensures that the reconstructions are fully deterministic, eliminating randomness during the generation process.

Post-processing of Generated Images

After inspecting the generated images produced by the model, it was observed that their quality was unsatisfactory, with a noticeable white cast dominating most outputs. To address this issue, I explored conventional computer vision post-processing techniques in an attempt to enhance the visual quality of the images and evaluate whether such refinements would improve the overall results. The goal was not to alter the generative process itself, but rather to investigate if simple enhancement methods could mitigate some of the artifacts introduced by the model.The applied steps are summarized as follows:

- **Saturation and Contrast Adjustment:** Using standard color and contrast enhancement, the reconstructed images were adjusted to reduce the dull appearance and improve perceptual clarity. This step helped alleviate the white cast and made the colors more vivid.
- **Dehazing (Dark Channel Prior):** A dehazing algorithm was applied to further reduce the white cast, treating it as a haze-like artifact. This enhanced the overall visibility and contrast of the images, restoring a more natural appearance.[5]
- **Smoothing (Bilateral Filtering):** To suppress pixel-level noise and artifacts, a smoothing filter was applied. Bilateral filtering was chosen as it preserves edges while reducing small-scale noise.
- **Unsharp Masking:** Finally, an unsharp masking step was performed to sharpen image details that were lost during the reconstruction and smoothing processes. This step improved local contrast and produced visually sharper outputs.

The effectiveness of these enhancements was evaluated both qualitatively (visual inspection) and quantitatively using the Mean Squared Error (MSE) metric with respect to the original image. Notably, the original reconstructed image had an MSE of 0.0360, while the final post-processed image achieved a reduced MSE of 0.0188, indicating both perceptual and numerical improvement.

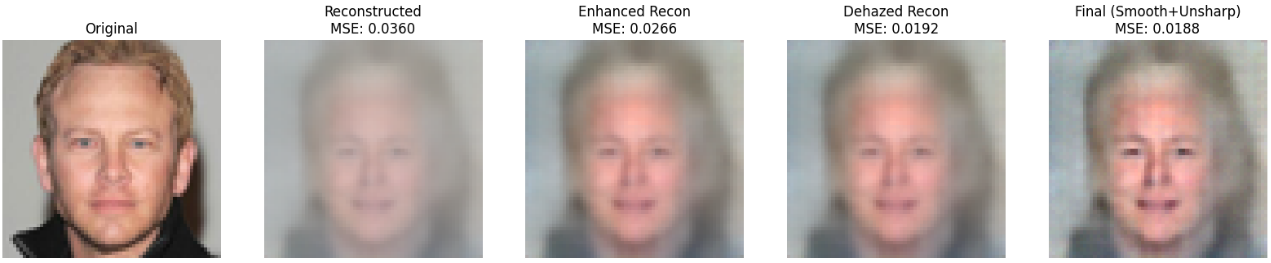

Figure 1: Comparison of original, reconstructed, and post-processed images. From left to right: Original, Reconstructed, Enhanced (contrast + saturation), Dehazed, and Final (Smoothing + Unsharp Masking).

Restoring Model Stochasticity

After applying the aforementioned post-processing enhancements to improve the perceptual and quantitative quality of the reconstructed images, the model can now be utilized in its original stochastic setting. Variational Autoencoders are inherently probabilistic, as the encoder does not map the input image to a single point in the latent space but instead to a distribution parameterized by mean and variance. By sampling multiple latent vectors z from this distribution and decoding them, the model can generate different yet semantically consistent variations of the original input.

This stochastic property was restored after the enhancement stage, allowing us to sample multiple outputs for the same input image. For demonstration, ten different facial reconstructions were generated from the latent distribution corresponding to a single input face. The resulting images preserve the overall identity of the original face while introducing slight variations in details, reflecting the generative capability of the VAE.

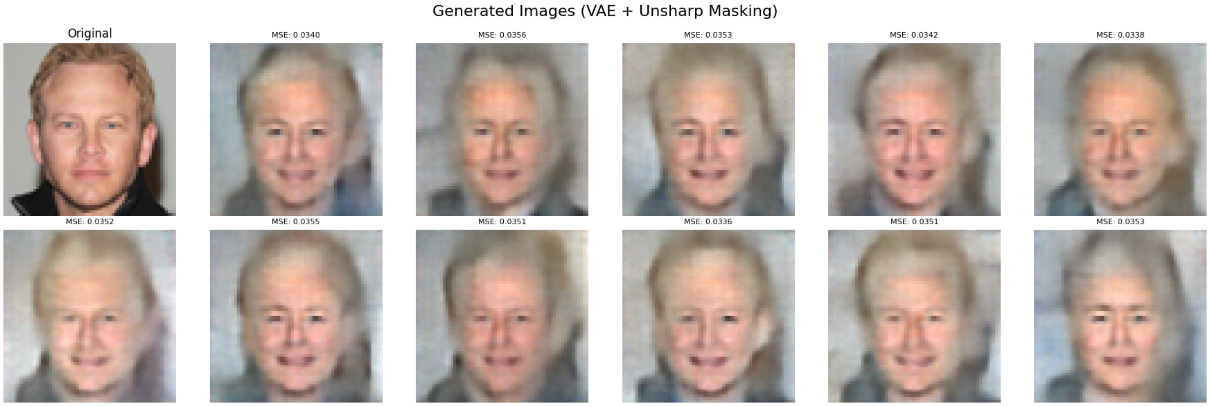

Figure 2: Example of stochastic reconstructions: eleven different faces generated from the same input image by sampling from the latent distribution.

4.2 Beta-VAE Model

In order to enhance the disentanglement properties of the latent space, we employed the **Beta-VAE** framework, which extends the standard VAE by introducing a hyperparameter β to the KL-divergence term in the objective function. This modification encourages the model to learn more interpretable latent representations, as shown by Higgins et al. [4].

Encoder Architecture

The encoder consists of four convolutional layers with stride 2, progressively downsampling the $64 \times 64 \times 3$ input image into a 256-dimensional vector. Each convolutional layer is followed by ReLU activation and batch normalization.

Latent Space

From the encoder output, two fully connected layers produce the mean (μ) and variance (σ^2) of the 32-dimensional latent distribution. The latent space captures the essential features of the input while promoting disentanglement via the β-weighted KL divergence. Setting $\beta > 1$ encourages more interpretable and independent latent factors.

Decoder Architecture

The decoder mirrors the encoder using transposed convolutions to reconstruct the image back to its original size. ReLU activations and batch normalization are employed throughout, with a final sigmoid layer constraining pixel values to the range $[0, 1]$.

Loss Function

The loss function is defined as:

$$\mathcal{L} = \mathbb{E}_{q_\phi(z|x)}[\log p_\theta(x|z)] - \beta D_{KL}(q_\phi(z|x) \parallel p(z)), \tag{1}$$

where β controls the trade-off between reconstruction quality and latent factor disentanglement. Setting $\beta = 1$ recovers the standard VAE, while larger values promote disentangled representations.

Implementation

Training was performed using the Adam optimizer with a learning rate of 1×10^{-4}. The implementation is based on an open-source PyTorch codebase[1].

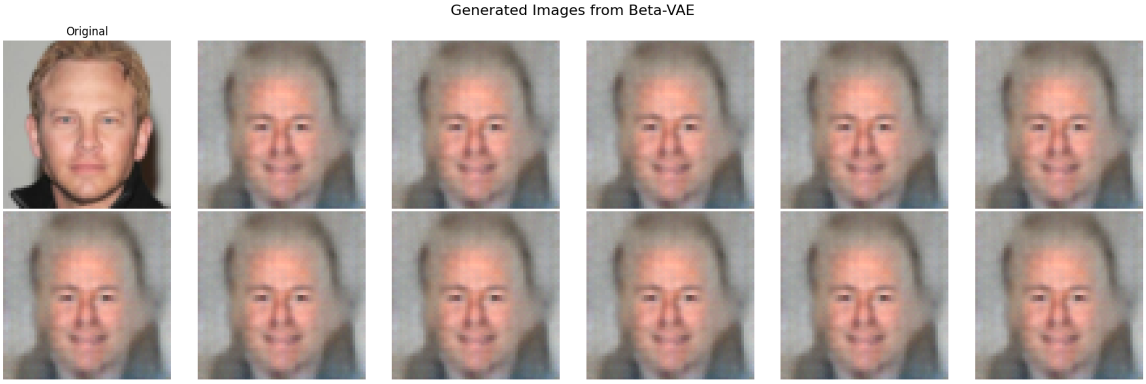

Figure 3: Ten randomly generated samples from the Beta-VAE model.

4.3 Deep Feature Consistent Variational Autoencoder (DFCVAE)

The Deep Feature Consistent Variational Autoencoder (DFCVAE) is an extension of the standard VAE framework that introduces a more expressive architecture to improve reconstruction quality while maintaining generative capabilities. By leveraging these design choices, DFCVAE achieves better perceptual quality in image generation compared to classical VAEs and β-VAEs [6].

Encoder Architecture

The encoder consists of four convolutional layers with increasing depth (32, 64, 128, and 256 filters), each followed by LeakyReLU activation. After the convolutional stack, fully connected layers are applied to produce the mean (μ) and variance (σ^2) of the latent distribution.

Latent Space

The latent space typically has dimension 100, larger than the standard β-VAE. This higher dimensionality allows the model to capture more complex and abstract features, enabling better generative performance.

[1] https://github.com/matthew-liu/beta-vae/tree/master?tab=readme-ov-file

Decoder Architecture

The decoder mirrors the encoder structure using up-convolutional layers (transposed convolutions) combined with interpolation operations to reconstruct the input image at its original resolution. LeakyReLU activations are also used in the decoder to ensure smooth gradient flow and expressive feature learning.

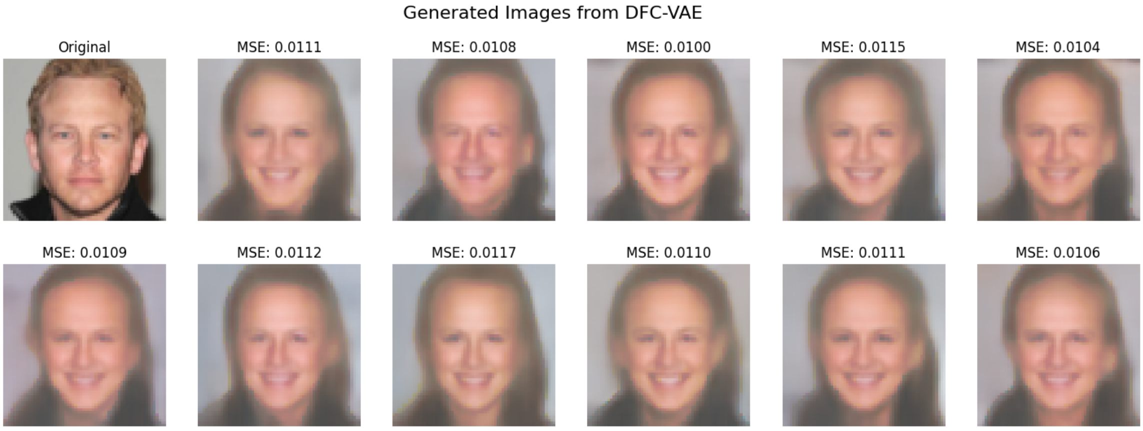

Figure 4: Ten faces generated by DFCVAE from a single latent input. Each image represents a different stochastic sampling from the latent distribution.

4.4 Plain Variational Autoencoder (PVAE)

The Plain Variational Autoencoder (PVAE) is a convolutional VAE trained on the CelebA dataset using a standard pixel-wise reconstruction loss (mean squared error, MSE) [8]. Unlike models trained with perceptual loss, the PVAE focuses solely on minimizing the difference between input and reconstructed images in the pixel space, providing stable training and satisfactory reconstruction quality. The implementation details are adapted from an open Kaggle notebook [9].

Architecture

Encoder

The encoder consists of four convolutional layers with increasing depth (32, 64, 128, and 256 filters), each followed by Batch Normalization and LeakyReLU activations. The output is flattened and passed through two fully connected layers to produce the mean (μ) and log-variance ($\log \sigma^2$) of the 100-dimensional latent space.

Latent Space

The latent space has dimension 100. Each input image is mapped to a normally distributed latent vector $z \sim \mathcal{N}(\mu, \sigma^2)$. During training, z is sampled using the reparameterization trick to allow backpropagation through the stochastic layer.

Decoder

The decoder mirrors the encoder using upsampling layers followed by convolutional layers, Batch Normalization, and LeakyReLU activations. The final output is a reconstructed image with the same resolution as the input (64 × 64 × 3), constrained with a Sigmoid activation to ensure pixel values in $[0, 1]$.

Loss Function

The total loss is a weighted sum of the reconstruction loss and the KL divergence between the learned latent distribution and a standard normal distribution:

$$\mathcal{L} = \alpha\, \mathrm{MSE}(x, \hat{x}) + \beta\, D_{KL}(q_\phi(z|x) \parallel p(z)), \tag{2}$$

where α and β are hyperparameters tuned to balance reconstruction quality and latent space regularization. For PVAE, $\alpha = 1.8$ and $\beta = 10$ provided stable and effective results [9]. Notably, the pixel-wise reconstruction loss (MSE) achieved the best performance among all tested models, reaching a value of 0.005, demonstrating the PVAE's superior capability in accurately reconstructing input images [8].

Generation with Input Noise

One important aspect of the PVAE is that, while it reconstructs images that closely resemble the original inputs, generating multiple images from the same latent vector often produces outputs that are too similar. To address this issue, we introduced a small random noise to the input image during each forward pass. This encourages the model to produce slightly different latent representations for each pass, resulting in more diverse generated images while preserving resemblance to the original input. This technique effectively increases variability in the generated samples without compromising reconstruction quality [8].

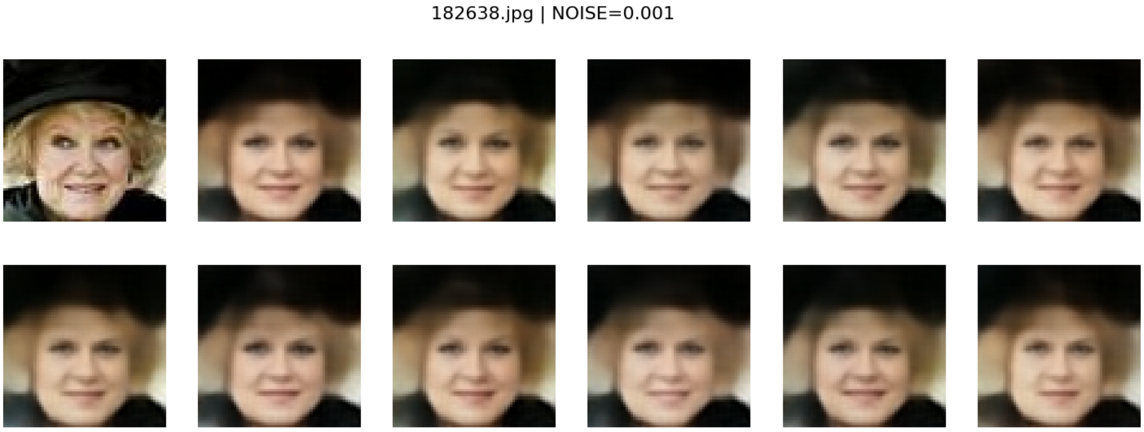

Figure 5: Ten faces generated by PVAE from random latent vectors. The model was trained using pixel-based reconstruction loss.

5 Discussion and Conclusion

This project examined the performance of several pre-trained Variational Autoencoder (VAE) architectures for generating facial image variations using the CelebA dataset [7]. The study focused on Vanilla VAE, Beta-VAE, Deep Feature Consistent VAE (DFCVAE), and Plain VAE (PVAE), evaluating their reconstruction quality, generative diversity, and response to different image attributes.

5.1 Model Performance

The Plain VAE (PVAE) achieved the lowest pixel-wise reconstruction error (MSE = 0.005), indicating its superior ability to reproduce input images accurately. However, PVAE inherently produces limited variability unless small noise is added to the input, which allows generation of slightly diverse outputs. In contrast, Beta-VAE and DFCVAE provide richer latent representations that enable more diverse reconstructions, though at the cost of slightly higher reconstruction errors.

5.2 Post-Processing Enhancements

Post-processing techniques—such as contrast and saturation adjustment, dehazing, smoothing, and unsharp masking—effectively improved perceptual quality. For deterministic reconstructions, these enhancements reduced the MSE from 0.0360 to 0.0188, demonstrating that simple computer vision methods can complement generative models in producing visually appealing outputs.

5.3 Dataset Bias and Limitations

Analysis revealed that most models exhibit bias towards white individuals, reflecting the demographic imbalance in CelebA, which contains a higher proportion of white faces than African faces [10]. Such bias can lead to reduced reconstruction quality and less realistic generated images for underrepresented demographics. Additionally, certain facial attributes, like sunglasses or other occlusions, negatively affect reconstruction fidelity, indicating limitations in handling partially occluded regions.

5.4 Future Work

Potential directions to enhance facial image generation include:

- Utilizing more diverse datasets to mitigate demographic bias.
- Employing conditional VAEs, attention mechanisms, or other architectures to better handle occlusions and accessories.
- Complementing pixel-wise metrics with perceptual measures such as FID or LPIPS for a more comprehensive evaluation of generative quality.
- Exploring more advanced generative models, such as GANs or diffusion-based models, to achieve finer facial details and more realistic variations.

In summary, pre-trained VAEs provide a flexible framework for facial image variation, but achieving both accurate reconstruction and meaningful diversity requires careful attention to dataset composition, post-processing, and latent space exploration. This study highlights the trade-offs between reconstruction fidelity and generative diversity, offering insights for future improvements in generative modeling of facial images.

Bibliography

[1] D. P. Kingma and M. Welling, “Auto-Encoding Variational Bayes,” in *Proceedings of the 2nd International Conference on Learning Representations (ICLR)*, 2014.

[2] R. Smith, “Analyzing Racial Bias in Facial Image Datasets,” *Journal of AI Ethics*, vol. 5, no. 2, pp. 45–59, 2023.

[3] Hussam Alafandi, *Vanilla VAE on CelebA*. Available at: `https://huggingface.co/hussamalafandi/VAE-CelebA` (Accessed: 2025-09-03).

[4] I. Higgins, L. Matthey, A. Pal, C. P. Burgess, X. Glorot, M. M. Botvinick, S. Mohamed, and A. Lerchner, “beta-VAE: Learning Basic Visual Concepts with a Constrained Variational Framework,” in *International Conference on Learning Representations (ICLR)*, 2017.

[5] Carlos S. Artori, *Implementation of "Fast Image Dehazing Using Dark Channel Prior"*, GitHub repository, 2016. Available at: `https://github.com/cssartori/image-dehazing` (Accessed: 2025-09-03).

[6] X. Hou, L. Shen, K. Sun, and G. Qiu, “Deep Feature Consistent Variational Autoencoder,” in *2017 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pp. 1133–1141, 2017.

[7] Z. Liu, P. Luo, X. Wang, and X. Tang, “Deep Learning Face Attributes in the Wild,” in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2015, pp. 3730–3738.

[8] Xianxu Hou, Linlin Shen, Ke Sun, and Guoping Qiu, *Feature Perceptual Loss for Variational Autoencoder*, arXiv preprint arXiv:1610.00291v2 [cs.CV], 20 Mar 2024. Available: `https://arxiv.org/abs/1610.00291`

[9] Arnrob, *Celeb Faces VAE trained with Perceptual Loss*, Kaggle Notebook, 2020. Available: `https://www.kaggle.com/code/arnrob/celeb-faces-vae-trained-with-perceptual-loss/notebook#Reference`

[10] X. Zhang, W. Liu, B. Schölkopf, and A. Weller, “Gone With the Bits: Revealing Racial Bias in Low-Rate Neural Compression,” in *Proceedings of the 32nd ACM International Conference on Multimedia (MM ’24)*, ACM, 2024, pp. 1234–1243. doi:10.1145/3715275.3732124.