This Emoji Does Not Exist (Yet): Learning to Create Emoji Faces with Diffusion Model

Senhao Cheng, Zijian Huang, Ziyu Liu, Runqiu Wang

EECS 553 Group 16, University of Michigan

Introduction

- **Motivation**: Emojis are essential tools in modern digital communication, enriching online interaction through emotional nuance and expressive symbolism [1]. However, the current emoji sets—especially the widely used **"yellow face"** series—are limited in variety and unable to reflect the full spectrum of human emotion.
- **Related work**: Diffusion models, especially denoising diffusion probabilistic models (DDPM) and their variants, have recently achieved remarkable success in generative modeling, demonstrating their ability to produce high-quality and diverse images across multiple complex visual domains [2]. Initial applications primarily involved general datasets such as CIFAR-10, CelebA-HQ, and ImageNet, where diffusion models showed competitive or superior performance compared to traditional generative adversarial networks (GANs) [2, 4].
  Further research has explored specialized visual domains, including the generation of stylized illustrations like anime and manga characters. Kamb and Ganguli (2024) demonstrated how diffusion models effectively captured structured visual styles, maintaining coherent artistic details while producing diverse and novel samples [5]. Similar advancements were reported by Wang et al. (2024), who utilized controllable diffusion models to generate consistent and stylistically coherent animation sequences, underscoring the model's capability for nuanced style transfer and structured content generation [**?**].
  Prior research has applied diffusion models to stylized domains such as anime and manga, demonstrating strong generative performance for structured visual content [5]. However, the emoji domain—especially realistic yellow-face emojis—remains underexplored.
- **Our goal**: In this project, we explore whether diffusion models can be applied to the domain of emoji generation. Our goal is to study (1) Can diffusion models generate high-quality new emoji samples? (2) If so, are they merely reproducing (memorizing) examples from the training set, or have they truly captured the underlying emoji distribution and can generate genuinely new designs?
- **Our work**:
  - For emoji generation, we **implement a diffusion model from scratch**, following the principles outlined in the Elucidated Diffusion Models (EDM) [4] framework, which provides a flexible and efficient framework for generative tasks, introducing noise scheduling, refined discretization, and high-order solvers.
  - For assessment, We evaluate our generated samples comprehensively, **assessing both image quality and model memorization**. We complement quantitative evaluations with human preference studies to better understand the perceptual quality and expressiveness of our synthesized emojis.

Data Construction

To support controllable emoji generation under low-resource settings, we construct a compositional emoji dataset featuring hybrid facial expressions generated from pairwise combinations of existing "yellow face" emojis, through programmatically accessing an online image synthesis platform. This significantly **increases the complexity of the data** (from 60 emojis to over 3000 emojis), making it possible for diffusion models to generate more novel emojis.

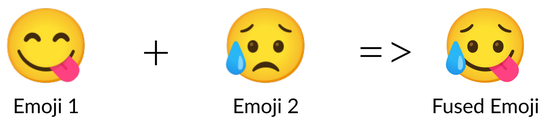

Figure 1. Emoji fusion example: combining two yellow face emojis to produce a novel expression.

Generation Methods

Diffusion models generate data by inverting a gradual noising process. Starting from a clean sample $x_0 \sim p_{\text{data}}$, the forward SDE

$$q(x_t \mid x_{t-1}) = \mathcal{N}\big(x_t; \sqrt{\alpha_t}\, x_{t-1}, (1-\alpha_t) I\big)$$

adds Gaussian noise with schedule $\{\alpha_t\}$. A neural network (score model) learns the reverse dynamics by predicting the gradient of the log-density $s_\theta(x, \sigma) \approx \nabla_x \log p_\sigma(x)$.

$$q(\lambda_t \mid \lambda_{t-1})$$

In implementation, we follow the **EDM** [4] framework, which separates the training and sampling process, and improves over traditional diffusion models by:

- Decoupling training/sampling for flexibility and stability,
- Preconditioning network inputs/outputs to normalize signal scales,
- High-order solvers (e.g., Heun's method) enabling few-step, high-fidelity sampling.

These improvements lead to faster generation and better quality under the same compute budget. Our implementation is aligned with the modular design proposed by Karras et al., allowing independent tuning of training schedule, solver, and noise scaling.

Evaluation Metrics

- To assess the quality of generated emojis, we use the following metrics:
- **Fréchet Inception Distance (FID):**

$$\text{FID} = \|\mu_r - \mu_g\|_2^2 + \text{Tr}\big(\Sigma_r + \Sigma_g - 2(\Sigma_r \Sigma_g)^{1/2}\big)$$

  - μ_r, Σ_r: mean and covariance of real images
  - μ_g, Σ_g: mean and covariance of generated images
  - All features are extracted using a pre-trained Inception-V3 network
- **Kernel Inception Distance (KID):**

$$\text{KID} = \frac{1}{n(n-1)} \sum_{i \neq i'} k(x_i, x_{i'}) + \frac{1}{m(m-1)} \sum_{j \neq j'} k(y_j, y_{j'}) - \frac{2}{nm} \sum_{i,j} k(x_i, y_j)$$

  - Kernel: $k(u, v) = \big(\frac{1}{d} u^\top v + c\big)^3$, where $d = 2048$
  - Also uses Inception-V3 features for comparison
- **Limitations of FID and KID:**
  - Designed for natural images
  - Inception-V3 is trained on ImageNet, which differs from stylized emoji data
- **Proposed Metric – MeanMSE:**

$$\text{MeanMSE} = \mathbb{E}_{x \sim p_\theta}\left[\min_{y \in \mathcal{D}} \|x - y\|^2\right] \approx \frac{1}{N} \sum_{i=1}^{N} \min_{y \in \mathcal{D}} \|x_i - y\|^2$$

  - Measures how close each generated emoji is to its nearest training example
  - Better suited for low-res, simple, stylized datasets

Results

- Demonstrate some image results, including succuessful ones and failure ones.

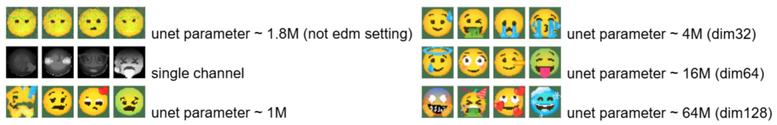

Figure 3. Examples of generated images

- FID and KID results for three model sizes over training.

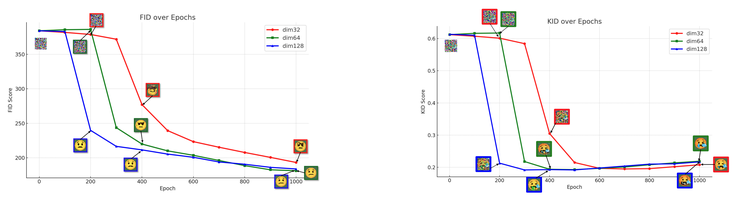

Figure 4. The 128-dimension model converges faster, but all three models reach similar final scores. While both metrics stabilize quickly, they do not accurately reflect actual image quality.

- MeanMSE, FID, and KID results on 128-dimensional model

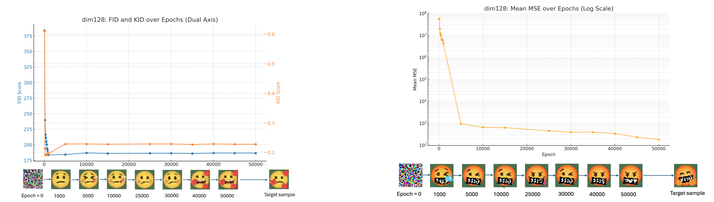

Figure 5. While FID and KID plateau early with no further improvement, MeanMSE continues to decrease over training—reflecting closer distance to originial dataset.

- Smallest MSE distribution analysis across different UNet embedding dimensions

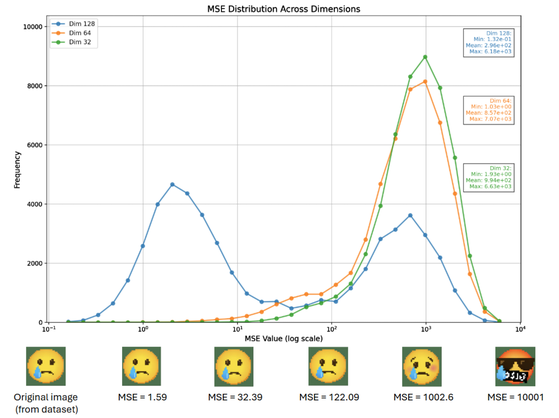

Figure 6. MSE distribution and visual examples. Histograms show smallest MSE values for different model sizes (128/64/32 dimensions). Sampled emojis demonstrate the correlation between lower MSE and improved visual fidelity.

Conclusion And Discussion

Diffusion models can rapidly generate high-quality 32×32 emojis—training takes ≈ 2 hours, and sampling thousands of images takes ≈ 10 minutes on a single A40. We observe strong FID/KID improvements and qualitatively novel and high quality emojis beyond the training set.

Our findings also suggest that mainstream metrics like **FID and KID** — designed for natural images — **may not be fully appropriate** for evaluating stylized, low-resolution data like emoji. Not only do these metrics struggle to capture per-sample novelty, but they are also **not a good measure of image generation quality**, especially under small dataset constraints.

Interestingly, we observe signs that the diffusion model actually generate novel emoji through understanding and recombining emoji components (e.g., tongue, sweat drops, halos), echoing recent findings on compositional creativity [3] in diffusion models. This opens up exciting directions for future work on interpretability—such as tracing generated components back to training examples using local gradients or MSE maps.

Finally, we speculate that if the model begins to generate truly novel emojis beyond the training set, we may observe a brief increase in MSE before it stabilizes again—signaling a shift from memorization to generalization. Unfortunately, due to time and computational constraints, we leave deeper investigation of this behavior to future work.

References

[1] Nerea Aldunate and Roberto González-Ibáñez. An integrated review of emoticons in computer-mediated communication, Dec 2016.

[2] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models, Dec 2020.

[3] Mason Kamb and Surya Ganguli. An analytic theory of creativity in convolutional diffusion models, December 2024.

[4] Tero Karras, Miika Aittala, Timo Aila, and Samuli Laine. Elucidating the design space of diffusion-based generative models, Oct 2022.

[5] Qiang Wang, Minghua Liu, Junjun Hu, Fan Jiang, and Mu Xu. Controllable longer image animation with diffusion models, May 2024.