Kaleido Diffusion: Improving Conditional Diffusion Models with Autoregressive Latent Modeling



Jiatao Gu^{†*}, Ying Shen^{&*}, Shuangfei Zhai[†], Yizhe Zhang[†], Navdeep Jaitly[†], Josh Susskind[†]

[†]Apple ^{\$}Virginia Tech * equal contribution

Motivation

Diffusion models, while adept at generating high-quality images from text, often produce limited visual diversity, especially under high classifier-free guidance settings. For instance, given a description, "a cat sat on the mat", existing text-to-image diffusion models predominantly produce image samples depicting cats with similar colors and patterns.

To tackle this, we introduce Kaleido - a new method that improves conditional diffusion model generation by incorporating autoregressive latent priors! This allows us generate much more diverse outputs even with high CFG just like a kaleidoscope.



Kaleido Diffusion

We propose Kaleido, a general framework that integrates an autoregressive prior with diffusion model to enhance image generation. Kaleido comprises two major components: an AR model that generates latent tokens as abstract representations, and a latent-augmented diffusion model that iteratively synthesizes images based on these latents together with the original condition.

Qualitative Results

Diversity of Generated Images





(1) Autoregressive Latent Modeling:

Given the original context c, Kaleido employs an autoregressive model $p_{\theta}(\boldsymbol{z}|\boldsymbol{c})$, to generate abstract discrete latents $\boldsymbol{z} = [\boldsymbol{z}_1, \dots, \boldsymbol{z}_N]$, serving as an intermediary representation for guiding the generation process. We explore various latents, including textual descriptions, bounding boxes, blobs, and abstract visual tokens.



(2) Latent-augmented Diffusion Models:

The diffusion model is conditioned on both the original text prompt c and the autoregressively generated discrete latents z for generating an image x. To capture the complex distribution of real images, Kaleido explicitly model "mode selection" through $p_{\theta}(\boldsymbol{z}|\boldsymbol{c})$ and leave $p_{\theta}(\boldsymbol{x}|\boldsymbol{z},\boldsymbol{c})$ to model other variations including local noise by applying diffusion steps.

Control from Latent Tokens

Input: a teddy bear wearing glasses A teddy bear wearing glasses and a sweater sits on a stool in a dimly lit room. The background features a chair and a wall, and the lighting is moody. The teddy bear appears to be deep in thought, with its chin re



Input: Two raccoons walk under a drizzly rain, and each of them holds an umbrella





nput: a photo of a blue Jay standing on a large basket of strawberrie





Input: a squirrel wearing a crown



Input: an owl with a knitted hat holding a board with "Thank You" written

The image generation follows a two-step procedure: $\boldsymbol{z} \sim p_{\theta}(\boldsymbol{z}|\boldsymbol{c}), \boldsymbol{x} \sim \tilde{p}_{\theta}(\boldsymbol{x}|\boldsymbol{z}, \boldsymbol{c})$, where CFG can be applied after z is sampled. From the perspective of score function, diffusion with CFG in Kaleido can be written as:

 $\nabla_{\boldsymbol{x}} \log \tilde{p}_{\theta}(\boldsymbol{x} | \boldsymbol{c}, \boldsymbol{z}) = \gamma \left[\nabla_{\boldsymbol{x}} \left(\log p_{\theta}(\boldsymbol{x} | \boldsymbol{c}) + \log p_{\theta}(\boldsymbol{z} | \boldsymbol{x}, \boldsymbol{c}) - \log p_{\theta}(\boldsymbol{x}) \right) \right] + \nabla_{\boldsymbol{x}} \log p_{\theta}(\boldsymbol{x}).$

Compared to standard diffusion process, the highlighted term above pushes the updating direction towards the sampled modes at each step, ensuring diverse generation as long as $p_{\theta}(\boldsymbol{z}|\boldsymbol{c})$ is diverse.

Quantitative Results

Compared with the baseline diffusion models (MDM) with various guidance scales, Kaleido consistently enhances the diversity of samples without compromising their quality across different CFG, evidenced by the general improvement in both FID and Recall.



Paper

Please check out our paper for more details!





Latent Editing

Autoregressive Generation: Diffusion Generation: In the image, a frog is seen sipping on a cup of coffee, seemingly enjoying a relaxing break. The frog is positioned on <u>a log</u>, with its eyes closed and a small smile on its face, as if it's savoring the flavor of the coffee. The cup of coffee is placed on a rock next to the frog, and the background features <u>a body of water</u>. The frog's green and yellow coloration stands out against the natural setting, making for a charming and whimsical scene. Diffusion Re-generation: Edit1: In the image, a frog is seen sipping on a cup of coffee, seemingly enjoying a relaxing preat. The frog is positioned es, with its eyes closed and a small smile on its face, as if it's savoring the flavor of the coffee. The cup of coffee is placed on a rock next to the frog, and the background features <mark>forest</mark>. The frog's green and yellow coloration stands out against the natural setting, making for a charming and whimsical scene. Edit2 **Diffusion Re-generation:** In the image, a frog is seen sipping on a cup of coffee, seemingly enjoying a relaxing break. The frog is positioned stones, with its eyes closed on cob and a small smile on its face, as if it's savoring the flavor of the coffee. The cup of coffee is placed on a rock next to the frog, and the background features forest. The frog's green and yellow coloration stands out against the natural setting, making for a charming and whimsical scene.

Input: a photo of a frog drinking coffee