Kaleido Diffusion: Improving Conditional Diffusion Models with Autoregressive Latent Modeling

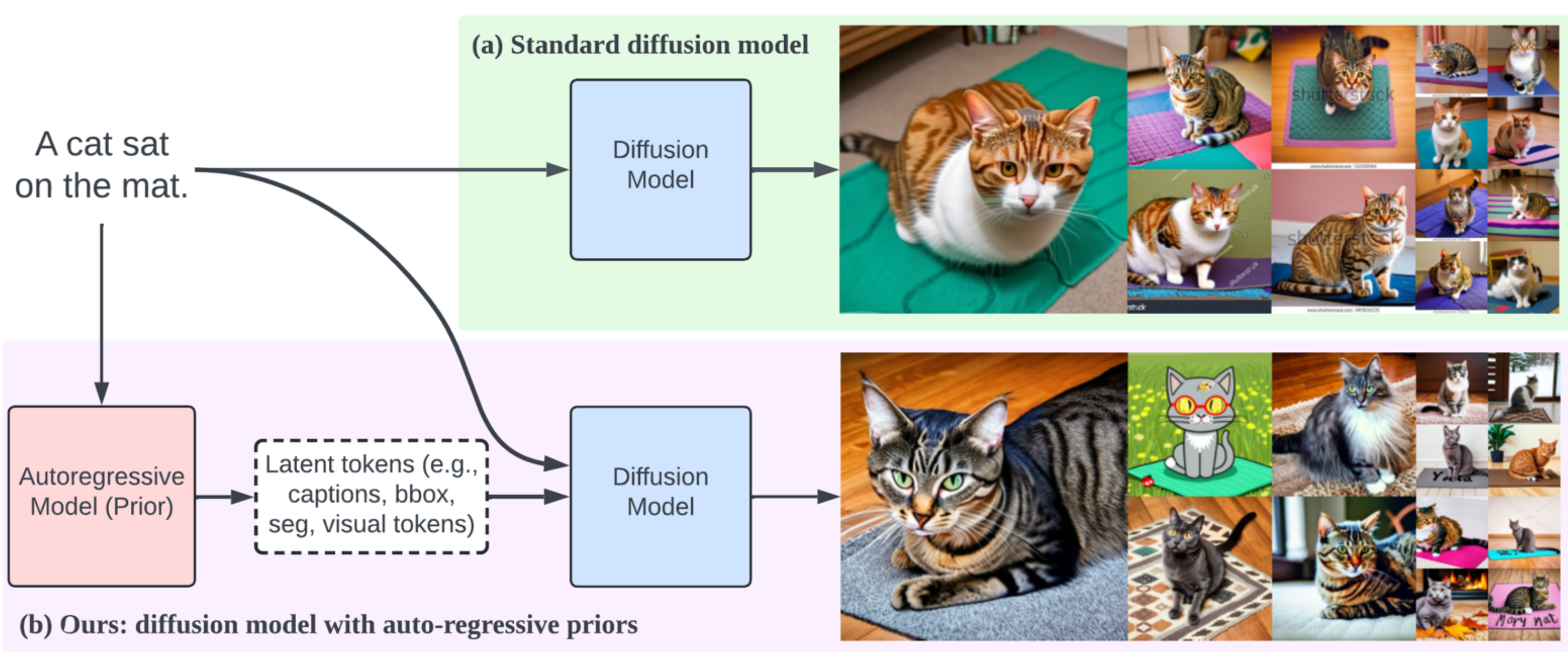
Jiatao Gu[†*], Ying Shen[◊*], Shuangfei Zhai[†], Yizhe Zhang[†], Navdeep Jaitly[†], Josh Susskind[†]

[†]Apple [◊]Virginia Tech [*] equal contribution

Motivation

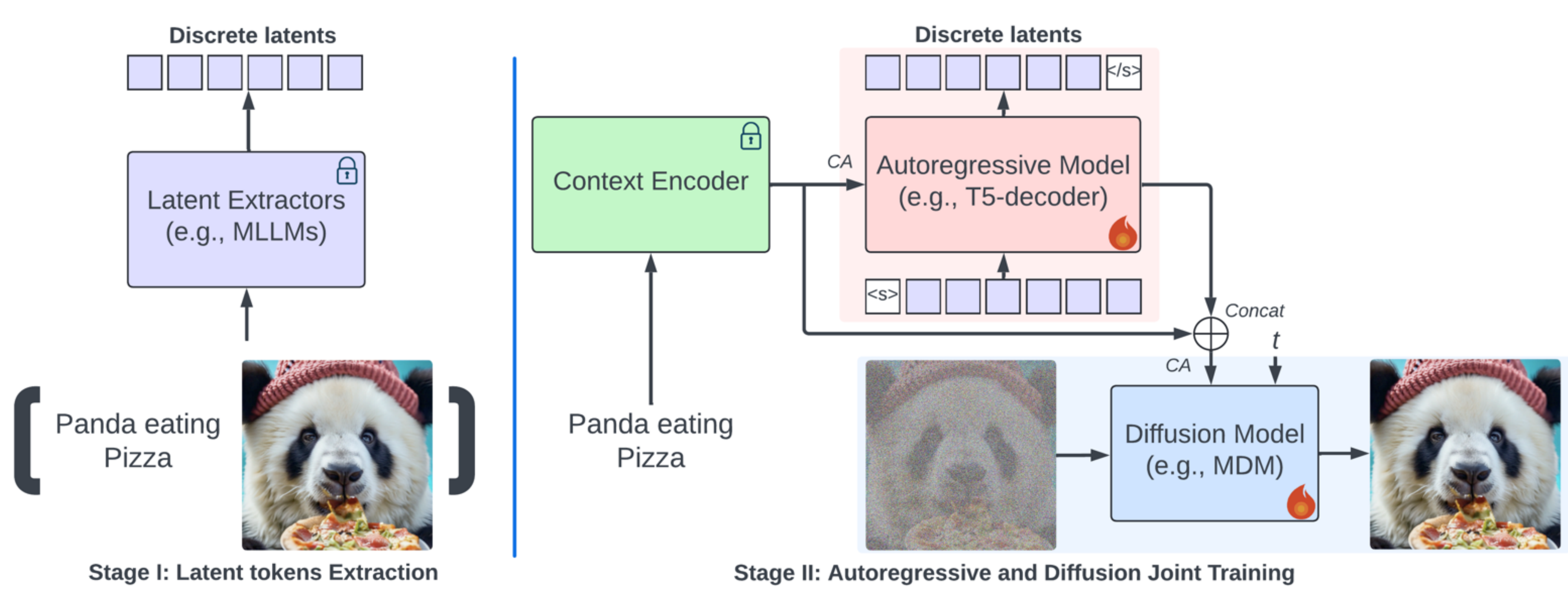
Diffusion models, while adept at generating high-quality images from text, often produce limited visual diversity, especially under high classifier-free guidance settings. For instance, given a description, "a cat sat on the mat", existing text-to-image diffusion models predominantly produce image samples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



(1) Autoregressive Latent Modeling:

Given the original context $\boldsymbol{c}$, Kaleido employs an autoregressive model $p_\theta(\boldsymbol{z}|\boldsymbol{c})$, to generate abstract discrete latents $\boldsymbol{z} = [\boldsymbol{z}_1, \ldots \boldsymbol{z}_N]$, 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 $\boldsymbol{c}$ and the autoregressively generated discrete latents $\boldsymbol{z}$ for generating an image $\boldsymbol{x}$. 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.

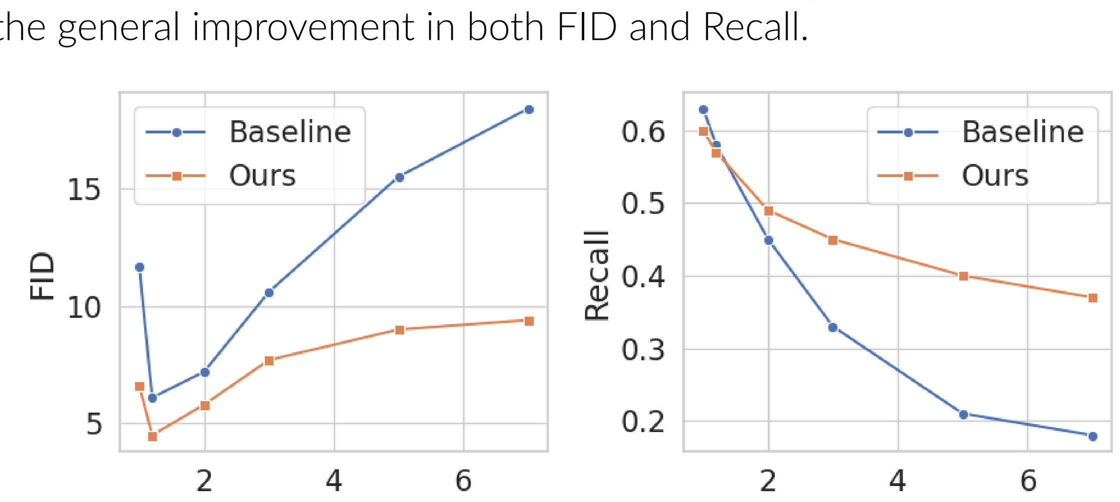
The image generation follows a two-step procedure: $\boldsymbol{z} \sim p_\theta(\boldsymbol{z}|\boldsymbol{c}), \boldsymbol{x} \sim \tilde{p}_\theta(\boldsymbol{x}|\boldsymbol{z},\boldsymbol{c})$, where CFG can be applied after $\boldsymbol{z}$ is sampled. From the perspective of score function, diffusion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

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

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.



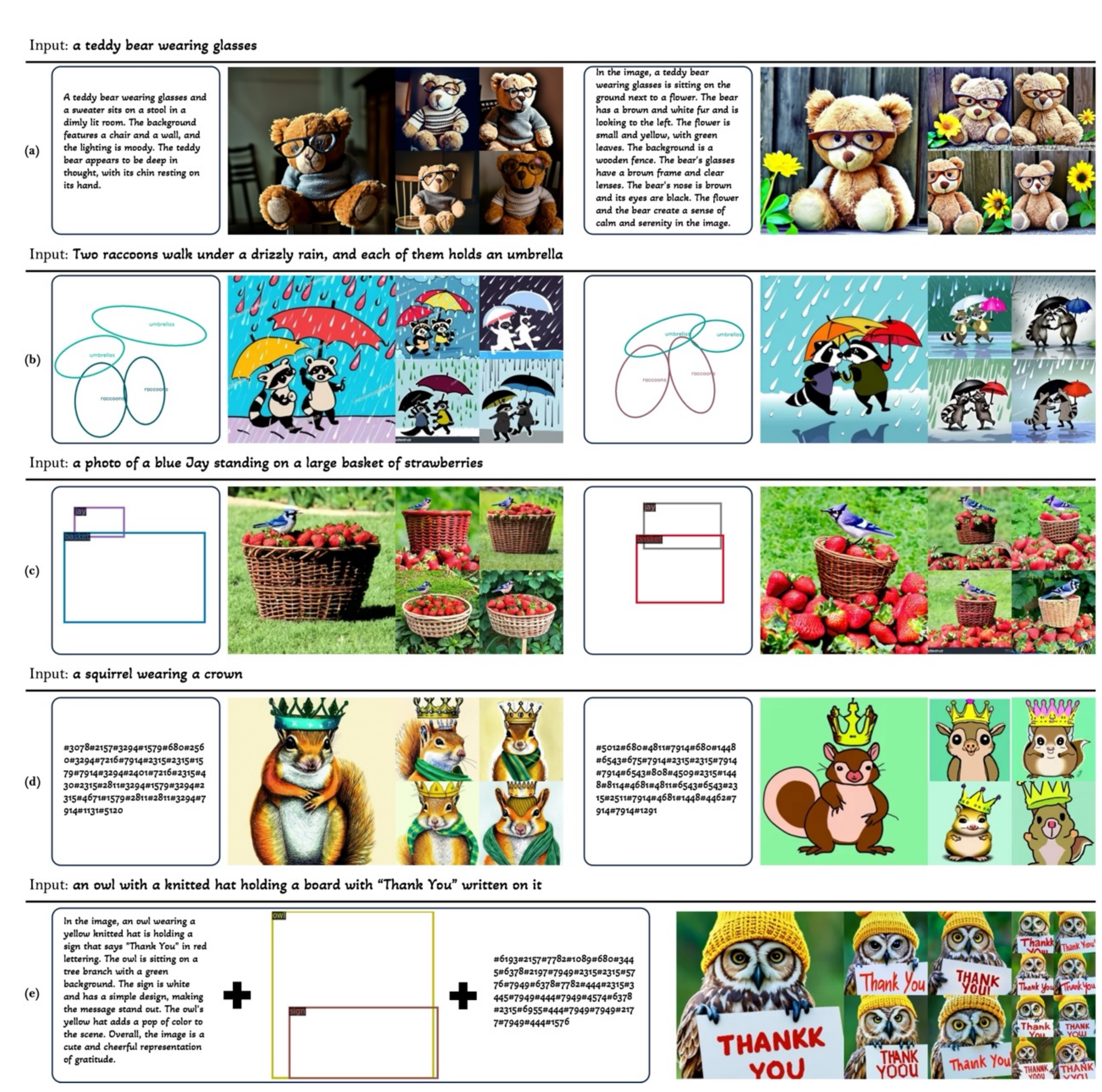
Paper

Please check out our paper for more details!

Qualitative Results

Diversity of Generated Images

Control from Latent Tokens



Latent Editing