JUNE 18-22, 2023
CVPR
VANCOUVER, CANADA

Unite and Conquer: Plug & Play Multi-Modal Synthesis using Diffusion Models

Nithin Gopalakrishnan Nair, Wele Gedara Chaminda Bandara and Vishal M Patel
Johns Hopkins University, Baltimore, MD, USA

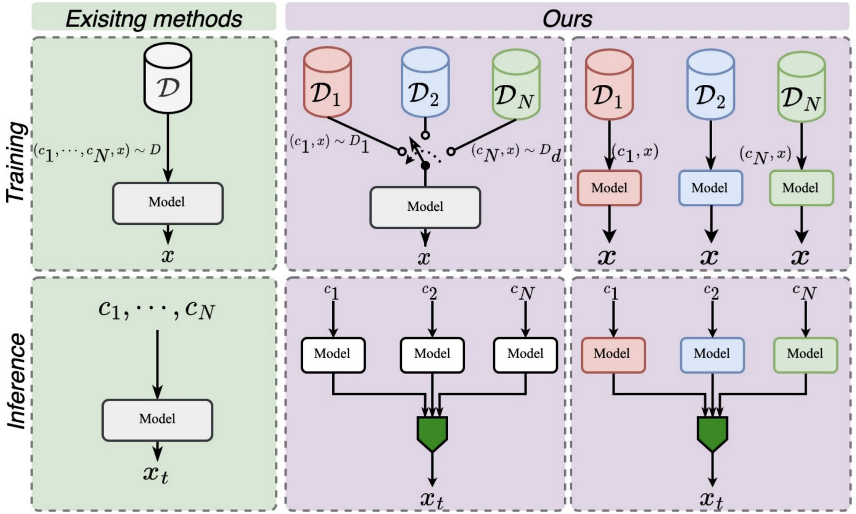
TUE-PM-186

Why do we need this?

- *Generating photos satisfying multiple constraints require paired data consisting of all modalities (i.e., conditions) and their corresponding output*
- *We propose a diffusion-based solution for image generation under the presence of multimodal priors without paired data.*
- *Our method is easily scalable and can be incorporated with off-the-shelf models to add additional constraints.*

How is it different?

- *Existing methods require training a model with all conditions.*
- *Our method just needs one at a time*
- *Sampling strategy that interpolates unconditional domains*

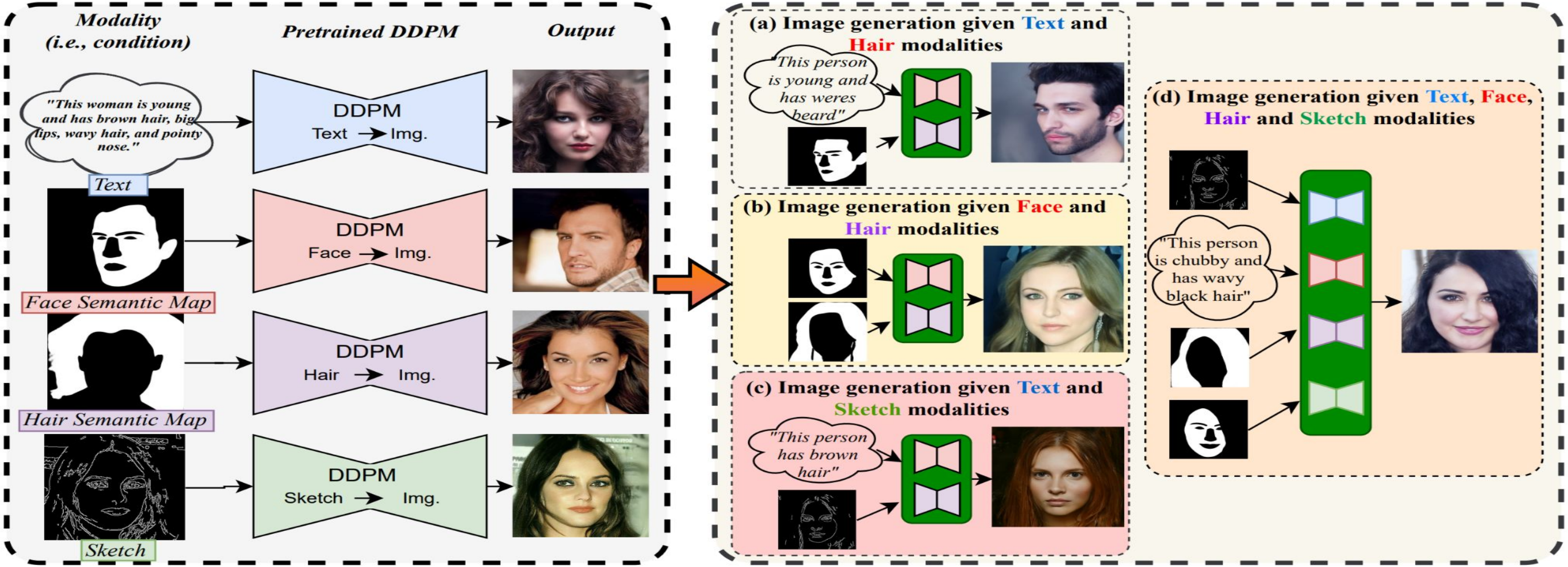


Label Combined OURS

JOHNS HOPKINS UNIVERSITY

JUNE 18-22, 2023 CVPR VANCOUVER, CANADA

What does it do?



JOHNS HOPKINS UNIVERSITY

JUNE 18-22, 2023 CVPR VANCOUVER, CANADA

How does it work?

- *Reverse sampling in a diffusion model can be written as,*

$$z_{t-1} \leftarrow \frac{1}{\sqrt{1-\beta_t}}(z_t - \beta_t s_\theta(z_t, x, t)) + \sigma_t^2 \boldsymbol{\eta} \qquad \boldsymbol{\eta} = \mathcal{N}(\mathbf{0}, \boldsymbol{I})$$

- *Here* s_θ *is the score function describing the diffusion process,* ϵ_θ *is the prediction of diffusion U-Net*

$$s_\theta(z_t, t) = \nabla_x \log P(z_t|x) = \frac{\epsilon_\theta(z_t, x, t)}{\sqrt{1-\bar{\alpha}_t}}$$

- *The effective unconditional density of the image space we are trying to model can be decomposed as a combination of multiple subspaces united by generalized product of experts.*

$$P(z) = \prod_{i=1}^{N} P_{\delta_i}^{a_i}(z|\phi) \qquad P(z|\mathbf{X}) = \frac{P(z)}{P(\mathbf{X})}\prod_{i=1}^{N} P(x_i|z) \approx K P(z)\frac{\prod_{i=1}^{N} P^{w_i}(z|x_i)}{\prod_{i=1}^{N} P^{w_i}(z)}$$

- *Utilizing this, the effective score becomes*

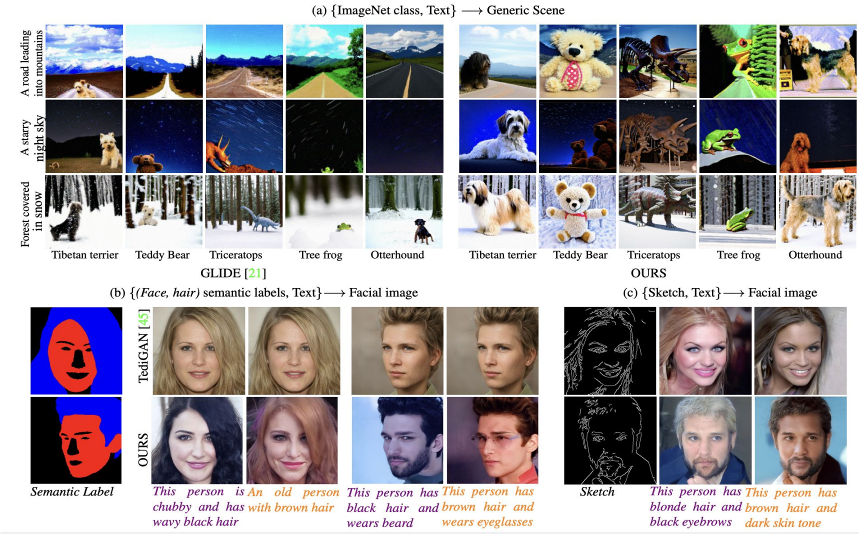
$$\epsilon_c = \sum_{i=1}^{N} w_i \epsilon_i(z_t, x_i, t) - (\sum_{i=1}^{N} w_i - 1)\sum_{j=1}^{N} a_j \epsilon_j(z_t, \phi, t). \qquad w_i \geq 1$$

JOHNS HOPKINS UNIVERSITY

JUNE 18-22, 2023 CVPR VANCOUVER, CANADA

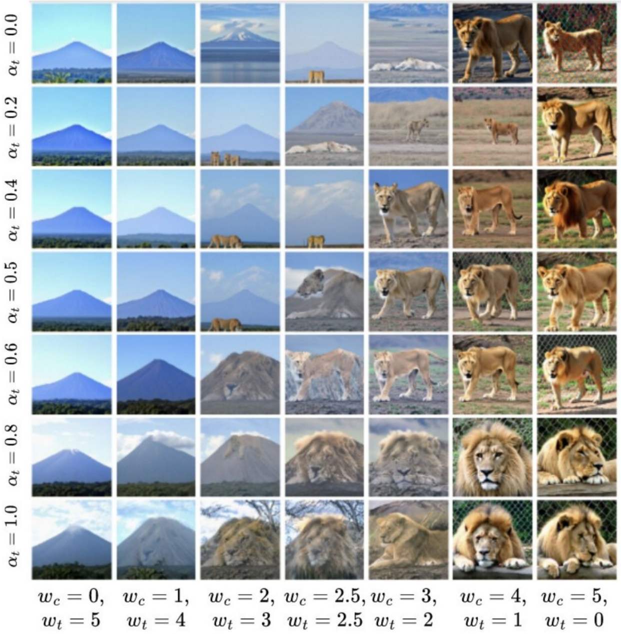
Applications of our method

- ➢ *Generating a composite scene consisting of a text based background and an ImageNet class by combining a pretrained text based generator and ImageNet class generator*
- ➢ *Generating faces satisfying semantic constraints and text descriptions*



Cross domain Interpolation

- Interpolation by varying strengths of unconditional model and conditions.
- Unite and Conquer can achieve any level of contextual interpolation across domains



Thank you!