Collaborative Diffusion
and Human-Machine Collaborative AIGC

Ziqi Huang 黄子琪

MMLab@NTU | S-Lab, Nanyang Technological University

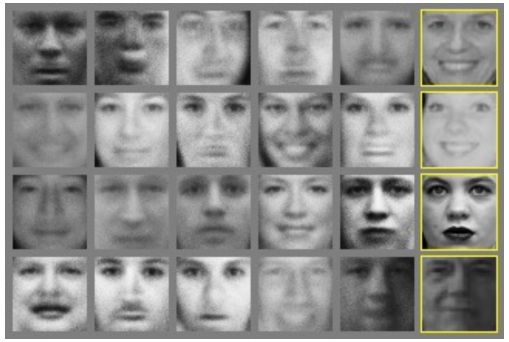
About Me

- Ziqi Huang 黄子琪
- Ph.D. student at MMLab@NTU
  - supervised by Prof. Ziwei Liu
  - Nanyang Technological University (NTU)
  - generative models, visual generation and manipulation
- Undergraduate
  - 2018-2022
  - Nanyang Technological University (NTU)

Overview

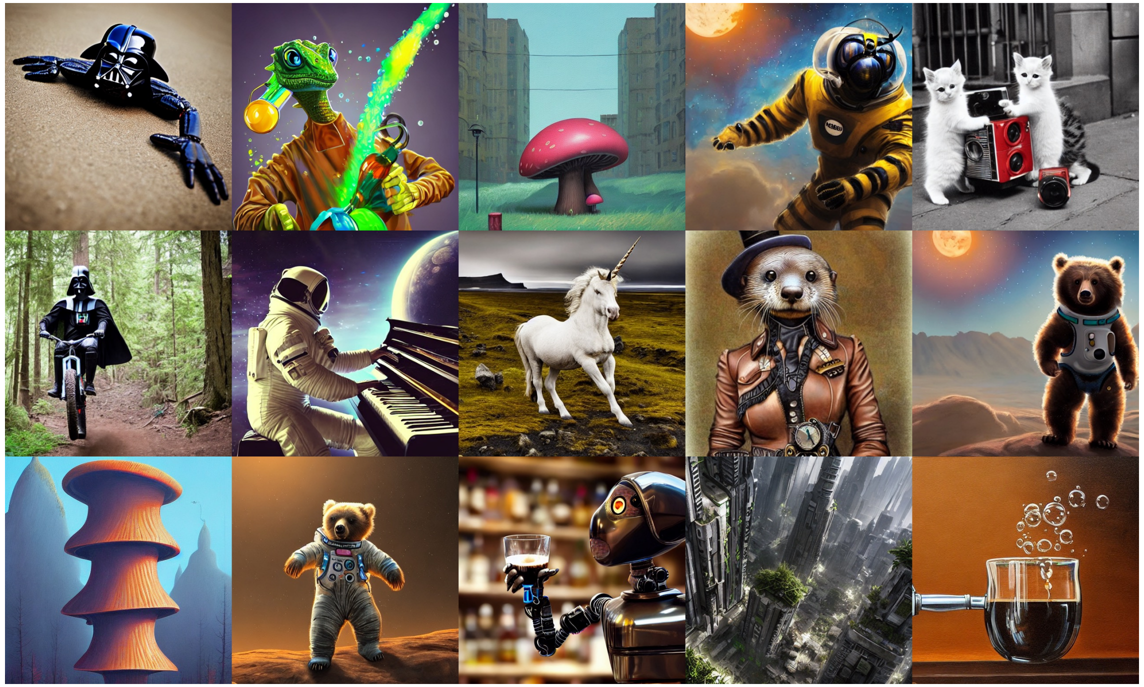
- Background: Generative AI, Diffusion Models
- Collaborative Diffusion for Multi-Modal Face Generation and Editing (CVPR 2023)
- Recent Works

Generative AI



GAN (2014)

StyleGAN2 (2020)



Stable Diffusion (2022)

Diffusion Models

Gradually denoise to image

Image  Noise

$\mathbf{x}_0$ $\mathbf{x}_{t-1}$ $\mathbf{x}_t$ $\mathbf{x}_T$

gradually adds Gaussian noise to the data

- Deep Unsupervised Learning using Nonequilibrium Thermodynamics (ICML 2015)
- Denoising Diffusion Probabilistic Models (NeurIPS 2020)
- Score-based generative modeling through stochastic differential equations (ICLR 2021)
- Diffusion Models Beat GANs on Image Synthesis (NeurIPS 2021)

Image Credit: CVPR 2022 Tutorial: Denoising Diffusion-based Generative Modeling: Foundations and Applications

Forward Process / Diffusion Process

Image  Noise

$\mathbf{x}_0$ $\mathbf{x}_{t-1}$ $\mathbf{x}_t$ $\mathbf{x}_T$

gradually adds Gaussian noise to the data

$$q(\mathbf{x}_{1:T}|\mathbf{x}_0) := \prod_{t=1}^{T} q(\mathbf{x}_t|\mathbf{x}_{t-1}),$$

$$q(\mathbf{x}_t|\mathbf{x}_{t-1}) := \mathcal{N}(\mathbf{x}_t; \sqrt{1-\beta_t}\mathbf{x}_{t-1}, \beta_t\mathbf{I}).$$

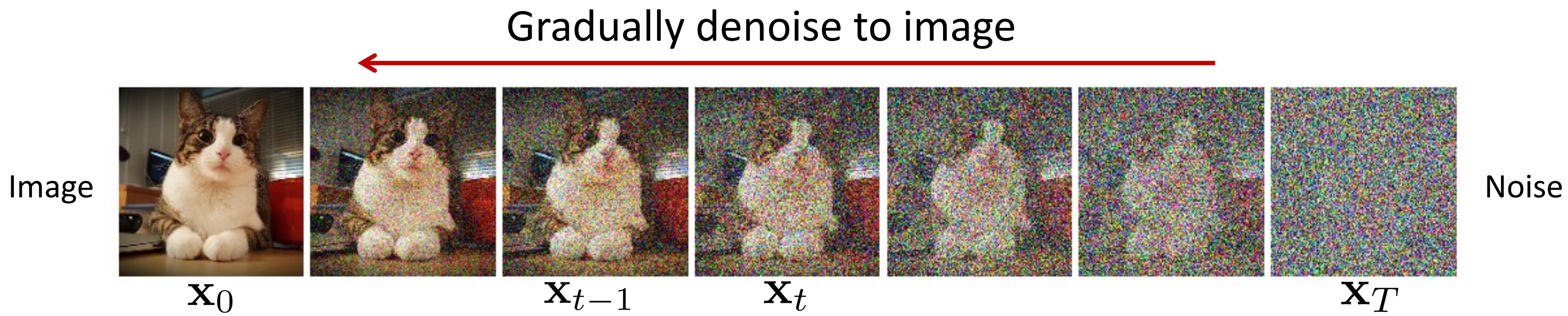
Direct sampling: $q(\mathbf{x}_t|\mathbf{x}_0) = \mathcal{N}(\mathbf{x}_t; \sqrt{\bar{\alpha}_t}\mathbf{x}_0, (1-\bar{\alpha}_t)\mathbf{I})$

$\bar{\alpha}_t := \prod_{s=1}^{t} \alpha_s$ and $\alpha_t := 1 - \beta_t$

$$\mathbf{x}_t(\mathbf{x}_0, \boldsymbol{\epsilon}) = \sqrt{\bar{\alpha}_t}\mathbf{x}_0 + \sqrt{1-\bar{\alpha}_t}\boldsymbol{\epsilon} \text{ for } \boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$$

Image Credit: CVPR 2022 Tutorial: Denoising Diffusion-based Generative Modeling: Foundations and Applications

Reverse Process (Generation)

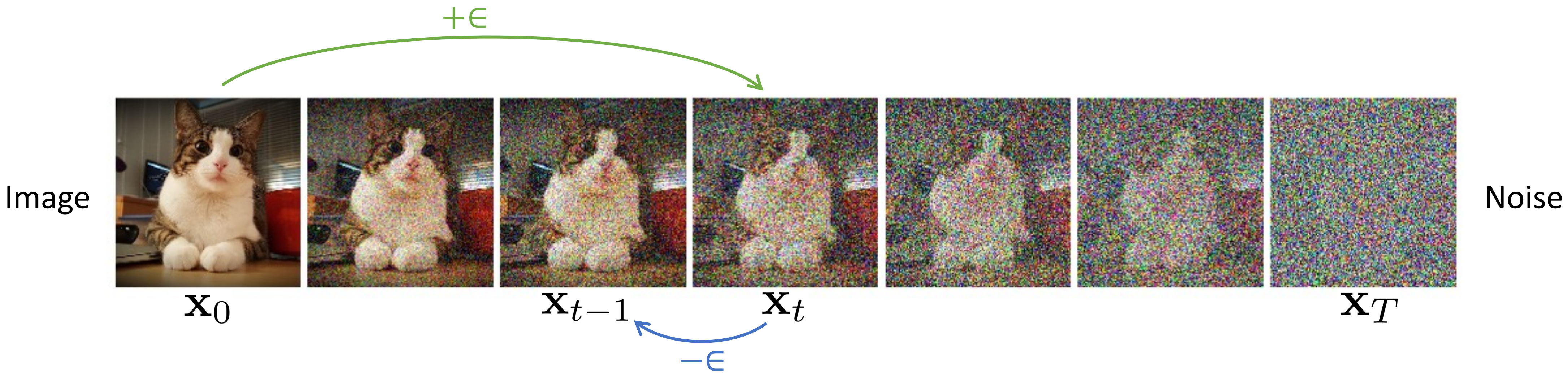


$$p_\theta(\mathbf{x}_{0:T}) \coloneqq p(\mathbf{x}_T) \prod_{t=1}^{T} p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t),$$

$$p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t) \coloneqq \mathcal{N}(\mathbf{x}_{t-1}; \boldsymbol{\mu}_\theta(\mathbf{x}_t, t), \boldsymbol{\Sigma}_\theta(\mathbf{x}_t, t))$$

Image Credit: CVPR 2022 Tutorial: Denoising Diffusion-based Generative Modeling: Foundations and Applications

Training & Sampling



Algorithm 1 Training

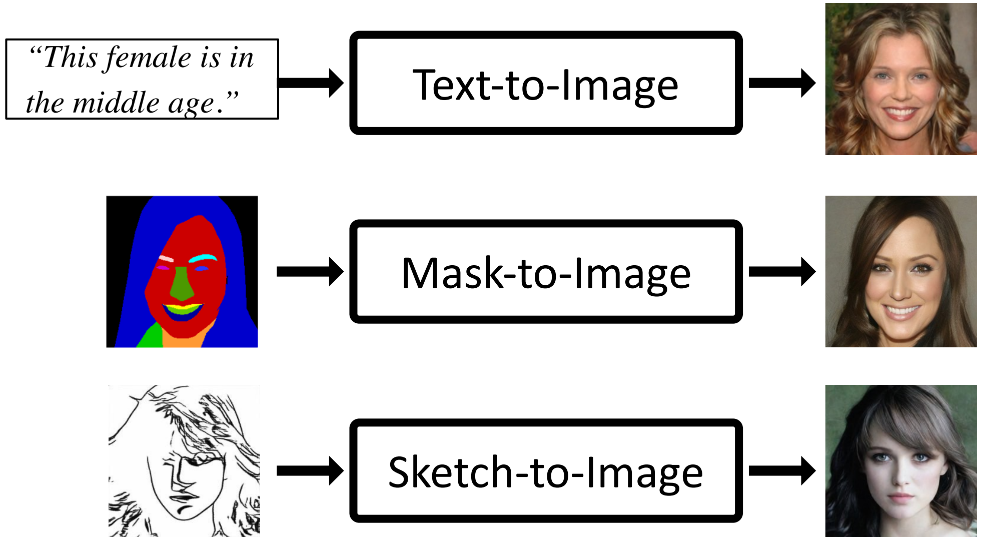
1: **repeat**
2: $\quad \mathbf{x}_0 \sim q(\mathbf{x}_0)$
3: $\quad t \sim \mathrm{Uniform}(\{1, \ldots, T\})$
4: $\quad \boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$
5: $\quad$ Take gradient descent step on
$$\nabla_\theta \left\| \boldsymbol{\epsilon} - \boldsymbol{\epsilon}_\theta\left(\boxed{\sqrt{\bar{\alpha}_t}\mathbf{x}_0 + \sqrt{1-\bar{\alpha}_t}\boldsymbol{\epsilon}}, t\right) \right\|^2$$
6: **until** converged

Algorithm 2 Sampling

1: $\mathbf{x}_T \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$
2: **for** $t = T, \ldots, 1$ **do**
3: $\quad \mathbf{z} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ if $t > 1$, else $\mathbf{z} = \mathbf{0}$
4: $\quad \mathbf{x}_{t-1} = \boxed{\frac{1}{\sqrt{\alpha_t}}\left(\mathbf{x}_t - \frac{1-\alpha_t}{\sqrt{1-\bar{\alpha}_t}}\boldsymbol{\epsilon}_\theta(\mathbf{x}_t, t)\right)} + \sigma_t \mathbf{z}$
5: **end for**
6: **return** $\mathbf{x}_0$

Image Credit: CVPR 2022 Tutorial: Denoising Diffusion-based Generative Modeling: Foundations and Applications

Uni-Modal Diffusion Models



......

Task Highlight

(A) Multi-Modal Face Generation

given multi-modal controls

"This female is in the middle age."

……

synthesize high-quality image consistent with the controls

Task Highlight

(B) Multi-Modal Face Editing

given input image

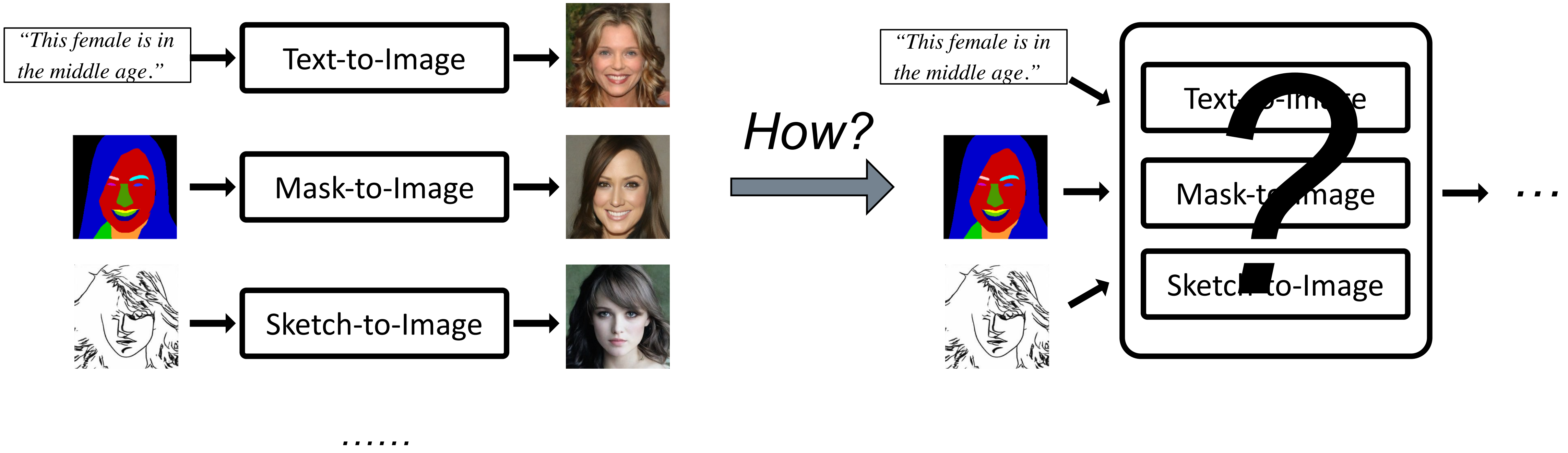
and target multi-modal conditions

“This man has beard of medium length. He is in his thirties.”

……

edit the image to 1) satisfy the target conditions while 2) preserving the facial identity

Multi-Modal Control



S-LAB FOR ADVANCED INTELLIGENCE

NANYANG TECHNOLOGICAL UNIVERSITY SINGAPORE

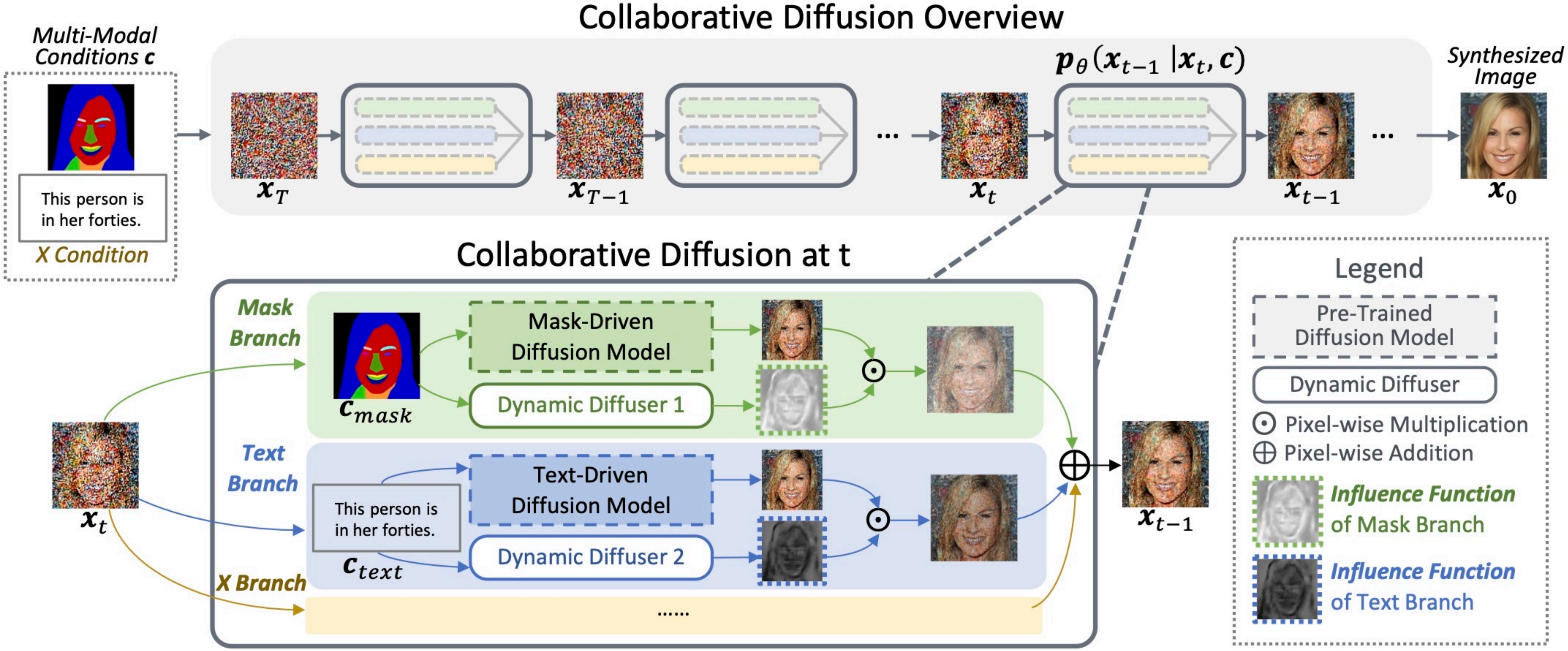
JUNE 18-22, 2023 CVPR VANCOUVER, CANADA

Collaborative Diffusion for Multi-Modal Face Generation and Editing

Ziqi Huang   Kelvin C.K. Chan   Yuming Jiang   Ziwei Liu

S-Lab, Nanyang Technological University

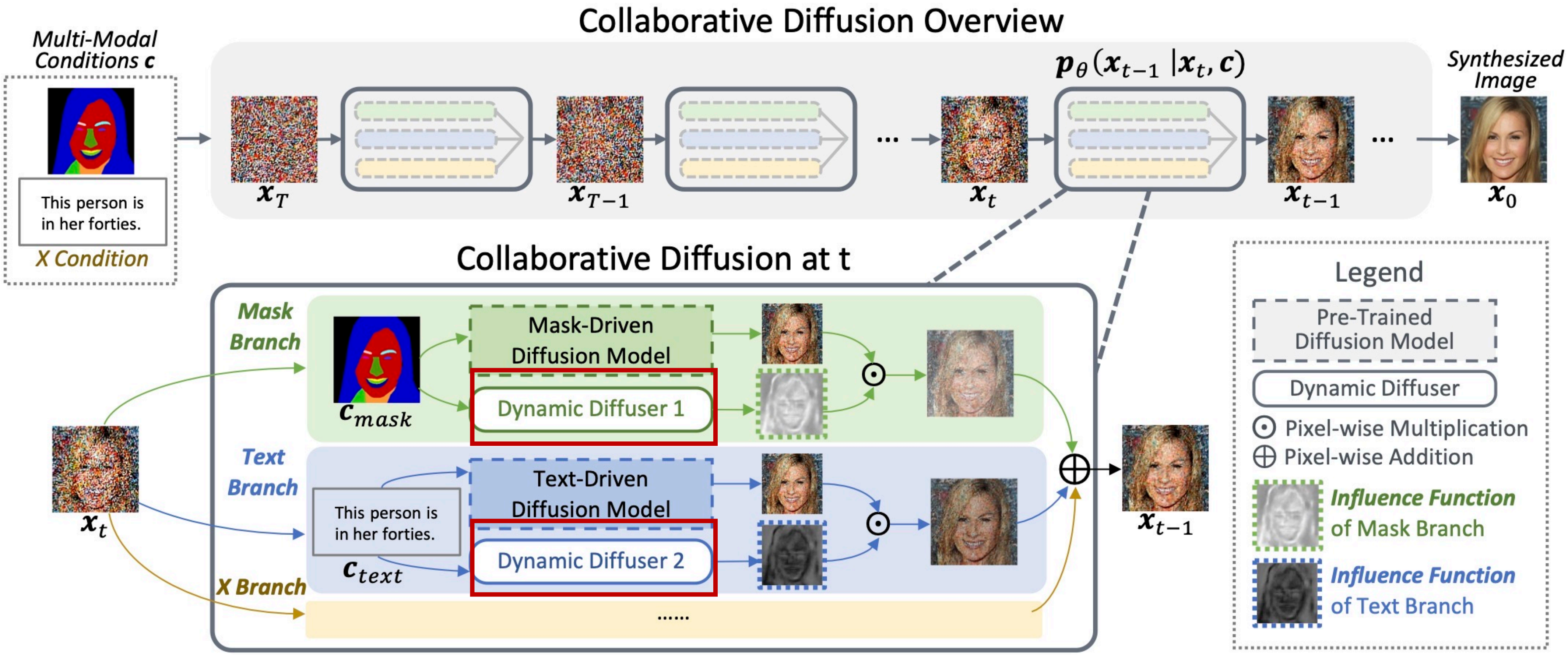
Collaborative Diffusion Framework



The framework consists of two components:

- **Collaborators**: pre-trained diffusion models (e.g. mask-driven, text-driven)
- **Dynamic Diffusers**: facilitate collaboration among different collaborators

Dynamic Diffuser



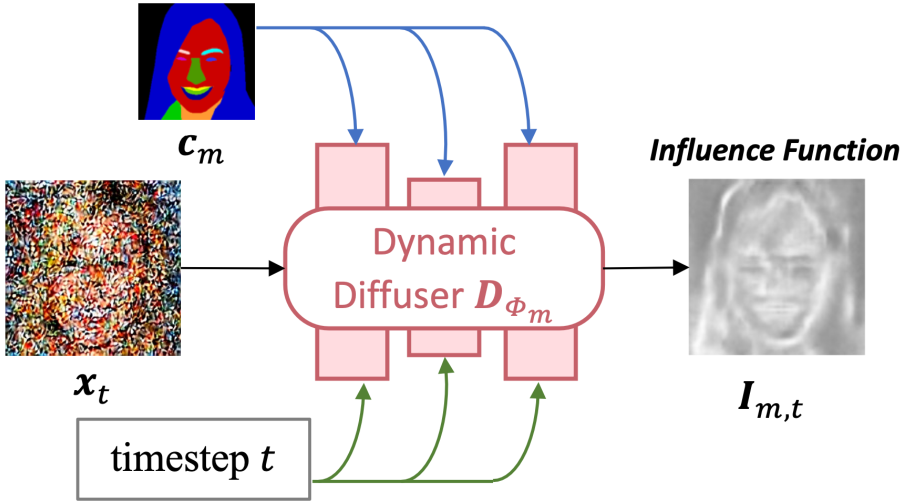
Collaborative Diffusion for Multi-Modal Face Generation and Editing (CVPR 2023)

Dynamic Diffuser

- *Dynamic Diffuser* predicts *Influence Functions* to determine <u>when, where, and how much each collaborator contributes</u>

$$\mathbf{I}_{m,t} = \mathbf{D}_{\phi_m}(\mathbf{x}_t, t, c_m)$$

$$\hat{\mathbf{I}}_{m,t,p} = \frac{\exp(\mathbf{I}_{m,t,p})}{\sum_{j=1}^{M} \exp(\mathbf{I}_{j,t,p})}$$

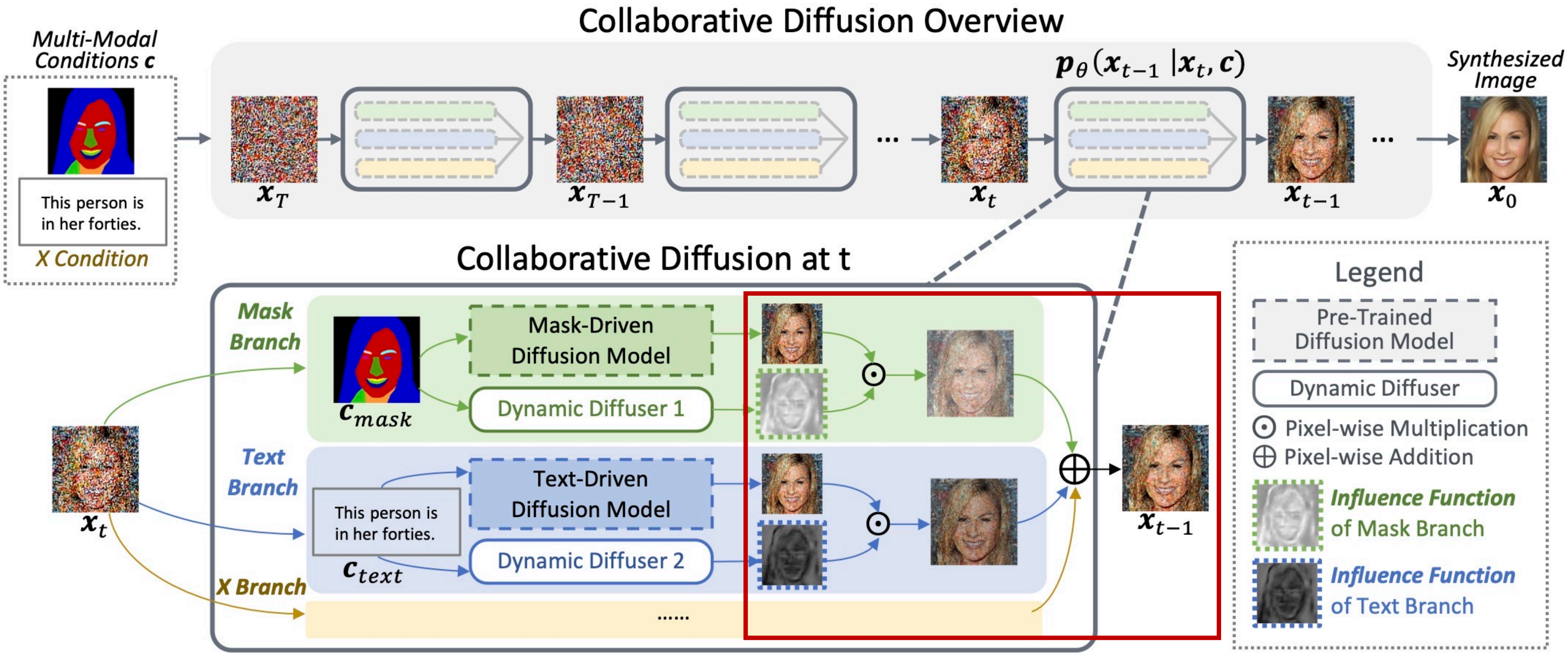


Dynamic Diffusers

- Dynamic Diffusers are lightweight.
- A dynamic diffuser is much smaller than a uni-modal conditional diffusion model.

Model Name	Number of Parameters
Mask-Driven Pre-trained Diffusion Model	403.6M
Text-Driven Pre-trained Diffusion Model	403.6M
Dynamic Diffuser for Mask Branch	13.1M
Dynamic Diffuser for Text Branch	13.1M

Multi-Modal Collaboration



Collaborative Diffusion for Multi-Modal Face Generation and Editing (CVPR 2023)

Multi-Modal Collaboration

- *Influence Functions* selectively enhance or suppress the contributions of the given modalities at each iterative step

$$\boldsymbol{\epsilon}_{pred,t} = \sum_{m=1}^{M} \hat{\mathbf{I}}_{m,t} \odot \boldsymbol{\epsilon}_{\theta_m}(x_t, t, c_m)$$

Algorithm: Training & Sampling

Algorithm 1 Dynamic Diffuser Training

repeat
  $\mathbf{x}_0, c_1, c_2, ..., c_M \sim q(\mathbf{x}_0, c_1, c_2, ..., c_M)$
  $t \sim \text{Uniform}(\{1, \ldots, T\})$
  $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$
  for $m = 1, ..., M$ **do**
    $\boxed{\boldsymbol{\epsilon}_{pred,m,t} = \boldsymbol{\epsilon}_{\theta_m}(\sqrt{\bar{\alpha}_t}\mathbf{x}_0 + \sqrt{1-\bar{\alpha}_t}\boldsymbol{\epsilon}, t, c_m)}$
    $\boxed{\mathbf{I}_{m,t} = \mathbf{D}_{\phi_m}(\sqrt{\bar{\alpha}_t}\mathbf{x}_0 + \sqrt{1-\bar{\alpha}_t}\boldsymbol{\epsilon}, t, c_m)}$
  end for
  $\hat{\mathbf{I}}_{m,t,p} = \frac{\exp(\mathbf{I}_{m,t,p})}{\sum_{j=1}^{M}\exp(\mathbf{I}_{j,t,p})}$, softmax at each pixel p
  $\boxed{\boldsymbol{\epsilon}_{pred,t} = \sum_{m=1}^{M} \hat{\mathbf{I}}_{m,t} \odot \boldsymbol{\epsilon}_{pred,m,t}}$
  Take gradient descent step on
    $\nabla_\phi \left\| \boldsymbol{\epsilon} - \boldsymbol{\epsilon}_{pred,t} \right\|^2$ where $\phi = \{\phi_m | m = 1, ..., M\}$
until converged

Pre-Trained Uni-Modal DM

Multi-Modal Collaboration

Algorithm 2 Collaborative Sampling

$\mathbf{x}_T \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$
for $t = T, \ldots, 1$ **do**
  $\mathbf{z} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ if $t > 1$, else $\mathbf{z} = \mathbf{0}$
  for $m = 1, ..., M$ **do**
    $\boxed{\boldsymbol{\epsilon}_{pred,m,t} = \boldsymbol{\epsilon}_{\theta_m}(\mathbf{x}_t, t, c_m)}$
    $\boxed{\mathbf{I}_{m,t} = \mathbf{D}_{\phi_m}(\mathbf{x}_t, t, c_m)}$
  end for
  $\hat{\mathbf{I}}_{m,t,p} = \frac{\exp(\mathbf{I}_{m,t,p})}{\sum_{j=1}^{M}\exp(\mathbf{I}_{j,t,p})}$, softmax at each pixel p
  $\boxed{\boldsymbol{\epsilon}_{pred,t} = \sum_{m=1}^{M} \hat{\mathbf{I}}_{m,t} \odot \boldsymbol{\epsilon}_{pred,m,t}}$
  $\mathbf{x}_{t-1} = \frac{1}{\sqrt{\alpha_t}}\left(\mathbf{x}_t - \frac{1-\alpha_t}{\sqrt{1-\bar{\alpha}_t}}\boldsymbol{\epsilon}_{pred,t}\right) + \sigma_t \mathbf{z}$
end for
return $\mathbf{x}_0$

Dynamic Diffusers predict *Influence Functions*

Collaborative Diffusion for Multi-Modal Face Generation and Editing (CVPR 2023)

Algorithm: Editing

Algorithm 3 Collaborative Editing

Require:
input image $\mathbf{x}_{input}$, target conditions $c_{m,target}$, diffusion models $\boldsymbol{\epsilon}_{\theta_m}$, dynamic diffusers $\mathbf{D}_{\phi_m}$, $(m = 1, \ldots, M)$,
interpolation scale α

for $m = 1, ..., M$ **do** ▷ Uni-Modal Editing
  $c_m = c_{m,target}$
  $\mathbf{x}_t = \sqrt{\bar{\alpha}_t}\mathbf{x}_{input} + \sqrt{1 - \bar{\alpha}_t}\boldsymbol{\epsilon}$
  $c_{m,opt} = \text{argmin}_{c_m} \mathbb{E}_{\boldsymbol{\epsilon},t} \|\boldsymbol{\epsilon} - \boldsymbol{\epsilon}_{\theta_m}(\mathbf{x}_t, t, c_m)\|^2$
  $\theta_{m,opt} = \text{argmin}_{\theta_m} \mathbb{E}_{\boldsymbol{\epsilon},t} \|\boldsymbol{\epsilon} - \boldsymbol{\epsilon}_{\theta_m}(\mathbf{x}_t, t, c_{m,opt})\|^2$
  $c_{m,int} = \alpha \cdot c_{m,target} + (1 - \alpha) \cdot c_{m,opt}$
end for
$\mathbf{x}_T \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ ▷ Collaborate the Uni-Modal Edits
for $t = T, \ldots, 1$ **do**
  $\mathbf{z} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ if $t > 1$, else $\mathbf{z} = \mathbf{0}$
  for $m = 0, ..., M$ **do** Pre-Trained Uni-Modal DM
    $\boxed{\boldsymbol{\epsilon}_{pred,m,t} = \boldsymbol{\epsilon}_{\theta_{m,opt}}(\mathbf{x}_t, t, c_{m,int})}$
    $\boxed{\mathbf{I}_{m,t} = \mathbf{D}_{\phi_m}(\mathbf{x}_t, t, c_{m,int})}$ *Dynamic Diffusers predict Influence Functions*
  end for
  $\hat{\mathbf{I}}_{m,t,p} = \frac{\exp(\mathbf{I}_{m,t,p})}{\sum_{j=1}^{M} \exp(\mathbf{I}_{j,t,p})}$, softmax at each pixel p
  $\boxed{\boldsymbol{\epsilon}_{pred,t} = \sum_{m=1}^{M} \hat{\mathbf{I}}_{m,t} \odot \boldsymbol{\epsilon}_{pred,m,t}}$ Multi-Modal Collaboration
  $\mathbf{x}_{t-1} = \frac{1}{\sqrt{\alpha_t}}\left(\mathbf{x}_t - \frac{1-\alpha_t}{\sqrt{1-\bar{\alpha}_t}}\boldsymbol{\epsilon}_{pred,t}\right) + \sigma_t \mathbf{z}$
end for
return $\mathbf{x}_0$

Imagic

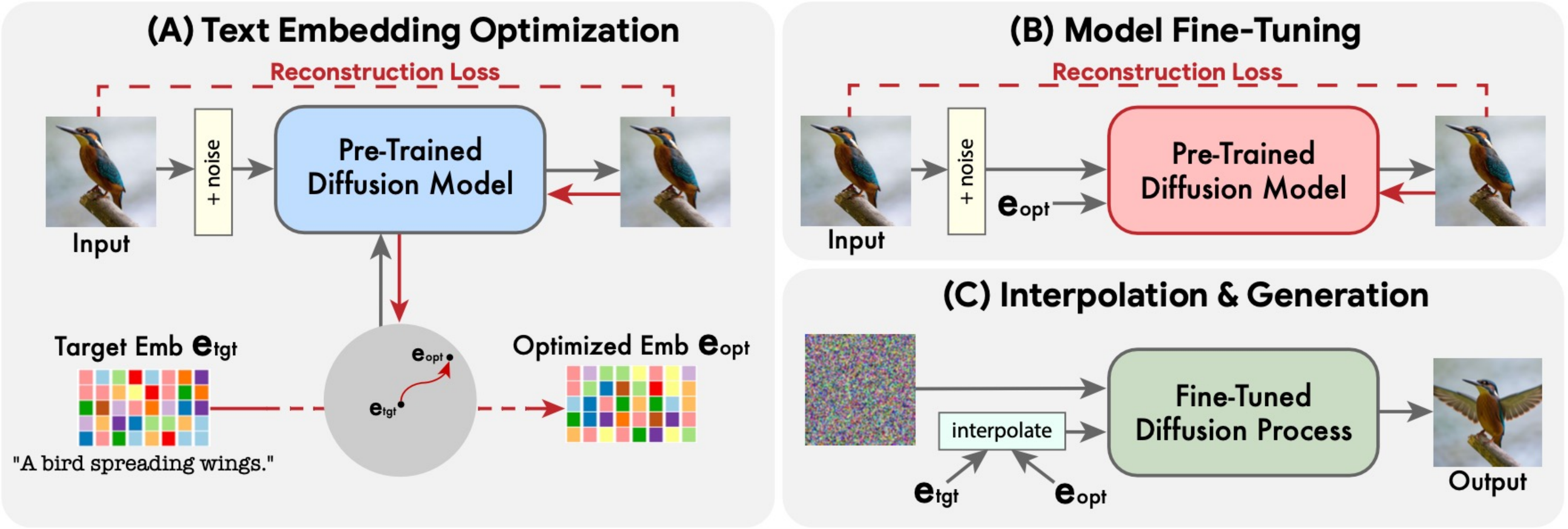
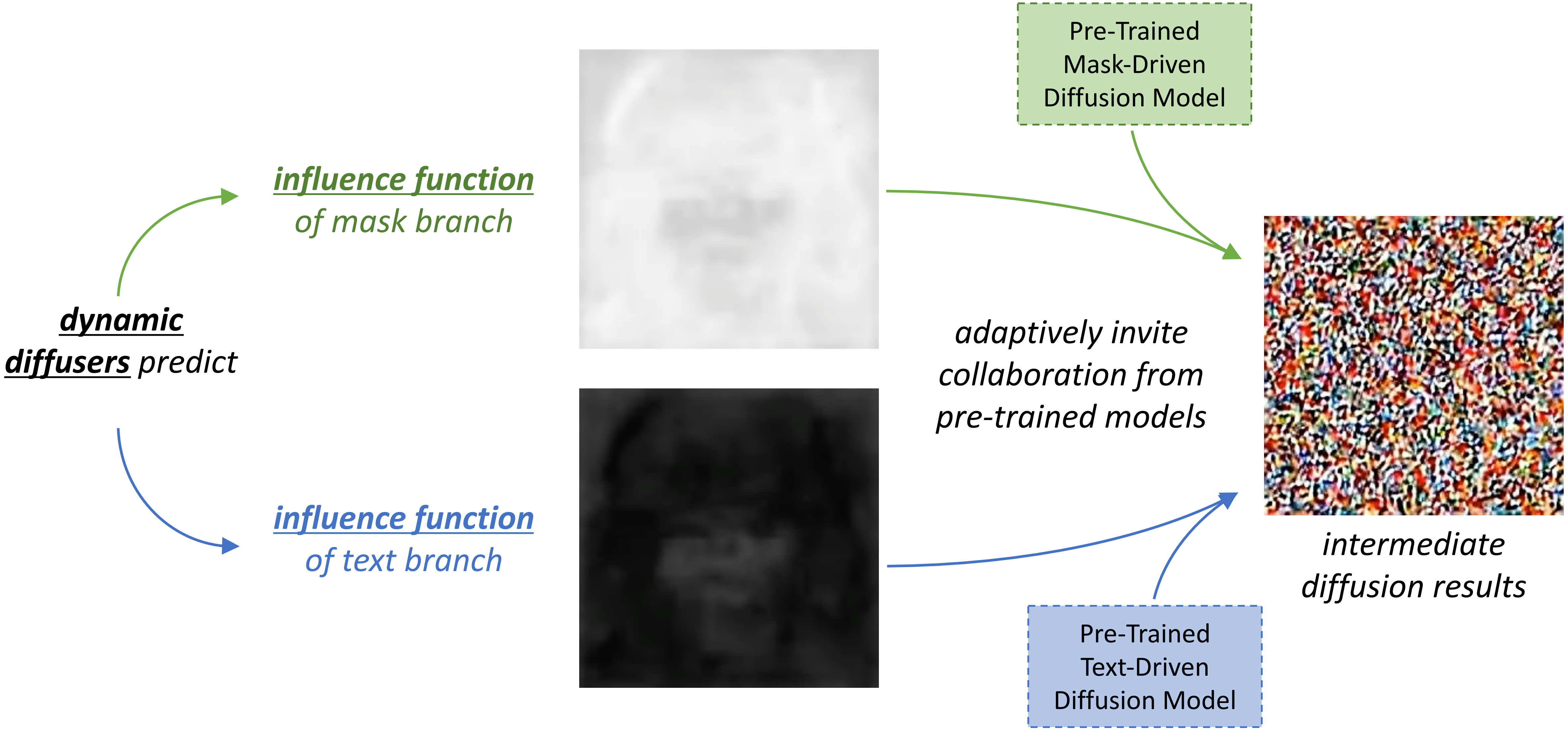


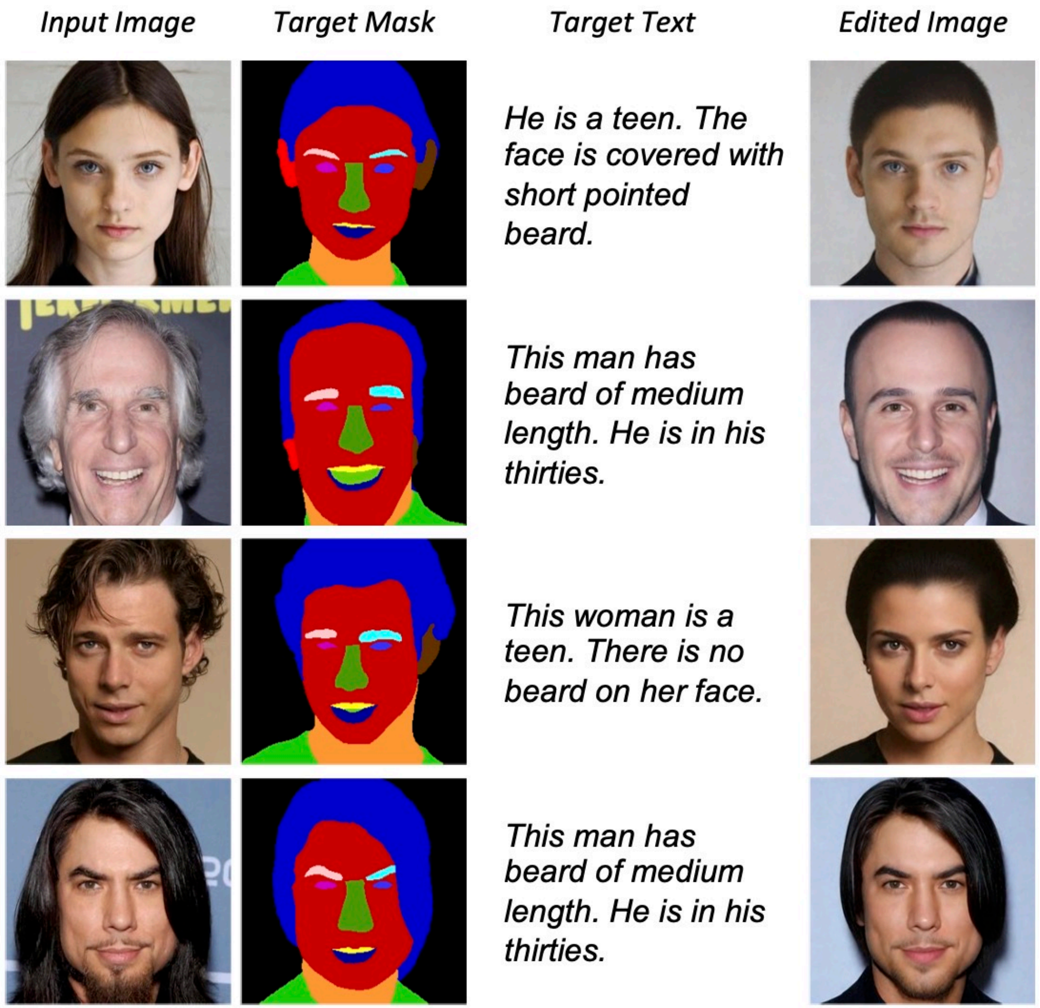
Figure 3. **Schematic description of *Imagic*.** *Given a real image and a target text prompt: (A) We encode the target text and get the initial text embedding* $\mathbf{e}_{tgt}$*, then optimize it to reconstruct the input image, obtaining* $\mathbf{e}_{opt}$*; (B) We then fine-tune the generative model to improve fidelity to the input image while fixing* $\mathbf{e}_{opt}$*; (C) Finally, we interpolate* $\mathbf{e}_{opt}$ *with* $\mathbf{e}_{tgt}$ *to generate the final editing result.*

Dynamic Diffusers predict Influence Functions



Visual Results

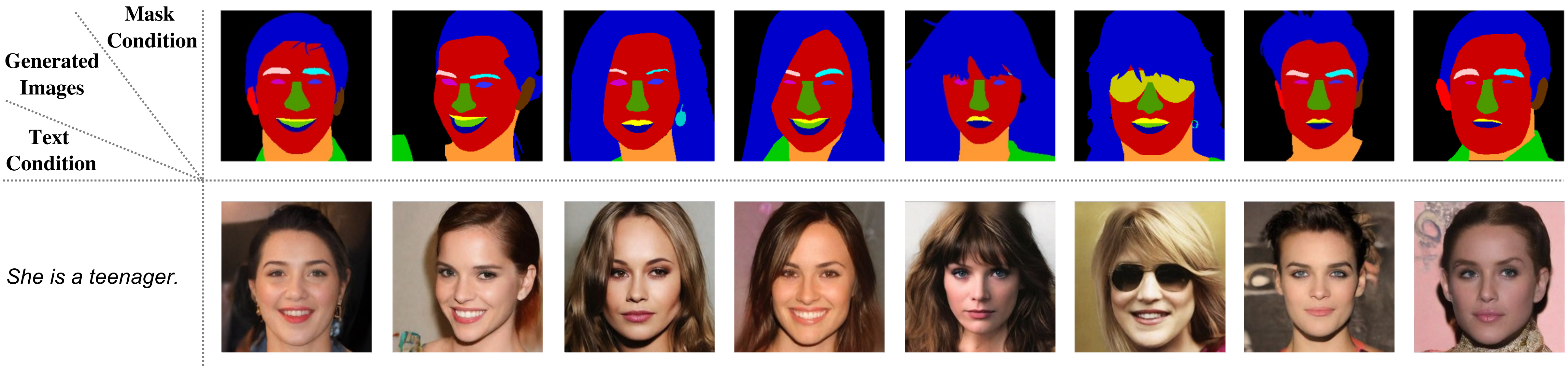
Face Generation



Face Editing


Mask Condition
Generated Images
Text Condition
He is a teen. The face is covered with short pointed beard.
He looks very old. He doesn't have any mustache at all.
She is a teenager.
This female is in the middle age.
This man has beard of medium length. He is in his thirties.
This woman looks very old.

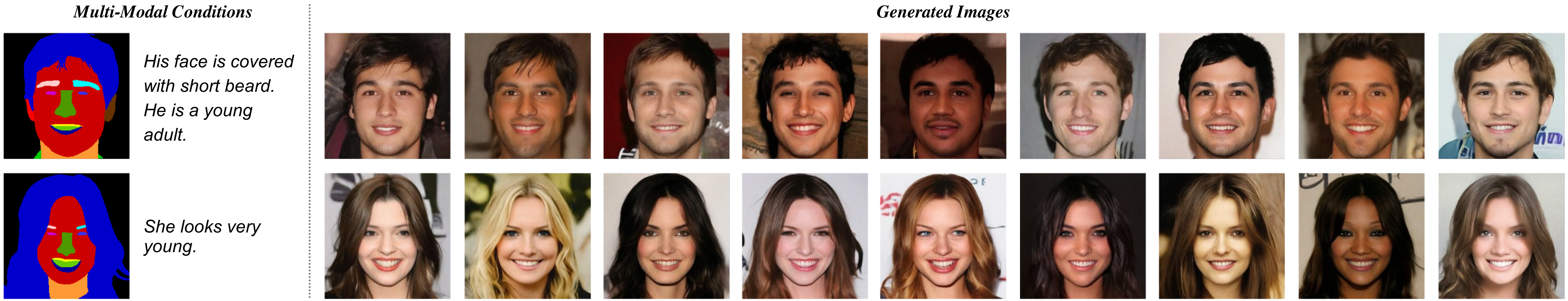
Visual Results: Generation



Collaborative Diffusion for Multi-Modal Face Generation and Editing (CVPR 2023)

Visual Results: Generation

Diversity of Synthesis Results



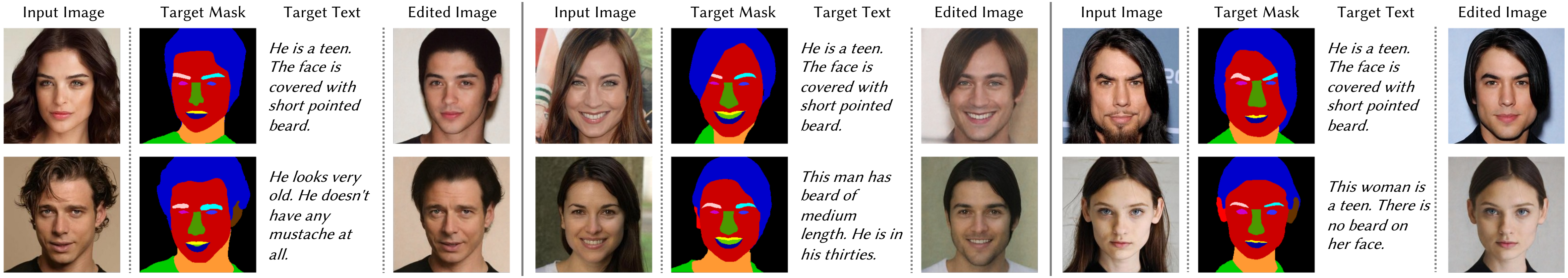
Quantitative Results of Face Generation

NANYANG TECHNOLOGICAL UNIVERSITY SINGAPORE | S-LAB FOR ADVANCED INTELLIGENCE

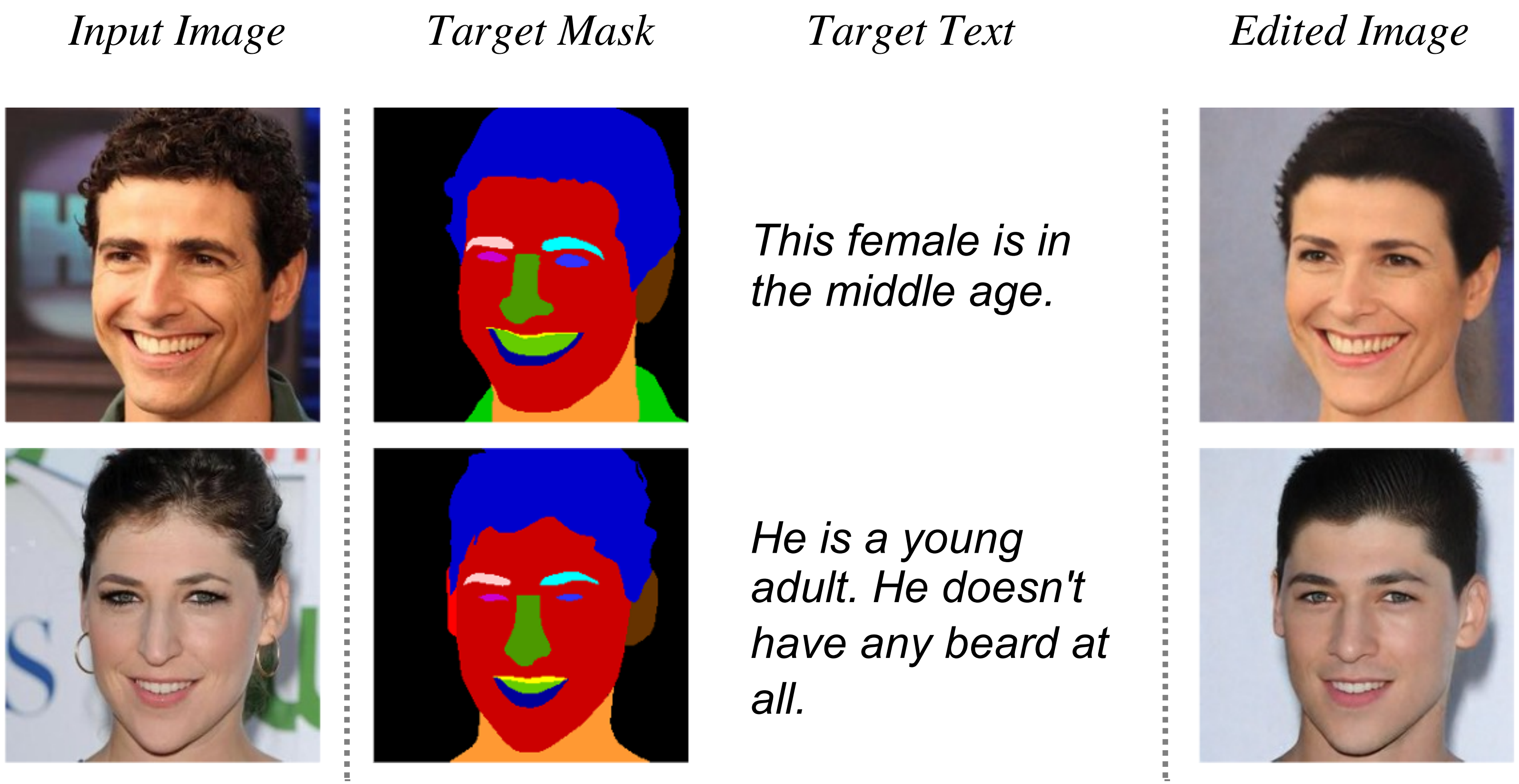
- Our method synthesizes images with better quality (lower FID), and higher consistency with the text and mask conditions.

Method	FID $\downarrow$	Text (%) $\uparrow$	Mask (%) $\uparrow$
TediGAN [74, 75]	157.81	24.27	72.19
Composable [41]	124.62	23.94	76.11
Ours	**111.36**	**24.51**	**80.25**

Visual Results: Editing

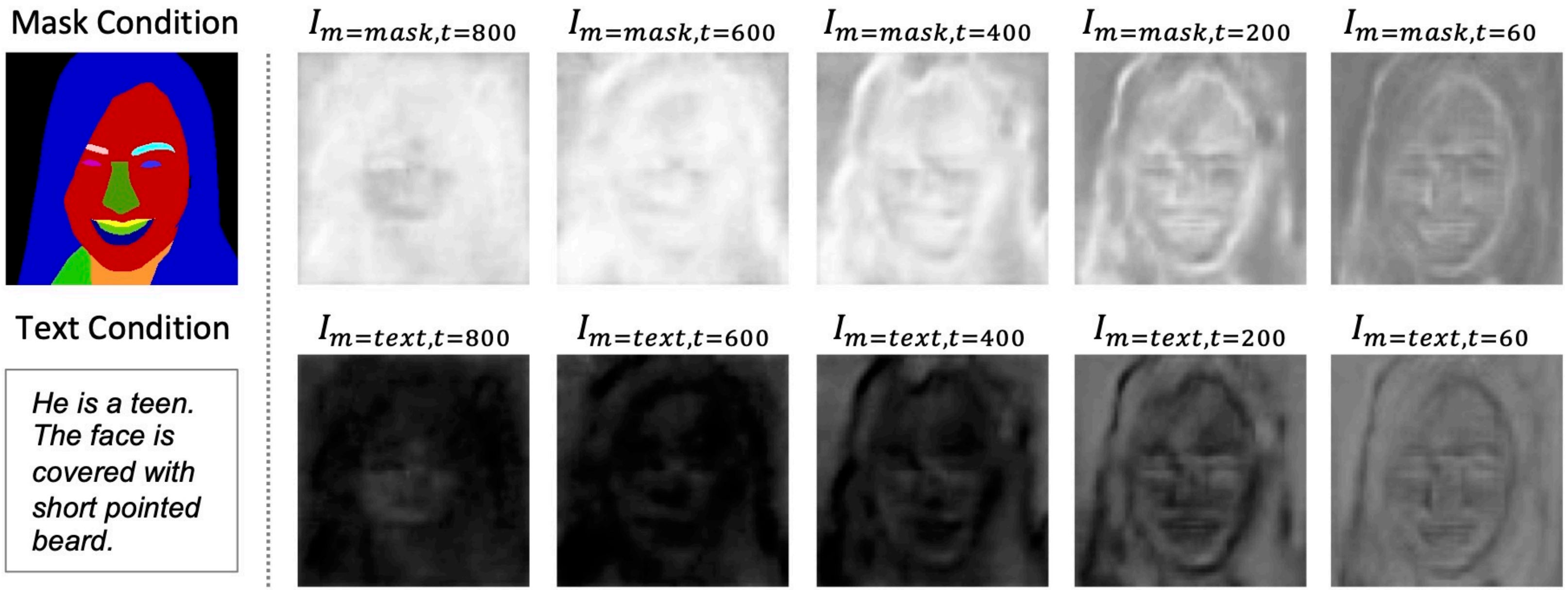


Visual Results: Editing



Observation on **Influence Functions**

- **Spatial Variations**:
  - Mask-to-image model: contours
  - Text-to-image model: skin textures and details
- **Temporal Variations**: Layout first, details later



Ablation Study

- Temporal or spatial suppression in influence variation introduces performance drops, which shows the necessity of influence functions' spatial-temporal adaptivity.

Method	**FID** $\downarrow$	**Text (%)** $\uparrow$	**Mask (%)** $\uparrow$
Ours w/o Spatial	117.81	24.36	80.08
Ours w/o Temporal	117.34	24.48	77.07
Ours	**111.36**	**24.51**	**80.25**

Summary

- In ***Collaborative Diffusion***, pre-trained uni-modal diffusion models collaboratively achieve multi-modal face generation and editing without being re-trained.
- ***Dynamic diffuser*** predicts the spatial-temporal ***influence functions*** to selectively enhance or suppress the contributions from each collaborator.
- Both **quantitative and qualitative results demonstrate the superiority** of Collaborative Diffusion in multi-modal face generation and editing.
- Our Collaborative Diffusion framework could be used to extend ***arbitrary uni-modal approach*** (*e.g.*, conditional motion and 3D generation) to the multi-modal paradigm.

Future Works

- Handle conflicts in multi-modal input
- Collaborate other forms of diffusion models
- Video generation

Related Works

- Adding Conditional Control to Text-to-Image Diffusion Models
- T2I-Adapter: Learning Adapters to Dig out More Controllable Ability for Text-to-Image Diffusion Models.

Recent Explorations

Recent Works: ***Relation Inversion***

Input

Exemplar Images

“vegetable is contained inside bag”

Output

Relation Prompt

<R>

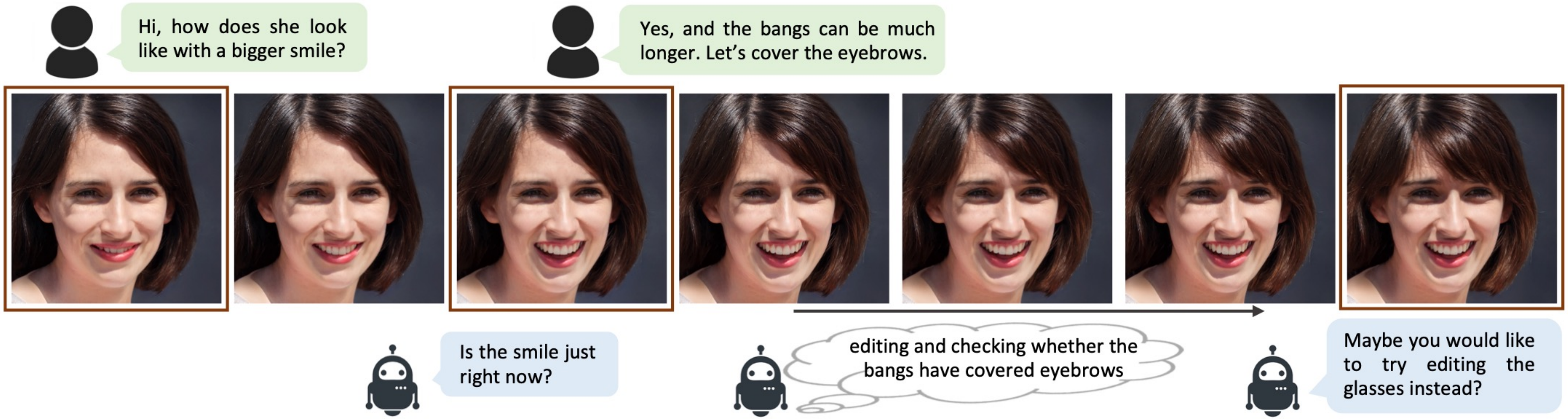
represent the co-existing relation in exemplar images

Application

Relation-Specific Text-to-Image Synthesis

“Spiderman <R> paper bag”

Recent Works: ***Talk-to-Edit***



Summary

- Human-Machine Collaborative
  - Multi-Modal Control
  - Multi-Round Interactions
- Future
  - Video Generation
  - Complexity & Quality & Controllability

S-LAB FOR ADVANCED INTELLIGENCE

NANYANG TECHNOLOGICAL UNIVERSITY SINGAPORE

JUNE 18-22, 2023 CVPR VANCOUVER, CANADA

Collaborative Diffusion
for Multi-Modal Face Generation and Editing

Paper: https://arxiv.org/abs/2304.10530

Code: https://github.com/ziqihuangg/Collaborative-Diffusion

Project Page: https://ziqihuangg.github.io/projects/collaborative-diffusion.html

Video: https://www.youtube.com/watch?v=inLK4c8sNhc

Q&A

Project Page *Code*