Learning to Generate Semantic Layouts for Higher Text-Image Correspondence in Text-to-Image Synthesis

ICCV23 PARIS

Minho Park* Jooyeol Yun* Seunghwan Choi Jaegul Choo
Korea Advanced Institute of Science and Technology (KAIST)

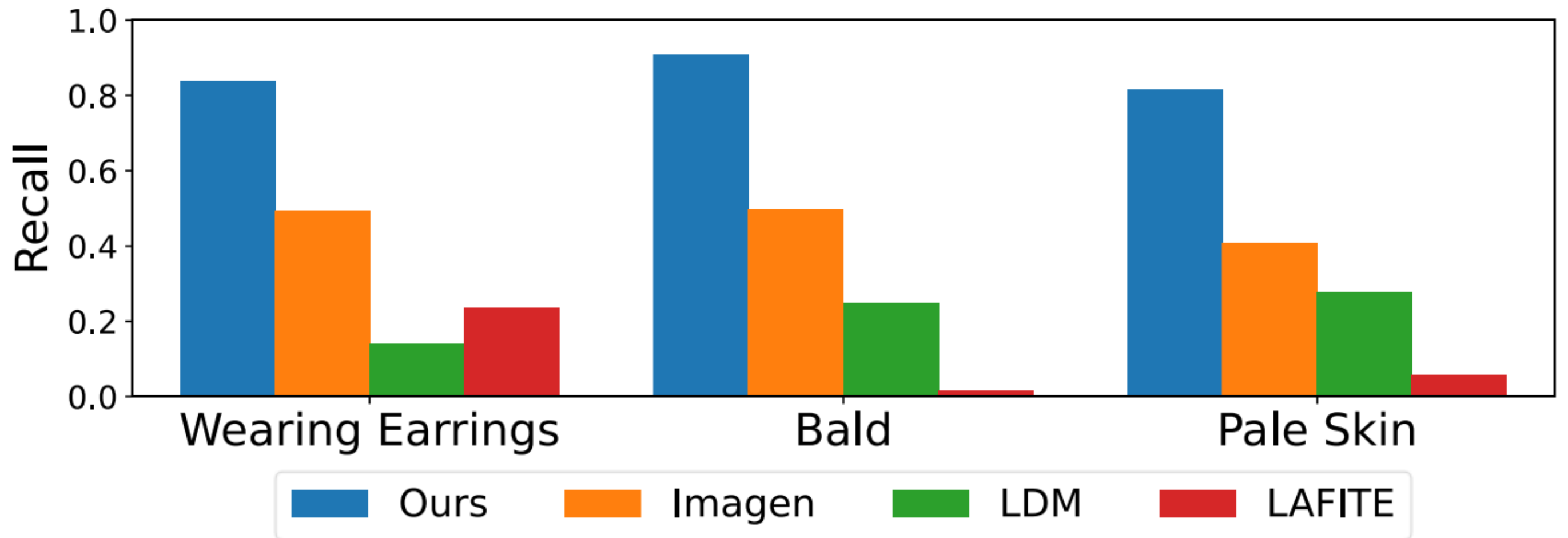
DAVIAN Data and Visual Analytics Lab

Project page

Contribution

- We define a ***Gaussian-categorical diffusion process*** for modeling joint image-layout distributions, which is the first approach to unify two diffusion processes for image-layout generation.
- Our experiments reveal that ***generating image-layout pairs can be a practical alternative to increase text-image correspondence*** in circumstances where collecting web-scale text-image pairs is infeasible.
- We present ***cross-modal outpainting***, which demonstrates that Gaussian-categorical diffusion models are also capable of modeling conditional distributions for semantic image synthesis and semantic segmentation.

Motivation



Recall of facial attributes specified in the text descriptions.

> “ ***Text-to-image generation approaches trained on small-scale dataset often fail to reflect text conditions.*** ”

Gaussian-categorical Diffusion Process

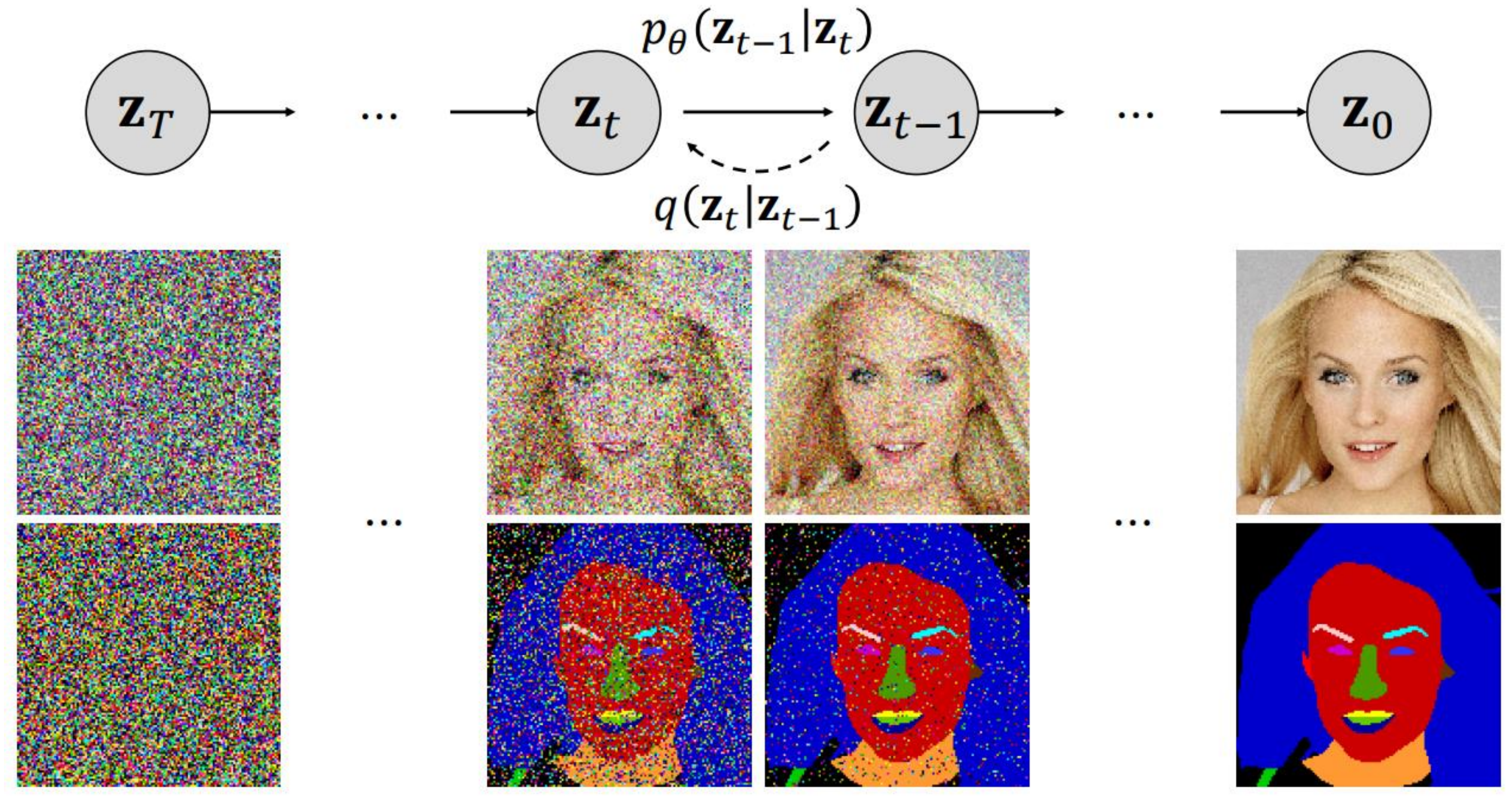
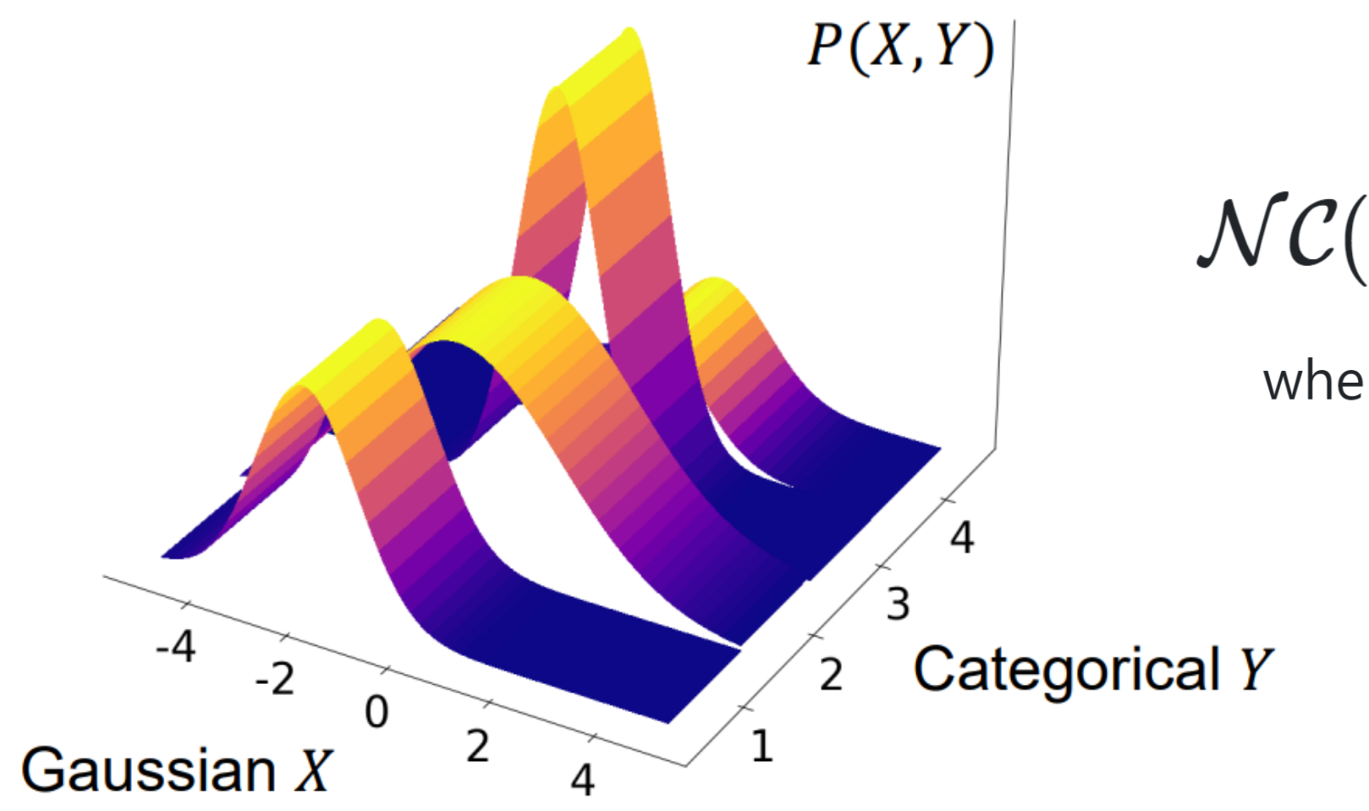

$p_\theta(\mathbf{z}_{t-1}|\mathbf{z}_t)$
$\mathbf{z}_T$
…
$\mathbf{z}_t$
$\mathbf{z}_{t-1}$
…
$\mathbf{z}_0$
$q(\mathbf{z}_t|\mathbf{z}_{t-1})$
…
…

Illustration of the Gaussian-categorical diffusion process
on the image-layout distribution of MM CelebA-HQ.

Gaussian-categorical distribution



$$\mathcal{NC}(\mathrm{x},\mathrm{y};\mu,\Sigma,\Theta)=\mathcal{C}(\mathrm{y};\Theta)\cdot\mathcal{N}(\mathrm{x};\mu_{\mathrm{y}},\Sigma_{\mathrm{y}})$$

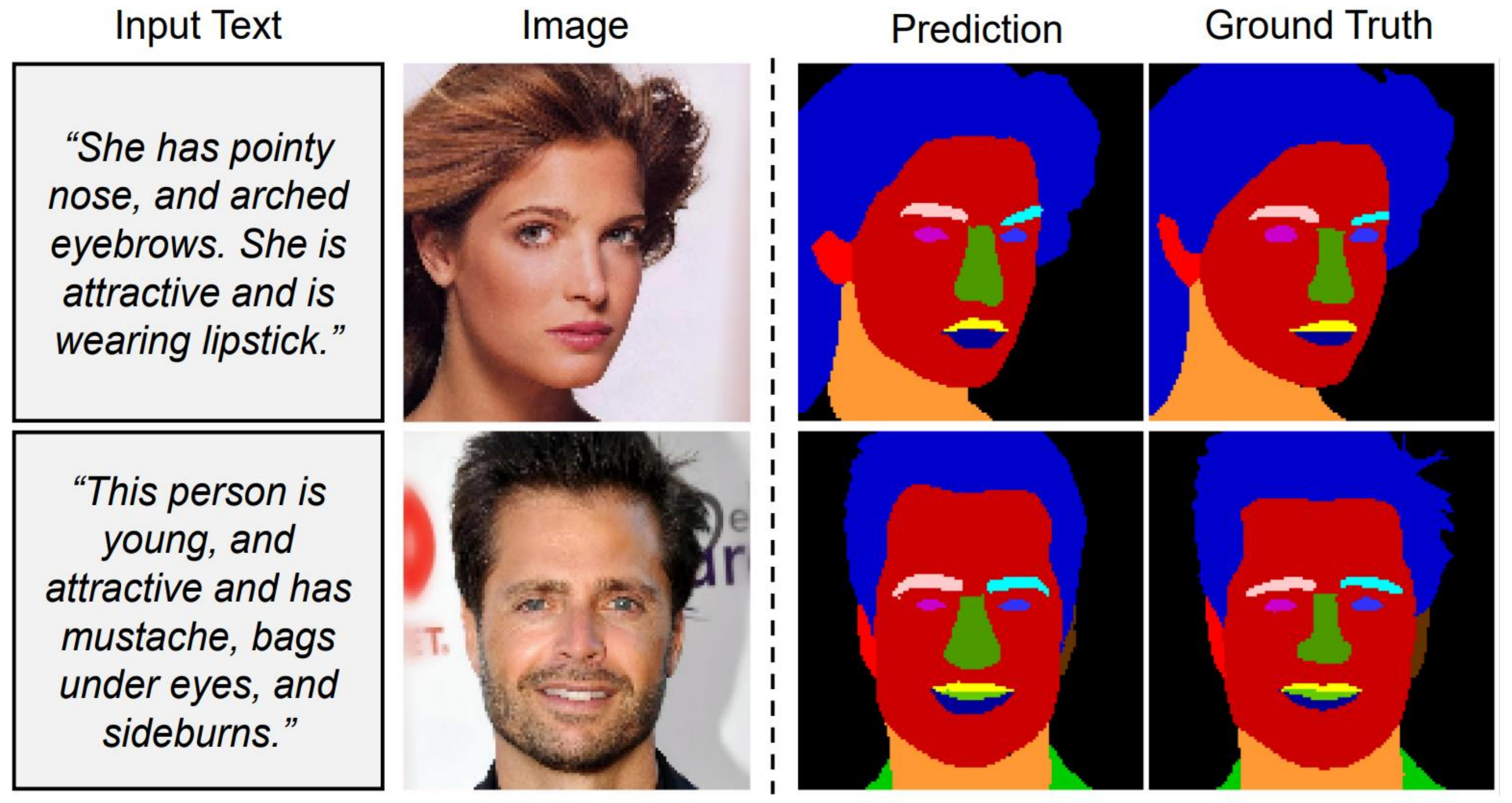
where $\mathrm{x}\in\mathbb{R}^N,\mathrm{y}\in\{1,2,\dots,K\}^M\subset\mathbb{R}^M$
$\mu\in\mathbb{R}^{S\times N},\Sigma\in\mathbb{R}^{S\times N\times N},\Theta\in\mathbb{R}^{M\times K},(S=K^M)$
$\mu_{\mathrm{y}}\in\mathbb{R}^N,\Sigma_{\mathrm{y}}\in\mathbb{R}^{N\times N}$

Final objective function

$$D_{KL}\big(q(\mathbf{z}_{t-1}|\mathbf{z}_t,\mathbf{z}_0)\parallel p_\theta(\mathbf{z}_{t-1}|\mathbf{z}_t)\big)=\underbrace{\mathbb{E}_q\left[\frac{1}{2\sigma_t^2}\|\tilde{\boldsymbol{\mu}}_t-\tilde{\boldsymbol{\mu}}_\theta(\mathbf{z}_t)\|^2\right]}_{\text{Gaussian Diffusion}}+\underbrace{D_{KL}\left(\tilde{\boldsymbol{\Theta}}_t\parallel\tilde{\boldsymbol{\Theta}}_\theta(\mathbf{z}_t)\right)}_{\text{Categorical Diffusion}}+C$$

* Detailed proofs for each step are provided in A.1.

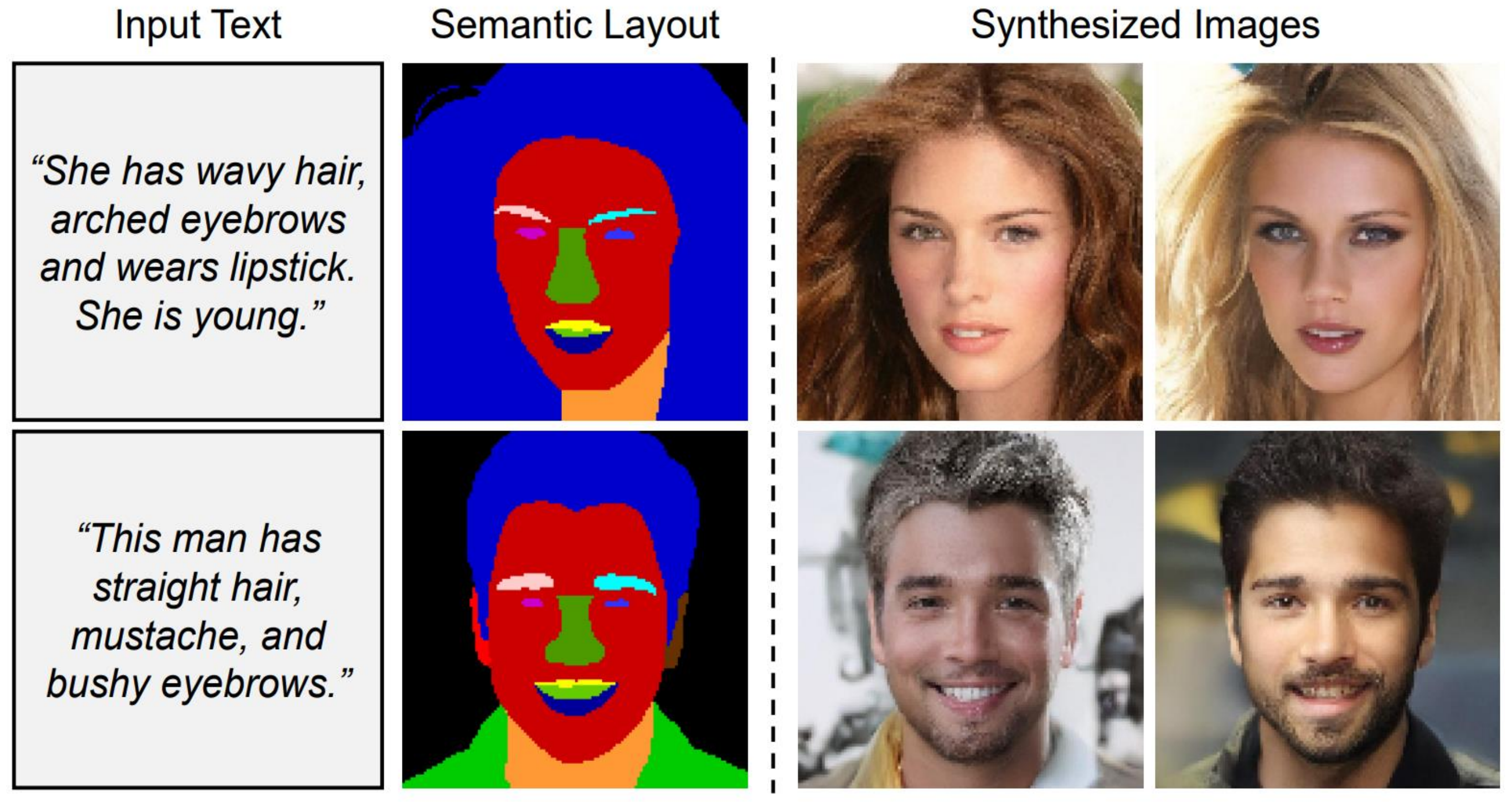
Cross-modal Outpainting



“She has pointy nose, and arched eyebrows. She is attractive and is wearing lipstick.”

“This person is young, and attractive and has mustache, bags under eyes, and sideburns.”

(a) Text-guided Image-to-Layout Generation



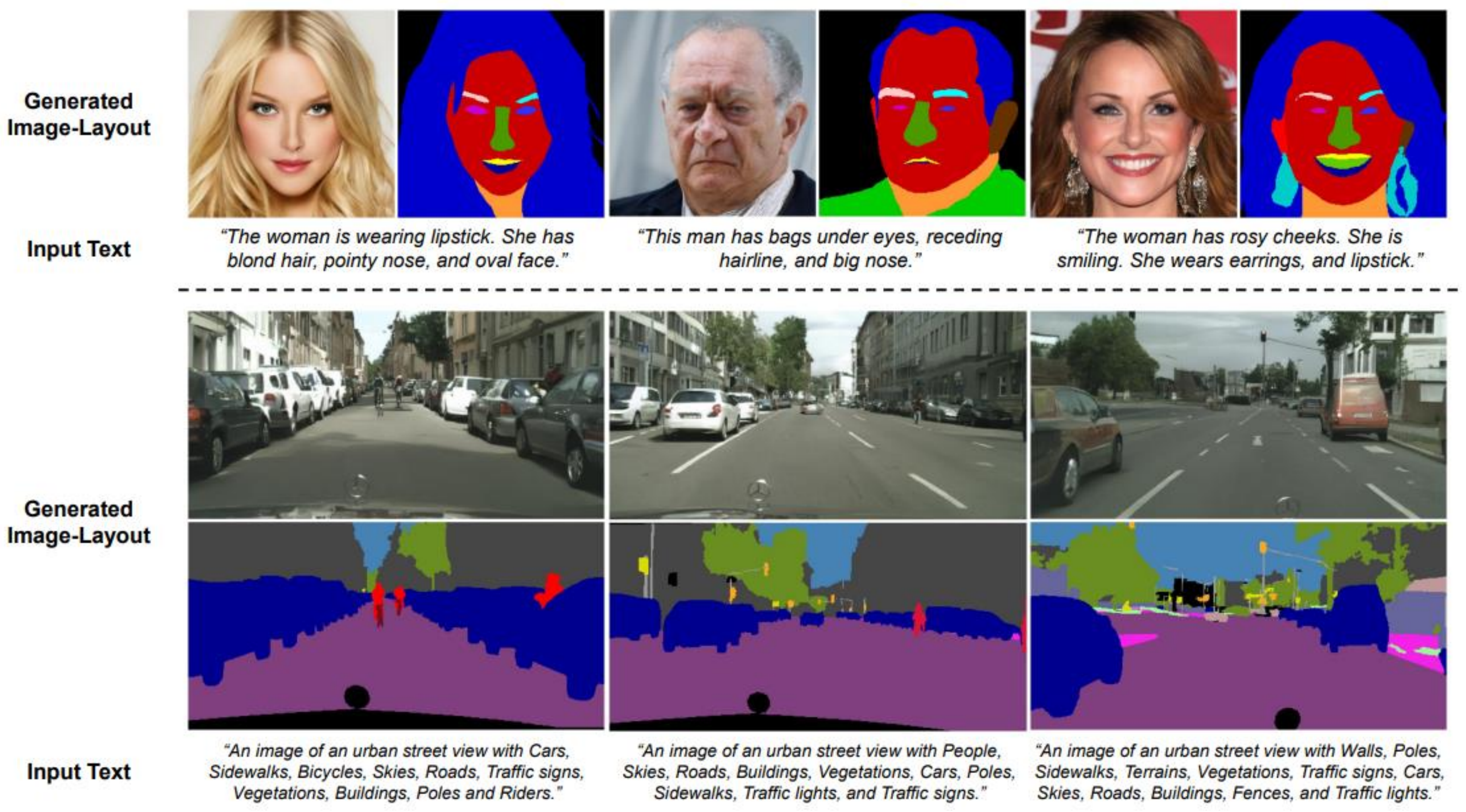
“She has wavy hair, arched eyebrows and wears lipstick. She is young.”

“This man has straight hair, mustache, and bushy eyebrows.”

(b) Text-guided Layout-to-Image Generation

> “ Since we model image-layout joint distribution, ***we can also model conditional distributions.*** ”

Qualitative Results



“The woman is wearing lipstick. She has blond hair, pointy nose, and oval face.”

“This man has bags under eyes, receding hairline, and big nose.”

“The woman has rosy cheeks. She is smiling. She wears earrings, and lipstick.”

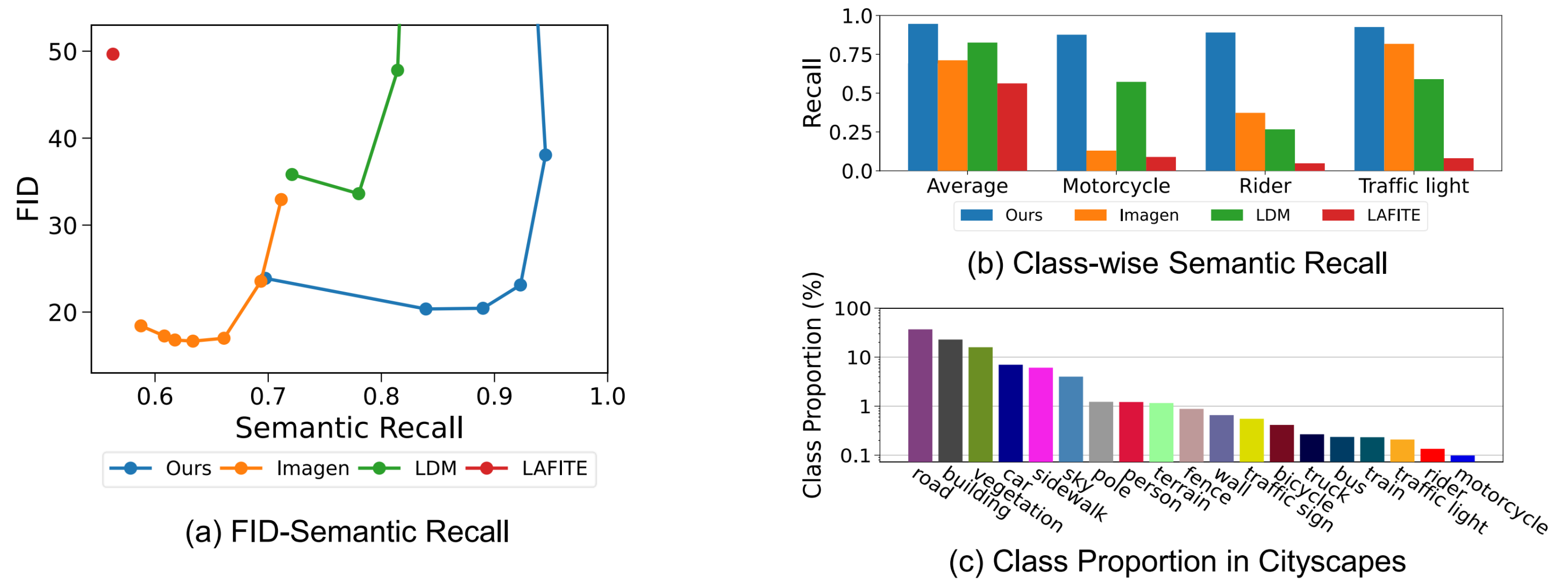
“An image of an urban street view with Cars, Sidewalks, Bicycles, Skies, Roads, Traffic signs, Vegetations, Buildings, Poles and Riders.”

“An image of an urban street view with People, Skies, Roads, Buildings, Vegetations, Cars, Poles, Sidewalks, Traffic lights, and Traffic signs.”

“An image of an urban street view with Walls, Poles, Sidewalks, Terrains, Vegetations, Traffic signs, Cars, Skies, Roads, Buildings, Fences, and Traffic lights.”

> “ GCDP generates ***aligned image-layout pairs*** from text descriptions. ”

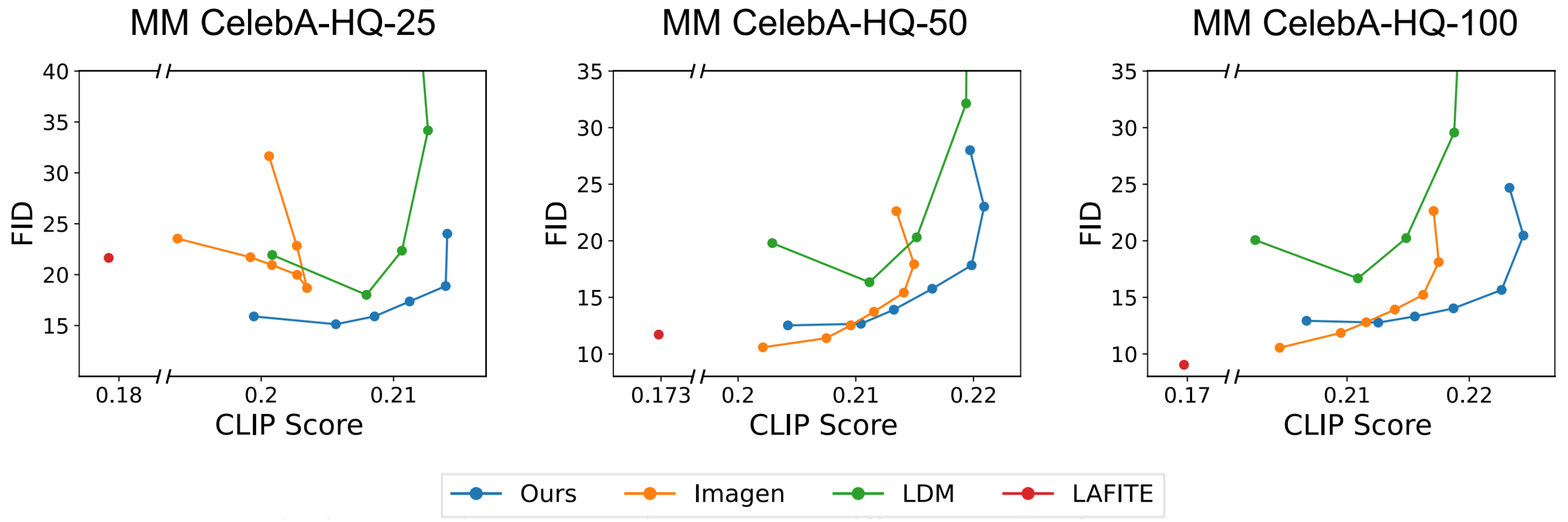
Quantitative Comparison



(a) FID-Semantic Recall

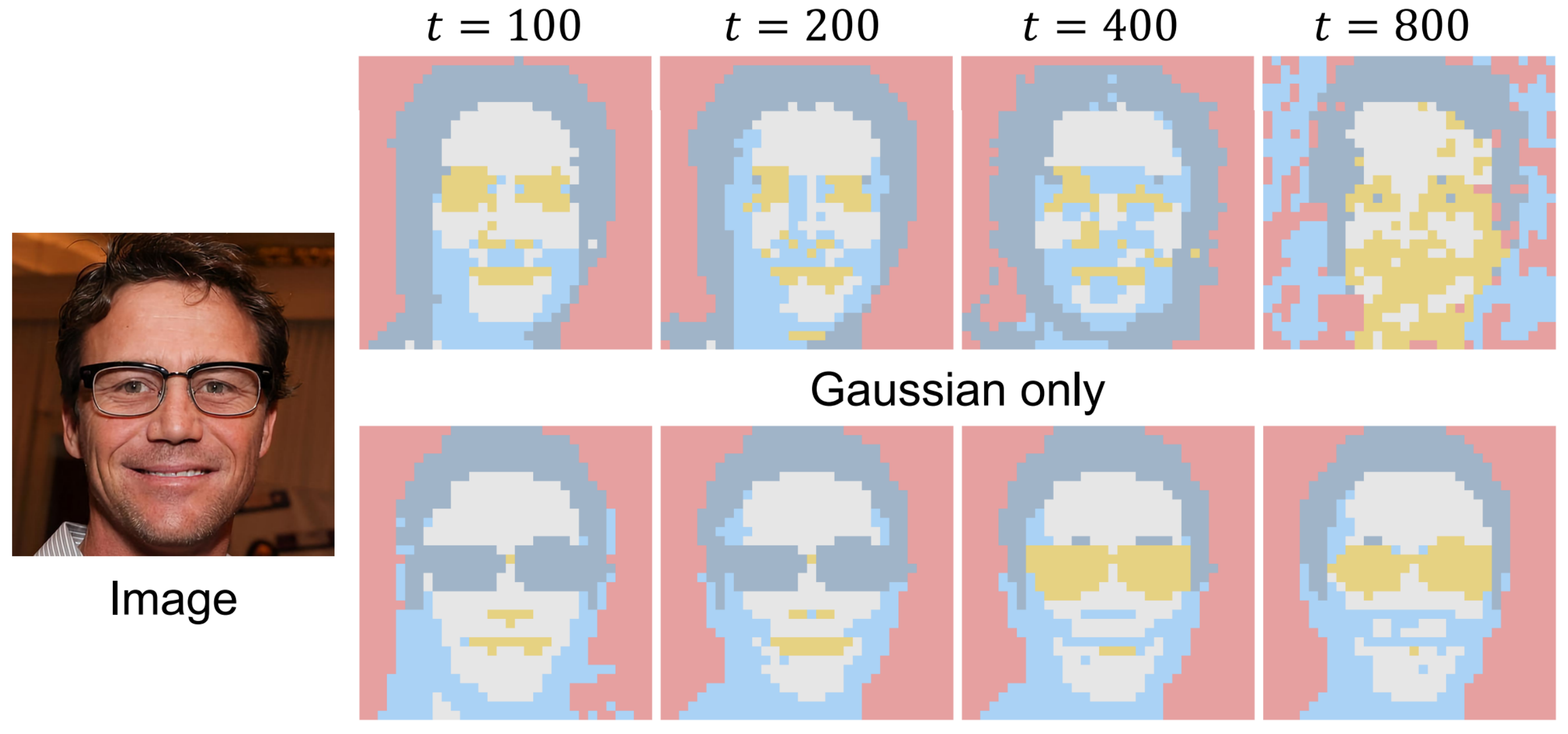
(b) Class-wise Semantic Recall

(c) Class Proportion in Cityscapes



> “ GCDP ***increase text-image correspondence*** maintaining image quality. ”

Visualizing the Internal Representation



Gaussian only

Gaussian-categorical

> “ GCDP has ***better understanding of semantic layout*** than Gaussian-only Diffusion Model. ”

Find us!

Jooyeol Yun

Minho Park