Appendix

Table of Contents

A.1. Derivation of the Gaussian-categorical diffusion process

A.2. Noise schedules of the Gaussian-categorical diffusion process

A.3. Comparison with Stable Diffusion

A.4. Visualizing the domain gaps in CLIP scores

A.5. Semantic Recall in Cityscapes

A.6. Quantitative results for cross-modal outpainting

A.7. Ablation study and additional baselines

A.8. Qualitative comparison

A.1. Derivation of the Gaussian-categorical diffusion process

In the following section, we provide detailed explanation of diffusion models including the categorical diffusion and the Gaussian-categorical diffusion.

A.1.1. Categorical diffusion process

In this section, our final goal is to derive the posterior $q\left(\mathbf{y}_{t-1} \,|\, \mathbf{y}_t, \mathbf{y}_0\right)$ of the categorical diffusion, given the forward noising process. The forward process of the categorical diffusion process is defined as follows:

$$\begin{aligned}
\forall t \in [1, 2, \dots T], \quad \alpha_t &\coloneqq 1 - \beta_t, && (1)\\
q\left(\mathbf{y}_t \,|\, \mathbf{y}_{t-1}\right) &\coloneqq \mathcal{C}(\mathbf{y}_t; (1-\beta_t)\mathbf{y}_{t-1} + \beta_t/K), && (2)\\
\mathbf{y}_t \in \{1, 2, ..., K\}^M \subset \mathbb{R}^M, &\quad \mathbb{1}[\mathbf{y}_t] \in \mathbb{R}^{M \times K}, && (3)
\end{aligned}$$

where β_t is the noise schedule for each timestep, K is the number of categories in the categorical distribution, and M is the number of categorical variables. $\mathbb{1}[\mathbf{y}_t]$ is the one-hot form of $\mathbf{y}_t$.

We will first prove $q\left(\mathbf{y}_t \,|\, \mathbf{y}_0\right) = \mathcal{C}(\mathbf{y}_t; \bar{\alpha}_t \mathbf{y}_0 + (1-\bar{\alpha}_t)/K)$ through mathematical induction. The base case $t = 1$ is evident though Equation (2) and let us assume the inductive case for $t-1$ where

$$q\left(\mathbf{y}_{t-1} \,|\, \mathbf{y}_0\right) \coloneqq \mathcal{C}(\mathbf{y}_{t-1}; \bar{\alpha}_{t-1}\mathbb{1}[\mathbf{y}_0] + (1-\bar{\alpha}_{t-1})/K) \quad \text{where } \bar{\alpha}_t \coloneqq \prod_{s=1}^{t} \alpha_s. \tag{4}$$

Then we can derive $q\left(\mathbf{y}_t \,|\, \mathbf{y}_0\right)$ as follows:

$$\begin{aligned}
q\left(\mathbf{y}_t \,|\, \mathbf{y}_0\right) &= \sum_{\mathbf{y}_{t-1}} q\left(\mathbf{y}_t \,|\, \mathbf{y}_{t-1}, \mathbf{y}_0\right) q\left(\mathbf{y}_{t-1} \,|\, \mathbf{y}_0\right) && (5)\\
&= \sum_{\mathbf{y}_{t-1}} q\left(\mathbf{y}_t \,|\, \mathbf{y}_{t-1}\right) q\left(\mathbf{y}_{t-1} \,|\, \mathbf{y}_0\right) && (6)\\
&= \sum_{\mathbf{y}_{t-1}} \left[\alpha_t \mathbb{1}[\mathbf{y}_{t-1}] + (1-\alpha_t)/K\right]_{\mathbf{y}_t} \left[\bar{\alpha}_{t-1}\mathbb{1}[\mathbf{y}_0] + (1-\bar{\alpha}_{t-1})/K\right]_{\mathbf{y}_{t-1}} && (7)\\
&= \sum_{\mathbf{y}_{t-1}} \left[\alpha_t \mathbb{1}[\mathbf{y}_t] + (1-\alpha_t)/K\right]_{\mathbf{y}_{t-1}} \left[\bar{\alpha}_{t-1}\mathbb{1}[\mathbf{y}_0] + (1-\bar{\alpha}_{t-1})/K\right]_{\mathbf{y}_{t-1}}. && (8)
\end{aligned}$$

where $[\mathbf{\Theta}]_{\mathbf{y}_t}$ denotes the probability of event $\mathbf{y}_t$ in the categorical distribution parameterized with $\mathbf{\Theta}$. By rewriting the summation as an inner product, we obtain

$$\begin{aligned}
q\left(\mathbf{y}_t \,|\, \mathbf{y}_0\right) &= \left[\alpha_t \mathbb{1}[\mathbf{y}_t] + (1-\alpha_t)/K\right] \cdot \left[\bar{\alpha}_{t-1}\mathbb{1}[\mathbf{y}_0] + (1-\bar{\alpha}_{t-1})/K\right] && (9)\\
&= \bar{\alpha}_t \mathbb{1}[\mathbf{y}_t] \cdot \mathbb{1}[\mathbf{y}_0] + (1-\alpha_t)\bar{\alpha}_{t-1}/K + (1-\bar{\alpha}_{t-1})\alpha_{t-1}/K + (1-\alpha_t)(1-\bar{\alpha}_{t-1})/K && (10)\\
&= \bar{\alpha}_t \mathbb{1}[\mathbf{y}_t] \cdot \mathbb{1}[\mathbf{y}_0] + (1-\bar{\alpha}_t)/K && (11)\\
&= \mathcal{C}(\mathbf{y}_t; \bar{\alpha}_t \mathbb{1}[\mathbf{y}_0] + (1-\bar{\alpha}_t)/K), && (12)
\end{aligned}$$

which is the t case of Equation (2). Through mathematical induction, we can conclude that $q\left(\mathbf{y}_t \,|\, \mathbf{y}_0\right) = \mathcal{C}(\mathbf{y}_t; \bar{\alpha}_t \mathbb{1}[\mathbf{y}_0] + (1-\bar{\alpha}_t)/K)$.

Next, we will derive the posterior $q\left(\mathbf{y}_{t-1} \,|\, \mathbf{y}_t, \mathbf{y}_0\right)$ using Bayes theorem as follows:

$$
\begin{aligned}
q\left(\mathbf{y}_{t-1} \,|\, \mathbf{y}_t, \mathbf{y}_0\right) &= \frac{q\left(\mathbf{y}_t \,|\, \mathbf{y}_{t-1}, \mathbf{y}_0\right) q\left(\mathbf{y}_{t-1} \,|\, \mathbf{y}_0\right)}{q\left(\mathbf{y}_t \,|\, \mathbf{y}_0\right)} && (13)\\
&= \frac{q\left(\mathbf{y}_t \,|\, \mathbf{y}_{t-1}\right) q\left(\mathbf{y}_{t-1} \,|\, \mathbf{y}_0\right)}{q\left(\mathbf{y}_t \,|\, \mathbf{y}_0\right)} && (14)\\
&= Z \left[\alpha_t \mathbb{1}[\mathbf{y}_{t-1}] + (1-\alpha_t)/K\right]_{\mathbf{y}_t} \left[\bar{\alpha}_{t-1} \mathbb{1}[\mathbf{y}_0] + (1-\bar{\alpha}_{t-1})/K\right]_{\mathbf{y}_{t-1}} && (15)\\
&= Z \left[\alpha_t \mathbb{1}[\mathbf{y}_t] + (1-\alpha_t)/K\right]_{\mathbf{y}_{t-1}} \left[\bar{\alpha}_{t-1} \mathbb{1}[\mathbf{y}_0] + (1-\bar{\alpha}_{t-1})/K\right]_{\mathbf{y}_{t-1}} && (16)\\
&= \mathcal{C}\big(\mathbf{y}_{t-1}; Z \left[\alpha_t \mathbb{1}[\mathbf{y}_t] + (1-\alpha_t)/K\right] \odot \left[\bar{\alpha}_{t-1} \mathbb{1}[\mathbf{y}_0] + (1-\bar{\alpha}_{t-1})/K\right]\big). && (17)
\end{aligned}
$$

Thus, the posterior $q\left(\mathbf{y}_{t-1} \,|\, \mathbf{y}_t, \mathbf{y}_0\right)$ is summarized as

$$
\begin{aligned}
q\left(\mathbf{y}_{t-1} \,|\, \mathbf{y}_t, \mathbf{y}_0\right) &= \mathcal{C}(\mathbf{y}_{t-1}; \widetilde{\mathbf{\Theta}}_t) && (18)\\
\widetilde{\mathbf{\Theta}}_t &\coloneqq Z[\alpha^c_t \mathbb{1}[\mathbf{y}_t] + (1-\alpha^c_t)/K] \odot [\bar{\alpha}^c_t \mathbb{1}[\mathbf{y}_0] + (1-\bar{\alpha}^c_{t-1})/K]. && (19)
\end{aligned}
$$

A.1.2. Gaussian-categorical diffusion process

We will derive the posterior $q\left(\mathbf{z}_{t-1} \,|\, \mathbf{z}_t, \mathbf{z}_0\right)$ of the Gaussian-categorical distribution, where the Gaussian distribution defined as follows:

$$
\begin{gathered}
X, Y \sim \mathcal{NC}\left(\mathbf{x}, \mathbf{y}; \boldsymbol{\mu}, \mathbf{\Sigma}, \mathbf{\Theta}\right), \qquad (20)\\
X = [X_1, X_2, ..., X_N] \in \mathbb{R}^N,\\
Y = [Y_1, Y_2, ..., Y_M] \in \{1, 2, ..., K\}^M,\\
\boldsymbol{\mu} \in \mathbb{R}^{S \times N}, \mathbf{\Sigma} \in \mathbb{R}^{S \times N \times N}, \mathbf{\Theta} \in \mathbb{R}^{M \times K}, \text{ and } S = K^M.
\end{gathered}
$$

$$
\mathcal{NC}(\mathbf{x}, \mathbf{y}; \boldsymbol{\mu}, \mathbf{\Sigma}, \mathbf{\Theta}) = \left(\prod_{i=1}^{M} \mathbf{\Theta}_{i, \mathbf{y}_i}\right) (2\pi)^{-\frac{N}{2}} \left|\mathbf{\Sigma}_{\mathbf{y}}\right|^{-\frac{1}{2}} \exp\Big(-\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu}_{\mathbf{y}})^\top \mathbf{\Sigma}_{\mathbf{y}}^{-1} (\mathbf{x} - \boldsymbol{\mu}_{\mathbf{y}})\Big). \qquad (21)
$$

and the forward noising process for the Gaussian-categorical diffusion is defined as

$$
\begin{aligned}
\forall t \in [1, 2, \dots, T], \quad \alpha^{\mathcal{N}}_t \coloneqq 1 - \beta^{\mathcal{N}}_t, \quad \alpha^c_t \coloneqq 1 - \beta^c_t, \quad \text{and} \quad \mathbf{z}_t \coloneqq (\mathbf{x}_t, \mathbf{y}_t), && (22)\\
q\left(\mathbf{z}_t \,|\, \mathbf{z}_{t-1}\right) \coloneqq \mathcal{NC}\left(\mathbf{z}_t; \left[\sqrt{1-\beta^{\mathcal{N}}_t}\mathbf{x}_{t-1}\right]_{\times S}, \left[\beta^{\mathcal{N}}_t \boldsymbol{I}\right]_{\times S}, (1-\beta^c_t)\mathbb{1}[\mathbf{y}_{t-1}] + \beta^c_t/K\right). && (23)
\end{aligned}
$$

We will first prove that $q\left(\mathbf{z}_t \,|\, \mathbf{z}_0\right) = \mathcal{NC}\Big(\mathbf{z}_t; \left[\sqrt{\bar{\alpha}^{\mathcal{N}}_t}\mathbf{x}_0\right]_{\times S}, \left[(1-\bar{\alpha}^{\mathcal{N}}_t)\boldsymbol{I}\right]_{\times S}, (1-\bar{\alpha}^c_t)\mathbb{1}[\mathbf{y}_0] + \bar{\alpha}^c_t/K\Big)$ where $\bar{\alpha}_t \coloneqq \prod_{s=1}^t \alpha_s$. We will prove this using mathematical induction, where the base case $t = 1$ is defined in Equation (23). Let us assume the inductive case for $t-1$,

$$
q\left(\mathbf{z}_{t-1} \,|\, \mathbf{z}_0\right) = \mathcal{NC}\Big(\mathbf{z}_{t-1}; \left[\sqrt{\bar{\alpha}^{\mathcal{N}}_{t-1}}\mathbf{x}_0\right]_{\times S}, \left[(1-\bar{\alpha}^{\mathcal{N}}_{t-1})\boldsymbol{I}\right]_{\times S}, (1-\bar{\alpha}^c_{t-1})\mathbb{1}[\mathbf{y}_0] + \bar{\alpha}^c_{t-1}/K\Big). \qquad (24)
$$

Then we can derive $q\left(\mathbf{z}_t \,|\, \mathbf{z}_0\right)$ as follows:

$$q\left(\mathbf{z}_t \,|\, \mathbf{z}_0\right) \tag{25}$$

$$= \int q\left(\mathbf{z}_t \,|\, \mathbf{z}_{t-1}, \mathbf{z}_0\right) q\left(\mathbf{z}_{t-1} \,|\, \mathbf{z}_0\right) d\mathbf{z}_{t-1} \tag{26}$$

$$= \int q\left(\mathbf{z}_t \,|\, \mathbf{z}_{t-1}\right) q\left(\mathbf{z}_{t-1} \,|\, \mathbf{z}_0\right) d\mathbf{z}_{t-1} \tag{27}$$

$$= \sum_{\mathbf{y}_{t-1}} \int \mathcal{NC}(\mathbf{z}_t; \left[\boldsymbol{\mu}_{t|t-1}\right]_{\times S}, \left[\boldsymbol{\Sigma}_{t|t-1}\right]_{\times S}, \boldsymbol{\Theta}_{t|t-1}) \cdot \mathcal{NC}(\mathbf{z}_{t-1}; \left[\boldsymbol{\mu}_{t-1|0}\right]_{\times S}, \left[\boldsymbol{\Sigma}_{t-1|0}\right]_{\times S}, \boldsymbol{\Theta}_{t-1|0}) d\mathbf{x}_{t-1}, \tag{28}$$

where $\boldsymbol{\Theta}_{i|j} := (1-\beta_i^c)\mathbb{1}[\mathbf{y}_j] + \beta_i^c/K$, and $[\mathbf{v}]_{\times S}$ indicates row-wise duplication of a vector $\mathbf{v}$ (*i.e.*, $[\mathbf{v}, \mathbf{v}, ..., \mathbf{v}]^T$). By decomposing the Gaussian-categorical into a Gaussian distribution and a categorical distribution, we can write the equation as follows:

$$q\left(\mathbf{z}_t \,|\, \mathbf{z}_0\right) \tag{29}$$

$$= \sum_{\mathbf{y}_{t-1}} \int \Big(\mathcal{C}(\mathbf{y}_t; \boldsymbol{\Theta}_{t|t-1}) \cdot \mathcal{N}(\mathbf{x}_t; \boldsymbol{\mu}_{t|t-1}, \boldsymbol{\Sigma}_{t|t-1})\Big) \cdot \Big(\mathcal{C}(\mathbf{y}_{t-1}; \boldsymbol{\Theta}_{t-1|0}) \cdot \mathcal{N}(\mathbf{x}_{t-1}; \boldsymbol{\mu}_{t-1|0}, \boldsymbol{\Sigma}_{t-1|0})\Big) d\mathbf{x}_{t-1} \tag{30}$$

$$= \sum_{\mathbf{y}_{t-1}} \mathcal{C}(\mathbf{y}_t; \boldsymbol{\Theta}_{t|t-1}) \cdot \mathcal{C}(\mathbf{y}_{t-1}; \boldsymbol{\Theta}_{t-1|0}) \int \mathcal{N}(\mathbf{x}_t; \boldsymbol{\mu}_{t|t-1}, \boldsymbol{\Sigma}_{t|t-1}) \cdot \mathcal{N}(\mathbf{x}_{t-1}; \boldsymbol{\mu}_{t-1|0}, \boldsymbol{\Sigma}_{t-1|0}) d\mathbf{x}_{t-1} \tag{31}$$

$$= \mathcal{C}(\mathbf{y}_t; \bar{\alpha}_t^c \mathbb{1}[\mathbf{y}_0] + (1-\bar{\alpha}_t^c)/K) \cdot \mathcal{N}(\mathbf{x}_t; \sqrt{\bar{\alpha}_t^{\mathcal{N}}}\mathbf{x}_0, (1-\bar{\alpha}_t^{\mathcal{N}})\boldsymbol{I}) \tag{32}$$

$$= \mathcal{NC}\Big(\mathbf{z}_t; \left[\sqrt{\bar{\alpha}_t^{\mathcal{N}}}\mathbf{x}_0\right]_{\times S}, \left[(1-\bar{\alpha}_t^{\mathcal{N}})\boldsymbol{I}\right]_{\times S}, (1-\bar{\alpha}_t^c)\mathbb{1}[\mathbf{y}_0] + \bar{\alpha}_t^c/K\Big), \tag{33}$$

where $\boldsymbol{\mu}_{i|j} := \sqrt{1-\beta_i^{\mathcal{N}}}\mathbf{x}_j$ and $\boldsymbol{\Sigma}_{i|j} := \beta_i^{\mathcal{N}}\boldsymbol{I}$. Through mathematical induction, we can conclude that $q\left(\mathbf{z}_t \,|\, \mathbf{z}_0\right) = \mathcal{NC}\Big(\mathbf{z}_t; \left[\sqrt{\bar{\alpha}_t^{\mathcal{N}}}\mathbf{x}_0\right]_{\times S}, \left[(1-\bar{\alpha}_t^{\mathcal{N}})\boldsymbol{I}\right]_{\times S}, (1-\bar{\alpha}_t^c)\mathbb{1}[\mathbf{y}_0] + \bar{\alpha}_t^c/K\Big)$.

Next, we will derive the posterior $q\left(\mathbf{z}_{t-1} \,|\, \mathbf{z}_t, \mathbf{z}_0\right)$ using Bayes theorem,

$$q\left(\mathbf{z}_{t-1} \,|\, \mathbf{z}_t, \mathbf{z}_0\right) = \frac{q\left(\mathbf{z}_t \,|\, \mathbf{z}_{t-1}, \mathbf{z}_0\right) q\left(\mathbf{z}_{t-1} \,|\, \mathbf{z}_0\right)}{q\left(\mathbf{z}_t \,|\, \mathbf{z}_0\right)} \tag{34}$$

$$= \frac{q\left(\mathbf{z}_t \,|\, \mathbf{z}_{t-1}\right) q\left(\mathbf{z}_{t-1} \,|\, \mathbf{z}_0\right)}{q\left(\mathbf{z}_t \,|\, \mathbf{z}_0\right)} \tag{35}$$

$$= \frac{\mathcal{NC}(\mathbf{z}_t; \left[\boldsymbol{\mu}_{t|t-1}\right]_{\times S}, \left[\boldsymbol{\Sigma}_{t|t-1}\right]_{\times S}, \boldsymbol{\Theta}_{t|t-1}) \cdot \mathcal{NC}(\mathbf{z}_{t-1}; \left[\boldsymbol{\mu}_{t-1|0}\right]_{\times S}, \left[\boldsymbol{\Sigma}_{t-1|0}\right]_{\times S}, \boldsymbol{\Theta}_{t-1|0})}{\mathcal{NC}(\mathbf{z}_t; \left[\boldsymbol{\mu}_{t|0}\right]_{\times S}, \left[\boldsymbol{\Sigma}_{t|0}\right]_{\times S}, \boldsymbol{\Theta}_{t|0})}. \tag{36}$$

We again decompose the Gaussian-categorical diffusion into a Gaussian distribution and a categorical distribution

$$
\begin{aligned}
&q\left(\mathbf{z}_{t-1} \,|\, \mathbf{z}_t, \mathbf{z}_0\right) && (37)\\
&= \frac{\Big(\mathcal{C}(\mathbf{y}_t; \boldsymbol{\Theta}_{t|t-1}) \cdot \mathcal{N}(\mathbf{x}_t; \boldsymbol{\mu}_{t|t-1}, \boldsymbol{\Sigma}_{t|t-1})\Big) \cdot \Big(\mathcal{C}(\mathbf{y}_{t-1}; \boldsymbol{\Theta}_{t-1|0}) \cdot \mathcal{N}(\mathbf{x}_{t-1}; \boldsymbol{\mu}_{t-1|0}, \boldsymbol{\Sigma}_{t-1|0})\Big)}{\mathcal{C}(\mathbf{y}_t; \boldsymbol{\Theta}_{t|0}) \cdot \mathcal{N}(\mathbf{x}_t; \boldsymbol{\mu}_{t|0}, \boldsymbol{\Sigma}_{t|0})} && (38)\\
&= \frac{\mathcal{C}(\mathbf{y}_t; \boldsymbol{\Theta}_{t|t-1}) \cdot \mathcal{C}(\mathbf{y}_{t-1}; \boldsymbol{\Theta}_{t-1|0})}{\mathcal{C}(\mathbf{y}_t; \boldsymbol{\Theta}_{t|0})} \cdot \frac{\mathcal{N}(\mathbf{x}_t; \boldsymbol{\mu}_{t|t-1}, \boldsymbol{\Sigma}_{t|t-1}) \cdot \mathcal{N}(\mathbf{x}_{t-1}; \boldsymbol{\mu}_{t-1|0}, \boldsymbol{\Sigma}_{t-1|0})}{\mathcal{N}(\mathbf{x}_t; \boldsymbol{\mu}_{t|0}, \boldsymbol{\Sigma}_{t|0})} && (39)\\
&= \mathcal{C}(\mathbf{y}_{t-1}; \widetilde{\boldsymbol{\Theta}}_t) \cdot \mathcal{N}(\mathbf{x}_{t-1}; \widetilde{\boldsymbol{\mu}}_t, \widetilde{\boldsymbol{\Sigma}}_t) && (40)\\
&= \mathcal{NC}(\mathbf{z}_{t-1}; \big[\widetilde{\boldsymbol{\mu}}_t\big]_{\times S}, \big[\widetilde{\boldsymbol{\Sigma}}_t\big]_{\times S}, \widetilde{\boldsymbol{\Theta}}_t), && (41)
\end{aligned}
$$

$$
\begin{aligned}
\widetilde{\boldsymbol{\mu}}_t &\coloneqq \frac{\sqrt{\bar{\alpha}^{\mathcal{N}}_{t-1}}\beta^{\mathcal{N}}_t}{1 - \bar{\alpha}^{\mathcal{N}}_t}\mathbf{x}_0 + \frac{\sqrt{\alpha^{\mathcal{N}}_t}(1 - \bar{\alpha}^{\mathcal{N}}_{t-1})}{1 - \bar{\alpha}^{\mathcal{N}}_t}\mathbf{x}_t, && (42)\\
\widetilde{\boldsymbol{\Sigma}}_t &\coloneqq \big((1 - \bar{\alpha}^{\mathcal{N}}_{t-1})\beta^{\mathcal{N}}_t / (1 - \bar{\alpha}^{\mathcal{N}}_t)\big)\boldsymbol{I}, && (43)\\
\widetilde{\boldsymbol{\Theta}}_t &\coloneqq Z[\alpha^{\mathcal{c}}_t \mathbb{1}[\mathbf{y}_t] + (1 - \alpha^{\mathcal{c}}_t)/K] \odot [\bar{\alpha}^{\mathcal{c}}_t \mathbb{1}[\mathbf{y}_0] + (1 - \bar{\alpha}^{\mathcal{c}}_{t-1})/K], && (44)
\end{aligned}
$$

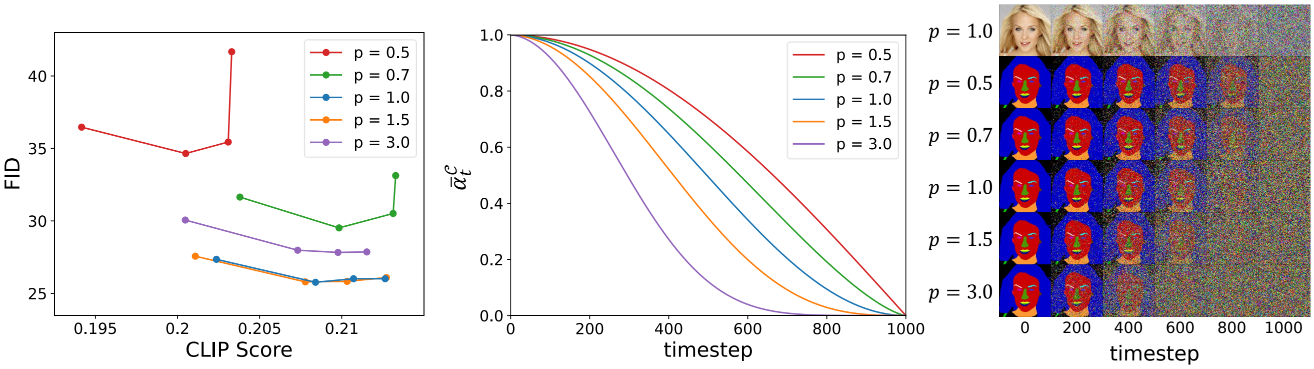
The posterior distribution $q\left(\mathbf{z}_{t-1} \,|\, \mathbf{z}_t, \mathbf{z}_0\right)$ can be summarized as follows:

$$
q\left(\mathbf{z}_{t-1} \,|\, \mathbf{z}_t, \mathbf{z}_0\right) = \mathcal{NC}\Big(\mathbf{z}_{t-1}; \big[\widetilde{\boldsymbol{\mu}}_t\big]_{\times S}, \big[\widetilde{\boldsymbol{\Sigma}}_t\big]_{\times S}, \widetilde{\boldsymbol{\Theta}}_t\Big), \qquad (45)
$$

where Z is a normalizing constant. We approximate the reverse process by matching $\widetilde{\boldsymbol{\mu}}_\theta(\mathbf{z}_t)$, $\widetilde{\boldsymbol{\Sigma}}_\theta(\mathbf{z}_t)$, and $\boldsymbol{\Theta}_\theta(\mathbf{z}_t)$.

Finally, minimizing the KL divergence term $D_{KL}\big(q\left(\mathbf{z}_{t-1} \,|\, \mathbf{z}_t, \mathbf{z}_0\right) \parallel p_\theta\left(\mathbf{z}_{t-1} \,|\, \mathbf{z}_t\right)\big)$ can be decomposed into two separate terms for the Gaussian variable and the categorical variable as follows:

$$
\begin{aligned}
&D_{KL}\big(q\left(\mathbf{z}_{t-1}\,|\,\mathbf{z}_t,\mathbf{z}_0\right)\parallel p_\theta\left(\mathbf{z}_{t-1}\,|\,\mathbf{z}_t\right)\big) && (46)\\
&=\int q\left(\mathbf{z}_{t-1}\,|\,\mathbf{z}_t,\mathbf{z}_0\right)\log\frac{q\left(\mathbf{z}_{t-1}\,|\,\mathbf{z}_t,\mathbf{z}_0\right)}{p_\theta\left(\mathbf{z}_{t-1}\,|\,\mathbf{z}_t\right)}d\mathbf{z}_{t-1} && (47)\\
&=\int \mathcal{NC}(\mathbf{z}_{t-1};\,\big[\widetilde{\boldsymbol{\mu}}_t\big]_{\times S},\,\big[\widetilde{\boldsymbol{\Sigma}}_t\big]_{\times S},\,\widetilde{\boldsymbol{\Theta}}_t)\log\frac{\mathcal{NC}(\mathbf{z}_{t-1};\,\big[\widetilde{\boldsymbol{\mu}}_t\big]_{\times S},\,\big[\widetilde{\boldsymbol{\Sigma}}_t\big]_{\times S},\,\widetilde{\boldsymbol{\Theta}}_t)}{\mathcal{NC}(\mathbf{z}_{t-1};\,\big[\widetilde{\boldsymbol{\mu}}_\theta(\mathbf{z}_t)\big]_{\times S},\,\big[\widetilde{\boldsymbol{\Sigma}}_\theta(\mathbf{z}_t)\big]_{\times S},\,\boldsymbol{\Theta}_\theta(\mathbf{z}_t))}d\mathbf{z}_{t-1} && (48)\\
&=\int \mathcal{C}(\mathbf{y}_{t-1};\widetilde{\boldsymbol{\Theta}}_t)\cdot\mathcal{N}(\mathbf{x}_{t-1};\widetilde{\boldsymbol{\mu}}_t,\widetilde{\boldsymbol{\Sigma}}_t)\log\frac{\mathcal{C}(\mathbf{y}_{t-1};\widetilde{\boldsymbol{\Theta}}_t)\cdot\mathcal{N}(\mathbf{x}_{t-1};\widetilde{\boldsymbol{\mu}}_t,\widetilde{\boldsymbol{\Sigma}}_t)}{\mathcal{C}(\mathbf{y}_{t-1};\boldsymbol{\Theta}_\theta(\mathbf{z}_t))\cdot\mathcal{N}(\mathbf{x}_{t-1};\widetilde{\boldsymbol{\mu}}_\theta(\mathbf{z}_t),\widetilde{\boldsymbol{\Sigma}}_\theta(\mathbf{z}_t))}d\mathbf{z}_{t-1} && (49)\\
&=\int \mathcal{C}(\mathbf{y}_{t-1};\widetilde{\boldsymbol{\Theta}}_t)\cdot\mathcal{N}(\mathbf{x}_{t-1};\widetilde{\boldsymbol{\mu}}_t,\widetilde{\boldsymbol{\Sigma}}_t)\log\frac{\mathcal{C}(\mathbf{y}_{t-1};\widetilde{\boldsymbol{\Theta}}_t)}{\mathcal{C}(\mathbf{y}_{t-1};\boldsymbol{\Theta}_\theta(\mathbf{z}_t))}d\mathbf{z}_{t-1}\\
&\quad+\int \mathcal{C}(\mathbf{y}_{t-1};\widetilde{\boldsymbol{\Theta}}_t)\cdot\mathcal{N}(\mathbf{x}_{t-1};\widetilde{\boldsymbol{\mu}}_t,\widetilde{\boldsymbol{\Sigma}}_t)\log\frac{\mathcal{N}(\mathbf{x}_{t-1};\widetilde{\boldsymbol{\mu}}_t,\widetilde{\boldsymbol{\Sigma}}_t)}{\mathcal{N}(\mathbf{x}_{t-1};\widetilde{\boldsymbol{\mu}}_\theta(\mathbf{z}_t),\widetilde{\boldsymbol{\Sigma}}_\theta(\mathbf{z}_t))}d\mathbf{z}_{t-1} && (50)\\
&=\int \mathcal{C}(\mathbf{y}_{t-1};\widetilde{\boldsymbol{\Theta}}_t)\cdot\log\frac{\mathcal{C}(\mathbf{y}_{t-1};\widetilde{\boldsymbol{\Theta}}_t)}{\mathcal{C}(\mathbf{y}_{t-1};\boldsymbol{\Theta}_\theta(\mathbf{z}_t))}d\mathbf{y}_{t-1}\\
&\quad+\int \mathcal{N}(\mathbf{x}_{t-1};\widetilde{\boldsymbol{\mu}}_t,\widetilde{\boldsymbol{\Sigma}}_t)\log\frac{\mathcal{N}(\mathbf{x}_{t-1};\widetilde{\boldsymbol{\mu}}_t,\widetilde{\boldsymbol{\Sigma}}_t)}{\mathcal{N}(\mathbf{x}_{t-1};\widetilde{\boldsymbol{\mu}}_\theta(\mathbf{z}_t),\widetilde{\boldsymbol{\Sigma}}_\theta(\mathbf{z}_t))}d\mathbf{x}_{t-1} && (51)\\
&=D_{KL}(\mathcal{C}(\mathbf{y}_{t-1};\widetilde{\boldsymbol{\Theta}}_t)\parallel\mathcal{C}(\mathbf{y}_{t-1};\boldsymbol{\Theta}_\theta(\mathbf{z}_t)))+D_{KL}(\mathcal{N}(\mathbf{x}_{t-1};\widetilde{\boldsymbol{\mu}}_t,\widetilde{\boldsymbol{\Sigma}}_t)\parallel\mathcal{N}(\mathbf{x}_{t-1};\widetilde{\boldsymbol{\mu}}_\theta(\mathbf{z}_t),\widetilde{\boldsymbol{\Sigma}}_\theta(\mathbf{z}_t))) && (52)\\
&=\mathbb{E}_q\left[\frac{1}{2\sigma_t^2}\left\|\widetilde{\boldsymbol{\mu}}_t-\widetilde{\boldsymbol{\mu}}_\theta(\mathbf{z}_t)\right\|^2\right]+D_{KL}(\widetilde{\boldsymbol{\Theta}}_t\parallel\boldsymbol{\Theta}_\theta(\mathbf{z}_t))+C && (53)
\end{aligned}
$$



(a) FID-CLIP Score

(b) Illustration of different noise schedules

Figure A.1. (a) FID-CLIP score pairs for different noise schedules $\beta^{\mathcal{C}}$. FID and CLIP scores are measured in 128×128 resolution. (b) The illustration of different noise schedules. A larger p indicates stronger noise near $t = 1000$.

A.2. Noise schedules of the Gaussian-categorical diffusion process

The Gaussian-categorical diffusion process can have different noise schedules for β^c and $\beta^{\mathcal{N}}$ as defined in Equation (23). In order to search for a reasonable noise schedule, we train the Gaussian-categorical diffusion model on different schedules for β^c, relative to the Gaussian noise schedule $\beta^{\mathcal{N}}$. Specifically, we fix $\beta^{\mathcal{N}}$ as the cosine noise schedule [14] and set β^{c} as a function of a p^{th} power of $\beta^{\mathcal{N}}$, in other words $\beta^{c} := (\beta^{\mathcal{N}})^{p}$, which are plotted in Figure A.1 (b). In Figure A.1, we present the FID-CLIP score of these results at the 128×128 resolution on the CelebA-HQ dataset [11]. Overall, choosing p near 1 is a good choice for achieving text-image correspondence. We leave further analysis on noise scheduling between different modalities as a future research topic.

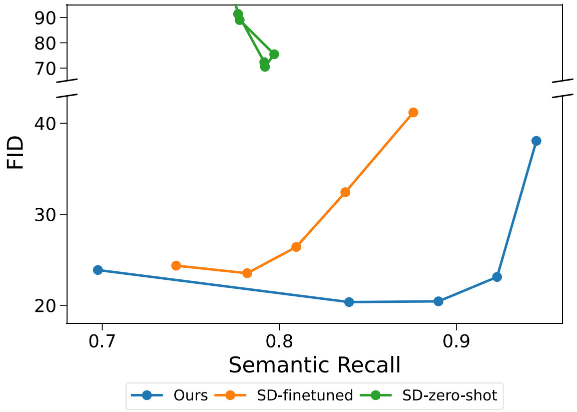


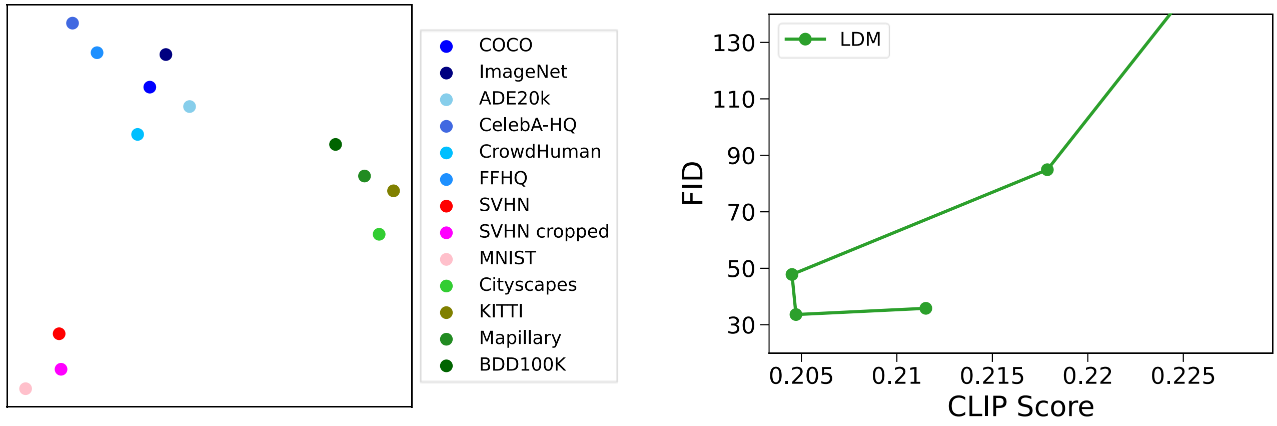
Figure A.2. FID-Semantic Recall of the Gaussian-categorical diffusion model compared to the results generated by the Stable Diffusion model finetuned on Cityscapes [4] (SD-finetuned) and zero-shot text-to-image generation of the pretrained Stable Diffusion (SD-zero-shot). We use the Stable Diffusion v1.4 for both zero-shot generation and finetuning.

A.3. Comparison with Stable Diffusion

Recently, finetuning a general-purpose text-to-image generation model using domain-specific datasets has shown great success in generating high-quality images with strong text-image correspondence. Specifically, the Stable Diffusion project provides a large pretrained Latent Diffusion Model (LDM) [16] trained on a web-scale dataset, the LAION 5B [17], that is capable of generating artistic images. In this section, we demonstrate the limitation of finetuning a generative model in cases of significant domain gaps. We finetune Stable Diffusion v1.4 using the Cityscapes dataset and report the FID-Semantic Recall pairs in Figure A.2. We also provide zero-shot text-to-image generation results for comparison. While finetuning stable diffusion can be effective in natural domains such as the MM CelebA-HQ, it should not be viewed as an all-encompassing solution for addressing issues in text-to-image generation. Neither finetuning Stable Diffusion nor zero-shot text-to-image generation exhibits a low FID or a high Semantic Recall for generating the urban scenes of Cityscapes [4]. Training a Gaussian-categorical diffusion model can be an effective approach for generating unique domains such as medical images or aerial photos.

A.4. Visualizing the domain gaps in CLIP scores

The CLIP score [7] is a reliable measure in most cases for evaluating the quality of text-to-image generation in natural domains such as the MS COCO [10]. However, in certain cases, the CLIP model [15] may have poor generalization abilities for specific domains with significant differences from its training data. Since the train dataset of CLIP is not publicly available for this analysis, we replace it with the MS COCO dataset which contains diverse images of different scenes. As shown in Figure A.3



(a) CLIP feature Visualization

(b) FID-CLIP Score in Cityscapes

Figure A.3. (a) Visualization of CLIP features from different datasets using t-SNE. While the CelebA-HQ dataset closely clusters with several large-scale image datasets such as the ImageNet and MS COCO dataset, urban scene datasets such as Cityscapes or BDD100K form distinct clusters. (b) CLIP scores display inconsistent trends when measured on the Cityscapes dataset.

(a), we plot the features from the CLIP image encoder [15] for different datasets using the t-SNE visualization technique [19]. Each point in Figure A.3 (a) represents the averaged CLIP features from a single dataset. While general image datasets such as the ImageNet [5], ADE20K [22], and the CelebA-HQ [11] are closely clustered to the MS COCO dataset, other datasets such as the urban scene datasets (*e.g.*, Cityscapes [4] and BDD100K [3]) or the number datasets (*e.g.*, MNIST [6] and SVHN [13]) form distinct clusters apart from the MS COCO dataset [10].

This indicates that the Cityscapes dataset may have a domain gap significantly large enough to render the CLIP score unreliable. As shown in Figure A.3 (b), FID-CLIP score pairs for the Latent Diffusion Model (LDM) [16] display inconsistent trends of increase and decrease as the guidance scale increases. Thus, we do not use the CLIP score to evaluate the Cityscapes text-to-image generation and instead use the Semantic Recall.

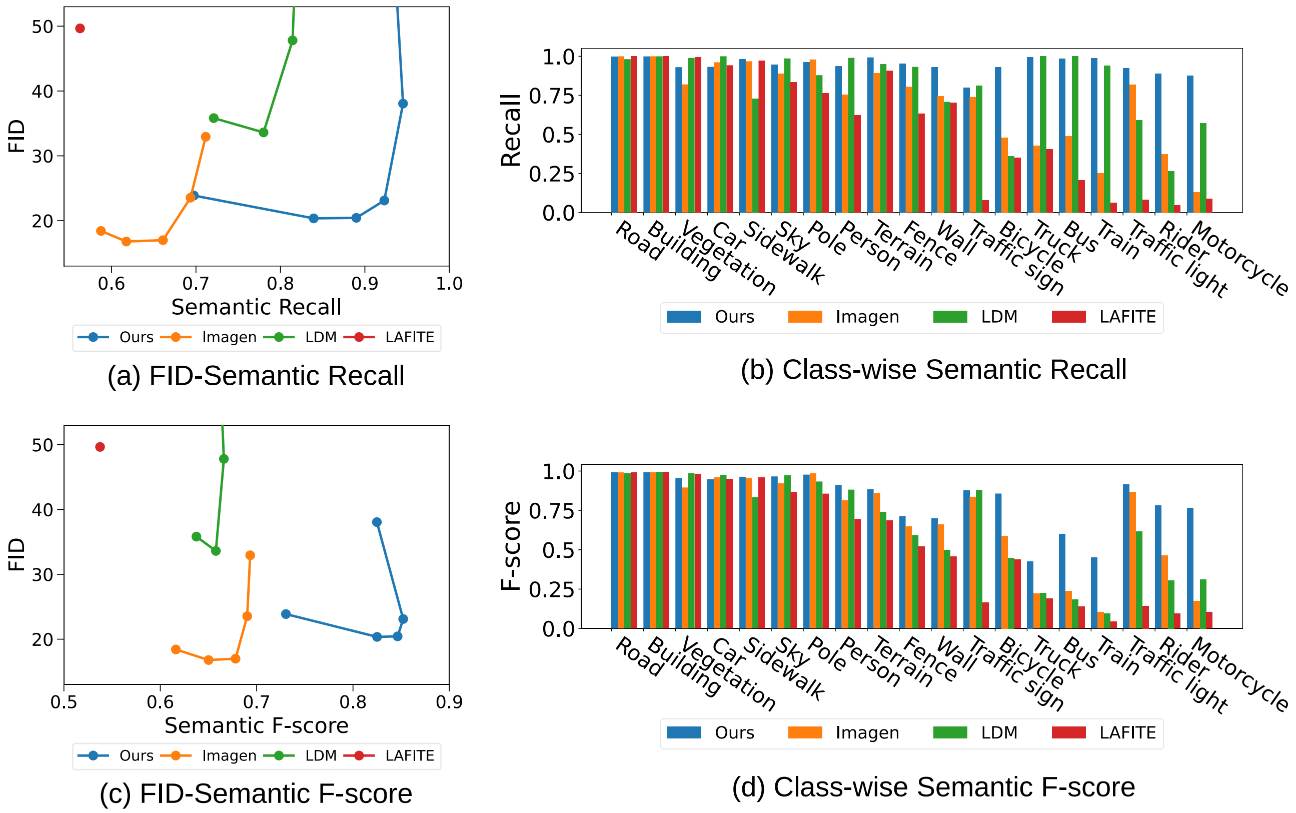
A.5. Semantic Recall in Cityscapes

To compensate for the limitations of the CLIP score when evaluating datasets with large domain gaps, we introduce the Semantic Recall which evaluates the generation of specific semantic categories specified in the test description. The Semantic Recall is the average ratio of correctly detected classes in the generated image to the total number of classes in the ground-truth layouts,

$$\text{Semantic Recall} := \frac{1}{\mid \mathcal{G} \mid} \sum_{x_i, y_i \in \mathcal{G}} \frac{\mid \text{Classes in } F(x_i) \cap \text{ Classes in } y_i \mid}{\mid \text{Classes in } y_i \mid},$$

where $\mathcal{G}$ is the set of generated image-layout pairs (x_i, y_i) and $| \cdot |$ indicates the cardinality of a given set. $F(\cdot)$ is the pretrained semantic segmentation model [20]. We provide full details of the Semantic Recall for each class in Figure A.4 (b). The Gaussian-categorical diffusion model is especially effective for generating less frequently encountered classes such as the *Motorcycle* and *Traffic light* classes.

In this section, we also report the Semantic F-score as an evaluation measure for the semantic accuracy of the generated image. The Semantic F-score is similar to the Semantic Recall but uses the F-score,



(a) FID-Semantic Recall

(b) Class-wise Semantic Recall

(c) FID-Semantic F-score

(d) Class-wise Semantic F-score

Figure A.4. (a) FID-Semantic Recall for the Cityscapes dataset and (b) detailed class-wise Semantic Recall. (c) FID-Semantic F-score for the Cityscapes dataset and (c) detailed class-wise Semantic Recall. Classes are sorted from the most occurring classes (left) to the least occurring (right). The Gaussian-categorical diffusion model outperforms existing baselines by a large margin in the Semantic F-score, indicating that our approach does not overly generate objects.

which takes both recall and precision into account as:

$$\text{Semantic F-score} := \frac{2}{\text{Semantic Recall}^{-1} + \text{Semantic Precision}^{-1}},$$

where Semantic Precision is calculated similarly to the Semantic Recall. While the Semantic Recall is useful for detecting the existence of certain objects, it may overcompensate for verbose generation. For instance, a text-to-image generation model that generates all semantic classes regardless of the text condition may achieve a high recall without understanding the text description. Therefore, we use the F-score to evaluate whether a text-to-image generation model precisely generates the classes specified in the text description. The results in Figure A.4 (c) demonstrate that the Gaussian-categorical diffusion model outperforms existing text-to-image in the Cityscapes [4] dataset, exhibiting a high F-score and a low FID. This suggests that our model does not overly generate semantic classes regardless of the text description.

A.6. Quantitative results for cross-modal outpainting

As demonstrated in the main paper, a well-trained Gaussian-categorical diffusion is capable of performing text-guided segmentation and layout-to-image generation. The key idea is to view an image or a layout as a masked image-layout pair and inpaint the masked modality using the RePaint technique [12]. The detailed algorithm following RePaint [12] is provided in Algorithm 1. We also compare the quantitative comparison of the results for segmentation and layout-to-image generation on the CelebA-HQ dataset [8,11] in Table A.1 and Table A.2. We train a segmentation (*i.e.*, Deeplab v3 [2]) and a layout-to-image generation model (*i.e.*, OASIS [18]) on the MM CelebA-HQ-25. While the Gaussian-categorical diffusion does not outperform models dedicated to segmentation or layout-to-image generation, it yields reasonable quantitative results which suggest that the Gaussian-categorical diffusion can serve as a generative prior for tasks other than text-to-image generation. Additionally, we find that training the Gaussian-categorical diffusion with a lower p value leans towards better layout-to-image generation while a higher p value leads to better segmentation performance. In this manner, extreme values of p (*i.e.*, $p = 0$ and $p \to \infty$) are equivalent to training a conditional generation model (*i.e.*, layout-to-image and semantic segmentation).

Algorithm 1 Cross-modal outpainting for conditional generation.

```
$\mathbf{z}_T \sim \mathcal{NC}(\mathbf{x}, \mathbf{y}; \mathbf{0}, \boldsymbol{I}, \boldsymbol{\Theta})$
$t \leftarrow T$
while $t > 0$ do
    $n \leftarrow N$
    while $n > 0$ do
        $\mathbf{z}^{\text{known}}_{t-1} \sim \mathcal{NC}(\mathbf{z}_{t-1}; [\boldsymbol{\mu}_{t-1|0}]_{\times S}, [\boldsymbol{\Sigma}_{t-1|0}]_{\times S}, \boldsymbol{\Theta}_{t-1|0})$    ▷ Apply noise to known area $\mathbf{z}^{\text{known}}$
        $\mathbf{z}^{\text{unknown}}_{t-1} \sim \mathcal{NC}(\mathbf{z}_{t-1}; [\widetilde{\boldsymbol{\mu}}_\theta(\mathbf{z}_t)]_{\times S}, [\widetilde{\boldsymbol{\Sigma}}_\theta(\mathbf{z}_t)]_{\times S}, \boldsymbol{\Theta}_\theta(\mathbf{z}_t))$    ▷ Denoise single step $\mathbf{z}_t$
        $\mathbf{z}_{t-1} = m \odot \mathbf{z}^{\text{known}}_{t-1} + (1 - m) \odot \mathbf{z}^{\text{unknown}}_{t-1}$    ▷ Update unknown area
        if $n < N$ and $t > 1$ then
            $\mathbf{z}_t \sim \mathcal{NC}(\mathbf{z}_t; [\boldsymbol{\mu}_{t|t-1}]_{\times S}, [\boldsymbol{\Sigma}_{t|t-1}]_{\times S}, \boldsymbol{\Theta}_{t|t-1})$    ▷ Resample timestep $t$
        end if
        $n \leftarrow n - 1$
    end while
    $t \leftarrow t - 1$
end while
```

Method	mIoU ↑
Deeplab v3 [2]	73.88
Ours $p = 0.5$	32.52
Ours $p = 1.0$	51.56
Ours $p = 3.0$	59.82

Table A.1. Quantitative results for semantic segmentation on the 25% of the MM CelebAMask-HQ dataset [9]. Segmentation predictions are generated by resampling noise 5 times for each timestep ($N = 5$).

Method	FID ↓	mIoU ↑
OASIS [18]	20.64	77.35
Ours $p = 0.5$	30.45	71.51
Ours $p = 1.0$	33.25	66.81
Ours $p = 3.0$	47.89	40.09

Table A.2. Quantitative results for layout-to-image generation on MM CelebAMask-HQ-25 dataset [9]. mIoU is measured between the input layout and the segmentation results of the generated image using a pretrained HRNet [20].

A.7. Ablation study and additional baselines

In this section, we provide results for different text-to-image generation approaches and compare them against our approach. First, we train a Gaussian diffusion model with an identical architecture as our model which generates images *without* the corresponding layouts. The visualization in Figure 8. of the main paper demonstrate that the internal features of this Gaussian-categorical diffusion model form distinct clusters compared to the Gaussian diffusion model.

Second, we present a text-to-image generation approach that leverages semantic segmentation labels during training. Given text inputs, we sequentially generate layouts from texts and then images from the generated layouts. Specifically, we train a categorical diffusion model [1] for text-to-layout generation and a layout-to-image synthesis model called SDM [21]. We train a modified version of SDM to incorporate text conditions to generate image from layouts.

To provide quantitative results, we report the FID-CLIP score pairs for the MM CelebA-HQ-25 in Figure A.5. Our approach effectively enhances the performance of the Gaussian diffusion model by simultaneously generating corresponding semantic layouts. Also, our simultaneous generation of images and layouts outperforms the sequential generation from text to layouts and then to images.

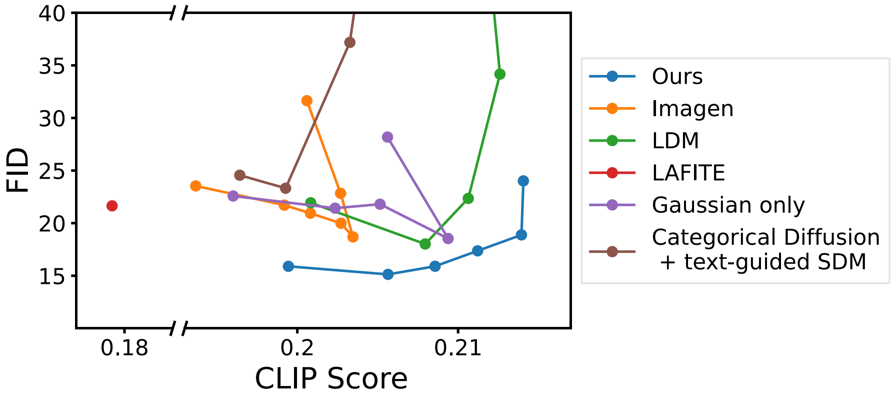
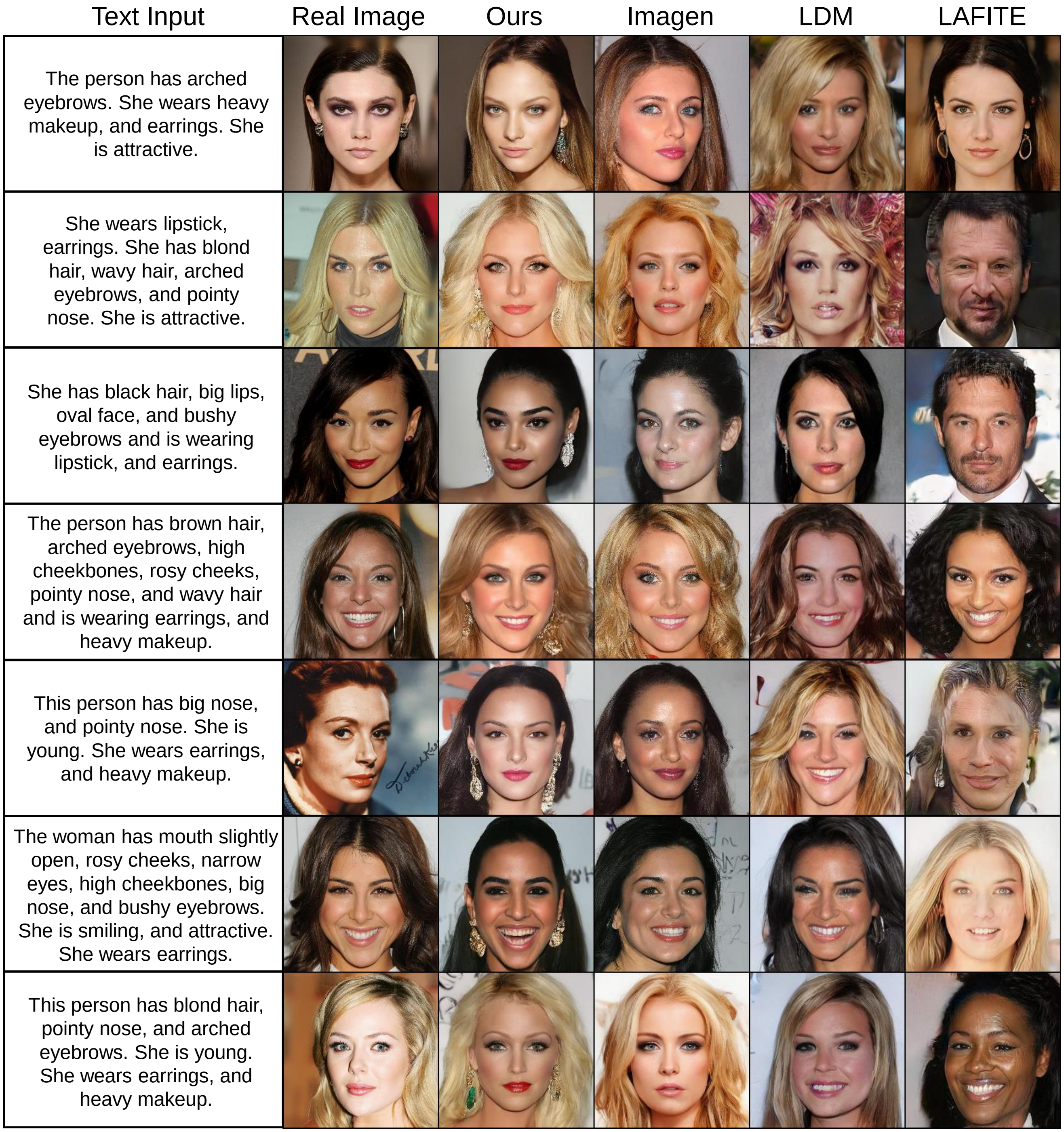


Figure A.5. FID-CLIP scores for the Gaussian diffusion on the MM CelebA-HQ-25 dataset, compared against existing approaches and the Gaussian-categorical diffusion.

A.8. Qualitative comparison

We provide the qualitative results from existing text-to-image generation models, and the Gaussian-categorical diffusion trained on MM CelebA-HQ-25 in the remaining supplementary material (Figure A.6, Figure A.7, and Figure A.8). Since diffusion-based models produce different results based on the guidance scale of the classifier-free guidance, we sample images from results exhibiting FID around 20. The guidance scales for each model to achieve an FID of 20 are 2, 10, and 10 for LDM, Imagen, and the Gaussian-categorical diffusion, respectively. We also provide uncurated results for generated image-layout pairs from the Gaussian-categorical diffusion model in Figure A.9 and Figure A.10.



The person has arched eyebrows. She wears heavy makeup, and earrings. She is attractive.

She wears lipstick, earrings. She has blond hair, wavy hair, arched eyebrows, and pointy nose. She is attractive.

She has black hair, big lips, oval face, and bushy eyebrows and is wearing lipstick, and earrings.

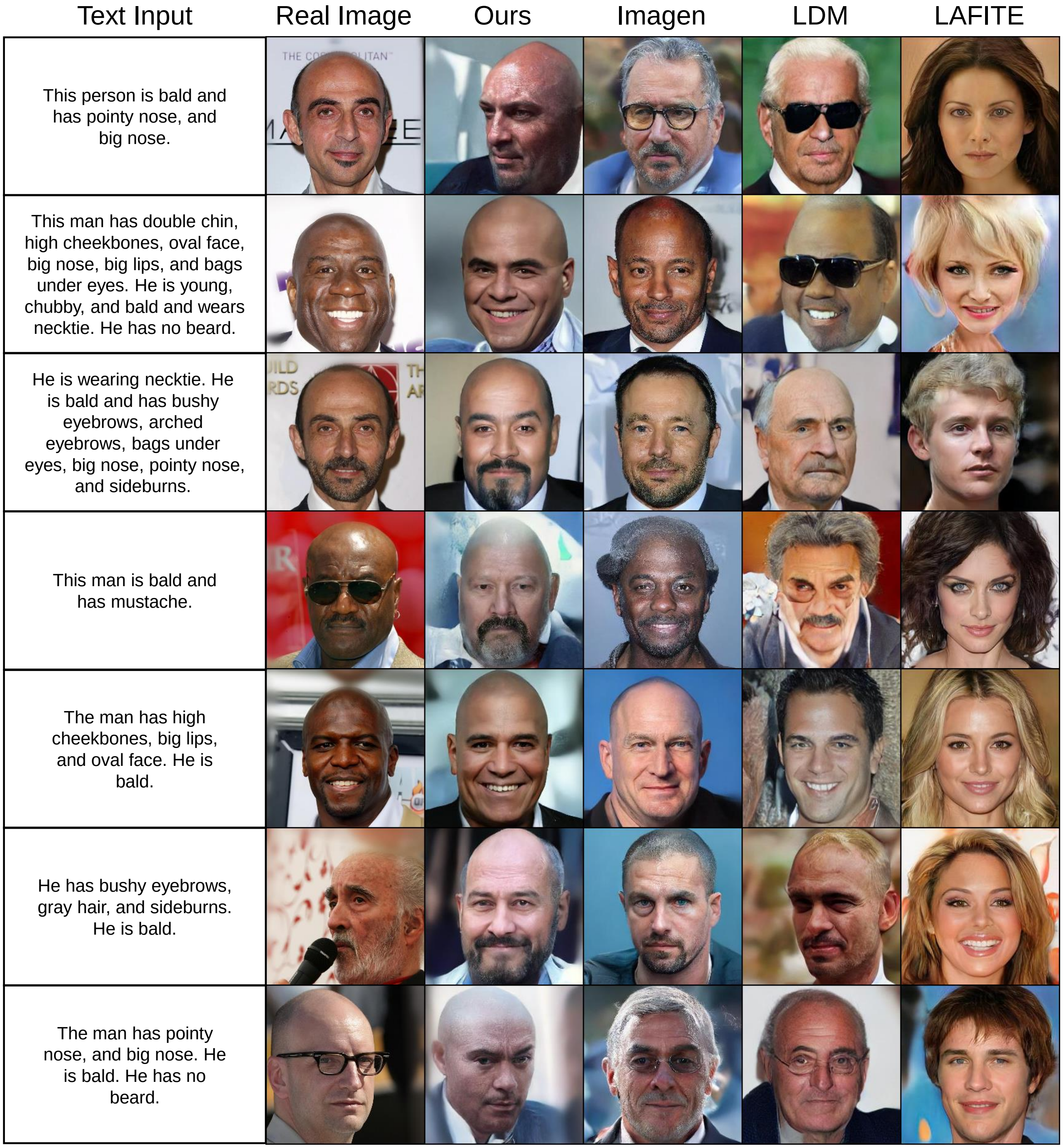
The person has brown hair, arched eyebrows, high cheekbones, rosy cheeks, pointy nose, and wavy hair and is wearing earrings, and heavy makeup.

This person has big nose, and pointy nose. She is young. She wears earrings, and heavy makeup.

The woman has mouth slightly open, rosy cheeks, narrow eyes, high cheekbones, big nose, and bushy eyebrows. She is smiling, and attractive. She wears earrings.

This person has blond hair, pointy nose, and arched eyebrows. She is young. She wears earrings, and heavy makeup.

Figure A.6. Qualitative comparison between the Gaussian-categorical diffusion model and existing text-to-image generation models on MM CelebA-HQ-25. We observe that existing models struggle to generate accessories such as earrings.



This person is bald and has pointy nose, and big nose.

This man has double chin, high cheekbones, oval face, big nose, big lips, and bags under eyes. He is young, chubby, and bald and wears necktie. He has no beard.

He is wearing necktie. He is bald and has bushy eyebrows, arched eyebrows, bags under eyes, big nose, pointy nose, and sideburns.

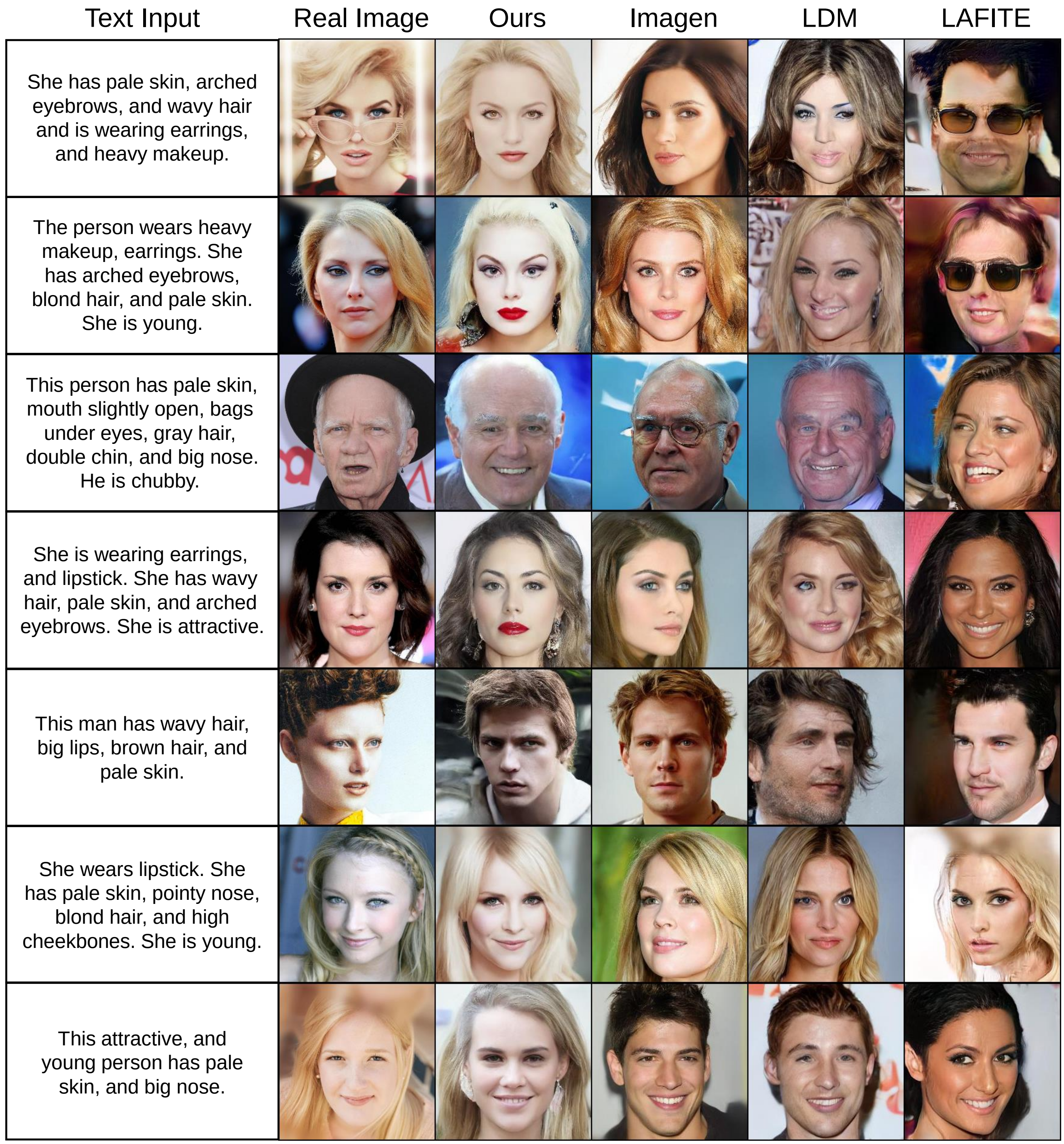
This man is bald and has mustache.

The man has high cheekbones, big lips, and oval face. He is bald.

He has bushy eyebrows, gray hair, and sideburns. He is bald.

The man has pointy nose, and big nose. He is bald. He has no beard.

Figure A.7. Qualitative comparison between the Gaussian-categorical diffusion model and existing text-to-image generation models on MM CelebA-HQ-25. We observe that existing models tend to generate hair even when given text conditions specifying baldness.



She has pale skin, arched eyebrows, and wavy hair and is wearing earrings, and heavy makeup.

The person wears heavy makeup, earrings. She has arched eyebrows, blond hair, and pale skin. She is young.

This person has pale skin, mouth slightly open, bags under eyes, gray hair, double chin, and big nose. He is chubby.

She is wearing earrings, and lipstick. She has wavy hair, pale skin, and arched eyebrows. She is attractive.

This man has wavy hair, big lips, brown hair, and pale skin.

She wears lipstick. She has pale skin, pointy nose, blond hair, and high cheekbones. She is young.

This attractive, and young person has pale skin, and big nose.

Figure A.8. Qualitative comparison between the Gaussian-categorical diffusion model and existing text-to-image generation models on MM CelebA-HQ-25. We observe that existing approaches often fail to appropriately generate colors of skin.

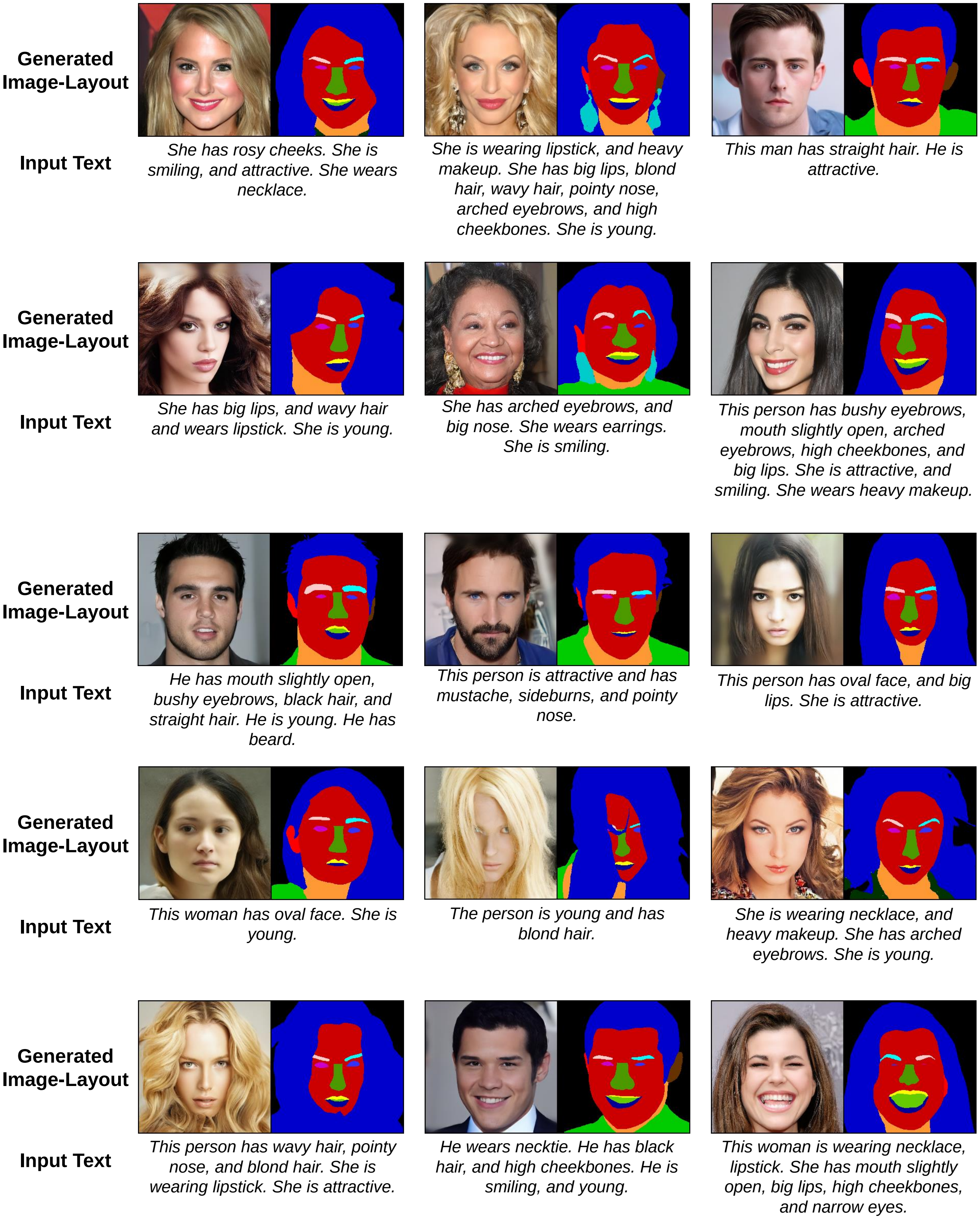
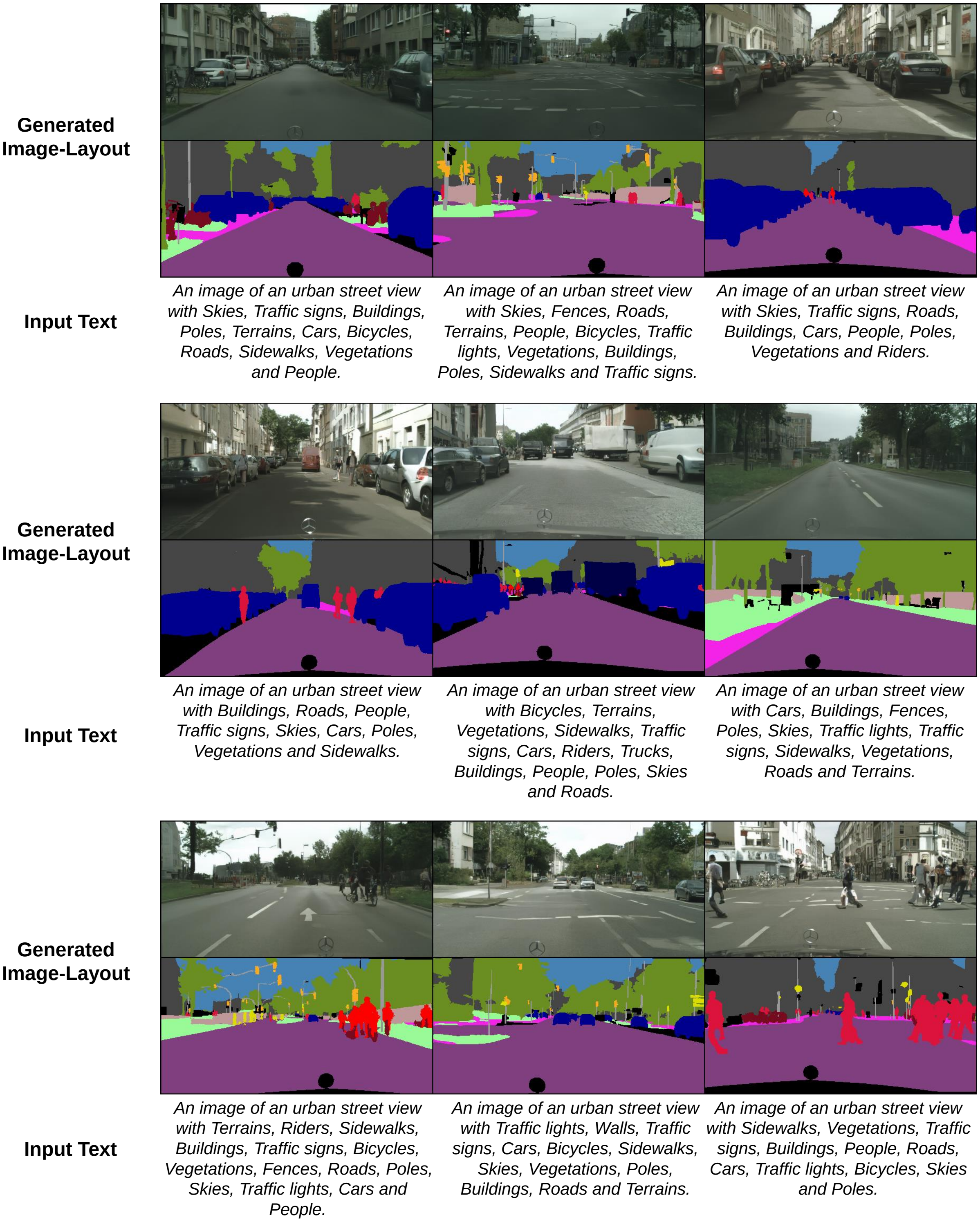

Generated Image-Layout
Input Text
She has rosy cheeks. She is smiling, and attractive. She wears necklace.
She is wearing lipstick, and heavy makeup. She has big lips, blond hair, wavy hair, pointy nose, arched eyebrows, and high cheekbones. She is young.
This man has straight hair. He is attractive.
Generated Image-Layout
Input Text
She has big lips, and wavy hair and wears lipstick. She is young.
She has arched eyebrows, and big nose. She wears earrings. She is smiling.
This person has bushy eyebrows, mouth slightly open, arched eyebrows, high cheekbones, and big lips. She is attractive, and smiling. She wears heavy makeup.
Generated Image-Layout
Input Text
He has mouth slightly open, bushy eyebrows, black hair, and straight hair. He is young. He has beard.
This person is attractive and has mustache, sideburns, and pointy nose.
This person has oval face, and big lips. She is attractive.
Generated Image-Layout
Input Text
This woman has oval face. She is young.
The person is young and has blond hair.
She is wearing necklace, and heavy makeup. She has arched eyebrows. She is young.
Generated Image-Layout
Input Text
This person has wavy hair, pointy nose, and blond hair. She is wearing lipstick. She is attractive.
He wears necktie. He has black hair, and high cheekbones. He is smiling, and young.
This woman is wearing necklace, lipstick. She has mouth slightly open, big lips, high cheekbones, and narrow eyes.

Figure A.9. Example image-layout pairs generated by the Gaussian-categorical diffusion trained on the MM CelebA-HQ-100 dataset.



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

An image of an urban street view with Skies, Fences, Roads, Terrains, People, Bicycles, Traffic lights, Vegetations, Buildings, Poles, Sidewalks and Traffic signs.

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

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

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

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

An image of an urban street view with Terrains, Riders, Sidewalks, Buildings, Traffic signs, Bicycles, Vegetations, Fences, Roads, Poles, Skies, Traffic lights, Cars and People.

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

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

Figure A.10. Example image-layout pairs generated by the Gaussian-categorical diffusion trained on the cityscapes dataset.

References

[1] Jacob Austin, Daniel D Johnson, Jonathan Ho, Daniel Tarlow, and Rianne van den Berg. Structured denoising diffusion models in discrete state-spaces. *NeurIPS*, 34:17981–17993, 2021. 11

[2] Liang-Chieh Chen, George Papandreou, Florian Schroff, and Hartwig Adam. Rethinking atrous convolution for semantic image segmentation. *arXiv preprint arXiv:1706.05587*, 2017. 10

[3] Xin Wang Wenqi Xian Yingying Chen, Fangchen Liu Vashisht Madhavan Trevor Darrell, Fisher Yu, and Haofeng Chen. Bdd100k: A diverse driving dataset for heterogeneous multitask learning. *arXiv preprint arXiv: 1805.04687*, 2018. 8

[4] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The cityscapes dataset for semantic urban scene understanding. In *CVPR*, 2016. 7, 8, 9

[5] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *CVPR*, pages 248–255. Ieee, 2009. 8

[6] Li Deng. The mnist database of handwritten digit images for machine learning research [best of the web]. *IEEE signal processing magazine*, 29(6):141–142, 2012. 8

[7] Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. Clipscore: A reference-free evaluation metric for image captioning. *arXiv preprint arXiv:2104.08718*, 2021. 7

[8] Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. Progressive growing of gans for improved quality, stability, and variation. In *ICLR*, 2018. 10

[9] Cheng-Han Lee, Ziwei Liu, Lingyun Wu, and Ping Luo. Maskgan: Towards diverse and interactive facial image manipulation. In *CVPR*, 2020. 10

[10] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *ECCV*. Springer, 2014. 7, 8

[11] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *ICCV*, December 2015. 7, 8, 10

[12] Andreas Lugmayr, Martin Danelljan, Andres Romero, Fisher Yu, Radu Timofte, and Luc Van Gool. Repaint: Inpainting using denoising diffusion probabilistic models. In *CVPR*, pages 11461–11471, 2022. 10

[13] Yuval Netzer, Tao Wang, Adam Coates, Alessandro Bissacco, Bo Wu, and Andrew Y Ng. Reading digits in natural images with unsupervised feature learning. *NIPS Workshop*, 2011. 8

[14] Alexander Quinn Nichol and Prafulla Dhariwal. Improved denoising diffusion probabilistic models. In *International Conference on Machine Learning*, pages 8162–8171. PMLR, 2021. 7

[15] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *ICML*, pages 8748–8763. PMLR, 2021. 7, 8

[16] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *CVPR*, 2022. 7, 8

[17] Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, et al. Laion-5b: An open large-scale dataset for training next generation image-text models. *arXiv preprint arXiv:2210.08402*, 2022. 7

[18] Vadim Sushko, Edgar Schönfeld, Dan Zhang, Juergen Gall, Bernt Schiele, and Anna Khoreva. Oasis: Only adversarial supervision for semantic image synthesis. *IJCV*, 130(12):2903–2923, 2022. 10

[19] Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *JMLR*, 9(11), 2008. 8

[20] Jingdong Wang, Ke Sun, Tianheng Cheng, Borui Jiang, Chaorui Deng, Yang Zhao, Dong Liu, Yadong Mu, Mingkui Tan, Xinggang Wang, et al. Deep high-resolution representation learning for visual recognition. *IEEE TPAMI*, 43(10):3349–3364, 2020. 8, 10

[21] Weilun Wang, Jianmin Bao, Wengang Zhou, Dongdong Chen, Dong Chen, Lu Yuan, and Houqiang Li. Semantic image synthesis via diffusion models. *arXiv*, 2022. 11

[22] Bolei Zhou, Hang Zhao, Xavier Puig, Tete Xiao, Sanja Fidler, Adela Barriuso, and Antonio Torralba. Semantic understanding of scenes through the ade20k dataset. *IJCV*, 127:302–321, 2019. 8