

Supplementary Materials: PROMOTE: Prior-Guided Diffusion Model with Global-Local Contrastive Learning for Exemplar-Based Image Translation

Anonymous Authors

A PROOF OF THEOREM 1

We model image representation in a hyperbolic space defined by Riemannian metric to better explore cross-domain visual correspondence. With this special Riemannian manifold, we can introduce the concept of distance between any two distributions p_{z_0} and p_{z_1} .

Formally, considering a Riemannian manifold \mathcal{M} and the tangent space at a certain location $T_z(\mathcal{M})$, we can leverage a metric tensor $G : T_z(\mathcal{M}) \times T_z(\mathcal{M}) \rightarrow \mathbb{R}$ on the tangent space $T_z(\mathcal{M})$ to quantify the length of tangent vector $v \in T_z(\mathcal{M})$ via $\|v\|_G := \sqrt{G(v, v)}$. This metric tensor allows us to compute the length of the curve γ in Fisher geometry, i.e. the integral of tangent vector length along γ :

$$\begin{aligned} L(\gamma) &:= \int_0^1 \sqrt{\langle \dot{\gamma}(t), \dot{\gamma}(t) \rangle_{G(z(t))}} dt \\ &= \int_0^1 \sqrt{\dot{z}(t)^T G(z(t)) \dot{z}(t)} dt. \end{aligned} \quad (1)$$

In local coordinates $z = (z^1, \dots, z^n)$, the curve $z(t)$ is a solution to the geodesic $\gamma(t)$ differential equation:

$$\ddot{z}^k + \sum_{i=1}^n \sum_{j=1}^n \Gamma_{ij}^k \dot{z}^i \dot{z}^j = 0, \quad k = 1, \dots, n, \quad (2)$$

where Γ_{ij}^k are the Christoffel symbols of the second kind, which can be calculated by the equation:

$$\sum_{k=1}^n g_{lk} \Gamma_{ij}^k = \frac{1}{2} \left(\frac{\partial}{\partial z^i} g_{jl} + \frac{\partial}{\partial z^j} g_{li} - \frac{\partial}{\partial z^l} g_{ij} \right). \quad (3)$$

Referring to Theorem A.1, the Fisher-Rao distance between p_{z_0} and p_{z_1} in Riemannian manifold \mathcal{M} is the infimum length of the differentiable curve γ joining them:

$$\begin{aligned} d_{FR}(z_0, z_1) &:= d_{FR}(p_{z_0}, p_{z_1}) := \inf_{\gamma} L(\gamma) \\ \text{s.t. } &\gamma(0) = p_{z_0}, \gamma(1) = p_{z_1}. \end{aligned} \quad (4)$$

THEOREM A.1. (Hopf-Rinow Theorem). *Let \mathcal{M} be a connected Riemannian manifold, and suppose that it is complete as a metric space (i.e., every Cauchy sequence converges). Then we have:*

(1) *For any two points $p, q \in \mathcal{M}$ there exists a length-minimizing geodesic between them (i.e., a curve whose length is equal to the Fisher-Rao distance $d_{FR}(p, q)$).*

(2) *For every point $z \in \mathcal{M}$ the exponential map \exp_p is defined all over the tangent space $T_z(\mathcal{M})$.*

$$D_{KL}(p_{z_0} || p_{z_1}) = \frac{1}{2} d_{FR}^2(p_{z_0}, p_{z_1}) + o(d_{FR}^2(p_{z_0}, p_{z_1})), \quad (5)$$

where $o(\cdot)$ denotes the limit quantity $\lim_{x \rightarrow 0} \frac{o(x)}{x} = 0$. Notably, computing Fisher-Rao distance between two distributions in a statistical manifold requires solving the geodesic differential equation (Eq. 2) and then calculating the integral (Eq. 1), which is difficult. So

we instead compute the Fisher-Rao distance on a manifold parameterized by a single real number since the geodesics are immediately given.

In the case of this manifold, Fisher matrix $G(z) = [g_{11}(z)]$, there is only one path connecting any two distributions and its length does not depend on the parameter selection. Therefore, the length of curve γ and the Fisher-Rao distance between p_{z_0} and p_{z_1} can be rewritten as Eq. 6 and Eq. 7 respectively:

$$L(\gamma) = \int_{z_0}^{z_1} \sqrt{g_{11}(z(t))} dt = \int_{z_0}^{z_1} \sqrt{g_{11}(z)} dz, \quad (6)$$

$$d_{FR}(z_0, z_1) = \left| \int_{z_0}^{z_1} \sqrt{g_{11}(z)} dz \right|. \quad (7)$$

The probability density function of a Gaussian distribution Gaussian is $p(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$, defined for $x \in \mathbb{R}$, and parametrised by the mean μ and variance σ . Thus $\partial_{\mu} l := \partial_{\mu} l(\mu, \sigma) = \frac{x-\mu}{\sigma^2}$ and $\partial_{\sigma} l := \partial_{\sigma} l(\mu, \sigma) = -\frac{1}{\sigma} + \frac{(x-\mu)^2}{\sigma^3}$. The elements of the Fisher matrix are:

$$g_{11} = \mathbb{E}[(\partial_{\mu} l)^2] = \mathbb{E}\left[\left(\frac{X-\mu}{\sigma^2}\right)^2\right] = \frac{\mathbb{E}[(X-\mu)^2]}{\sigma^4} = \frac{1}{\sigma^2}, \quad (8)$$

$$\begin{aligned} g_{12} = g_{21} &= \mathbb{E}[(\partial_{\mu} l)(\partial_{\sigma} l)] = \mathbb{E}\left[\left(\frac{X-\mu}{\sigma^2}\right)\left(-\frac{1}{\sigma} + \frac{(X-\mu)^2}{\sigma^3}\right)\right] \\ &= -\frac{\mathbb{E}[X-\mu]}{\sigma^3} + \frac{\mathbb{E}[(X-\mu)^3]}{\sigma^5} = 0, \end{aligned} \quad (9)$$

$$\begin{aligned} g_{22} &= \mathbb{E}[(\partial_{\sigma} l)^2] = \mathbb{E}\left[\left(-\frac{1}{\sigma} + \frac{(X-\mu)^2}{\sigma^3}\right)^2\right] \\ &= \frac{1}{\sigma^2} - \frac{2\mathbb{E}[(X-\mu)^2]}{\sigma^4} + \frac{\mathbb{E}[(X-\mu)^4]}{\sigma^6} = \frac{2}{\sigma^2}, \end{aligned} \quad (10)$$

where we set the odd centralised moments of the Gaussian random variable to 0 and let $\mathbb{E}[(X-\mu)^2] = \sigma^2$ and $\mathbb{E}[(X-\mu)^4] = 3\sigma^4$. Thus the Fisher matrix denoted as:

$$G = \begin{bmatrix} \frac{1}{\sigma^2} & 0 \\ 0 & \frac{2}{\sigma^2} \end{bmatrix}. \quad (11)$$

The Poincaré half-plane model $\mathcal{H} := \{z = x + iy \in \mathbb{C} | \text{Im}(z) > 0\}$ equipped with the hyperbolic metric is employed to find the geodesics instead of solving geodesic differential equations, denoted as:

$$G_{\mathcal{H}}(x, y) = \begin{bmatrix} \frac{1}{y^2} & 0 \\ 0 & \frac{2}{y^2} \end{bmatrix}. \quad (12)$$

In this case, the geodesics in manifold are vertical half-lines and half-circles centred at $y = 0$, and the geodesic distance between

points $a, b \in \mathcal{H}$ is given by:

$$\begin{aligned} d_{\mathcal{H}}(a, b) &= \log \frac{|a - \bar{b}| + |a - b|}{|a - \bar{b}| - |a - b|} \\ &= \operatorname{arccosh}\left(1 + \frac{|a - b|^2}{2\operatorname{Im}(a)\operatorname{Im}(b)}\right) \\ &= 2\operatorname{arctanh}\left|\frac{a - b}{a - \bar{b}}\right|. \end{aligned} \quad (13)$$

Then we can combine the Fisher-Rao distance in the Gaussian manifold with the distance in the Poincaré half-plane based on 12 and 13:

$$d_{FR}((\mu_1, \sigma_1), (\mu_2, \sigma_2)) = \sqrt{2}d_{\mathcal{H}}\left(\frac{\mu_1}{\sqrt{2}} + i\sigma_1, \frac{\mu_2}{\sqrt{2}} + i\sigma_2\right). \quad (14)$$

Therefore, the Fisher-Rao distance is:

$$\begin{aligned} d_{FR}((\mu_1, \sigma_1), (\mu_2, \sigma_2)) &= 2\sqrt{2}\operatorname{arctanh}\left(\sqrt{\frac{(\mu_1 - \mu_2)^2 + 2(\sigma_1 - \sigma_2)^2}{(\mu_1 - \mu_2)^2 + 2(\sigma_1 + \sigma_2)^2}}\right). \end{aligned} \quad (15)$$

B DERIVATION OF POSTERIOR

Now we model $\tilde{q}(z_{t-1}|z_t, z_0)$ via Bayes theorem:

$$\begin{aligned} \tilde{q}(z_{t-1}|z_t, z_0) &= \frac{\tilde{q}(z_{t-1}, z_t, z_0)}{\tilde{q}(z_t, z_0)} = \frac{\tilde{q}(z_t|z_{t-1}, z_0) \cdot \tilde{q}(z_{t-1}|z_0)}{\tilde{q}(z_t|z_0)} \\ &= \frac{\tilde{q}(z_t|z_{t-1}) \cdot \tilde{q}(z_{t-1}|z_0)}{\tilde{q}(z_t|z_0)} \\ &= \frac{N(z_t; \sqrt{\alpha_t}z_{t-1} + (1 - \sqrt{\alpha_t})r, \beta_t\eta^2\mathbf{I})}{N(z_t; \sqrt{\alpha_t}z_0 + \sqrt{1 - \alpha_t}r, (1 - \alpha_t)\eta^2\mathbf{I})} \\ &= \frac{N(z_{t-1}; \sqrt{\alpha_{t-1}}z_0 + \sqrt{1 - \alpha_{t-1}}r, (1 - \alpha_{t-1})\eta^2\mathbf{I})}{N(z_t; \sqrt{\alpha_t}z_0 + \sqrt{1 - \alpha_t}r, (1 - \alpha_t)\eta^2\mathbf{I})}, \end{aligned} \quad (16)$$

then we substitute the Gaussian probability density function $p(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$:

$$\begin{aligned} \tilde{q}(z_{t-1}|z_t, z_0) &= \frac{\exp\left\{-\left[\frac{(z_t - \sqrt{\alpha_t}z_{t-1} - (1 - \sqrt{\alpha_t})r)^2}{2\beta_t\eta^2}\right]\right\}}{\exp\left\{-\left[\frac{(z_t - \sqrt{\alpha_t}z_0 - \sqrt{1 - \alpha_t}r)^2}{2(1 - \alpha_t)\eta^2}\right]\right\}} \\ &\quad \frac{\exp\left\{-\left[\frac{(z_t - \sqrt{\alpha_{t-1}}z_0 - \sqrt{1 - \alpha_{t-1}}r)^2}{2(1 - \alpha_{t-1})\eta^2}\right]\right\}}{\exp\left\{-\left[\frac{(z_t - \sqrt{\alpha_t}z_0 - \sqrt{1 - \alpha_t}r)^2}{2(1 - \alpha_t)\eta^2}\right]\right\}} \\ &= \exp\left\{-\frac{\frac{\sqrt{\alpha_t}\delta_{t-1}z_t + \sqrt{\alpha_{t-1}}\beta_t z_0}{\delta_t} + \frac{2\delta_{t-1}\beta_t - \sqrt{\alpha_t}\delta_{t-1}\eta}{\delta_t}r}{2\frac{1 - \alpha_{t-1}\beta_t}{1 - \alpha_t}\eta^2}\right\} \\ &:= N(z_{t-1}; \tilde{\mu}_q(z_t, z_0), \tilde{\Sigma}_q(z_t, z_0)), \end{aligned}$$

where $\tilde{\mu}_q(z_t, z_0) = \frac{\sqrt{\alpha_t}\delta_{t-1}z_t + \sqrt{\alpha_{t-1}}\beta_t z_0}{\delta_t} + \frac{2\delta_{t-1}\beta_t - \sqrt{\alpha_t}\delta_{t-1}\eta}{\delta_t}r$,

$\delta_t = 1 - \alpha_t$, and $\tilde{\Sigma}_q(z_t, z_0) = \frac{\delta_{t-1}\beta_t}{\delta_t}\eta^2$. (17)

C DERIVATION OF DIFFUSION LOSS

Now we further model $\tilde{p}_\theta(z_{t-1}|z_t) = N(\tilde{\mu}_\theta, \tilde{\Sigma}_\theta)$ and attempt to minimize the KL-Divergence between $\tilde{q}(z_{t-1}|z_t, z_0)$ and $\tilde{p}_\theta(z_{t-1}|z_t)$. Following conventional DDPM, the variance $\tilde{\Sigma}_\theta = \tilde{\Sigma}_q$ and the KL-Divergence denoted as:

$$\begin{aligned} D_{KL}(\tilde{q}(z_{t-1}|z_t, z_0) || \tilde{p}_\theta(z_{t-1}|z_t)) &= D_{KL}(N(z_{t-1}; \tilde{\mu}_q, \tilde{\Sigma}) || N(z_{t-1}; \tilde{\mu}_\theta, \tilde{\Sigma}_\theta)) \\ &= \frac{1}{2} \left[\log \frac{|\tilde{\Sigma}_\theta|}{|\tilde{\Sigma}_q|} - d + \operatorname{tr}(\tilde{\Sigma}_\theta^{-1}\tilde{\Sigma}_\theta) + (\tilde{\mu}_\theta - \tilde{\mu}_q)^T \tilde{\Sigma}_q^{-1}(\tilde{\mu}_\theta - \tilde{\mu}_q) \right] \\ &= \frac{1}{2} \left[\log 1 - d + d + (\tilde{\mu}_\theta - \tilde{\mu}_q)^T \tilde{\Sigma}_q^{-1}(\tilde{\mu}_\theta - \tilde{\mu}_q) \right] \\ &= \frac{1}{2} [(\tilde{\mu}_\theta - \tilde{\mu}_q)^T \tilde{\Sigma}_q^{-1}(\tilde{\mu}_\theta - \tilde{\mu}_q)] \\ &= \frac{1}{2} [(\tilde{\mu}_\theta - \tilde{\mu}_q)^T (\tilde{\sigma}_q^2 \mathbf{I})^{-1}(\tilde{\mu}_\theta - \tilde{\mu}_q)] \\ &= \frac{1}{2\tilde{\sigma}_q^2} [\|\tilde{\mu}_\theta - \tilde{\mu}_q\|_2^2]. \end{aligned} \quad (18)$$

Then we set $\rho_t = \frac{2\delta_{t-1}\beta_t - \sqrt{\alpha_t}\delta_{t-1}\eta}{\delta_t}$, and parameterize the true noise by establishing a neural network ϵ_θ and transform the complex mean difference optimization into minimizing the difference between predicted noise and true noise ϵ , obtaining the proposed prior denoising loss:

$$\begin{aligned} \mathcal{L}_{\text{diff}} &= \|\tilde{\mu}_q(z_t, z_0) - \tilde{\mu}_\theta(z_t)\|^2 \\ &= \|\rho_t r + \epsilon - \epsilon_\theta(\sqrt{\alpha_t}z_0 + \sqrt{1 - \alpha_t}\eta\epsilon + \sqrt{1 - \alpha_t}r, t)\|^2, \end{aligned} \quad (19)$$

D LIMITATIONS

Although our PROMOTE generally outperforms previous methods, it does not sufficiently preserve the texture and details of the example (e.g., facial wrinkles, human eye and lipstick colors) during image translation, as illustrated in Fig. 1. In the future we will explore how to construct superior ground truth to achieve more effective self-supervised training, and attempt to deploy existing text-to-image large models to exemplar-based image translation task to obtain more realistic and controllable generated images.

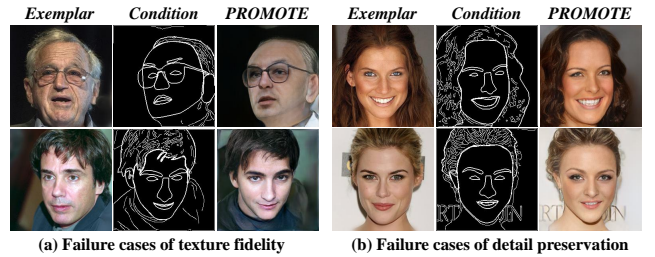


Figure 1: Several failure results on CelebA-HQ dataset.

E ADDITIONAL QUALITATIVE RESULTS

More qualitative results on seven datasets are presented in Fig. 2 and Fig. 3. We can observe that the images translated by our PROMOTE possess the style of exemplars while well maintaining the structure of conditional images across all benchmarks.

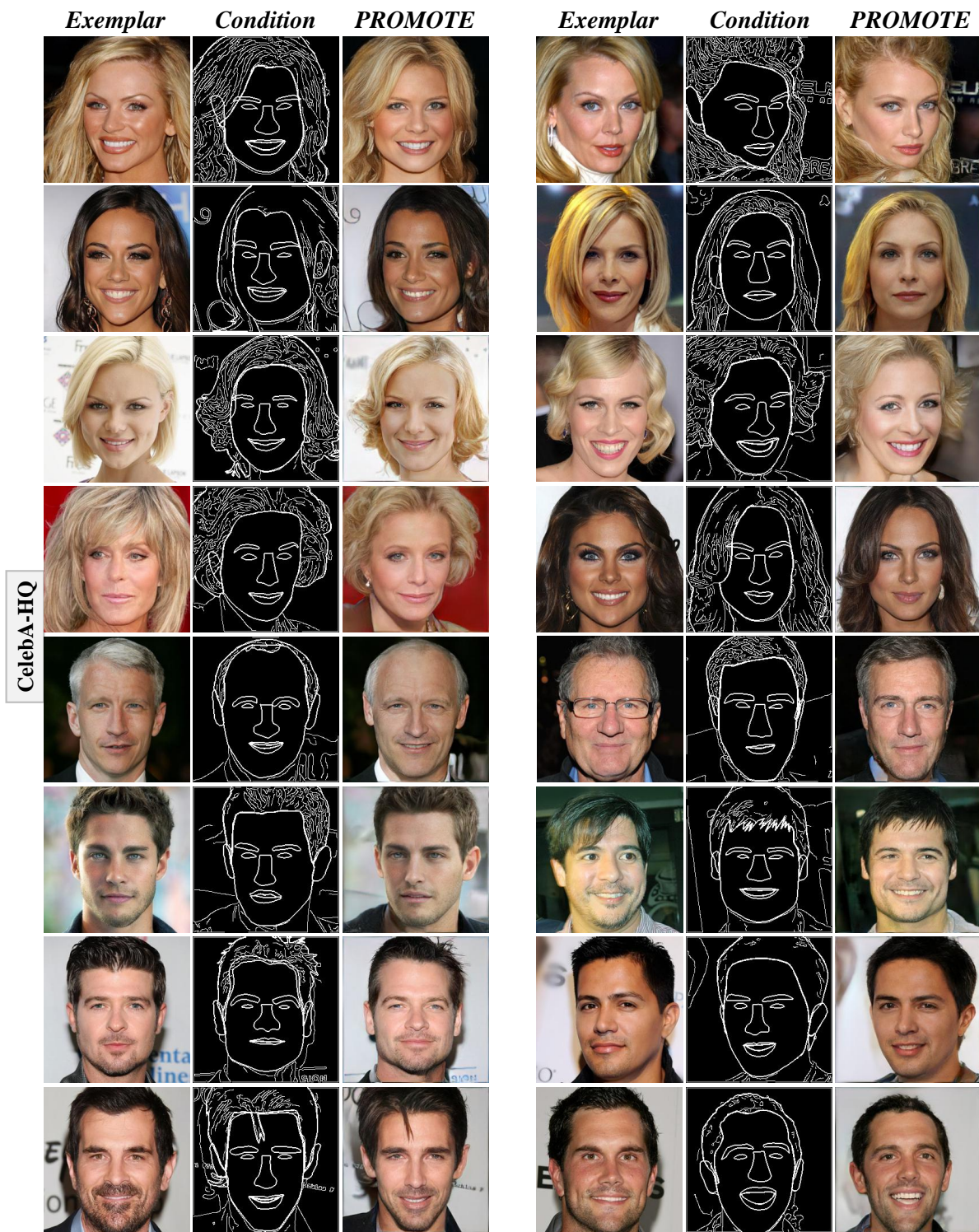


Figure 2: Additional qualitative results on CelebA-HQ dataset.

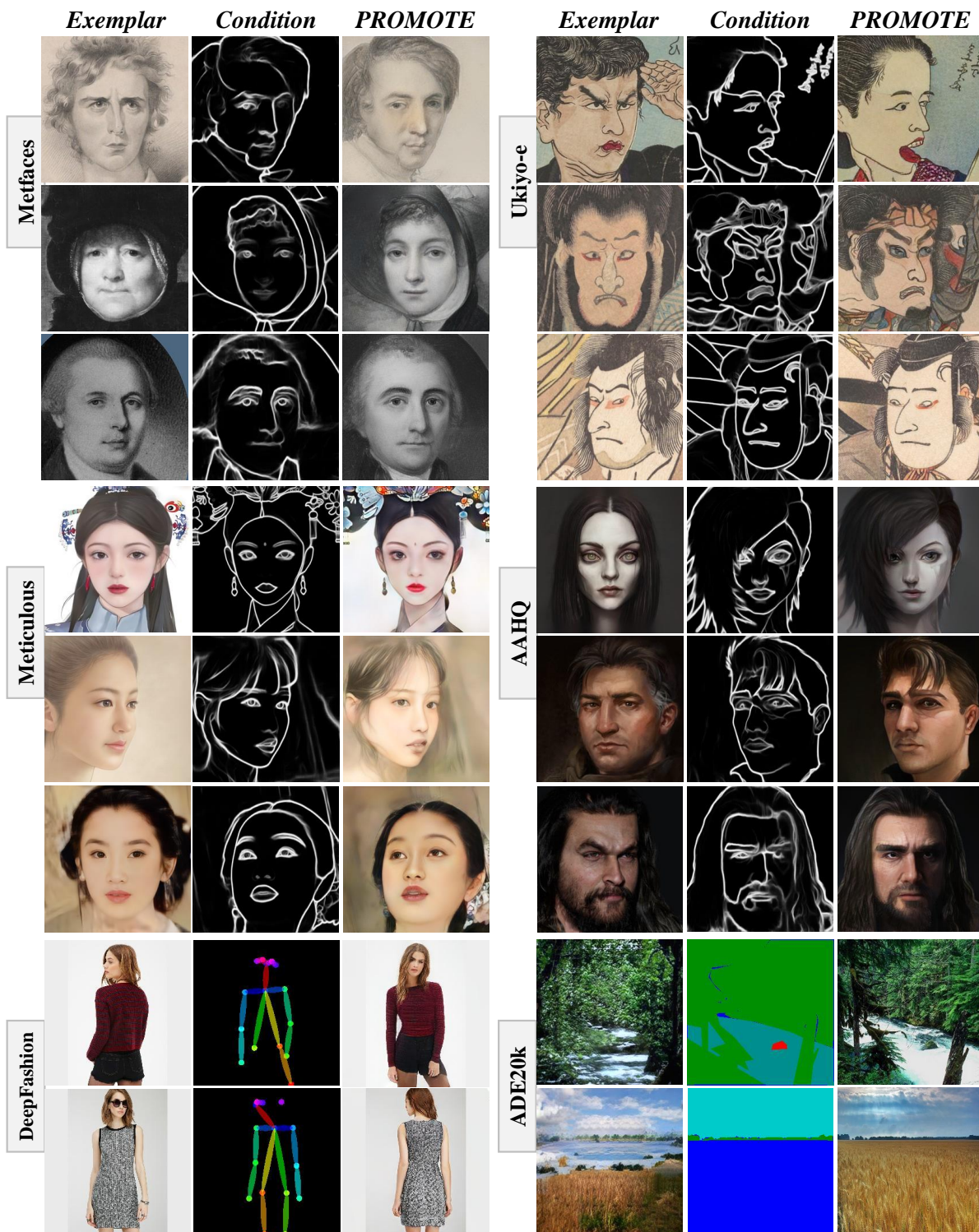


Figure 3: Additional qualitative results on Metfaces, Meticulous, Ukiyo-e, AAHQ, DeepFashion and ADE20k datasets.