We thank the reviewers for fruitful comments. Here, we respond to the major comments. For the minor points like presentation issues, we will fix them in the camera-ready (and *we do not mention them here due to space limitation.*)

Reviewer #1 (i) We could not compare our method to [21], because [21] was published in this April and their code was not available. (ii) As the reviewer mentioned, a couple of images, say c_{smile} and $c_{\text{non-smile}}$, were used as conditions for smiling and non-smiling images, respectively. Our model is trained using $\{(x, c_{\text{smile}}) \mid x \in X_{\text{smile}}\}$ and $\{(x, c_{\text{non-smile}}) \mid x \in X_{\text{non-smile}}\}$, where X_{smile} and $X_{\text{non-smile}}$ are sets consisting of smiling and not-smiling images in CelebA, respectively. Our model then learns the difference between $x \in X_{\text{smile}}$ and $x \in X_{\text{non-smile}}$ in the embedded space, in which c_{smile} and $c_{\text{not-smile}}$ are only used to distinguish whether $x \in X_{\text{smile}}$ or $x \in X_{\text{non-smile}}$. In contrast, our CHC experiment demonstrates a situation in which the condition image (i.e., the initial microstructure) of each image has more diversity. (iii) We used Inception-V3 to calculate FID for CHC. (iv) We apologize that the description "make a part of z follow the non-informative distribution $\mathcal{N}(0, 1)$" in line 125 of our submitted manuscript was incorrect. It should be modified as "make a part of z follow $\mathcal{N}(f_\mu(b_\mu), \text{diag}(\exp f_\sigma(b_\sigma)))$ that is independent of c".

Reviewer #2 (i) We trained FUNS using CelebA with Male-Female conditions. An interpolated result is shown in Fig. A. We can see that z is successfully disentangled into the condition dependent (z^{sp}) and independent (z^{iv}) parts.

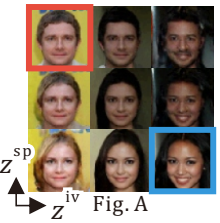

Fig. A

(ii) We trained a ResNet to predict smile/non-smile for CelebA, and then, evaluated the prediction accuracy for images generated by FUNS, PUNet, and VUNet, as well as the real test set of CelebA. For CHC, we trained in-house U-Net to predict the initial condition, and then, evaluated the prediction error using L_2 distance between the ground truth and the prediction. The results are summarized in Table A. Note that the prediction error for the CHC prediction is 22.84 when the ground truth are randomly shuffled. We see that every model can generate images that depend on the conditions in terms of the prediction accuracy. PUNet and VUNet achieved higher prediction accuracy compared to Real/FUNS in both CelebA and CHC. This may be because PUNet and VUNet generated similar images that are easy to predict the condition. As the c-LPIPS scores suggest, PUNet and VUNet generated less diverse images compared to Real/FUNS.

(Table A)	CelebA		CHC	
	Acc.	c-LPIPS	Err.	c-LPIPS
Real	0.924	0.284	9.54	0.169
FUNS	0.973	0.262	9.67	0.157
PUNet	1.000	0.180	9.48	0.108
VUNet	0.977	0.146	9.46	0.118

(iii) We have not tried to train FUNS with CelebA-HQ. It is uncertain if FUNS can be successfully trained using HQ images (further techniques might be required). However, as shown in Table 1 of our paper, FUNS has advantages over VAE-based I2I models (incl. PUNet, VUNet) in the image quality (FID) and diversity (LPIPS). The main advantage of flow-based models over GAN-based ones is, as discussed in [14], that they have invertible mappings between images and latent codes, which will be useful for downstream tasks, e.g., Gómez-Bombarelli, R., et al., *ACS Cent. Sci.*, **4**, (2018), 268–276.

(Table B)	Err.	$\dim(z^{\text{sp}})$
$\mathcal{L}^{\text{flow}}$	9.56	4,096
$+\mathcal{L}^{\text{recons}}$	9.51	4,040
$+\mathcal{L}^{\text{squeeze}}$	9.58	1,095
$+\mathcal{L}^{\text{entropy}}$	9.67	550

(iv) We carried out an ablation study to show the effect of each loss term (only with CHC dataset due to time limitation). We first trained FUNS using only $\mathcal{L}^{\text{flow}}$, and then, other loss terms are added sequentially. The results are shown in Table B, in which L_2 errors (explained in Table A) and the number of non-zero elements in M ($\dim(z^{\text{sp}})$) are reported. The lower error means that the generated images are more related to the respective conditions. From Table B, if we train FUNS by using only $\mathcal{L}^{\text{flow}}$, $\dim(z^{\text{sp}}) = 4096$, which is in fact equal to the whole latent variable (therefore then $\dim(z^{\text{iv}}) = 0$), suggesting that all the latent variables are dependent on the condition c. In contrast, by adding proposed loss terms, $\dim(z^{\text{sp}})$ decreases to 550 (then $\dim(z^{\text{iv}}) = 3546$), while the prediction error is almost maintained. This result suggests that the input image x is successfully disentangled in the latent space into condition-dependent/independent parts. The impact of dropping $\mathcal{L}^{\text{entropy}}$ is also illustrated in Fig. 5 of our manuscript.

Reviewer #3 (i) We modified our manuscript to clarify the basic idea in the earlier part of the paper according to the reviewer's comment. (ii) As the reviewer pointed out, it is interesting to apply variational inference to train the encoder-decorder in our model, which will be described as a future work. (iii) As shown in Table B, 550 of latent variables were unmasked (therefore the remaining 3546 latents were finally masked) for CHC dataset. In more detail, let $z^{(l)}$ be the latent variables at level l ($z^{(1)} \in \mathbb{R}^{32\times32\times2}$, $z^{(2)} \in \mathbb{R}^{16\times16\times4}$, $z^{(3)} \in \mathbb{R}^{8\times8\times8}$, and $z^{(4)}, z^{(5)} \in \mathbb{R}^{4\times4\times16}$), the numbers of unmasked latents were 2, 2, 200, 181, and 165 for $l = 1, \ldots, 5$, respectively. It means that the latents in higher resolution layers are more likely to be masked. (iv) We generated CelebA samples using FUNS with the condition of Smile by varying only a part of $z^{(l)}$, while the remaining $z^{(l)}$ are fixed to $f_\mu(b_\mu^{(l)})$. Fig. B shows the results: (top) $z^{(1)}$, $z^{(2)}$, and $z^{(3)}$ are sampled. (middle) $z^{(4)}$ is sampled. (bottom) $z^{(5)}$ is sampled. We can see that larger diversity is captured by latent variables in lower resolution layers ($z^{(4)}$ and $z^{(5)}$). Latents in high resolution layers ($z^{(1)}$, $z^{(2)}$, and $z^{(3)}$) control very subtle facial expressions. (v) Please refer to response (ii) to Reviewer 2 for the experimental results that evaluate the proposed model in terms of the prediction accuracy by pretrained classifiers. (vi) We will correct grammatical issues through an English proofreading service.

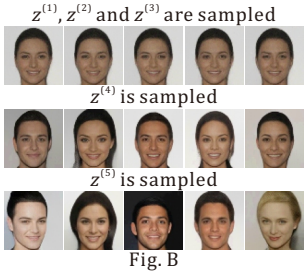

Fig. B