A. Derivation of approximation

In the main paper, we proposed iterative latent variable refinement (ILVR), where each transition of the generative process is matched with a given reference image. Condition in each transition was replaced with a local condition based on our approximation, as suggested in Eq.7 of the main text.

Before detailed derivations of the approximation (Eq.7), we review notations used in the main text. With pre-defined hyperparameter $\overline{\alpha}_t$, latent variable x_t can be sampled in closed-form: $x_t \sim q(x_t|x_0)$ (Eq.2). Trained model $\epsilon_\theta(x_t, t)$ predicts noise added in x_t, conditioned with time step t.

From the property of the forward process that latent variable x_t can be sampled from x_0 in closed-form, denoised data x_0 can be approximated with model prediction $\epsilon_\theta(x_t, t)$:

$$x_0 \approx f_\theta(x_t,\ t) = (x_t - \sqrt{1-\overline{\alpha}_t}\,\epsilon_\theta(x_t, t))/\sqrt{\overline{\alpha}_t} \quad \text{(A)}$$

Below is a derivation of Eq.7, where we approximated each conditioned Markov transition. We denote ϕ_N as ϕ and $f_\theta(x_t,\ t)$ as $f(x_t)$ for brevity. From Eq. A, each conditional Markov transition with given reference image y can be approximated as follows:

$$\begin{aligned}
& p_\theta(x_{t-1}|x_t, \phi(x_0) = \phi(y)) \\
\approx\ & p_\theta(x_{t-1}|x_t, \phi(f(x_{t-1})) = \phi(y)) \\
\approx\ & \mathbb{E}_{q(y_{t-1}|y)}[p_\theta(x_{t-1}|x_t, \phi(f(x_{t-1})) = \phi(f(y_{t-1})))].
\end{aligned}$$

With linear property of operation ϕ and Eq. A, we have

$$\begin{aligned}
& \mathbb{E}_{q(y_{t-1}|y)}[p_\theta(x_{t-1}|x_t, \phi(f(x_{t-1})) = \phi(f(y_{t-1})))] \\
=\ & \mathbb{E}_{q(y_{t-1}|y)}[p_\theta(x_{t-1}|x_t, \phi(x_{t-1}) = \phi(y_{t-1}), \\
& \phi(\epsilon_\theta(x_{t-1})) = \phi(\epsilon_\theta(y_{t-1})))] \\
\approx\ & \mathbb{E}_{q(y_{t-1}|y)}[p_\theta(x_{t-1}|x_t, \phi(x_{t-1}) = \phi(y_{t-1}))].
\end{aligned}$$

As shown in Eq.8 and Algorithm 1 of the main text, we first compute unconditional proposal x'_{t-1}, then refine it by ensuring $\phi(x_{t-1}) = \phi(y_{t-1})$. Therefore,

$$\begin{aligned}
& \mathbb{E}_{q(y_{t-1}|y)}[p_\theta(x_{t-1}|x_t, \phi(x_{t-1}) = \phi(y_{t-1}))] \\
=\ & \mathbb{E}_{q(y_{t-1}|y)}[p_\theta(\phi(y_{t-1}) + (I - \phi)(x'_{t-1}) \\
& |x_t, \phi(x_{t-1}) = \phi(y_{t-1}))] \\
=\ & \mathbb{E}_{q(y_{t-1}|y)}[p_\theta(x'_{t-1}|x_t)] \\
=\ & p_\theta(x'_{t-1}|x_t) \\
=\ & p_\theta(x_{t-1}|x_t, \phi(x_{t-1}) = \phi(y_{t-1})).
\end{aligned}$$

N	HR	Nearest	Bicubic	PULSE	ILVR
16 ↓	5.25	17.56	8.09	4.34	**4.06**
64 ↓	5.25	14.15	12.45	4.10	**4.02**

Table A: **NIQE comparison on generation quality.** Lower is better. Scores measured with generated images from reference images downscaled by a factor of 16 and 64. ILVR exhibits the highest perceptual quality.

CycleGAN [20]	MUNIT [5]	CUT [12]	Ours
85.9	104.4	**76.2**	**79.8**

Table B: **FID comparison on image translation.** FID measured with images translated from test set of AFHQ-dog. ILVR is comparable to a state-of-the-art model.

B. Additional evaluations

B.1. Generation quality

We provide additional qualitative and quantitative evaluations on the generation quality of ILVR. We evaluate images generated from low-resolution (LR) images downsampled by a factor of 16 and 64. Here, we compare ILVR with bicubic interpolation and PULSE [9], a super-resolution study that leverages pre-trained StyleGAN [7]. PULSE finds a latent vector that generates an image that downscales to the given LR image. We used publicly available StyleGAN2 [8] model[1] trained at 256×256. Combining loss function from PULSE and StyleGAN2, we search for latent vectors with a loss as follows:

$$\begin{aligned}
L_{total} =& ||\phi(G(z)) - \phi(y)||_2^2 \\
& + GEOCROSS(v_1, ..., v_{14}) + \alpha L_{noise}, \quad \text{(B)}
\end{aligned}$$

where each term refers to mean square error (MSE), geodesic cross loss [9], and noise regularization [8], respectively. MSE ensures generated image $G(z)$ and reference image y to match at low-resolution space. The geodesic cross loss ensures the latent vectors $v_1, ..., v_{14}$ remain in the learned latent space. Noise regularization L_{noise} discourages signal sneaking into the noise maps of StyleGAN2. We chose $\alpha = 5e^3$. Refer to StyleGAN2 literature for details on the noise regularization. We inherited latent vector initialization and learning rate schedule from StyleGAN2.

Fig. A presents additional qualitative results. ILVR and PULSE both show high-quality images generated from extremely downscaled images. Table. A shows NIQE [10] score, which is a no-reference metric that measures the perceptual quality of an image. ILVR shows higher perceptual quality, even better than the original 256^2 reference images (HR). We measured NIQE with reference images in Fig. B.

[1] https://github.com/rosinality/stylegan2-pytorch

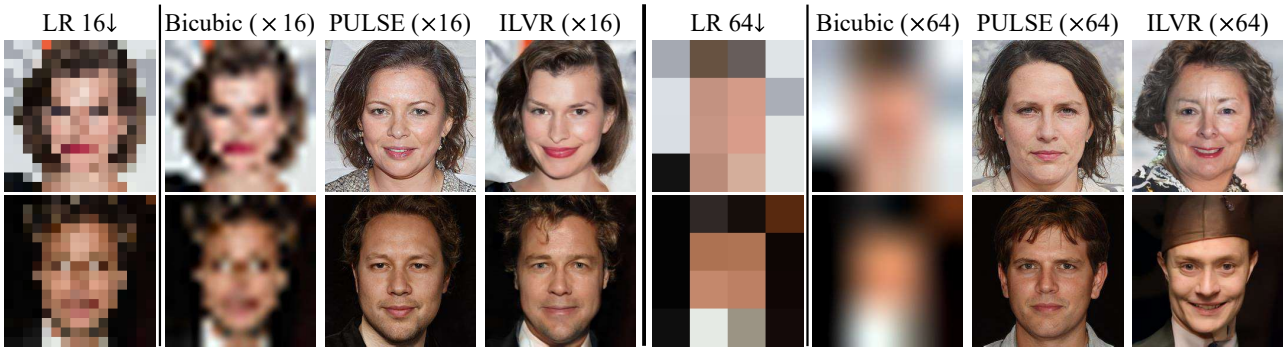

Figure A: **Qualitative comparison on generation quality.** Images generated from reference images downsampled by a factor of 16 and 64. From LR images, ILVR generates faces with detailed features.

Figure B: Images used for NIQE score.

B.2. Image translation

We compare Frechét inception distance (FID) [3] with image translation models on cat-to-dog (AFHQ [1] dataset) translation. Table. B shows the results. FID scores are calculated with the test set from AFHQ [1]. ILVR presents comparable performance to CUT [12], which is a state-of-the-art on cat-to-dog translation. Note that ILVR requires a model trained only on dog images, unlike the other models trained on both cat and dog images. We expect our result to broaden the applicability of DDPM to such image translation tasks.

B.3. Additional samples

Fig. C shows samples generated with publicly available guided-diffusion [2] trained on LSUN [17] datasets. We present additional editing with scribbles in Fig. D.

C. Implementation details

We trained unconditional DDPM with publicly available PyTorch implementation.[2]

C.1. Low-pass filters

We used bicubic downsampling and upsampling with correctly implemented function [14]. In Fig. E, we compare generated samples where the same noises were added through the generative process, only differing resizing kernels. Among kernels, images are almost identical, suggesting that our method is robust to kernel choice.

C.2. Datasets and training

Here we describe datasets and training details. For all datasets, we trained at 256^2 resolution with a batch size 8.

FFHQ [7] consists of 70,000 high-resolution face images. We trained a model for 1.2M steps.

METFACES [6] consists of 1,000 high-resolution portrait images. To avoid overfitting, we fine-tuned a model pre-trained on FFHQ [7], for 20k steps.

AFHQ [1] consists of 15,000 high-resolution animal face images, which are equally split into three categories: dog, cat, and wild. We trained on the train set of dog category, then used test sets of three categories as reference images to demonstrate multi-domain image translation.

Places365 [19] consists of 10M images of over 400 scene categories. We trained a model on a waterfall category, which consists of 5,000 images. We used this model to paint-to-image task.

LSUN Church [17] consists of 126,227 images of churches. We trained a model for 1M steps.

Paintings used for paint-to-image task are collected from the web.

C.3. Architecture

We trained the same neural network architecture as Ho *et al*. [4], which is U-Net [13] based on Wide ResNet [18]. Details include group normalization [16], self-attention blocks at 16×16 resolution, sinusoidal positional embedding [15], and a fixed linear variance schedule $\beta_1, ..., \beta_T$.

C.4. Evaluation

In Table 1 of the main text, we calculated FID scores on 50,000 real images and 50,000 generated images using code[3] of PyTorch framework.

[2] https://github.com/rosinality/denoising-diffusion-pytorch

[3] https://github.com/mseitzer/pytorch-fid

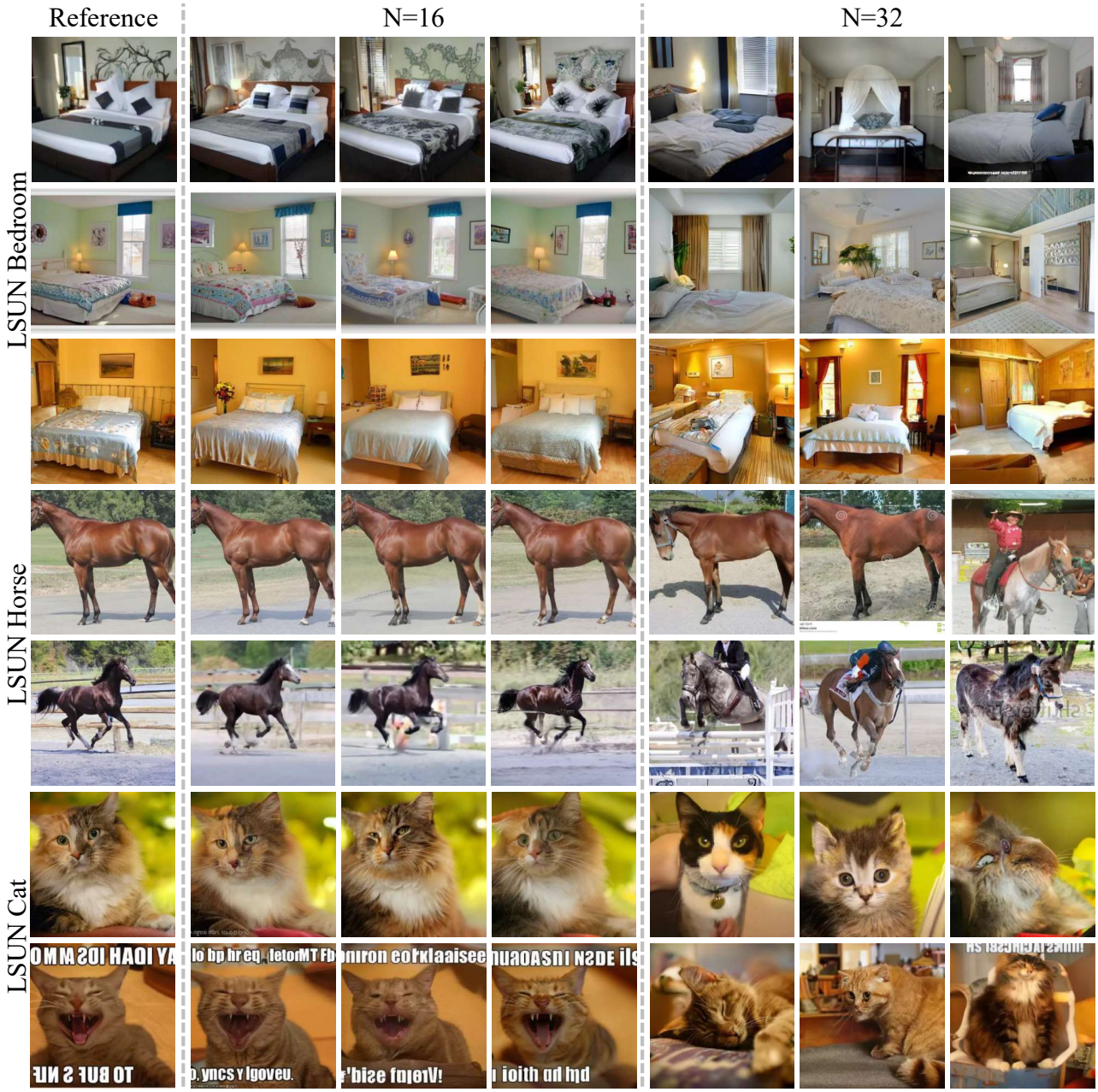

Figure C: **ILVR samples with guided-diffusion [2].** Publicly available guided-diffusion trained on LSUN Bedroom, Horse, and Cat datasets. For efficiency, samples are generated with 250 steps using uniform stride, following IDDPM [11]. Conditions are given in factor N=16,64 from time step 250 to 100. Samples share either coarse or fine semantics from the references.

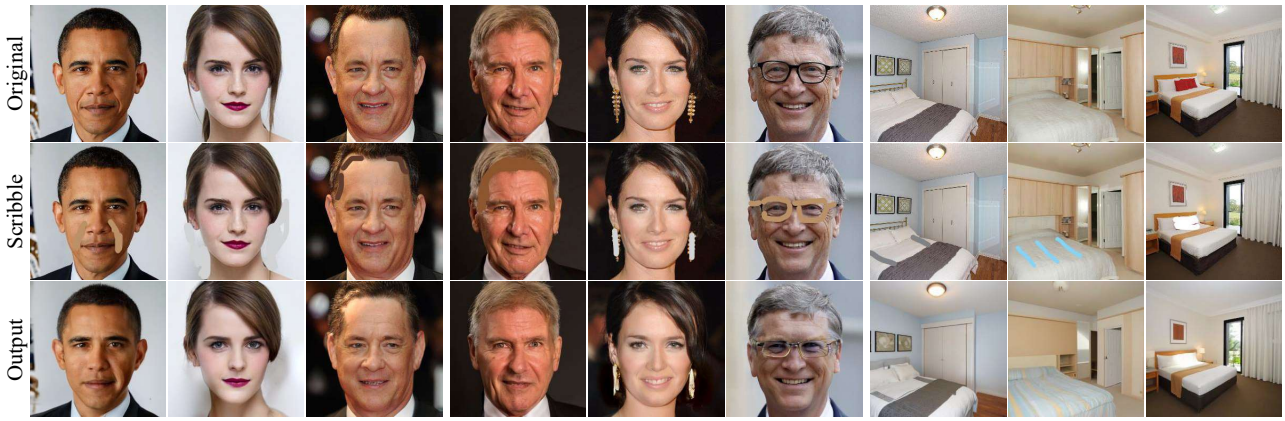

Figure D: **Additional editing with scribbles.** Faces generated with our reproduced model trained on FFHQ [7]. Bedrooms generated with publicly available model [2] trained on LSUN Bedroom [17].

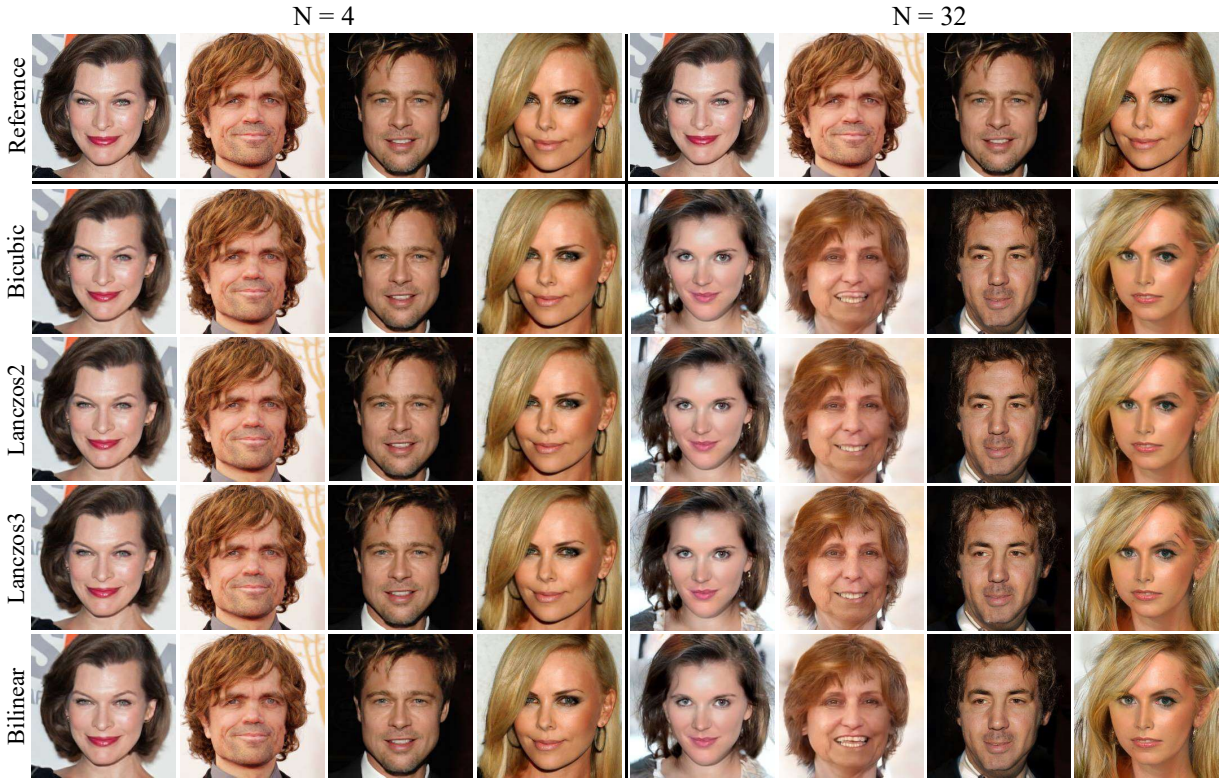

Figure E: **Ablation on low-pass filters.** First column set: samples from downsampling factor N=4; Second column set: samples from downsampling factor N=32. Samples are generated with bicubic, lanczos2, lanczos3, bilinear interpolation for downsampling and upsampling. There is only a minor difference among filters, such as the exact position of teeth and hair.

References

[1] Yunjey Choi, Youngjung Uh, Jaejun Yoo, and Jung-Woo Ha. Stargan v2: Diverse image synthesis for multiple domains. In *CVPR*, 2020. 2

[2] Prafulla Dhariwal and Alex Nichol. Diffusion models beat gans on image synthesis. *arXiv preprint arXiv:2105.05233*, 2021. 2, 3, 4

[3] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. In *NeurIPS*, 2017. 2

[4] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *arXiv preprint arXiv:2006.11239*, 2020. 2

[5] Xun Huang, Ming-Yu Liu, Serge Belongie, and Jan Kautz. Multimodal unsupervised image-to-image translation. In *Proceedings of the European conference on computer vision (ECCV)*, pages 172–189, 2018. 1

[6] Tero Karras, Miika Aittala, Janne Hellsten, Samuli Laine, Jaakko Lehtinen, and Timo Aila. Training generative adversarial networks with limited data. *NeurIPS*, 2020. 2

[7] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *CVPR*, 2019. 1, 2, 4

[8] Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Analyzing and improving the image quality of stylegan. In *CVPR*, 2020. 1

[9] Sachit Menon, Alexandru Damian, Shijia Hu, Nikhil Ravi, and Cynthia Rudin. Pulse: Self-supervised photo upsampling via latent space exploration of generative models. In *CVPR*, 2020. 1

[10] Anish Mittal, Rajiv Soundararajan, and Alan C Bovik. Making a "completely blind" image quality analyzer. *IEEE Signal processing letters*, 20(3):209–212, 2012. 1

[11] Alex Nichol and Prafulla Dhariwal. Improved denoising diffusion probabilistic models. *arXiv preprint arXiv:2102.09672*, 2021. 3

[12] Taesung Park, Alexei A Efros, Richard Zhang, and Jun-Yan Zhu. Contrastive learning for unpaired image-to-image translation. In *ECCV*, 2020. 1, 2

[13] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, 2015. 2

[14] Assaf Shocher. Resizeright. https://github.com/assafshocher/ResizeRight, 2018. 2

[15] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. *arXiv preprint arXiv:1706.03762*, 2017. 2

[16] Yuxin Wu and Kaiming He. Group normalization. In *Proceedings of the European conference on computer vision (ECCV)*, pages 3–19, 2018. 2

[17] Fisher Yu, Ari Seff, Yinda Zhang, Shuran Song, Thomas Funkhouser, and Jianxiong Xiao. Lsun: Construction of a large-scale image dataset using deep learning with humans in the loop. *arXiv preprint arXiv:1506.03365*, 2015. 2, 4

[18] Sergey Zagoruyko and Nikos Komodakis. Wide residual networks. *arXiv preprint arXiv:1605.07146*, 2016. 2

[19] Bolei Zhou, Agata Lapedriza, Aditya Khosla, Aude Oliva, and Antonio Torralba. Places: A 10 million image database for scene recognition. *IEEE transactions on pattern analysis and machine intelligence*, 2017. 2

[20] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *ICCV*, 2017. 1