Structure Matters: Tackling the Semantic Discrepancy in Diffusion Models for Image Inpainting
Supplementary Material

Haipeng Liu[1] Yang Wang[1*] Biao Qian[1] Meng Wang[1] Yong Rui[2]
[1]Hefei University of Technology, China [2]Lenovo Research, China
hpliu_hfut@hotmail.com, yangwang@hfut.edu.cn,
{hfutqian,eric.mengwang}@gmail.com, yongrui@lenovo.com

Due to page limitation of the main body, as indicated, the supplementary material offers more details on the ideal reverse state $\tilde{y}^*_{t-1}$, further discussion on the threshold Δ, additional quantitative results and more visual results with higher resolution, which are summarized below:

- More derivation details on the ideal reverse state $\tilde{y}^*_{t-1}$, as mentioned in Sec.2.2.2 of the main body (Sec.1).
- More *intuition* on the threshold Δ involved in the adaptive resampling strategy, as mentioned in Sec.2.4.2 of the main body (Sec.2).
- Visualization of the denoised results with *higher resolution* for IR-SDE [5] and StrDiffusion during the denoising process, as mentioned in Sec.3.2 of the main body (Sec.3).
- Additional quantitative results for the comparison with state-of-the-arts, as mentioned in Sec.3.3 of the main body (Sec.4).
- Additional visual results about the ablation study about the progressive sparsity for the structure over time, as mentioned in Sec.3.4.1 of the main body (Sec.5).

1. More details on the Ideal Reverse State $\tilde{y}^*_{t-1}$

Due to page limitation, we offer more derivation details from Eq.(7) to Eq.(11) in the main body. Based on the Eq.(7) of the main body, the optimal reverse state is naturally acquired by minimizing the negative log-likelihood:

$$\begin{aligned}\tilde{y}^*_{t-1} &= \arg\min_{y_{t-1}}[-\log q(y_{t-1}|y_t, y_0, x_{t-1}, x_0)] \\ &= \arg\min_{y_{t-1}}[-\log \frac{q(y_{t-1}|y_0)}{q(x_{t-1}|x_0)}],\end{aligned} \tag{1}$$

where $\tilde{y}^*_{t-1}$ denotes the ideal state reversed from $\tilde{y}_t$ under the structure guidance. To solve the above objective, we compute its gradient as:

$$\begin{aligned}&\nabla_{\tilde{y}^*_{t-1}}\left\{-\log q(\tilde{y}^*_{t-1}|y_t, y_0, x^*_{t-1}, x_0)\right\} \\ &= \nabla_{\tilde{y}^*_{t-1}}\left\{-\log \frac{q(\tilde{y}^*_{t-1}|y_0)}{q(x^*_{t-1}|x_0)}\right\} \\ &= -\nabla_{\tilde{y}^*_{t-1}}\log q(\tilde{y}^*_{t-1}|y_0) + \nabla_{x^*_{t-1}}\log q(x^*_{t-1}|x_0) \\ &= \frac{\tilde{y}^*_{t-1} - \mu_y - e^{-\overline{\theta}_{t-1}}(y_0-\mu_y)}{1-e^{-2\overline{\theta}_{t-1}}} \\ &\quad - \frac{x^*_{t-1} - \mu_x - e^{-\overline{\delta}_{t-1}}(x_0-\mu_x)}{1-e^{-2\overline{\delta}_{t-1}}},\end{aligned} \tag{2}$$

where the texture μ_y is the masked version of its initial state y_0 and θ_t is time-dependent parameter that characterizes the speed of the mean-reversion, the structure μ_x is the masked version of its initial state x_0 and δ_t is time-dependent parameter that characterizes the speed of the mean-reversion, $\overline{\theta}_{t-1} = \int_0^{t-1}\theta_z \mathrm{d}z$ and $\overline{\delta}_{t-1} = \int_0^{t-1}\delta_z \mathrm{d}z$. Setting Eq.(2) to be zero, we can get $\tilde{y}^*_{t-1}$ as:

$$\begin{aligned}\tilde{y}^*_{t-1} =& \frac{(1-e^{-2\overline{\theta}_{t-1}})(x^*_{t-1}-\mu_x)}{1-e^{-2\overline{\delta}_{t-1}}} \\ &- \frac{(1-e^{-2\overline{\theta}_{t-1}})e^{-\overline{\delta}_{t-1}}(x_0-\mu_x)}{1-e^{-2\overline{\delta}_{t-1}}} \\ &+ e^{-\overline{\theta}_{t-1}}(y_0-\mu_y) + \mu_y,\end{aligned} \tag{3}$$

where x^*_{t-1} is the ideal state reversed from x_t for the structure, given as:

$$\begin{aligned}x^*_{t-1} =& \frac{1-e^{-2\overline{\delta}_{t-1}}}{1-e^{-2\overline{\delta}_t}}e^{-\delta'_t}(x_t-\mu_x) \\ &+ \frac{1-e^{-2\delta'_t}}{1-e^{-2\overline{\delta}_t}}e^{-\overline{\delta}_{t-1}}(x_0-\mu_x) + \mu_x.\end{aligned} \tag{4}$$

*Yang Wang is the corresponding author.

To simplify the notation, $\delta'_t = \int_{t-1}^{t} \delta_i \mathrm{d}i$, we can derive the ideal reverse state $\tilde{y}^*_{t-1}$ as:

$$
\begin{aligned}
\tilde{y}^*_{t-1} =& \frac{1-e^{-2\overline{\theta}_{t-1}}}{1-e^{-2\overline{\delta}_t}} e^{-\delta'_t}(x_t-\mu_x) \\
&+ \frac{(1-e^{-2\overline{\theta}_{t-1}})(e^{-2\overline{\delta}_t}-e^{-2\delta'_t})}{(1-e^{-2\overline{\delta}_{t-1}})(1-e^{-2\overline{\delta}_t})} e^{-\overline{\delta}_{t-1}}(x_0-\mu_x) \\
&+ e^{-\overline{\theta}_{t-1}}(y_0-\mu_y)+\mu_y.
\end{aligned}
\tag{5}
$$

Since $\overline{\delta}_t = \overline{\delta}_{t-1} + \delta'_t$, we can reformulated the second term in Eq.(5) as follows:

$$
\begin{aligned}
&\frac{(1-e^{-2\overline{\theta}_{t-1}})(e^{-2\overline{\delta}_t}-e^{-2\delta'_t})}{(1-e^{-2\overline{\delta}_{t-1}})(1-e^{-2\overline{\delta}_t})} e^{-\overline{\delta}_{t-1}}(x_0-\mu_x) \\
=&\frac{(1-e^{-2\overline{\theta}_{t-1}})(e^{-2(\overline{\delta}_{t-1}+\delta'_t)}-e^{-2\delta'_t})}{(1-e^{-2\overline{\delta}_{t-1}})(1-e^{-2\overline{\delta}_t})} e^{\delta'_t-\overline{\delta}_t}(x_0-\mu_x) \\
=&\frac{(1-e^{-2\overline{\theta}_{t-1}})(e^{-2\overline{\delta}_{t-1}}-1)}{(1-e^{-2\overline{\delta}_{t-1}})(1-e^{-2\overline{\delta}_t})} e^{-2\delta'_t} e^{\delta'_t-\overline{\delta}_t}(x_0-\mu_x) \\
=&-\frac{(1-e^{-2\overline{\theta}_{t-1}})}{(1-e^{-2\overline{\delta}_t})} e^{-\delta'_t} e^{-\overline{\delta}_t}(x_0-\mu_x) \\
=&-\left(\frac{(1-e^{-2\overline{\theta}_{t-1}})}{(1-e^{-2\overline{\delta}_t})} e^{-\delta'_t}\right) e^{-\overline{\delta}_t}(x_0-\mu_x).
\end{aligned}
\tag{6}
$$

Based on the above, we have

$$
\begin{aligned}
\tilde{y}^*_{t-1} =& \underbrace{\left(\frac{1-e^{-2\overline{\theta}_{t-1}}}{1-e^{-2\overline{\delta}_t}} e^{-\delta'_t}\right)(x_t-\mu_x)}_{\text{Consistency term for masked regions}} \\
&- \underbrace{\left(\frac{1-e^{-2\overline{\theta}_{t-1}}}{1-e^{-2\overline{\delta}_t}} e^{-\delta'_t}\right) e^{-\overline{\delta}_t}(x_0-\mu_x)}_{\text{Balance term for masked regions}} \\
&+ \underbrace{e^{-\overline{\theta}_{t-1}}(y_0-\mu_y)}_{\text{Semantics term for masked regions}} + \underbrace{\mu_y}_{\text{Unmasked regions}}.
\end{aligned}
\tag{7}
$$

2. More Intuition on the Threshold Δ in the Adaptive Resampling Strategy

The specific threshold Δ in the adaptive resampling strategy is utilized to evaluate the semantic correlation between the structure and texture during the inference denoising process. Specifically, when the score value S from the discriminator D is smaller than the threshold Δ, *i.e.*, $S < \Delta$, we will perform the resampling operation for the structure to enhance the semantic correlation for desirable results. A naive strategy is to mutually select a fixed value of Δ,

Algorithm 1: Adaptive Resampling Strategy

Input: the noise version of masked texture y_T and the noise version of masked structure x_T, trained noise-prediction networks $\tilde{\epsilon}_\phi$ and $\tilde{\epsilon}_\varphi$ for the texture and structure, the timestep T, the discriminator D, the maximum number of iterations U

Output: The denoised inpainted result y_0

for *t* = *T*, . . . , *1* **do**
  Denoised structure $x_{t-1} = x_t - (\mathrm{d}x_t)_{\tilde{\epsilon}_\varphi(x_t,t)}$
  Denoised texture $y_{t-1} = y_t - (\mathrm{d}y_t)_{\tilde{\epsilon}_\phi(y_t,x_{t-1},t)}$
  Obtain the threshold $\Delta = D(y_{t-1}, x_{t-1}, t-1)$
  for *u* = *1*, . . . , *U* **do**
    $\tilde{x}_t = x_{t-1} + (\mathrm{d}x_{t-1})$
    $\tilde{x}_{t-1} = \tilde{x}_t - (\mathrm{d}\tilde{x}_t)_{\tilde{\epsilon}_\varphi(\tilde{x}_t,t)}$
    $\tilde{y}_{t-1} = y_t - (\mathrm{d}y_t)_{\tilde{\epsilon}_\phi(\tilde{y}_t,\tilde{x}_{t-1},t)}$
    Obtain the score $S = D(\tilde{y}_{t-1}, \tilde{x}_{t-1}, t-1)$
    if $S < \Delta$ **then**
      Update $x_{t-1} = \tilde{x}_{t-1}$
      Update $y_{t-1} = \tilde{y}_{t-1}$
    else
      break
    end
  end
end
return the denoised results y_0

which, however, is inflexible, since the semantic correlation actually *varies* greatly as the inference denoising process progresses. Unlike the previous work [4] that aims to condition the denoising process for image inpainting task via the resampling strategy, our adaptive resampling strategy actually serves as a by-product; the goal is to refine the correlation between the structure and texture as possible. To this end, we present to exploit the structure x_{t-1} and the texture y_{t-1} without the adaptive resampling strategy in the t-th timestep, to serve as the inputs for the discriminator D, leading to a score value, which is treated as the threshold Δ. Under such case, the semantic correlation between the structure and texture can be always boosted to yield better denoised results. Based on the threshold Δ, the whole adaptive resampling strategy is summarized in Algorithm 1.

3. Visualization of the Denoised Results with Higher Resolution

As mentioned in Sec.3.2 of the main body, due to page limitation, we further provide more denoised results with *higher resolution* on the Places2, PSV and CelebA datasets; see Fig.1. The results show that, unlike the denoised results from IR-SDE [5] that always address the *clear* semantic dis-

Table 1. Comparison of quantitative results (*i.e.*, PSNR, SSIM, and FID) under varied mask ratios on CelebA with irregular mask dataset. ↑: Higher is better; ↓: Lower is better. The best results are reported with **boldface**.

Metrics		PSNR↑			SSIM↑			FID↓		
Method	**Venue**	**0-20%**	**20-40%**	**40-60%**	**0-20%**	**20-40%**	**40-60%**	**0-20%**	**20-40%**	**40-60%**
PIC [9]	CVPR' 19	33.67	26.48	21.58	0.978	0.934	0.865	2.340	6.430	14.22
MAT [3]	CVPR' 22	35.31	27.76	23.22	0.984	0.946	0.888	0.900	2.550	**4.600**
CMT [2]	ICCV' 23	35.92	28.24	23.78	0.986	0.952	0.900	0.840	2.540	5.230
ICT [6]	CVPR' 21	33.27	26.40	21.84	0.979	0.939	0.877	1.870	5.610	12.42
BAT [8]	MM' 21	34.63	26.91	22.26	0.983	0.944	0.883	1.060	3.750	7.300
*RePaint** [4]	CVPR' 22	36.23	29.01	23.92	0.991	0.969	0.912	0.790	2.530	5.030
IR-SDE [5]	ICML' 23	36.01	28.85	23.76	0.991	0.966	0.910	0.870	2.840	5.700
StrDiffusion (Ours)	-	**36.44**	**29.31**	**24.50**	**0.994**	**0.971**	**0.923**	**0.660**	**2.400**	4.950

crepancy between the masked and unmasked regions (see Fig.1(a)), for StrDiffusion, such discrepancy progressively degraded until *vanished*, yielding the consistent semantics (see Fig.1(b)), which are consistent with our analysis in the main body.

4. Additional Quantitative Results

Evaluation metric. We adopt three metrics to evaluate the inpainted results below: 1) peak signal-to-noise ratio (PSNR); 2) structural similarity index (SSIM) [7]; and 3) Fréchet Inception Score (FID) [1]. PSNR and SSIM are used to compare the low-level differences over pixel level between the generated image and ground truth. FID evaluates the perceptual quality by measuring the feature distribution distance between the synthesized and real images.

As indiacted in the main body, we further exhibit additional quantitative results under varied mask ratios on CelebA with irregular mask dataset; see Table.1. It is observed that our StrDiffusion enjoys a much smaller FID score, together with larger PSNR and SSIM than the competitors, confirming that StrDiffusion effectively addresses semantic discrepancy between the masked and unmasked regions, while yielding the reasonable semantics. Notably, RePaint* and IR-SDE still remain the large performance margins (at most 1.0% for PSNR, 0.2% for SSIM and 5.2% for FID) compared to StrDiffusion, owing to the semantic discrepancy in the denoised results incurred by the dense texture. Albeit ICT and BAT focus on the guidance of the structure similar to StrDiffusion, they suffer from a performance loss due to the semantic discrepancy between the structure and texture, which *confirms* our proposal in Sec.2.2 of the main body — *the progressively sparse structure provides the time-dependent guidance for texture denoising process*.

5. Additional Visual Results for the Ablation Study in Sec.3.4.1 of the Main Body

The ablation study in Sec.3.4.1 of the main body aims to verify why the semantic sparsity of the structure should be strengthened over time. In this section, we further exhibit additional visual results by performing the experiments on the PSV and Places2 datasets; see Fig.2. It is observed that our **gray2edge** (Fig.2(d)) exhibits better consistency with meaningful semantics in the inpainted results against others, especially for **edge2gray** (Fig.2(c)), implying the *benefits* of strengthening the sparsity of the structure over time. Notably, for **gray2gray**, the discrepancy issue in the denoised results still suffers (Fig.2(b)), while **edge2edge** receives the poor semantics (Fig.2(a)), which attributes to their *invariant* semantic sparsity over time. Such fact *confirms* our proposals in Sec.3.4.1 of the main body.

References

[1] 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. *Advances in neural information processing systems*, 30, 2017.

[2] Keunsoo Ko and Chang-Su Kim. Continuously masked transformer for image inpainting. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 13169–13178, 2023.

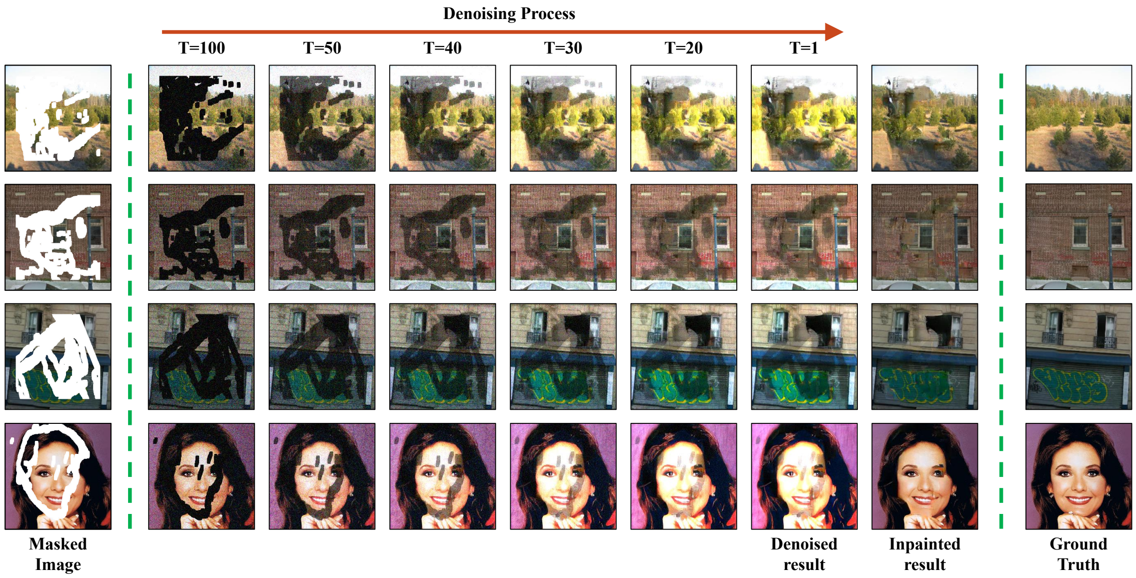
[3] Wenbo Li, Zhe Lin, Kun Zhou, Lu Qi, Yi Wang, and Jiaya Jia. Mat: Mask-aware transformer for large hole image inpainting. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10758–10768, 2022.

[4] Andreas Lugmayr, Martin Danelljan, Andres Romero, Fisher Yu, Radu Timofte, and Luc Van Gool. Repaint: Inpainting using denoising diffusion probabilistic models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11461–11471, 2022.

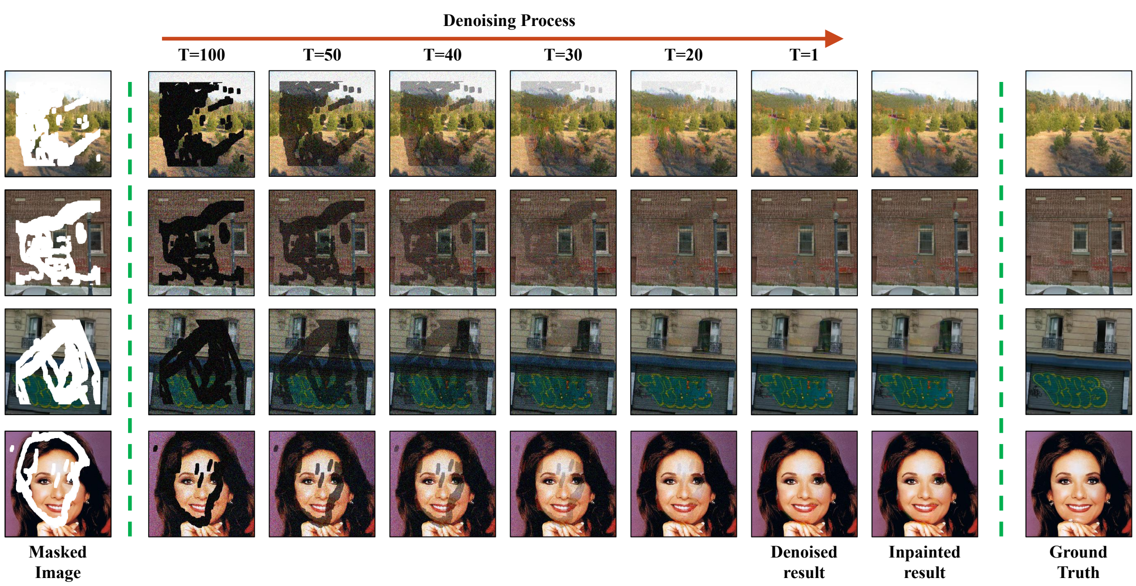
[5] Ziwei Luo, Fredrik K Gustafsson, Zheng Zhao, Jens Sjölund, and Thomas B Schön. Image restoration with mean-reverting stochastic differential equations. *arXiv preprint arXiv:2301.11699*, 2023.

[6] Ziyu Wan, Jingbo Zhang, Dongdong Chen, and Jing Liao. High-fidelity pluralistic image completion with transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4692–4701, 2021.

[7] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to



(a) IR-SDE



(b) StrDiffusion (Ours)

Figure 1. Visualization of the denoised results for IR-SDE (a) and StrDiffusion (b) in the varied timesteps during the denoising process, as an *extension* of Fig.6 in the main body.

structural similarity. *IEEE transactions on image processing*, 13(4):600–612, 2004.

[8] Yingchen Yu, Fangneng Zhan, Rongliang Wu, Jianxiong Pan, Kaiwen Cui, Shijian Lu, Feiying Ma, Xuansong Xie, and Chunyan Miao. Diverse image inpainting with bidirectional and autoregressive transformers. In *Proceedings of the 29th ACM International Conference on Multimedia*, pages 69–78, 2021.

[9] Chuanxia Zheng, Tat-Jen Cham, and Jianfei Cai. Pluralistic image completion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1438–1447, 2019.

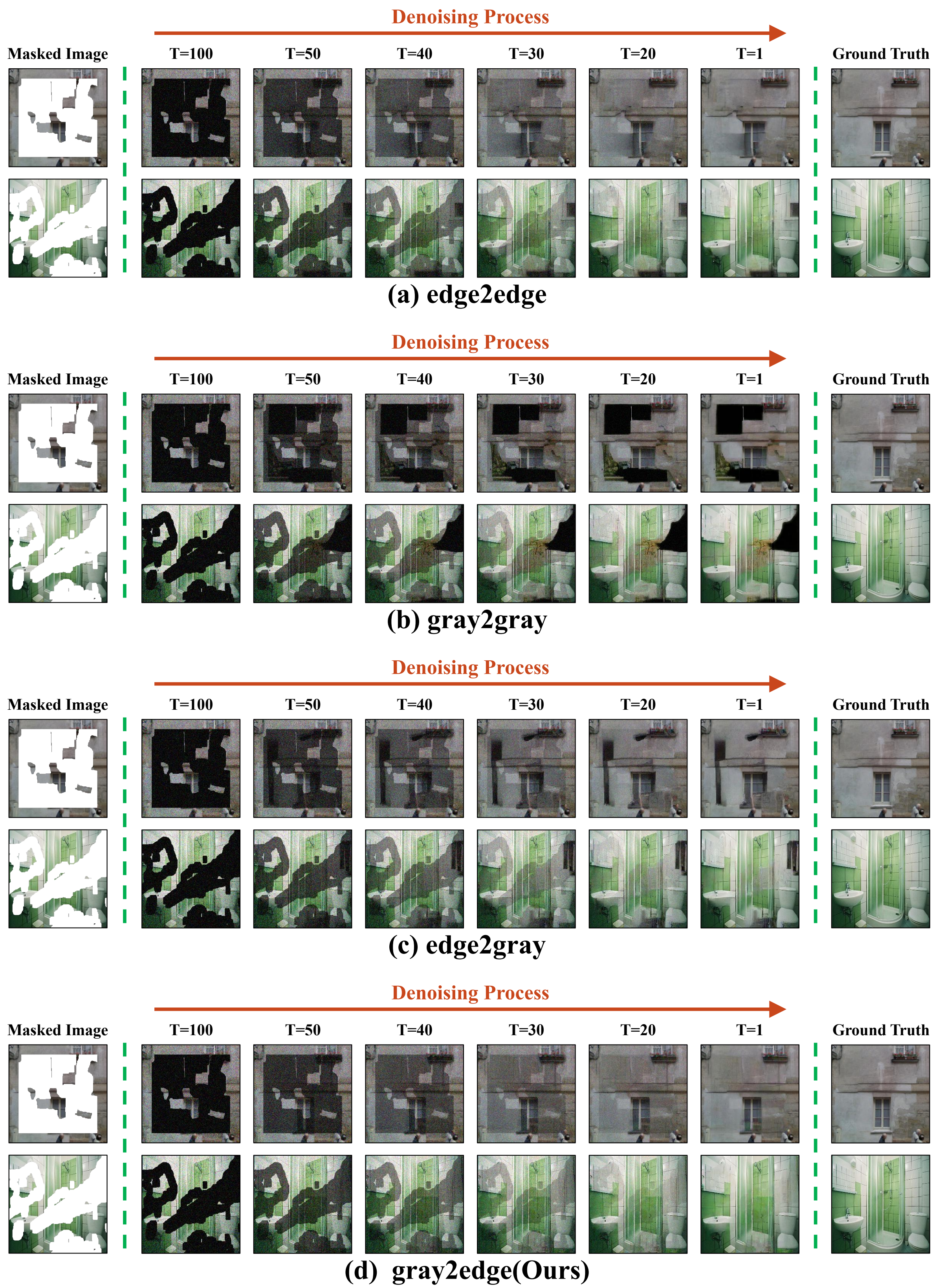


Figure 2. Additional visual results for the ablation study about the progressive sparsity for the structure over time, as an *extension* of Fig.8 in the main body.