LatentPaint: Image Inpainting in Latent Space with Diffusion Models

Ciprian Corneanu*
Amazon

Raghudeep Gadde*
Amazon

Aleix M Martinez
Amazon

Figure 1. We propose *LatentPaint*, an information propagation mechanism that can condition any existing diffusion model for inpainting. Shown here are a few examples of faces and indoor scenes with semantic inputs.

Abstract

Image inpainting using diffusion models is generally done using either preconditioned models, i.e. image conditioned models fine-tuned for the painting task, or postconditioned models, i.e. unconditioned models repurposed for the painting task at inference time. Preconditioned models are fast at inference time but extremely costly to train. Postconditioned models do not require any training but are slow during inference, requiring multiple forward and backward passes to converge to a desirable solution. Here, we derive an approach that does not require expensive training, yet is fast at inference time. To solve the costly inference computational time, we perform the forward-backward fusion step on a latent space rather than the image space. This is solved with a newly proposed propagation module in the diffusion process. Experiments on a number of domains demonstrate our approach attains or improves state-of-the-art results with the advantages of preconditioned and postconditioned models and none of their disadvantages.

1. Introduction

Image inpainting infers missing parts in an image based on available regions specified by a binary mask. To achieve this goal, inpainting approaches use generative models modified to condition on the available image regions in order to produce high-quality inferences. In this paper, we address the problem of image inpainting using Diffusion Models (DM).

There are two approaches to image inpainting with diffusion models: (a) *preconditioned* and (b) *postconditioned*.

Preconditioning is when we build inpainting into the model during training. That is, rather than training a generative model to learn the domain's distribution $p(x)$, a con-

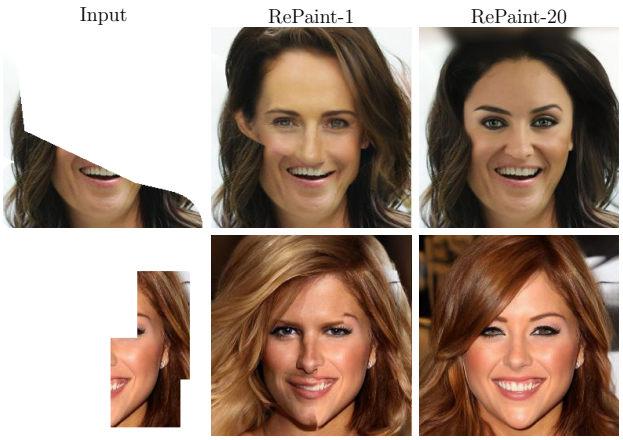

Figure 2. **Illustrating the importance of information propagation for inpainting.** RePaint [20] with no resampling (second column) and with 20 resampling steps (third column). To obtain harmonized results, resampling increases the runtime by 20x.

ditional model $p(x|y)$ is trained instead. A popular technique to achieve this, is by concatenating the masked image (the condition) to random noise and training the model to produce samples similar to those in the training distribution [27]. During inference, when the trained preconditioned model is applied on a masked image, it produces an output that infers the missing (masked out) parts. This approach works well but it requires the model to be specifically trained on every domain we wish it apply to, which is a very expensive task.

Postconditioning does not require the above specific training. Instead, the idea is to use a unconditional DM as a generative prior to harmonize information between masked and unmasked pixels. This is achieved by performing forward diffusion of unmasked pixels and reverse diffusion of masked pixels [20]. A fusion image is constructed by choosing pixel values, based on an input mask, from the appropriate forward-backward diffusion process. Because fusing in the pixel space produces semantically inconsistent results, the diffusion has to be repeated multiple times. The problem is illustrated in Fig. 2. For RePaint-1 [20] no repetition is done. In RePaint-20, 20 additional passes allow the reverse diffusion to correct the inconsistencies and converge to a desirable result. Unfortunately, computational cost depends linearly on the number of passes.

For the image inpainting task, propagating information from the conditioned (or unmasked) pixels to the inferred (or masked) pixels is important to produce coherent, harmonious and semantically consistent images. Both preconditioned and postconditioned models propagate information implicitly, either through the convolution [11, 16, 39], convolution and transformer layers in the diffusion model [27] or through a complex scheduling of forward/reverse diffusion steps [20]. Particularly of interest are the visual self-attention mechanisms in the transformer blocks. However, self-attention propagates information at a coarse level. The propagation technique presented in this paper, propagates information at finer, i.e. pixel level. Further, improvements in the empirical results indicate that latent paint complements the information propagated through other layers of the diffusion model. In this work, we propose a new approach that is as cheap as preconditioned models at inference and as cheap as postconditioned models in training. Specifically, *we derive LatentPaint, a conditioning mechanism that works in latent spaces rather than image space*. In this way, we are able to condition a pretrained unconditional DM with *minimal training* and sample from it with *no extra computational cost during inference*.

We demonstrate the accuracy and advantages of the proposed method on three visual domains: faces, bedrooms, and livingrooms. We also perform several ablative experiments on CelebA-HQ to validate the advantages of individual components. Our empirical results validate that the proposed approach betters several recent state-of-art techniques for inpainting along with a fast runtime.

2. Related Work

Historically, approaches to inpainting use patch similarity to propagate information from the conditioned to the inferred regions of the image [2–4].

Since the introduction of GANs [6], most of the existing methods follow a standard configuration using an encoder-decoder architecture as the main inpainting generator, adversarial training, and tailored losses that aim at photo-realism [24]. Multiple works have produced impressive results using this approach [10, 17, 22, 26, 40].

In order to generalize to both local and global context various architectural designs are proposed such as dilated convolutions to expand the receptive field [11], dedicated discriminators to encourage global and local consistency independently, partial convolutions [16] and gated convolutions [39] to guide the convolution kernel according to irregular masks, contextual attention [38] to leverage on global information, edges maps [7, 21, 32, 33] or semantic segmentation maps [9, 23] to further guide the generation and Fourier convolutions [31] to include both global and local information efficiently. Among the losses proposed, pixel-wise and adversarial losses dominate [11, 17, 18, 21, 24, 24, 26, 34–36, 38, 38–40, 42].

Some of the first notable diffusion models were also applied to image inpainting [29,30]. However, only qualitative results were shown and no specific inpainting approaches were provided. More recently, [1, 20] repurposed unconditional diffusion models for this task.

Inpainting using diffusion models can also be done in image-to-image translation frameworks by training an image-conditional diffusion model [27, 28]. Unlike both

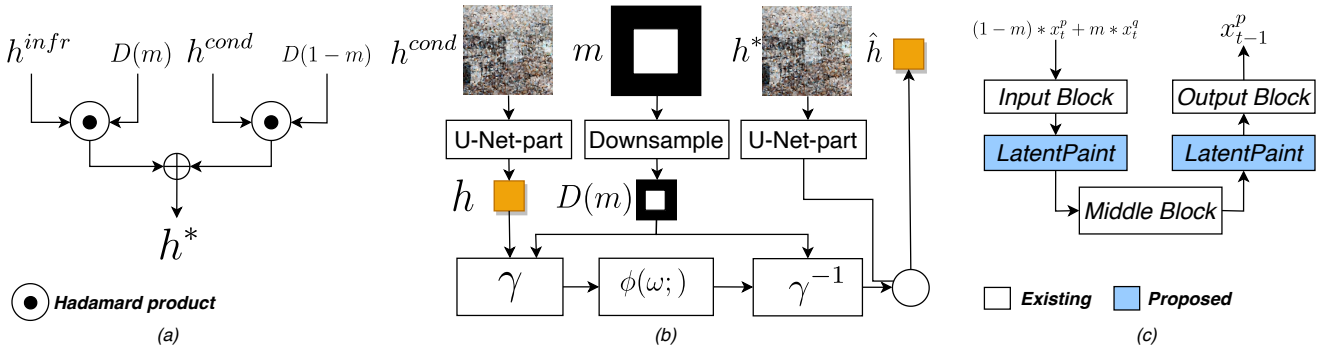

Figure 3. **LatentPaint.** LatentPaint is an easy add-on to a diffusion model. It consists of two parts: the Latent Space Conditioning (a) and the Explicit Propagation (b). With this addition, a pretrained unconditional diffusion model gets conditioned for inpainting. LatentPaint can be plugged into any U-Net like diffusion model. For example in (c) we illustrate how Latent Paint is applied twice between the existing Input, Middle and Output blocks of a diffusion U-Net. The proposed approach is detailed in Alg. 1.

these concurrent works, we leverage an unconditional diffusion model and only condition through the reverse diffusion process itself. It allows our approach to generalize to any mask shape for free-form inpainting.

3. Method

This section is divided into two main parts. We start with a summary of the diffusion steps needed to derive our approach, followed by a detailed description of our solution.

3.1. Denoising diffusion models

DM are generative models that given observations from a distribution x learn its true distribution $p(x)$. They are a generalization of variational autoencoders with a hierarchical set of latents that form a chain following the Markov property. The latents' dimension equal data's. The latent encoder is a predefined linear Gaussian model specially chosen such that the distribution of the latent at the final step is a standard Gaussian.

Intuitively, a DM performs a gradual noisification of an image x_0 over a predefined set of steps T. The input image is progressively corrupted by adding Gaussian noise at each step until eventually it becomes completely identical to pure Gaussian noise. This process is also called the *forward process*. The goal is to produce a model capable of learning the conditionals of the so called *reverse process* that takes pure Gaussian noise and produces a valid image over the same number of steps in the opposite direction.

Following the Markov property the posterior of the generative process, i.e. the forward process, can be written as:

$$q(x_{1:T}|x_0) = \prod_{t=1}^{T} q(x_t|x_{t-1}). \tag{1}$$

By definition each latent variable is a standard normal distribution centered around its previous hierarchical latent with mean $\mu_t(x_t) = \sqrt{\alpha_t}x_{t-1}$, and variance $\Sigma_t(x_t) = (1-\alpha_t)\mathbf{I}$,

$$q(x_t|x_{t-1}) = \mathcal{N}(x_t; \sqrt{\alpha_t}x_{t-1}, (1-\alpha_t)\mathbf{I}). \tag{2}$$

The joint distribution of the generative process, a.k.a. *the reverse process* is defined by:

$$p(x_{0:T}) = p(x_T) \prod_{t=1}^{T} p_\theta(x_{t-1}|x_t), \tag{3}$$

where

$$p(x_T) = \mathcal{N}(x_T; 0, \mathbf{I}). \tag{4}$$

Notice that the forward process $q(x_{1:T}|x_0)$ is not parameterized by θ, as it is completely modeled as Gaussians with defined mean and variance parameters at each timestep. Therefore, in a DM, we are only interested in learning conditionals $p_\theta(x_{t-1}|x_t)$, so that we can simulate new data. After optimizing the DM, the sampling procedure is as simple as sampling Gaussian noise from $p(x_T)$ and iteratively running the denoising transitions $p_\theta(x_{t-1}|x_t)$ for T steps to generate a novel x_0.

Like any variational autoencoder, the DM can be optimized by maximizing the ELBO [14], which can be derived as:

$$\log p(x) \geq \mathbb{E}_{q(x_{1:T}|x_0)} \left[\log \frac{p(x_{0:T})}{q(x_{1:T}|x_0)} \right] \tag{5}$$

$$= \underbrace{\mathbb{E}_{q(x_1|x_0)} \left[\log p_\theta(x_0|x_1) \right]}_{\text{reconstruction term}} \tag{6}$$

$$- \underbrace{D_{\mathrm{KL}}(q(x_T|x_0)||p(x_T))}_{\text{prior matching term}} \tag{7}$$

$$- \sum_{t=2}^{T} \underbrace{\mathbb{E}_{q(x_t|x_0)} \left[D_{\mathrm{KL}}(q(x_{t-1}|x_t, x_0)||p_\theta(x_{t-1}|x_t)) \right]}_{\text{denoising matching term}} \tag{8}$$

Algorithm 1 LatentPaint: Image inpainting with postconditioning in the latent space.

Input: diffusion model $(\mu_\theta(.), \Sigma_\theta(.))$, input image x_0, number of diffusion steps T , mask m and set of latents $\mathcal{H}$.
$x_T \leftarrow$ sample from $\mathcal{N}(0, \mathbf{I})$
for t from T to 1 **do**
  ▷ ***Step 1. Infer denoised image (p-sample)***
  $\mu^*, \Sigma^* \leftarrow \mu_\theta(x_t), \Sigma_\theta(x_t)$
  $x_{t-1}^{infr} \leftarrow$ sample from $\mathcal{N}(\mu^*, \Sigma^*)$
  ▷ ***Step 2. Compute noisy condition (q-sample)***
  $\mu, \Sigma \leftarrow \mu_q(x_0), \Sigma_q(x_0)$
  $x_{t-1}^{cond} \leftarrow$ sample from $\mathcal{N}(\mu, \Sigma)$
  ▷ ***Step 3. Post-conditioning in latent space***
  for h in $\mathcal{H}$ **do**
    ▷ *Latent-space conditioning*
    $h^* = h^{infr} \odot (1 - D(m)) + h^{cond} \odot D(m)$
    ▷ *Explicit propagation*
    $\hat{h} = \gamma^{-1}[\phi[\omega; \gamma(D(m), h^{cond})]]$
  end for
  ▷ ***Step 4. Compose condition with inferrence***
  $x_{t-1} = x_{t-1}^{infr} \odot (1 - m) + x_{t-1}^{cond} \odot m$
end for
return x_0

In our case, because we can set the variances of the two Gaussians to match exactly, optimizing the KL Divergence of the denoising matching term reduces to minimizing the difference between the means of the two distributions:

$$\begin{aligned} &\arg\min_\theta D_{\mathrm{KL}}(q(x_{t-1}|x_t, x_0)||p_\theta(x_{t-1}|x_t)) \\ &= \arg\min_\theta \frac{1}{2\sigma_q^2(t)} \left[\|\mu_\theta - \mu_q\|_2^2\right], \end{aligned} \tag{9}$$

where we have written μ_q as shorthand for $\mu_q(x_t, x_0)$, and μ_θ as shorthand for $\mu_\theta(x_t, t)$ for brevity.

Therefore, optimizing a DM boils down to learning a neural network to predict the original ground-truth image from an arbitrarily noisified version of it. Furthermore, minimizing the summation term of the derived ELBO objective across all noise levels can be approximated by minimizing the expectation over all timesteps:

$$\arg\min_\theta \mathbb{E}_{t\sim U\{2,T\}} \left[\mathbb{E}_{q(x_t|x_0)} \left[D_{\mathrm{KL}}(q(x_{t-1}|x_t, x_0)||p_\theta(x_{t-1}|x_t))\right]\right], \tag{10}$$

which can then be optimized using stochastic samples over timesteps.

3.2. LatentPaint

Let $\mathcal{D} = (x, m)$ be a dataset of observations where $x \in \mathbb{R}^{W\times H\times C}$ are color images with width W, height H and number of channels C and $m \in \mathbb{Z}_2^{W\times H\times 1}$ are binary matrices. Each sample (x, m) drawn from $\mathcal{D}$ represents an image with two distinct regions, a known region we would like to condition on, $x^{cond} = x \odot m$, and the complementary region we want to infer, $x^{infr} = x \odot (1-m)$. The goal of inpainting is to learn a function $p(x^{infr}|x^{cond})$ capable of producing realistic samples $x^* = x^{cond} + x^{infr}$ relative to the initial distribution $p(x)$.

We propose architectural additions to denoising diffusion models for achieving high-quality inpainting without preconditioning.

Specifically, we introduce the *Explicit Propagation* (EP) module that propagates information from the conditioned pixels to the inferred pixels in a latent space. This can be directly plugged into any existing DM and multiple instances at multiple locations in the DM can be utilized.

Latent Space Conditioning. Motivated by [1], we merge latent representations of the condition and the inferred signal at all levels of the DM. The latent representations of the condition (h^{cond}) are obtained by passing $q(x_T|x_0^{cond})$ through the denoising network. The computed conditional latent representations are combined with the corresponding latent representations of inferred signal using the input mask:

$$h^* = h^{infr} \odot (1 - D(m)) + h^{cond} \odot D(m) \tag{11}$$

$D(m)$, the mask at various resolutions is obtained by following the usual downsampling operations.[1] Note that due to average pooling, the downsampled mask is not binary anymore. We repeat the above form of merging representations at all possible locations.[2] See Algo 1 for step-by-step details.

Explicit propagation. The Explicit Propagation's (EP) main function is to properly propagate information between the condition and the inferred regions of the latent during inference. It consists of a set of transformations of the latent representation. More specifically, the latent $h^* \in \mathbb{R}^{\bar{W}\times\bar{H}\times\bar{C}}$ is transformed in the following way:

$$\hat{h} = \gamma^{-1}[\phi[\omega; \gamma(D(m), h^{cond})], \tag{12}$$

where $\bar{W}$ and $\bar{H}$ determine the size of the latent, with $\bar{W} < W$ and $\bar{H} < H$, and W, H are the width and height of the input image, and $\bar{C}$ the number of channels.

This creates a new latent $\hat{h}$ of the same dimension, $\hat{h} \in \mathbb{R}^{\bar{W}\times\bar{H}\times\bar{C}}$ which is passed to the downstream computations in the DM.

[1] For example average pooling like in guided diffusion [5].

[2] In the guided diffusion models [5], we do this after every block in inputblocks, middleblock and outputblocks.

In (12) the terms have the following meaning: D is a simple downsampling that brings the input mask $m \in \mathbb{R}^{W \times H \times C}$ to the size of the latent such that $D(m) \in \mathbb{R}^{\bar{W} \times \bar{H} \times 1}$ [3]. γ is a mask-wise max-pooling which takes the downsampled binary mask $D(m)$ and max-pools from the condition h^{cond} values for the inferred and condition region. ϕ is a function with parameters ω that learns a proper non-linear combination between the representations of the two regions where $\phi \rightarrow \mathbb{R}^{\bar{C} \times 1 \times 2}$[4]. Following this, a mask-wise unpooling γ^{-1} brings the representation back to its original size. The total number of trainable parameters introduced by this module is tiny in comparison to the parameters of the DM (less than 1%).

For a depiction of LatentPaint refer to Fig. 3. For a step-by-step implementation to Alg. 1.

4. Experiments

This section provides extensive experimental results and comparison with state-of-the-art methods.

4.1. Experimental Setup

Datasets. We evaluate our proposal in three visual domains: faces, bedrooms, and living rooms. *CelebA-HQ* [13] is a high-quality dataset of human faces. Following prior works [31, 44], we evaluate our technique on 2000 images with three types of masks: thin, medium, and thick. This is in contrast to RePaint [20] which computes scores on a small set of 100 samples. For bedrooms, we use 500 images from the *LSUN-Bedrooms* validation dataset [37]. Both Faces and Bedrooms contain images at 256x256-pixel resolution. To evaluate inpainting at high resolutions, we curated a dataset of living rooms at 512x512-pixel resolution and used it to train a Latent Diffusion Model (LDM). The masks for living rooms and bedrooms are obtained through semantic segmentation [19] of photos that are not part of the training set. Importantly, faces are a single object distribution whereas bedrooms and living rooms are scenes containing multiple objects. Inpainting requires techniques to fill in missing parts for faces and add multiple objects in coherent fashion for scenes. The goal of painting techniques is to produce coherent and consistent regions of pixels. Other mask types such as painting every alternate pixel do not help in evaluating image consistency. As a result, our experiments are mainly focused on previously mentioned thin, medium and thick masks following [31].

Baseline Methods. We compare the proposed technique with a variety of DM techniques, including LDM, Stable Diffusion [27], and RePaint [20]. LDM and Stable Diffusion are preconditioned latent diffusion models. Stable Diffusion is a large scale variant of LDM trained with more data and text conditioned. RePaint [20] is a postconditioned diffusion model in the pixel space. For a fair comparison, we incorporate latent space conditioning and the explicit propagation module to the diffusion model of [20]. Comparing the proposed LatentPaint to RePaint and LDM demonstrates the positive effect of the proposed propagation modules (see Table 1). We also compare our approach against several state-of-the-art inpainting methods: the autoregressive method DSI [25], and the GAN methods AOT [41], LaMa [31], MADF [45], and COMODGAN [44]. Comparison to these techniques will help in understanding how DM in general perform for inpainting against other generative models. Where available, we use publicly available pretrained models. Where not available, we train according to published procedures.

[3] Other techniques, i.e. bilinear or bicubic interpolation could be applied with similar effect.

[4] Implemented as a multi-layer perceptron

Method	LPIPS↓	FID↓	Runtime ↓	Type	Steps
LDM	0.081	12.51	9.00	PRE	250
RePaint [20]	0.075	7.74	535.00	POST	4570
Ours	0.068	7.06	55.00	-	250

Table 1. **Quality vs. efficiency in diffusion models.** Comparison between the proposed method and existing DM on CelebA-HQ 'thick'. For each evaluation metric the arrow indicates if less or more is better. Runtime is measured in seconds for producing one 256x256-pixel image on a single NVIDIA V-100 GPU. Note that RePaint and ours are DM in pixel space. PRE: preconditioned, POST: postconditioned.

Evaluation Metrics. For quantitative comparisons, we evaluate using *LPIPS* [43] which is a learned distance metric based on the deep feature space of AlexNet, *Fréchet Inception Distance (FID) [8]* a popular deep metric for perceptual rationality [12] which measures the distance between the distributions of real and synthetic image features, *Precision* which corresponds to the average sample quality [15], and *Recall* which measures the coverage of the sample distribution [15]which is an indicator of diversity.

4.2. Results and Discussion

Quality vs. Efficiency. Table 1 shows a quality vs. efficiency benchmark for DM. Notice how the improvement in quality that RePaint provides over LDM, as shown by lower FID, comes at the great expense of efficiency. Postconditioned models are ten times more costly to sample from and require almost 20x diffusion steps to harmonize results. Our proposal on the other hand gets the best of the two worlds without their disadvantages.

General Quantitative Benchmark. Next we want to compare performance of the proposed method against the state of the art for image inpainting by using faces as a

Figure 4. **Examples of inpainted faces with random condition.** Several state-of-the-art methods are shown for easy comparison with the proposed algorithm. The first column is the condition. Each other column shows samples from a different method as indicated on top.

	Thick				Medium				Thin				
Methods	LPIPS↓	FID↓	P ↑	R ↑	LPIPS↓	FID↓	P ↑	R ↑	LPIPS↓	FID↓	P ↑	R ↑	Type
MADF [41]	0.102	17.26	0.63	0.70	0.065	10.86	0.71	**0.72**	0.044	20.29	0.65	**0.71**	GAN
COMODGAN [41]	0.078	12.14	0.71	0.69	0.057	11.27	0.74	0.70	0.043	11.19	0.75	**0.71**	GAN
AOT [41]	0.110	16.69	0.64	0.70	0.076	11.28	0.71	**0.72**	0.049	9.49	0.79	0.69	GAN
LaMa [31]	**0.062**	8.03	0.82	0.67	**0.044**	7.91	0.80	0.69	0.035	8.47	0.80	0.69	GAN
DSI [25]	0.084	10.10	0.78	0.66	0.058	8.99	0.80	0.68	0.047	10.14	0.80	0.67	AR
LDM	0.081	12.51	0.72	0.66	0.065	14.57	0.74	0.64	0.072	19.50	0.68	0.70	DM
SD 1.5[5]	0.103	14.65	0.66	**0.71**	0.081	14.12	0.66	0.71	0.070	13.83	0.70	0.68	DM
RePaint [20]	0.075	7.74	0.82	0.67	0.050	7.51	**0.82**	0.71	0.032	7.16	**0.82**	**0.71**	DM
Ours	0.068	**7.06**	**0.83**	0.68	0.048	**6.91**	**0.82**	**0.72**	**0.029**	**6.84**	**0.82**	**0.71**	DM

Table 2. **Results on CelebA.** Comparison between the proposed method and the state-of-the-art. Following [31] three versions of masking are evaluated in ascending order of the percentage of inpainting: thin, medium and thick. For each evaluation metric we indicate with an arrow if less or more is better. Best result is marked in bold. Runtime comparison between different inpainting methods is measured in seconds for producing a 256x256-pixel image on a single NVIDIA V-100 GPU. The different types of generative models are GAN, AR, and DM.

test-bed. Table 2 shows results on the CelebA-HQ dataset. Among all non-diffusion model techniques for inpainting, LaMa, a GAN based approach is the fastest requiring only few milliseconds to produce a result. Additionally it is also the best performing, as shown by low LPIPS and FID. RePaint beats the state-of-the-art at the cost of an explosion in computational cost as it takes almost nine minutes to produce a sample. This is because of the several forward-backward diffusion steps required to harmonize the inpainting result. Notice how our method is able to achieve better performance both in LPIPS and FID and similar Precision and Recall compared to RePaint in a fraction of time. This is consistent across all types of masking studied. Stable diffusion, a foundational text-to-image model, trained on LAION-5B dataset, has high generalization capability and is the most diverse as illustrated by the high Recall when thick masking is applied.

Finally, Table 4 compares FID scores of RePaint and the proposed method on complex scenes. Overall scenes are more challenging than faces. Nevertheless, the advantage shown by our method over RePaint on faces replicates for bedrooms and livingrooms as well.

Figure 5. **Examples of inpainted faces conditioned on hair.** Top row is the condition and bottom row are samples produced by the proposed method.

Ablation Study. We investigate the empirical contributions of individual components of the proposed approach. We compare against the baseline technique of [20] which performs conditioning in the pixel space. Our first observation, similar to [1], is that enabling latent space conditioning improves the consistency of inpainting straight away. Adding explicit propagation further improves inpainting results; see Table 3. Note that adding these components to a baseline U-Net for diffusion only increases the runtime by a few milliseconds. This is negligible when compared to the total runtime of reverse diffusion. We have observed empirically that adding more than two explicit propagation modules does not bring any other significant gains.

Qualitative Results. We complement the qualitative benchmark with several sets of visual examples. In Fig. 4 we compare against several state-of-the-art methods, including the best performing GAN based methods and RePaint. Notice the quality produced by our samples, and the overall coherence between the condition and the inference. We show examples of extreme masking, where almost all the face is to be inferred. In the first examples, methods like DSI struggle to paint a structurally coherent face. LaMa's sample quality is poor, especially in the mouth region and almost all the competitors seem to struggle in reconstructing the right hand. For the second example, the faces is slightly tilted to the left. Notice how both RePaint and especially COMODGAN have problems respecting the pose on the right side of the face. In the case of LaMa the inferred half of the face has a shift in illumination. In the last example both LaMa and COMODGAN have problems in the mouth region, DSI shows overall low quality while RePaint and Ours provide appealing, high quality results.

Methods	LPIPS↓	FID↓
RePaint w/o resampling [20]	0.479	14.55
Ours w/ latent space conditioning	0.341	13.13
Ours w/ one explicit propagation module	0.127	9.81
Ours w/ two explicit propagation module	0.068	7.06

Table 3. **Ablation study on CelebA.** Evaluating various settings of our proposal. All evaluation performed with thick masking.

Methods	Livingrooms	LSUN-Bedrooms
RePaint-10 [20]	24.16	18.92
Ours	21.87	16.18

Table 4. **Results on LSUN-Bedrooms and curated livingrooms dataset.** FID scores of RePaint and LatentPaint around semantic masks of sofa and bed.

In Fig. 5 we show examples of how our method conditions on hair. Notice the high quality rendered faces. Refer to the supplementary material for more examples of semantic conditioning with faces.

Finally, in Fig. 6 we show examples of how our proposal inpaints rooms. There are two sets of samples, the first with conditioning the bedrooms model on beds and the second with conditioning sofas in livingrooms. Rooms are way more challenging than faces, coming with diverse scenes composition and containing large number of objects to align. Our proposal produces coherent backgrounds, structurally coherent scenes and at least in the case of bedrooms, highly diverse backgrounds. Some general quality problems persist as shown by the higher FID in Table 4. A particular interesting phenomenon is further commented in Sec. 4.3. It is important to mention that quality and diversity of inpainting samples depends directly on the unconditional DM training and currently the sampling quality of models

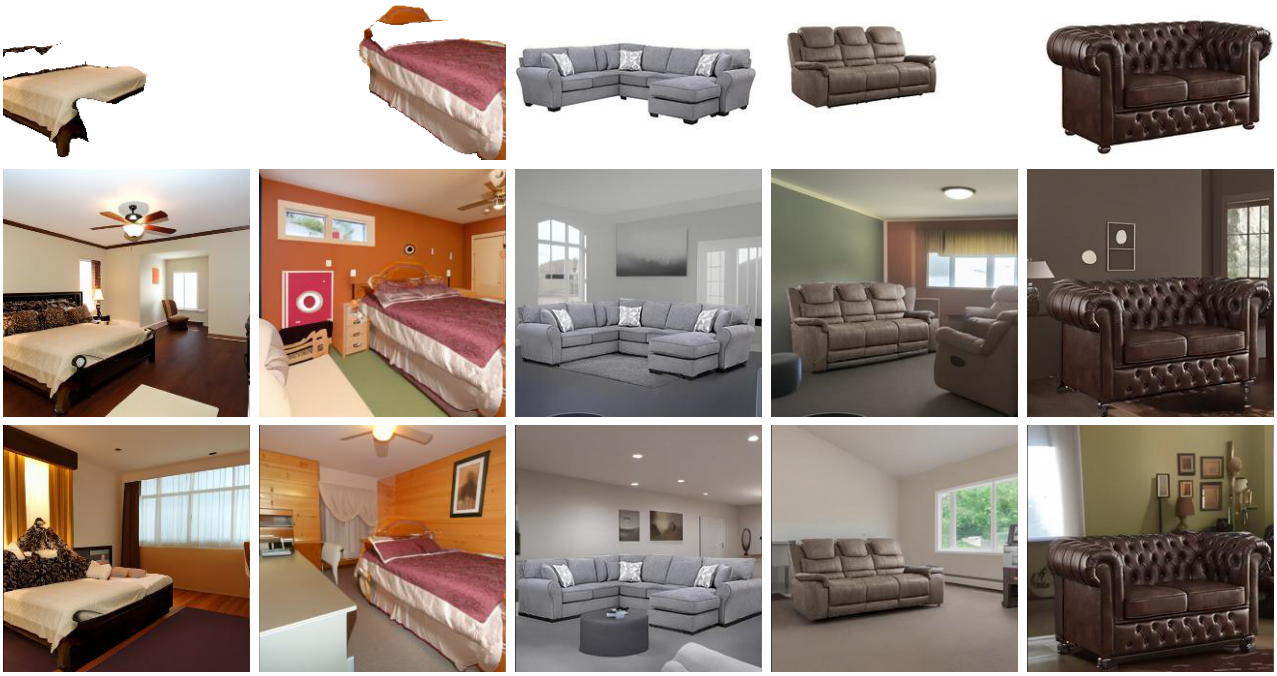

Figure 6. **Examples of inpainted rooms.** The first row is the condition. The next two rows show inpainting examples using LatentPaint. Left most two columns are bedrooms, the rest livingrooms.

on rooms does not match those of faces. More qualitative results comparing samples from all techniques are provided in supplementary.

4.3. Limitations

Our technique is not without limitations. Particular to complex scenes semantically conditioned on particular objects preserving identity can be challenging. We call this *object expansion*. For example, given a sofa (see Fig. 7) inpainting extends the sofa, adding additional legs or cushions that change it's identity. Object expansion is not specific to the proposed technique and *all* prior techniques suffer from this issue. This particular problem is more evident in scenes than in faces.

Figure 7. **Limitations.** Two examples of object expansion. Sometimes this can occur when inpainting scenes conditioned on an object. This is not a problem specific to our proposal. All methods have this problem when inpainting scenes.

5. Conclusions

We have presented LatentPaint, a simple but efficient method that of conditioning generic diffusion models for inpainting. We take the best of preconditioned and postconditioned model to derive an approach with their advantages but none of the disadvantages. First, due to its strong inductive bias, which optimizes only a small number of parameters of a carefully designed information propagation mechanism, our algorithm can be applied to existing foundational models at the cost of only a fraction of the fine-tuning steps usually required by existing preconditioned inpainting. Second, our approach samples in a fraction of the runtime compared to postcondition methods. Third, our method produces high-quality images, on-par or better than previous algorithms. We demonstrated this on faces, an intensively studied problem, and on bedrooms and living rooms two very challenging domains. Our proposal does not come without limitations though. At least in the case of complex scenes, "object expansion" remains a persistent problem for all current state-of-the-art methods, which unfortunately LatentPaint does not solve.

References

[1] Omri Avrahami, Ohad Fried, and Dani Lischinski. Blended latent diffusion. *ACM Transactions on Graphics*, 42(4):1–11, 2023. 2, 4, 7

[2] Coloma Ballester, Marcelo Bertalmio, Vicent Caselles, Guillermo Sapiro, and Joan Verdera. Filling-in by joint interpolation of vector fields and gray levels. *IEEE Transactions on Image Processing*, 10(8):1200–1211, 2001. 2

[3] Marcelo Bertalmio, Guillermo Sapiro, Vincent Caselles, and Coloma Ballester. Image inpainting. In *Proceedings of the 27th Annual Conference on Computer Graphics and Interactive Techniques*, pages 417–424, 2000. 2

[4] Marcelo Bertalmio, Luminita Vese, Guillermo Sapiro, and Stanley Osher. Simultaneous structure and texture image inpainting. *IEEE Transactions on Image Processing*, 12(8):882–889, 2003. 2

[5] Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. *Advances in Neural Information Processing Systems*, 34:8780–8794, 2021. 4

[6] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. *Advances in Neural Information Processing Systems*, 27, 2014. 2

[7] Xiefan Guo, Hongyu Yang, and Di Huang. Image inpainting via conditional texture and structure dual generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 14134–14143, 2021. 2

[8] 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. 5

[9] Seunghoon Hong, Xinchen Yan, Thomas S Huang, and Honglak Lee. Learning hierarchical semantic image manipulation through structured representations. *Advances in Neural Information Processing Systems*, 31, 2018. 2

[10] Zheng Hui, Jie Li, Xiumei Wang, and Xinbo Gao. Image fine-grained inpainting. *arXiv preprint arXiv:2002.02609*, 2020. 2

[11] Satoshi Iizuka, Edgar Simo-Serra, and Hiroshi Ishikawa. Globally and locally consistent image completion. *ACM Transactions on Graphics*, 36(4):1–14, 2017. 2

[12] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-image translation with conditional adversarial networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1125–1134, 2017. 5

[13] Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. Progressive growing of GANs for improved quality, stability, and variation. In *International Conference on Learning Representations*, 2018. 5

[14] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. In *International Conference on Learning Representations*, 2013. 3

[15] Tuomas Kynkäänniemi, Tero Karras, Samuli Laine, Jaakko Lehtinen, and Timo Aila. Improved precision and recall metric for assessing generative models. *Advances in Neural Information Processing Systems*, 32, 2019. 5

[16] Guilin Liu, Fitsum A Reda, Kevin J Shih, Ting-Chun Wang, Andrew Tao, and Bryan Catanzaro. Image inpainting for irregular holes using partial convolutions. In *Proceedings of the European Conference on Computer Vision*, pages 85–100, 2018. 2

[17] Hongyu Liu, Bin Jiang, Yibing Song, Wei Huang, and Chao Yang. Rethinking image inpainting via a mutual encoder-decoder with feature equalizations. In *Proceedings of the European Conference on Computer Vision*, pages 725–741, 2020. 2

[18] Hongyu Liu, Bin Jiang, Yi Xiao, and Chao Yang. Coherent semantic attention for image inpainting. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4170–4179, 2019. 2

[19] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin Transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10012–10022, 2021. 5

[20] 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. 2, 5, 6, 7

[21] Kamyar Nazeri, Eric Ng, Tony Joseph, Faisal Qureshi, and Mehran Ebrahimi. Edgeconnect: Structure guided image inpainting using edge prediction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*, 2019. 2

[22] Evangelos Ntavelis, Andrés Romero, Siavash Bigdeli, Radu Timofte, Zheng Hui, Xiumei Wang, Xinbo Gao, Chajin Shin, Taeoh Kim, Hanbin Son, et al. Aim 2020 challenge on image extreme inpainting. In *Proceedings of the European Conference on Computer Vision Workshops*, pages 716–741, 2020. 2

[23] Evangelos Ntavelis, Andrés Romero, Iason Kastanis, Luc Van Gool, and Radu Timofte. SESAME: semantic editing of scenes by adding, manipulating or erasing objects. In *Proceedings of the European Conference on Computer Vision*, pages 394–411, 2020. 2

[24] Deepak Pathak, Philipp Krahenbuhl, Jeff Donahue, Trevor Darrell, and Alexei A Efros. Context Encoders: Feature learning by inpainting. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2536–2544, 2016. 2

[25] Jialun Peng, Dong Liu, Songcen Xu, and Houqiang Li. Generating diverse structure for image inpainting with hierarchical VQ-VAE. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10775–10784, 2021. 5, 6

[26] Yurui Ren, Xiaoming Yu, Ruonan Zhang, Thomas H Li, Shan Liu, and Ge Li. Structureflow: Image inpainting via structure-aware appearance flow. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 181–190, 2019. 2

[27] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022. 2, 5

[28] Chitwan Saharia, William Chan, Huiwen Chang, Chris Lee, Jonathan Ho, Tim Salimans, David Fleet, and Mohammad Norouzi. Palette: Image-to-image diffusion models. In *ACM SIGGRAPH 2022 Conference Proceedings*, pages 1–10, 2022. 2

[29] Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *International Conference on Machine Learning*, pages 2256–2265, 2015. 2

[30] Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. In *International Conference on Learning Representations*, 2020. 2

[31] Roman Suvorov, Elizaveta Logacheva, Anton Mashikhin, Anastasia Remizova, Arsenii Ashukha, Aleksei Silvestrov, Naejin Kong, Harshith Goka, Kiwoong Park, and Victor Lempitsky. Resolution-robust large mask inpainting with fourier convolutions. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 2149–2159, 2022. 2, 5, 6

[32] Wei Xiong, Jiahui Yu, Zhe Lin, Jimei Yang, Xin Lu, Connelly Barnes, and Jiebo Luo. Foreground-aware image inpainting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5840–5848, 2019. 2

[33] Shunxin Xu, Dong Liu, and Zhiwei Xiong. E2I: Generative inpainting from edge to image. *IEEE Transactions on Circuits and Systems for Video Technology*, 31(4):1308–1322, 2020. 2

[34] Jie Yang, Zhiquan Qi, and Yong Shi. Learning to incorporate structure knowledge for image inpainting. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 12605–12612, 2020. 2

[35] Raymond A Yeh, Chen Chen, Teck Yian Lim, Alexander G Schwing, Mark Hasegawa-Johnson, and Minh N Do. Semantic image inpainting with deep generative models. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5485–5493, 2017. 2

[36] Zili Yi, Qiang Tang, Shekoofeh Azizi, Daesik Jang, and Zhan Xu. Contextual residual aggregation for ultra high-resolution image inpainting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7508–7517, 2020. 2

[37] 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. 5

[38] Jiahui Yu, Zhe Lin, Jimei Yang, Xiaohui Shen, Xin Lu, and Thomas S Huang. Generative image inpainting with contextual attention. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5505–5514, 2018. 2

[39] Jiahui Yu, Zhe Lin, Jimei Yang, Xiaohui Shen, Xin Lu, and Thomas S Huang. Free-form image inpainting with gated convolution. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4471–4480, 2019. 2

[40] Yanhong Zeng, Jianlong Fu, Hongyang Chao, and Baining Guo. Learning pyramid-context encoder network for high-quality image inpainting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1486–1494, 2019. 2

[41] Yanhong Zeng, Jianlong Fu, Hongyang Chao, and Baining Guo. Aggregated contextual transformations for high-resolution image inpainting. *IEEE Transactions on Visualization and Computer Graphics*, 2022. 5, 6

[42] Yu Zeng, Zhe Lin, Jimei Yang, Jianming Zhang, Eli Shechtman, and Huchuan Lu. High-resolution image inpainting with iterative confidence feedback and guided upsampling. In *Proceedings of the European Conference on Computer Vision*, pages 1–17, 2020. 2

[43] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 586–595, 2018. 5

[44] Shengyu Zhao, Jonathan Cui, Yilun Sheng, Yue Dong, Xiao Liang, Eric I Chang, and Yan Xu. Large scale image completion via co-modulated generative adversarial networks. In *International Conference on Learning Representations*, 2021. 5

[45] Manyu Zhu, Dongliang He, Xin Li, Chao Li, Fu Li, Xiao Liu, Errui Ding, and Zhaoxiang Zhang. Image inpainting by end-to-end cascaded refinement with mask awareness. *IEEE Transactions on Image Processing*, 30:4855–4866, 2021. 5