ZoomLDM: Latent Diffusion Model for multi-scale image generation

Srikar Yellapragada* Alexandros Graikos* Kostas Triaridis
Prateek Prasanna Rajarsi R. Gupta Joel Saltz Dimitris Samaras

Stony Brook University

Abstract

Diffusion models have revolutionized image generation, yet several challenges restrict their application to large-image domains, such as digital pathology and satellite imagery. Given that it is infeasible to directly train a model on 'whole' images from domains with potential gigapixel sizes, diffusion-based generative methods have focused on synthesizing small, fixed-size patches extracted from these images. However, generating small patches has limited applicability since patch-based models fail to capture the global structures and wider context of large images, which can be crucial for synthesizing (semantically) accurate samples. To overcome this limitation, we present ZoomLDM, a diffusion model tailored for generating images across multiple scales. Central to our approach is a novel magnification-aware conditioning mechanism that utilizes self-supervised learning (SSL) embeddings and allows the diffusion model to synthesize images at different 'zoom' levels, i.e., fixed-size patches extracted from large images at varying scales. ZoomLDM synthesizes coherent histopathology images that remain contextually accurate and detailed at different zoom levels, achieving state-of-the-art image generation quality across all scales and excelling in the data-scarce setting of generating thumbnails of entire large images. The multi-scale nature of ZoomLDM unlocks additional capabilities in large image generation, enabling computationally tractable and globally coherent image synthesis up to 4096×4096 *pixels and* $4\times$ *super-resolution. Additionally, multi-scale features extracted from ZoomLDM are highly effective in multiple instance learning experiments.* [1]

Accepted to The IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) 2025

Yellapragada, Srikar, et al. "ZoomLDM: Latent Diffusion Model for multi-scale image generation." arXiv preprint arXiv:2411.16969 (2024). ***Links:*** *CVPR, pdf*

Keywords: *Vision and Learning, AI for Health*

*Equal contribution. Correspondence to srikary@cs.stonybrook.edu

[1] Code is available at github.com/cvlab-stonybrook/ZoomLDM.

1. Introduction

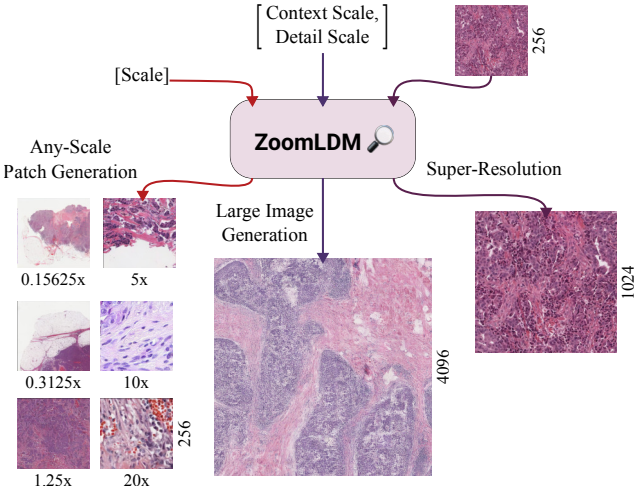

Figure 1. ZoomLDM can generate synthetic image patches at multiple scales (left). It can generate large images that preserve spatial context (center) and perform super-resolution (right), without any additional training. Large images from prior work [17, 26] suffer from blurriness and lack of global context.

Diffusion models have achieved remarkable success in photorealistic image synthesis [3], benefiting from the availability of vast multi-modal datasets [5, 41] and sophisticated conditioning techniques [20, 36]. Latent Diffusion models (LDMs) [39] have further advanced high-resolution image generation by introducing a two-step process that first compresses the images with a learned encoder and then trains the generative diffusion model in that encoder's latent space. In the natural image domain, LDMs like Stable Diffusion XL [36], which generates 1024×1024 images, have made high-resolution generation fast and cheap. Although such models demonstrate the potential of further scaling image diffusion to larger sizes, large-image domains such as digital histopathology and satellite imagery are beyond their feasible scope as images there are typically in the gigapixel scale (e.g. $32,000 \times 32,000$ pixels).

Apart from scale, large-image domains also lack paired image-annotation data with sufficient detail, which has been

key to the success of text-to-image diffusion models. Without access to a conditioning signal during training and inference, the performance of diffusion models degrades significantly [32]. At the same time, obtaining annotations for large images can be complex as it is both a laborious process for specialized fields, such as medical images, and often ambiguous since annotators can describe different features at different scales. A satellite image text caption corresponding to *'water'*, when viewed from up close, can turn into both the *'a lake'* and *'a river'* when viewed from further away, making it necessary to annotate at both levels.

Previous works have tried to address the issues of large image sizes and conditioning but are limited in applicability. Harb et al. [18] introduced a pixel-level diffusion model that can accommodate multiple scales (named magnifications) in medical images but lacked conditioning - a crucial element for achieving better image quality and enabling downstream tasks [11, 31, 49]. Graikos et al. [17] utilized embeddings from self-supervised learning (SSL) models to mitigate the need for costly annotations in large-image domains, but only trained a model to generate small patches. Recognizing that none of these methods can tackle the important problem of *controllable high-quality large-image synthesis*, we propose a unified solution, ZoomLDM.

To address large image sizes, we propose training a scale-conditioned diffusion model that learns to generate images at different 'zoom' levels, which correspond to magnifications in histopathology images (Fig. 1 (a)). By conditioning the model on the scale, we control the level of detail contained within each generated pixel. To control generation, we also incorporate a conditioning signal from a self-supervised learning (SSL) encoder. While SSL encoders are great at producing meaningful representations for images, using them in this multi-scale setting is nontrivial as they are usually trained to extract information from patches at a single scale. To share information across scales, we introduce the idea of a cross-magnification latent space; a shared latent space where the embeddings of all scales lie. We implement this with a trainable *summarizer* module that processes the array of SSL embeddings that describe an image, projecting them to the shared latent space that captures dependencies across all magnifications.

We train ZoomLDM on multi-scale histopathology using SSL embeddings from state-of-the-art image encoders as guidance. We find that sharing model weights across all scales significantly boosts the generation quality for scales where data is limited. To eliminate our model's reliance on SSL embeddings when sampling new images, we also train a Conditioning Diffusion Model (CDM) that generates conditions given a scale. This combined approach enables us to synthesize novel high-quality images at all scales.

With a multi-scale model, we hypothesize that jointly sampling images across scales would be beneficial for creating coherent images at multiple scales. However, this is challenging because each scale requires its own level of detail, and these details must be aligned across scales. To that end, we propose a novel joint multi-scale sampling approach that exploits ZoomLDM's multi-scale nature. Our cross-magnification latent space provides the necessary detail across scales, enabling large image generation and super-resolution without additional training. This approach effectively constructs a coherent image pyramid, making super-resolution and high-quality large image generation feasible. Our method surpasses previous approaches [17, 26], which struggled in generating either local details or global structure, and presents the first practical 4096×4096 image generation paradigm in histopathology (see supplementary for a comprehensive evaluation).

Finally, we probe ZoomLDM to show that features extracted from our model are highly expressive and suitable for multiple instance learning (MIL) tasks in digital histopathology. Prior work [7, 27] has demonstrated the effectiveness of multi-scale features for MIL, but these methods required training separate encoders for each scale. In contrast, ZoomLDM offers an efficient solution by enabling seamless multi-scale feature extraction using a single model. We condition ZoomLDM with UNI[9], a SoTA SSL model, and extract intermediate features from the denoiser at multiple magnifications for MIL. As expected, fusing ZoomLDM features from multiple scales outperforms using SoTA encoders in our MIL experiments, displaying the efficacy of its multi-scale representations. Surprisingly, our features from just the $20\times$ magnification alone surpass UNI features. We hypothesize that by learning to generate at multiple scales, ZoomLDM has learned to produce more informative features.

Our contributions are the following:

- We present **ZoomLDM**, the first multi-scale conditional latent diffusion model that generates images at multiple scales, achieving state-of-the-art synthetic image quality.
- We introduce a cross-magnification latent space, implemented with a trainable summarizer module, which provides conditioning across scales, allowing ZoomLDM to capture dependencies across magnifications.
- We propose a novel joint multi-scale sampling approach for generating large images that retain both global context and local fidelity, making us the first to efficiently synthesize good quality histopathology image samples of up to 4096×4096 pixels.
- We probe the learned multi-scale representations of ZoomLDM and demonstrate their usefulness by surpassing SoTA encoders on multiple instance learning tasks.

2. Related Work

Diffusion models: Since their initial introduction to image generation in Ho et al. [21], diffusion models have be-

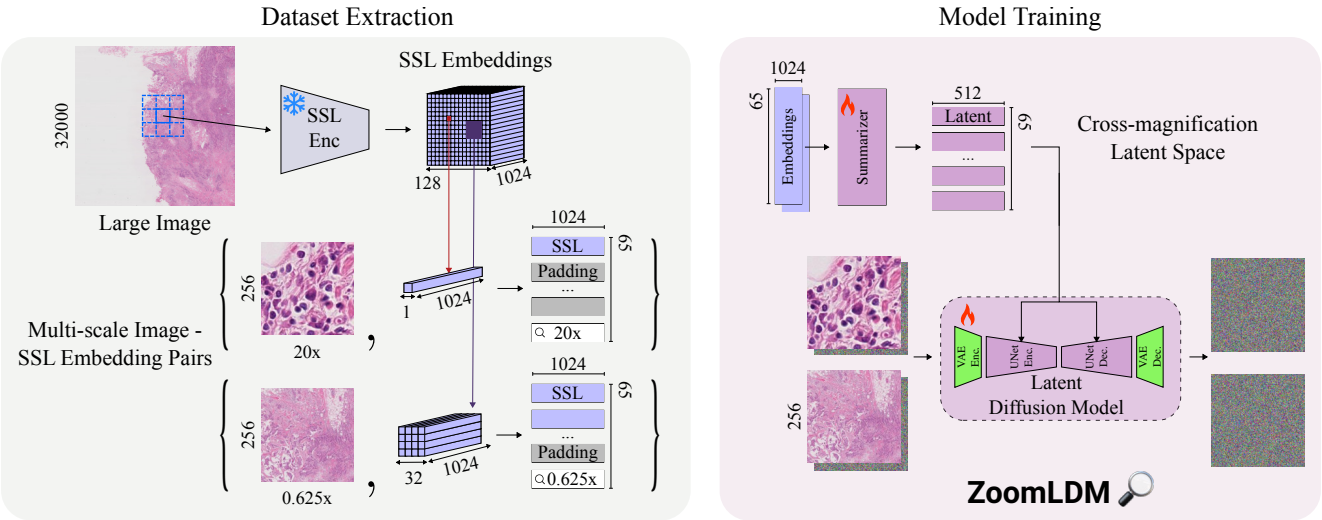

Figure 2. Overview of our approach. Left: We extract 256 × 256 patches from large images at the initial scale (20× for pathology) and generate SSL embedding matrices using pretrained encoders. The large image is then progressively downsampled by a factor of 2, with patches at each scale paired with the SSL embeddings of all overlapping initial-scale patches. Right: The SSL embeddings and magnification level are fed to the Summarizer, which projects them into the cross-magnification Latent space. The diffusion model is trained to generate 256 × 256 patches conditioned on the Summarizer's output.

come the dominant generative models for images. Several works have been pivotal; notably class conditioning [31] which highlighted the importance of guidance during training and sampling and its extensions with classifier [11] and classifier-free guidance [20]. Latent Diffusion Models (LDMs) [39] proposed a training the diffusion model in a Variational Autoencoder (VAE) latent space, compressing the input images by a factor of up to ×8 and enabling high-resolution and computationally practical image generation. Denoising Diffusion Implicit Models (DDIM) [44] accelerated the sampling process further, making diffusion models the preferred alternative over all previous generative model approaches (GANs, Normalizing Flows).

Diffusion Models in Large-Image Domains: Despite advances in the domain of natural images, training generative models directly at the gigapixel resolution of large image domains remains infeasible. Proposed alternatives generate images in a coarse-to-fine process by chaining models in a cascading manner [35, 40]. This has led to synthesizing images of up to 1024 × 1024 resolution at the cost of increased parameter count and slower inference speed. Recently, PixArt-Σ [6] introduced an efficient transformer architecture that enables image generation of up to $4k$ using a weak-to-strong training strategy.

In the context of histopathology, previous works have focused on training fixed-size, patch diffusion models [29, 30, 48, 49], with similar approaches in satellite data [13, 42]. Patch models were used to extrapolate to large images in [2], where a pre-generated segmentation mask guides the patch model over the large image, and [17] where a patch model is conditioned on SSL embeddings that smoothly vary across the large image, synthesizing appearance locally. Both methods fail to understand global structures and rely on external sources of information for guidance.

More closely related to our work, [18] trains a pathology diffusion model conditioned on image scales. However, limited evaluations and the absence of a conditioning mechanism restrict its applicability. A different approach by Le et al. [26] utilized an infinite-dimensional diffusion model that is resolution-free, meaning that it can be trained on arbitrarily large images. Their model can be scaled for up to 4096 × 4096 generation, but the final results are usually blurry and lack details.

3. Method

3.1. Unified Multi-Scale Training

We train ZoomLDM to generate fixed-size 256 × 256 patches extracted at different scales of large images. To guide generation, we introduce a novel conditioning mechanism allowing the model to learn multi-scale dependencies. Figure 2 provides an overview of our multi-scale training.

We begin by extracting 256 × 256 image patches from a large image at full resolution. Since there are no descriptive patch-level annotations in large-image domains, we resort to pre-trained SSL encoders to provide detailed descriptors in place of human labels, as in [17]. The SSL encoders in these domains are usually trained on patches from these large images – for histopathology, we utilize UNI [7], an image encoder trained on 224 × 224 px 20× patches. Af-

ter extracting patches $\boldsymbol{I}^1$ at the initial scale (=1) and SSL embeddings $\boldsymbol{e}$, we end up with a dataset of $\{\boldsymbol{I}_i^1, \boldsymbol{e}_i\}$ pairs.

We downsample the large image by a factor of 2 and repeat the patch extraction process, getting a new set of patches at the next zoom level. But, as previously mentioned, we cannot directly use the SSL encoder on images from different scales – e.g., UNI is only trained on $20\times$ images. Therefore, for scales above the first, we utilize the embeddings corresponding to the region contained within the context of the current-scale patch as conditioning. This means that we pair $\boldsymbol{I}^2$ patches with the embeddings of all the $\boldsymbol{I}^1$ images that they contain, giving us a dataset of $\{\boldsymbol{I}_i^2 \left(\begin{smallmatrix} \boldsymbol{e}_1 & \boldsymbol{e}_2 \\ \boldsymbol{e}_3 & \boldsymbol{e}_4 \end{smallmatrix}\right)_i\}$ pairs.

By repeating this process, we construct a dataset of (image, embeddings) pairs for all scales, which we want to utilize as our training data for a latent diffusion model. The issue is that the number of SSL embeddings for an image size increases exponentially as we increase scale. This leads to significant computational overhead, primarily due to the quadratic complexity of cross-attention mechanisms used to condition diffusion models. Additionally, conditioning the generation of 256×256 images with a massive number of embeddings is redundant, given that if we have a total of 8 scales then we will be using a $128 \times 128 \times D$ condition to generate a single $256 \times 256 \times 3$ patch.

To address this issue, we introduce the idea of a learned cross-magnification latent space, shared across embeddings of all scales. To implement this, we train a "*Summarizer*" transformer, jointly with the diffusion denoiser, that processes the SSL embeddings extracted alongside every image. The information contained in the embeddings is "summarized" in conjunction with an embedding of the image scale, extracting the essential information needed by the LDM to synthesize patches accurately.

The variable number of tokens (embeddings) in the summarizer input is transformed into a fixed-sized set of conditioning tokens. We utilize padding and pooling to provide a fixed-size output with which we train the LDM. The magnification embedding added to the input makes the summarizer scale-aware, allowing it to adapt to the appropriate level of detail required at different scales. The output of the Summarizer then serves as conditioning input for the LDM, enabling the model to generate high-quality patches with scale-adaptive conditioning.

Conditioning Diffusion Model. Our image synthesis pipeline requires a set of SSL embeddings and the desired magnification level, which involves extracting the conditioning information from reference real large-images. This becomes impractical when direct access to training data is unavailable. To address this, we train a second diffusion model, the Conditioning Diffusion Model (CDM), which learns to sample from the distribution of the learned cross-magnification latent space after training the LDM.

Rather than training a diffusion model to model the distribution of the SSL embeddings, which is as complex as learning the distribution of images, we learn the output of the Summarizer, as it captures the most relevant information for synthesizing an image at a given magnification. This approach allows the CDM to model a more refined, task-specific latent space. By also conditioning the CDM on scale, we enable magnification-aware novel image synthesis, which we show can generate high-quality, non-memorized images at the highest scale, even if the amount of data is incredibly scarce (2500 images at $0.15625\times$ magnification).

3.2. Joint Multi-Scale Sampling

One of the biggest challenges in large-image domains is synthesizing images that contain local details and exhibit global consistency. Due to their immense sizes, we cannot directly train a model on the full gigapixel images, and training on individual scales will either lead to loss of detail or contextually incoherent results.

We propose a multi-scale training pipeline intrinsically motivated by the need to sample images from multiple scales jointly. By drawing samples jointly, we can balance the computational demands of generating large images by separating the global context generation, which is offset by synthesizing an image at a coarser scale, and the synthesis of fine local details, which is done at the lowest level.

We develop a joint multi-scale sampling approach that builds upon ZoomLDM's multi-scale nature and enables us to generate large images of up to 4096×4096 pixels. The key to our approach is providing 'self-guidance' to the model by guiding the generation of the lowest scales using the so-far-generated global context. To implement this guidance we build upon a recent diffusion inference algorithm [16], which enables fast conditional inference.

Inference Algorithm An image at scale $s+1$ corresponds to four images at the previous scale s since, during training, we downsample the large images by a factor of 2 at every scale. We want to jointly generate the four patches at the smaller scale $\boldsymbol{x}_i^s,\ i = 1, \ldots, 3$ and the single image at the next level $\boldsymbol{x}^{s+1}$. The relationship between these images is known; we can recover $\boldsymbol{x}^{s+1}$ by multiplying with a linear downsampling operator $\boldsymbol{A}$:

$$\boldsymbol{x}^{s+1} = \boldsymbol{A} \begin{pmatrix} \boldsymbol{x}_1^s & \boldsymbol{x}_2^s \\ \boldsymbol{x}_3^s & \boldsymbol{x}_4^s \end{pmatrix}. \tag{1}$$

We use the above matrix notation to denote the spatial arrangement of images. The algorithm proposed in [16] introduces a method to sample an image from a diffusion model given a linear constraint. Given that our multi-scale images are related by a linear constraint, we use a modified version

of this algorithm to perform joint sampling across magnifications. We first provide a brief overview and then present the modifications necessary for joint multi-scale sampling.

Since we use an LDM, we perform the denoising in the VAE latent space and require the Dec and Enc networks to map from latents $\boldsymbol{z}$ to images $\boldsymbol{x}$ and back. The algorithm requires a linear operator $\boldsymbol{A}$ (and its transpose $\boldsymbol{A}^T$) and a pixel-space measurement $\boldsymbol{y}$ that we want our final sample $\boldsymbol{z}_0$ to match, minimizing $C = ||\boldsymbol{A}Dec(\boldsymbol{z}_0) - \boldsymbol{y}||_2^2$. In every step t of the diffusion process, the current noisy latent $\boldsymbol{z}_t$ is used to estimate the final 'clean' latent $\hat{\boldsymbol{z}}_0(\boldsymbol{z}_t)$, by applying the denoiser model $\boldsymbol{\epsilon}_\theta(\boldsymbol{z}_t)$ and Tweedie's formula [12]. In the typical DDIM [44] sampling process, the next diffusion step is predicted as

$$\boldsymbol{z}_{t-1} = \sqrt{\bar{\alpha}_t}\hat{\boldsymbol{z}}_0(\boldsymbol{z}_t)\sqrt{1-\bar{\alpha}_t}\boldsymbol{\epsilon}_\theta(\boldsymbol{z}_t) + \tilde{\beta}_t\boldsymbol{\epsilon}_t. \quad (2)$$

The algorithm of [16] proposes minimizing the $C(\boldsymbol{z}_t) = ||\boldsymbol{A}Dec(\hat{\boldsymbol{z}}_0(\boldsymbol{z}_t)) - \boldsymbol{y}||_2^2$ w.r.t. $\boldsymbol{z}_t$ at every timestep t before performing the DDIM step. To do that it first computes an error direction as

$$\boldsymbol{e} = \nabla\hat{\boldsymbol{z}}_0||\boldsymbol{A}Dec(\hat{\boldsymbol{z}}_0(\boldsymbol{z}_t)) - \boldsymbol{y}||_2^2. \quad (3)$$

This error direction and the current noisy sample $\boldsymbol{z}_t$ are used to compute the gradient $\boldsymbol{g} = \nabla_{\boldsymbol{z}_t}C(\boldsymbol{z}_t) = \nabla_{\boldsymbol{z}_t}||\boldsymbol{A}\hat{\boldsymbol{z}}_0(\boldsymbol{z}_t) - \boldsymbol{y}||_2^2$ using a finite difference approximation and the current noisy sample $\boldsymbol{z}_t$ is updated:

$$\boldsymbol{g} \approx \left[\hat{\boldsymbol{z}}_0(\boldsymbol{z}_t + \delta\boldsymbol{e}) - \hat{\boldsymbol{z}}_0(\boldsymbol{z}_t)\right]/\delta, \quad (4)$$

$$\boldsymbol{z}_t \leftarrow \boldsymbol{z}_t + \lambda\boldsymbol{g}. \quad (5)$$

Efficient Joint Sampling We make two significant modifications to this algorithm to perform the joint multi-scale sampling. First, since we do not have access to a real measurement $\boldsymbol{y}$, which corresponds to the higher scale image $\boldsymbol{x}^{s+1}$, we use the estimate of the image $Dec(\hat{\boldsymbol{z}}^{s+1})$ to guide the generation of z^s. Second, we propose a more efficient way of computing error direction (Eq. 3), which does not require memory and time-intensive backpropagations. To jointly sample images from scales $s+4$ and s we need to generate $16\times16+1$ total images, which would be infeasible with the previous error computation.

To avoid the backpropagation during (Eq. 3) we propose computing a numerical approximation of $\boldsymbol{e}$. Similar to Eq. 5 we utilize finite differences and compute

$$\boldsymbol{e} \approx \left[Enc(Dec(\hat{\boldsymbol{z}}_0) + \zeta\boldsymbol{e}_{img}) - Enc(Dec(\hat{\boldsymbol{z}}_0))\right]/\zeta \quad (6)$$

where $\boldsymbol{e}_{img} = \boldsymbol{A}^T(\boldsymbol{A}Dec(\hat{\boldsymbol{x}}_0(\boldsymbol{x}_t)) - \boldsymbol{y})$. This eliminates the need to backpropagate through the decoder without significantly sacrificing the quality of the images generated. We provide a detailed background of the conditional inference algorithm and how our approximation reduces computation in the supplementary material.

4. Experiments

In this section, we showcase the experiments conducted to validate the effectiveness of our method. We train the unified latent diffusion model, ZoomLDM, on patches from eight different magnifications in histopathology. We evaluate the quality of synthetic samples using both real and CDM-sampled conditions. Further, we exploit the multi-scale nature of ZoomLDM to demonstrate its strength in generating good quality high-resolution images across scales, and its utility in super-resolution (SR) and multiple instance learning (MIL) tasks.

4.1. Setup

4.1.1. Implementation details

We train the LDMs on 3 NVIDIA H100 GPUs, with a batch size 200 per GPU. We use the training code and checkpoints provided by [39]. Our LDM configuration consists of a VQ-f4 autoencoder and a U-Net model pre-trained on ImageNet. We set the learning rate at 10^{-4} with a warmup of 10,000 steps. The Summarizer is implemented as a 12-layer Transformer, modeled after ViT-Base. For the CDM, we train a Diffusion Transformer [34] on the outputs of the summarizer. We utilize DDIM sampling [44] with 50 steps for both models and apply classifier-free guidance [20] sampling with a scale of 2.0 to create synthetic images. See supplemental for more details on the Summarizer and CDM.

4.1.2. Dataset

We select 1,136 whole slide images (WSI) from TCGA-BRCA [4]. Using the code from DSMIL[27], we extract 256×256 patches at eight different magnifications: $20\times$, $10\times$, $5\times$, $2.5\times$, $1.25\times$, $0.625\times$, $0.3125\times$, and $0.15625\times$. Each patch is paired with its corresponding base resolution ($20\times$) region—for instance, a 256×256 pixel patch at 5x magnification is paired with a 1024×1024 pixel region at $20\times$. We then process the $20\times$ regions through the UNI encoder [8] to produce an embedding matrix for each patch.

The dimensions of this embedding matrix vary based on the patch's magnification level. For example, a $5\times$ patch corresponding to a $20\times$ region of size 1024×1024 results in an embedding matrix of dimensions $4\times4\times1024$. As discussed previously, to avoid redundancy in large embedding matrices, we average pool embeddings larger than 8×8 to 8×8 (magnifications $1.25 \times$ and lower).

In the supplementary, we also provide results for training ZoomLDM on satellite images. We use a similar training setting, replacing the WSIs from pathology with NAIP [45] tiles and the SSL encoder with DINO-v2 [33], showing the wider applicability of the proposed model.

4.2. Image quality

For every histopathology magnification, we generate 10,000 256×256 px patches using ZoomLDM and evaluate their

Table 1. FID of patches generated from ZoomLDM across different magnifications, compared with single magnification models. ZoomLDM achieved better FID scores than SoTA, with particularly significant improvements at lower scales.

Magnification	20×	10×	5×	2.5×	1.25×	0.625×	0.3125×	0.15625×
# Training patches	12 Mil	3 Mil	750k	186k	57k	20k	7k	2.5k
ZoomLDM	**6.77**	**7.60**	**7.98**	**10.73**	**8.74**	**7.99**	**8.34**	**13.42**
SoTA	6.98 [17]	7.64 [49]	9.74 [17]	20.45	39.72	58.98	66.28	106.14
CDM	9.04	10.05	14.36	19.68	14.06	13.46	14.40	26.09

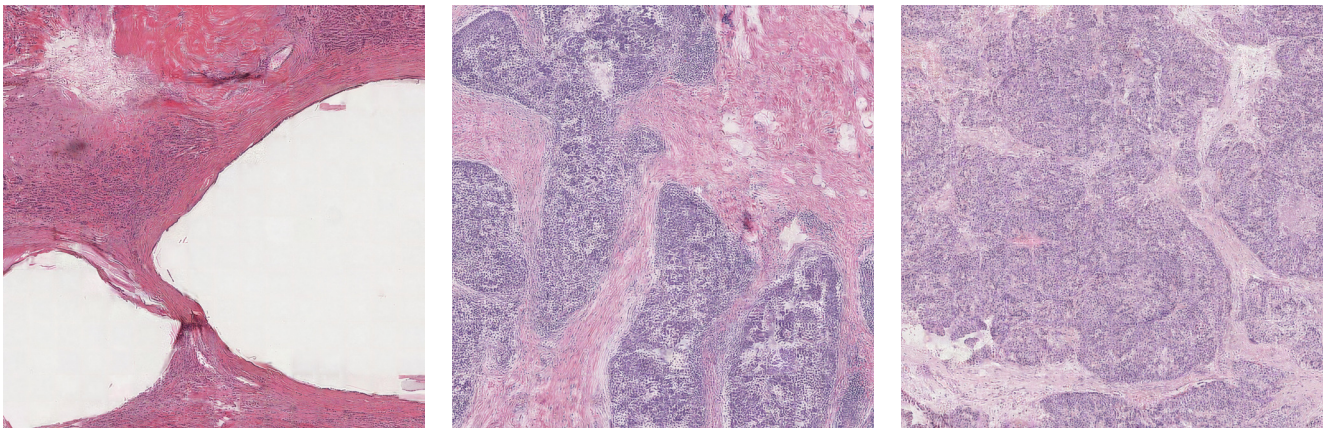

Figure 3. Large Images (4096 × 4096) generated from ZoomLDM. Our large image generation framework is the first to generate 4k pathology images with **local details** and **global consistency**, all within reasonable inference time. We provide more 4k examples and comparisons in the supplementary.

quality using the Fréchet Inception Distance (FID) [19]. For 20×, 10× and 5× magnifications, we compare against the state-of-the-art (SoTA) works of [17, 49]. For lower magnifications, we train standalone models specifically for patches from those magnifications, keeping the architecture consistent with ZoomLDM.

As indicated in Table 1, ZoomLDM achieves superior performance across all magnifications compared to the SoTA models. We see larger improvements for magnifications below 2.5×, where the data scarcity severely impacts the model's ability to synthesize diverse, high-quality images. This highlights the advantage of our unified architecture and conditioning approach. By leveraging data and conditioning across all magnifications, we allow the low-density data regions to benefit from the insights that the model gains from the entire dataset, improving both model performance and efficiency.

Novel image synthesis: For FID comparisons above, images were generated by randomly sampling SSL embeddings for different magnifications from the dataset. However, this approach is not always practical as it requires access to the dataset of embeddings at all times. To address this, we use the Conditioning Diffusion Model to draw samples from the shared cross-magnification latent space and generate new images conditioned on these latents (CDM row in Table 1). Despite the slight increase in FID – an expected outcome since the CDM cannot perfectly capture the true learned conditioning latent space, we still observe that the generated samples outperform the baselines in the data-scarce settings. We believe that this further emphasizes the importance of our shared cross-magnification latent space, by showing that we can model its distribution and capture all scales effectively. In supplementary we show synthetic images at 0.15625× and with their closest neighbors in the dataset to demonstrate the absence of memorization.

Table 2. CLIP and Crop FID values (lower is better) for our large image generation experiments. ZoomLDM outperforms previous works on 1024 × 1024 generation. While we lack in 4096 × 4096 FIDs, we provide qualitative examples in the supplementary that highlight the fundamental differences that emerge when scaling up the three methods. Inference time for a single image shows that our method is the only practical approach for 4k image generation.

Method	1024 × 1024			4096 × 4096		
	Time / img	CLIP FID	Crop FID	Time / img	CLIP FID	Crop FID
Graikos et al. [17]	60 s	7.43	15.51	4 h	2.75	**11.30**
∞-Brush [26]	30 s	3.74	17.87	12 h	**2.63**	14.76
ZoomLDM	28 s	**1.23**	**14.94**	8 m	6.75	18.90

4.3. Large image generation

In Section 3.2, we presented an algorithm for jointly sampling images at multiple scales. We perform experiments

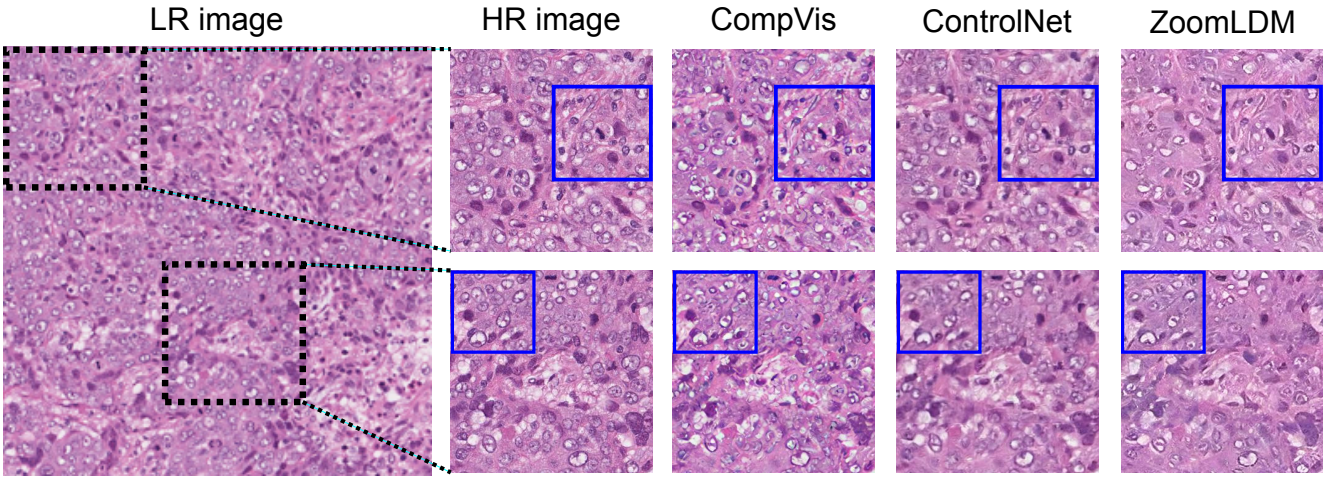

Figure 4. We showcase $4\times$ super-resolution results ($256 \times 256 \rightarrow 1024 \times 1024$). Samples generated by other methods [39, 52] exhibit artifacts, inconsistencies, and blurriness that are not present in our outputs. Specifically, in blue boxes, we can observe that CompVis[39] generates fine scale artifacts, while ControlNet[52] produces generally blurry outputs. ZoomLDM produces a sharp output, generating details generally consistent with the ground truth image.

Table 3. Super-resolution results on TCGA-BRCA [4] and BACH [1] using ZoomLDM and other diffusion-based baselines. Using ZoomLDM with the proposed condition inference achieves the best performance.

Method	Conditioning	TCGA BRCA					BACH				
		SSIM ↑	PSNR ↑	LPIPS↓	CONCH ↑	UNI ↑	SSIM ↑	PSNR ↑	LPIPS↓	CONCH ↑	UNI ↑
Bicubic	-	**0.653**	**24.370**	0.486	0.871	0.524	**0.895**	**34.690**	0.180	0.969	**0.810**
CompVis [39]	LR image	0.563	21.926	0.247	0.946	0.565	0.723	27.278	0.206	0.954	0.576
ControlNet [52]	LR image	0.543	21.980	0.252	0.874	0.563	0.780	27.339	0.276	0.926	0.721
ZoomLDM	Uncond	0.591	23.217	0.260	0.936	0.680	0.739	29.822	0.235	0.965	0.741
	GT emb	0.599	23.273	0.250	0.946	0.672	0.732	29.236	0.245	0.974	0.753
	Infer emb	0.609	23.407	**0.229**	**0.957**	**0.719**	0.779	30.443	**0.173**	**0.974**	0.808

on generating $20\times$ histopathology images jointly with other magnifications in two settings: Sampling $20\times$ with $5\times$, generating 1024×1024 images and sampling $20\times$ with $1.25\times$, giving 4096×4096 samples. We employ bicubic interpolation as the downsampling operator $\boldsymbol{A}$, where for $5\times$ and $1.25\times$, we downsample by $4\times$ and $16\times$, respectively.

In Table 2, we showcase CLIP FID and Crop FID values, adopted from [26], and compare our large-image generation method against existing state-of-the-art approaches. CLIP FID downsamples the full image and extracts features from a CLIP [37] model, whereas Crop FID extracts 256×256 crops from the large images and computes FID using the conventional Inception features [43].

On 1024×1024 generation we easily outperform existing approaches with similar or smaller sampling times. While, on 4096×4096 generation, we find that our method lags in two quality metrics but offers a reasonable inference time per image (8min vs $>$ 4hrs). However, regarding the 4096×4096 results, we find fundamental differences between our synthesized images (Figure 3) and those of [17, 26] (see supplementary). We particularly find that the local patch-based model of Graikos et al. [17] completely fails to capture the global context in the generated images. While it generates great quality patches and stitches them together over the 4096×4096 canvas, the overall image does not resemble a realistic pathology image. On the other hand, ∞-Brush [26] captures the global image structures but produces blurry results. In contrast, ZoomLDM balances local details and global structure, producing images that not only exhibit high fidelity but also maintain overall realism across the entire 4096×4096 canvas. We are the first to generate 4k pathology images with both detail and global coherency under a tractable computational budget.

4.4. Super-resolution

Our joint multi-scale sampling allows us to sample multiple images from different magnifications simultaneously. However, a question arises of whether we could also use ZoomLDM in super-resolution, where the higher-scale image is given and the details need to be inferred. We provide a solution for super-resolution with ZoomLDM using a straightforward extension of our joint sampling algorithm.

The main challenge we need to overcome is the absence of conditioning. Given only an image at a magnification other than $20\times$, we cannot obtain SSL embeddings, which are extracted from a $20\times$-specific encoder. Nevertheless, we discover an interesting inversion property of our model, which allows us to infer the conditioning given an image and its magnification. Similar to textual inversion [15], and more recently prompt tuning [10], we can optimize the SSL input to the summarizer to obtain a set of embeddings that generate images that resemble the one provided. We discuss the inversion approach in the supplementary material in more detail, along with inversion examples.

Once we have obtained a set of plausible conditioning embeddings, we can run our joint multi-scale sampling algorithm, fixing the measurement $\boldsymbol{y}$ to the real image we want to super-resolve. To test ZoomLDM's capabilities, we construct a simple testbed of $4\times$ super-resolution on in-distribution and out-of-distribution images from TCGA-BRCA and BACH [1] respectively. As baselines, we use bicubic interpolation, a naive super-resolution-specific LDM trained on OpenImages [25] (CompVis), and a ControlNet [52] trained on top of ZoomLDM.

In Table 3 and Figure 4, we present the results of our experiments. We find that SSIM and PSNR are slightly misleading as they favor the blurry bicubic images, but also point out some significant inconsistencies in the LDM and the ControlNet outputs. For better comparisons, we also compute LPIPS [53] and CONCH [28] similarity, which downsamples the image to 224×224 as well as UNI similarity, which we consider on a per 256×256 patch-level. In most perceptual metrics, we find ZoomLDM inference to be the best-performing while remaining faithful to the input image. Interestingly, we discover that using the embedding inversion that infers the conditions from the low-res given image performs better than providing the real embeddings.

Table 4. AUC for BRCA subtyping and HRD prediction. Features extracted from ZoomLDM outperform SoTA vision encoders.

Features	Mag	Subtyping	HRD
Phikon [14]	$20\times$	93.81	76.88
UNI [8]	$20\times$	94.09	81.79
CTransPath [47]	$5\times$	93.11	85.37
ZoomLDM	$20\times$	94.49	85.25
	$5\times$	94.09	86.26
	Multi-scale ($20\times$ + $5\times$)	**94.91**	**88.03**

4.5. Multiple Instance Learning

Multiple instance learning (MIL) tasks benefit from multi-scale information, as different magnifications reveal distinct and complementary features. Prior work [7, 27] that demonstrated this behavior required training separate encoders for each scale. We hypothesize that ZoomLDM offers an efficient solution by enabling seamless multi-scale feature extraction.

To validate this hypothesis, we utilize ZoomLDM as a feature extractor and apply a MIL approach for slide-level classification tasks of Breast cancer subtyping and Homologous Recombination Deficiency (HRD) prediction - both of which are binary classification tasks. For each patch in the WSI, we extract features from ZoomLDM's U-Net output block 3 at a fixed timestep $t = 100$, conditioned on UNI embeddings. We employ a 10-fold cross-validation strategy for subtyping, consistent with the data splits from HIPT [7], and a 5-fold cross-validation for HRD prediction, reporting performance on a held-out test split as per SI-MIL [24]. We compare ZoomLDM's features to those from SoTA encoders—Phikon [14], CTransPath [47], and UNI [8], using the ABMIL method [22, 23].

As expected, the results in Table 4 show that ZoomLDM's multi-scale features (fusing $20\times$ and $5\times$ outperform SoTA encoders in both tasks. This improvement highlights the effectiveness of ZoomLDM's cross-magnification latent space in capturing multi-scale dependencies. Surprisingly, even in a single magnification setting, ZoomLDM outperforms all SoTA encoders. This result suggests that by learning to generate across scales, ZoomLDM learns to produce features that can be aware of the cross-magnification long-range dependencies, and therefore exceed the capabilities of those produced by SSL encoders for downstream MIL tasks.

5. Conclusion

We presented ZoomLDM, the first conditional diffusion model capable of generating images across multiple scales with state-of-the-art synthetic image quality. By introducing a cross-magnification latent space, implemented with a trainable summarizer module, ZoomLDM effectively captures dependencies across magnifications. Our novel joint multi-scale sampling approach allows for efficient generation of large, high-quality and structurally coherent histopathology images up-to 4096×4096 pixels while preserving both global structure and fine details.

In addition to synthesis, ZoomLDM demonstrates its utility as a powerful feature extractor in multiple instance learning experiments. The multi-scale representations learned by our model outperform SoTA SSL encoders in slide-level classification tasks, enabling more accurate subtyping, prognosis prediction, and biomarker identification. Furthermore, our Condition Diffusion Model demonstrates the potential to integrate diverse input sources such as text or RNA sequences, paving the way for realistic synthetic datasets for training and evaluating pathologists as well as controlled datasets for quality assurance. ZoomLDM is

a step toward achieving foundation generative models in histopathology, with the potential to shed light on tumor heterogeneity, refine cancer gradings, and enrich our understanding of cancer's various manifestations.

Acknowledgements This research was partially supported by NSF grants IIS-2123920, IIS-2212046, NIH grants 1R01CA297843-01, 3R21CA258493-02S1 and NCI awards 1R21CA25849301A1, UH3CA225021.

References

[1] Guilherme Aresta, Teresa Araújo, Scotty Kwok, Sai Saketh Chennamsetty, Mohammed Safwan, Varghese Alex, Bahram Marami, Marcel Prastawa, Monica Chan, Michael Donovan, Gerardo Fernandez, Jack Zeineh, Matthias Kohl, Christoph Walz, Florian Ludwig, Stefan Braunewell, Maximilian Baust, Quoc Dang Vu, Minh Nguyen Nhat To, Eal Kim, Jin Tae Kwak, Sameh Galal, Veronica Sanchez-Freire, Nadia Brancati, Maria Frucci, Daniel Riccio, Yaqi Wang, Lingling Sun, Kaiqiang Ma, Jiannan Fang, Ismael Kone, Lahsen Boulmane, Aurélio Campilho, Catarina Eloy, António Polónia, and Paulo Aguiar. Bach: Grand challenge on breast cancer histology images. *Medical Image Analysis*, 56:122–139, 2019. 7, 8

[2] Marco Aversa, Gabriel Nobis, Miriam Hägele, Kai Standvoss, Mihaela Chirica, Roderick Murray-Smith, Ahmed Alaa, Lukas Ruff, Daniela Ivanova, Wojciech Samek, et al. Diffinfinite: Large mask-image synthesis via parallel random patch diffusion in histopathology. In *Thirty-seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2023. 3

[3] James Betker, Gabriel Goh, Li Jing, Tim Brooks, Jianfeng Wang, Linjie Li, Long Ouyang, Juntang Zhuang, Joyce Lee, Yufei Guo, et al. Improving image generation with better captions. *Computer Science. https://cdn. openai. com/papers/dall-e-3. pdf*, 2(3):8, 2023. 1

[4] JN Cancer Genome Atlas Research Network et al. The cancer genome atlas pan-cancer analysis project. *Nat. Genet*, 45 (10):1113–1120, 2013. 5, 7

[5] Soravit Changpinyo, Piyush Sharma, Nan Ding, and Radu Soricut. Conceptual 12m: Pushing web-scale image-text pre-training to recognize long-tail visual concepts. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3558–3568, 2021. 1

[6] Junsong Chen, Chongjian Ge, Enze Xie, Yue Wu, Lewei Yao, Xiaozhe Ren, Zhongdao Wang, Ping Luo, Huchuan Lu, and Zhenguo Li. Pixart-σ: Weak-to-strong training of diffusion transformer for 4k text-to-image generation, 2024. 3

[7] Richard J Chen, Chengkuan Chen, Yicong Li, Tiffany Y Chen, Andrew D Trister, Rahul G Krishnan, and Faisal Mahmood. Scaling vision transformers to gigapixel images via hierarchical self-supervised learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16144–16155, 2022. 2, 3, 8, 12, 13

[8] Richard J Chen, Tong Ding, Ming Y Lu, Drew FK Williamson, Guillaume Jaume, Bowen Chen, Andrew Zhang, Daniel Shao, Andrew H Song, Muhammad Shaban, et al. A general-purpose self-supervised model for computational pathology. *arXiv preprint arXiv:2308.15474*, 2023. 5, 8

[9] Richard J Chen, Tong Ding, Ming Y Lu, Drew FK Williamson, Guillaume Jaume, Bowen Chen, Andrew Zhang, Daniel Shao, Andrew H Song, Muhammad Shaban, et al. Towards a general-purpose foundation model for computational pathology. *Nature Medicine*, 2024. 2, 13, 18

[10] Hyungjin Chung, Jong Chul Ye, Peyman Milanfar, and Mauricio Delbracio. Prompt-tuning latent diffusion models for inverse problems. In *Forty-first International Conference on Machine Learning*, 2024. 8, 15

[11] Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. *Advances in neural information processing systems*, 34:8780–8794, 2021. 2, 3

[12] Bradley Efron. Tweedie's formula and selection bias. *Journal of the American Statistical Association*, 106(496):1602–1614, 2011. 5

[13] Miguel Espinosa and Elliot J Crowley. Generate your own scotland: Satellite image generation conditioned on maps. *arXiv preprint arXiv:2308.16648*, 2023. 3

[14] Alexandre Filiot, Ridouane Ghermi, Antoine Olivier, Paul Jacob, Lucas Fidon, Alice Mac Kain, Charlie Saillard, and Jean-Baptiste Schiratti. Scaling self-supervised learning for histopathology with masked image modeling. *medRxiv*, pages 2023–07, 2023. 8

[15] Rinon Gal, Yuval Alaluf, Yuval Atzmon, Or Patashnik, Amit Haim Bermano, Gal Chechik, and Daniel Cohen-or. An image is worth one word: Personalizing text-to-image generation using textual inversion. In *The Eleventh International Conference on Learning Representations*, 2023. 8, 15, 16

[16] Alexandros Graikos, Nebojsa Jojic, and Dimitris Samaras. Fast constrained sampling in pre-trained diffusion models. *arXiv preprint arXiv:2410.18804*, 2024. 4, 5, 14, 15

[17] Alexandros Graikos, Srikar Yellapragada, Minh-Quan Le, Saarthak Kapse, Prateek Prasanna, Joel Saltz, and Dimitris Samaras. Learned representation-guided diffusion models for large-image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8532–8542, 2024. 1, 2, 3, 6, 7, 12, 15, 19, 23

[18] Robert Harb, Thomas Pock, and Heimo Müller. Diffusion-based generation of histopathological whole slide images at a gigapixel scale. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 5131–5140, 2024. 2, 3

[19] 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. 6

[20] Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*, 2022. 1, 3, 5, 15

[21] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020. 2, 15

[22] Maximilian Ilse, Jakub Tomczak, and Max Welling. Attention-based deep multiple instance learning. In *International conference on machine learning*, pages 2127–2136. PMLR, 2018. 8

[23] Jakub R Kaczmarzyk, Joel H Saltz, and Peter K Koo. Explainable ai for computational pathology identifies model limitations and tissue biomarkers. *ArXiv*, pages arXiv–2409, 2024. 8

[24] Saarthak Kapse, Pushpak Pati, Srijan Das, Jingwei Zhang, Chao Chen, Maria Vakalopoulou, Joel Saltz, Dimitris Samaras, Rajarsi R Gupta, and Prateek Prasanna. Si-mil: Taming deep mil for self-interpretability in gigapixel histopathology. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11226–11237, 2024. 8

[25] Alina Kuznetsova, Hassan Rom, Neil Alldrin, Jasper Uijlings, Ivan Krasin, Jordi Pont-Tuset, Shahab Kamali, Stefan Popov, Matteo Malloci, Alexander Kolesnikov, Tom Duerig, and Vittorio Ferrari. The open images dataset v4: Unified image classification, object detection, and visual relationship detection at scale. *IJCV*, 2020. 8

[26] Minh-Quan Le, Alexandros Graikos, Srikar Yellapragada, Rajarsi Gupta, Joel Saltz, and Dimitris Samaras. ∞-brush: Controllable large image synthesis with diffusion models in infinite dimensions, 2024. 1, 2, 3, 6, 7, 12, 19, 23

[27] Bin Li, Yin Li, and Kevin W Eliceiri. Dual-stream multiple instance learning network for whole slide image classification with self-supervised contrastive learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 14318–14328, 2021. 2, 5, 8

[28] Ming Y Lu, Bowen Chen, Drew FK Williamson, Richard J Chen, Ivy Liang, Tong Ding, Guillaume Jaume, Igor Odintsov, Long Phi Le, Georg Gerber, et al. A visual-language foundation model for computational pathology. *Nature Medicine*, 30:863–874, 2024. 8

[29] Puria Azadi Moghadam, Sanne Van Dalen, Karina C Martin, Jochen Lennerz, Stephen Yip, Hossein Farahani, and Ali Bashashati. A morphology focused diffusion probabilistic model for synthesis of histopathology images. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 2000–2009, 2023. 3

[30] Gustav Müller-Franzes, Jan Moritz Niehues, Firas Khader, Soroosh Tayebi Arasteh, Christoph Haarburger, Christiane Kuhl, Tianci Wang, Tianyu Han, Teresa Nolte, Sven Nebelung, et al. A multimodal comparison of latent denoising diffusion probabilistic models and generative adversarial networks for medical image synthesis. *Scientific Reports*, 13 (1):12098, 2023. 3

[31] Alexander Quinn Nichol and Prafulla Dhariwal. Improved denoising diffusion probabilistic models. In *International Conference on Machine Learning*, pages 8162–8171. PMLR, 2021. 2, 3

[32] Alexander Quinn Nichol, Prafulla Dhariwal, Aditya Ramesh, Pranav Shyam, Pamela Mishkin, Bob Mcgrew, Ilya Sutskever, and Mark Chen. Glide: Towards photorealistic image generation and editing with text-guided diffusion models. In *International Conference on Machine Learning*, pages 16784–16804. PMLR, 2022. 2

[33] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023. 5, 12

[34] William Peebles and Saining Xie. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4195–4205, 2023. 5, 13

[35] Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. Sdxl: Improving latent diffusion models for high-resolution image synthesis. *arXiv preprint arXiv:2307.01952*, 2023. 3

[36] Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. SDXL: Improving latent diffusion models for high-resolution image synthesis. In *The Twelfth International Conference on Learning Representations*, 2024. 1

[37] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021. 7

[38] Caleb Robinson, Le Hou, Kolya Malkin, Rachel Soobitsky, Jacob Czawlytko, Bistra Dilkina, and Nebojsa Jojic. Large scale high-resolution land cover mapping with multi-resolution data. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 12726–12735, 2019. 12

[39] 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. 1, 3, 5, 7

[40] Chitwan Saharia, Jonathan Ho, William Chan, Tim Salimans, David J Fleet, and Mohammad Norouzi. Image super-resolution via iterative refinement. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(4):4713–4726, 2022. 3

[41] Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, et al. Laion-5b: An open large-scale dataset for training next generation image-text models. *Advances in Neural Information Processing Systems*, 35:25278–25294, 2022. 1

[42] Ahmad Sebaq and Mohamed ElHelw. Rsdiff: Remote sensing image generation from text using diffusion model. *arXiv preprint arXiv:2309.02455*, 2023. 3

[43] Maximilian Seitzer. pytorch-fid: FID Score for PyTorch. https://github.com/mseitzer/pytorch-fid, 2020. Version 0.3.0. 7

[44] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. In *International Conference on Learning Representations*, 2020. 3, 5

[45] USGS. National agriculture imagery program (NAIP), 2023. https://www.usgs.gov/centers/

eros/science/usgs-eros-archive-aerial-photography-national-agriculture-imagery-program-naip. 5, 12

[46] Jianyi Wang, Zongsheng Yue, Shangchen Zhou, Kelvin CK Chan, and Chen Change Loy. Exploiting diffusion prior for real-world image super-resolution. *International Journal of Computer Vision*, pages 1–21, 2024. 17, 18

[47] Xiyue Wang, Sen Yang, Jun Zhang, Minghui Wang, Jing Zhang, Junzhou Huang, Wei Yang, and Xiao Han. Transpath: Transformer-based self-supervised learning for histopathological image classification. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part VIII 24*, pages 186–195. Springer, 2021. 8

[48] Xuan Xu, Saarthak Kapse, Rajarsi Gupta, and Prateek Prasanna. Vit-dae: Transformer-driven diffusion autoencoder for histopathology image analysis. *arXiv preprint arXiv:2304.01053*, 2023. 3

[49] Srikar Yellapragada, Alexandros Graikos, Prateek Prasanna, Tahsin Kurc, Joel Saltz, and Dimitris Samaras. Pathldm: Text conditioned latent diffusion model for histopathology. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 5182–5191, 2024. 2, 3, 6

[50] Zongsheng Yue, Jianyi Wang, and Chen Change Loy. Efficient diffusion model for image restoration by residual shifting. *arXiv preprint arXiv:2403.07319*, 2024. 17, 18

[51] Zongsheng Yue, Jianyi Wang, and Chen Change Loy. Resshift: Efficient diffusion model for image super-resolution by residual shifting. *Advances in Neural Information Processing Systems*, 36, 2024. 17, 18

[52] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models, 2023. 7, 8

[53] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *CVPR*, 2018. 8

ZoomLDM: Latent Diffusion Model for multi-scale image generation

Supplementary Material

We organize the supplementary as follows:

6 ZoomLDM on satellite images
7 Ablation on SSL encoder and Summarizer
8 Experiment details:
8.1 Summerizer-CDM training details
8.2 Joint sampling
8.3 Image inversion
9 Additional Details
9.1 More super-resolution baselines
9.2 Data efficiency and memorization
9.3 Patches from all scales
9.4 Generated large images
9.5 Comparison to previous works

6. ZoomLDM on satellite images

In the main text, we focused on the digital histopathology domain and how our multi-scale diffusion model can prove useful in generation and downstream tasks. However, gigapixel images also concern the remote sensing domain, where satellite images regularly are in the range of 10000×10000 pixels. To show the wide applicability of our multi-scale approach, we trained ZoomLDM on satellite images from the NAIP dataset [45], specifically using NAIP images from the Chesapeake subset of [38]. NAIP images are at 1m resolution – the distance between pixel centers is 1m. We follow the same dataset preparation approach and extract 256×256 patches at four different scales with pixels corresponding to 1m, 2m, 4m, and 8m resolutions. For the SSL encoder, we resort to a pre-trained DINOv2 model [33], which has been known to perform well across many modalities, including satellite.

In Table 5, we provide the per-resolution FID numbers our model achieves. Similarly to histopathology, we observe that training a cross-scale model benefits the scales where there is not enough data to train a single-scale model on (8m resolution in this case). We also showcase patches generated by ZoomLDM at all four resolutions in Figure 13. We present examples from the satellite ZoomLDM variant in 8.2 and 9.3.

In Table 6, we provide the FID numbers for large satellite image generation (1024×1024). Our satellite ZoomLDM model achieves significantly better results on crop FID while achieving similar CLIP FID; this showcases our ability to synthesize high-quality images that simultaneously maintain global consistency.

Resolution	1m	2m	4m	8m
# Training patches	365 k	94 k	25 k	8.7 k
ZoomLDM	**10.93**	**7.77**	**7.34**	**8.46**
SoTA model	11.5 [17]	23.61	37.52	65.45

Table 5. NAIP FID values obtained by ZoomLDM versus training a state-of-the-art diffusion model on a single resolution. Having a shared model across multiple scales improves the generation quality for the data-scarce scales. For resolutions $>$1m we retrain the model of [17] on the samples from that resolution only.

Method	CLIP FID	Crop FID
Graikos et al. [17]	6.86	43.76
∞-Brush [26]	**6.32**	48.65
ZoomLDM	7.90	**13.25**

Table 6. CLIP and Crop FID values (lower is better) for large (1024×1024) satellite images. ZoomLDM outperforms previous works while also maintaining a reasonable inference time.

7. Ablation on SSL encoder and Summarizer

We retrain ZoomLDM with (i) a weaker SSL encoder (HIPT [7]) and (ii) both a weaker SSL encoder and a simpler summarizer network (CNN vs ViT). Table 7 shows that replacing UNI with HIPT degrades performance and further replacing the ViT summarizer network with a simple 4-layer CNN leads to a greater decline.

When comparing the downstream performance of the denoiser features on a multiple-instance learning task (MIL) we also see a decrease in performance when using a 'weaker' conditioning encoder. We believe that training a diffusion model

SSL	Summarizer	FID across magnifications ↓								MIL (AUC) ↑	
		20×	10×	5×	2.5×	1.25×	0.625×	0.3125×	0.15625×	Subtyping	HRD
HIPT [7]	CNN	18.88	16.75	19.31	16.01	14.45	14.21	15.44	18.47	86.20	72.44
HIPT [7]	ViT	13.49	14.42	15.84	13.32	14.32	12.31	16.25	19.90	87.26	75.92
UNI [9]	ViT	**6.77**	**7.60**	**7.98**	**10.73**	**8.74**	**7.99**	**8.34**	**13.42**	**94.49**	**85.25**

Table 7. Ablation on SSL encoder and summarizer network architecture. Using a weaker SSL encoder or summarizer leads to worse performance in both generation and downstream discriminative tasks.

conditioned on SSL representations complements the discriminative SSL pre-training with the newly learned generative features. In all our experiments, improved image quality leads to better downstream task performance. Additionally, the SSL encoders used in MIL are usually trained on a single magnification, making our approach a potential way to fuse features across different scales effectively.

8. Experiment details

8.1. Summarizer-CDM training details

Summarizer: We train the Summarizer jointly with the LDM. The Summarizer processes the SSL embeddings extracted alongside the image patches and projects them to a latent space that is shared across all scales (cross-magnification latent space). By training jointly with the LDM the Summarizer learns to compress the SSL embeddings into a representation useful for making images.

We pre-process the SSL embedding matrices via element-wise normalization. The Summarizer receives 64 SSL embeddings (or fewer SSL embeddings with appropriate padding to 64 tokens) concatenated with a learned magnification embedding as input. The network consists of a 12-layer Transformer encoder with a hidden dimension of 512, followed by a LayerNorm operation to normalize the output. The 65×512 dimensional output is then fed to the U-Net denoiser via cross-attention.

CDM: To avoid reliance on real images to extract the SSL embeddings required for sampling, we train a Conditioning Diffusion Model (CDM). The CDM is trained to draw samples from the learned cross-magnification latent space. After training the LDM and Summarizer jointly, we train the CDM with the denoising objective to sample from the 65×512 output. See Figure 5 for an overview of the Summarizer and CDM.

We implement the CDM as a Diffusion Transformer [34]. We use the DiT-Base architecture, consisting of 12 layers and a hidden size of 768. We use an MLP to project the output back to the exact channel dimensions as the input. We use a constant learning rate of 10^{-4}, following the implementation of [34]. We present samples generated by the CDM in Figure 12.

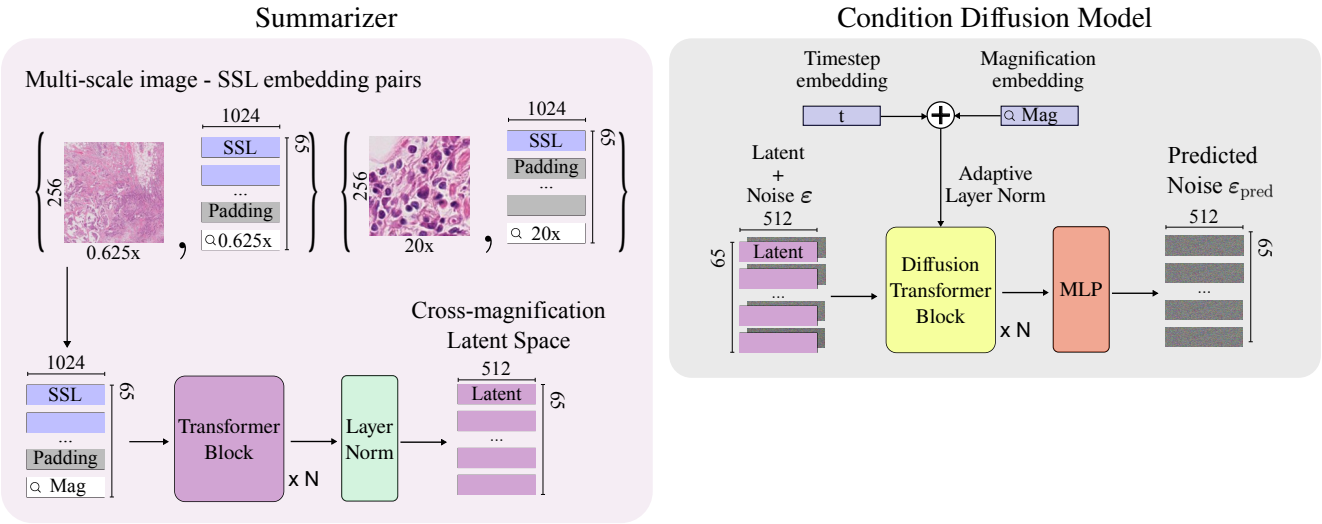

Figure 5. Overview of the Summarizer and Condition Diffusion Model.

8.2. Joint Sampling

In this section, we present an overview of the joint sampling algorithm. By jointly generating an image that depicts the global context and images that produce local details we are able to synthesize large images at the highest resolution that maintain global coherency. We achieve that by simultaneously generating patches i with high-resolution details $\boldsymbol{x}^i = Dec(\boldsymbol{z}^i)$ and a lower-resolution context $\boldsymbol{x}^L = Dec(\boldsymbol{z}^L)$ that globally guides the structure of the patches.

Our joint sampling method is based on a recent fast sampling algorithm for diffusion models under linear constraints, presented in [16]. The full algorithm is shown in Algorithm 1. We make two key changes to the inference algorithm to perform joint multi-scale sampling: (i) We replace the constraint $\boldsymbol{y}$ with the current estimate of the lower-scale image $Dec(\hat{\boldsymbol{z}}_0^L)$ and (ii) we replace the expensive backpropagation step required in computing the error $\boldsymbol{e}$ with a less memory-intensive approximation using forward passes through the encoder and the decoder.

Utilizing intermediate steps Instead of having access to a measurement $\boldsymbol{y}$ we only have access to the current estimate of the context image. That image is in practice a subsampled version of the spatially arranged patches $\boldsymbol{x}^i$. To relate the two, we rearrange $\boldsymbol{x}_i$ and apply a linear subsampling operator $\boldsymbol{A}$, such as bicubic interpolation. This operator is used to compute the difference between the current synthesized patches and the current context and will be used to update the content of the patch images.

Avoiding backpropagation For latent diffusion models, the original algorithm relies on computing the difference between the context and the patches which it then backpropagates through the decoder to get the direction towards which this error is minimized. However, when we synthesize 4k images, we end up with 256 high-resolution patches, and backpropagating becomes prohibitively memory-intensive. To that end, we propose a modification to the sampling algorithm that replaces the backpropagation step with forward passes through the encoder and decoder.

To produce the high-resolution images, we want to sample z_t under the guidance of the lower-scale image, minimizing a constraint $C(\boldsymbol{z}_t) = ||\boldsymbol{A}Dec(\hat{\boldsymbol{z}}_0(\boldsymbol{z}_t)) - Dec(\hat{\boldsymbol{z}}_0^L)||_2^2$. Algorithm 1 requires us to compute the direction $\boldsymbol{e}$ of $\hat{\boldsymbol{z}}_0$ towards which the constraint C is minimized and uses it to update the current diffusion latent as

$$\boldsymbol{g} = \frac{\hat{\boldsymbol{z}}_0(\boldsymbol{z}_t + \delta \boldsymbol{e}) - \hat{\boldsymbol{z}}_0(\boldsymbol{z}_t)}{\delta} \tag{7}$$

$$z_t' = z_t + \lambda g. \tag{8}$$

However, to calculate $\boldsymbol{g}$ we need $\boldsymbol{e} = \frac{\partial C}{\partial \hat{\boldsymbol{z}}_0}$ which we can calculate by backpropagating through the decoder model. Since this is computationally burdensome, we apply the chain rule to get

$$\boldsymbol{e} = \frac{\partial C}{\partial \hat{\boldsymbol{z}}_0} = \left(\frac{\partial Dec(\hat{\boldsymbol{z}}_0)}{\partial \hat{\boldsymbol{z}}_0}\right)^T \frac{\partial C}{\partial Dec(\hat{\boldsymbol{z}}_0)} = \left(\frac{\partial Dec(\hat{\boldsymbol{z}}_0)}{\partial \hat{\boldsymbol{z}}_0}\right)^T \boldsymbol{e}_{img}, \quad \boldsymbol{e}_{img} = \boldsymbol{A}^T(\boldsymbol{A}Dec(\hat{\boldsymbol{z}}_0(\boldsymbol{z}_t)) - Dec(\hat{\boldsymbol{z}}_0^L)) \tag{9}$$

The LDM VAEs that we use (VQ-VAE or KL-VAE) are trained in a way that forces the Jacobian of the Decoder to be approximately orthogonal, through vector quantization or minimizing the KL divergence between the predicted posterior and an isotropic Gaussian. For orthogonal Jacobians Eq. 9 can be simplified into:

$$\boldsymbol{e} = \left(\frac{\partial Dec(\hat{\boldsymbol{z}}_0)}{\partial \hat{\boldsymbol{z}}_0}\right)^T e_{img} \approx \frac{\partial \hat{\boldsymbol{z}}_0}{\partial Dec(\hat{\boldsymbol{z}}_0)} e_{img} \tag{10}$$

and assuming that the VAE has learned to reconstruct images perfectly, it can be written as:

$$\boldsymbol{e} \approx \frac{\partial \hat{\boldsymbol{z}}_0}{\partial Dec(\hat{\boldsymbol{z}}_0)} e_{img} \approx \frac{\partial Enc(Dec(\hat{\boldsymbol{z}}_0))}{\partial Dec(\hat{\boldsymbol{z}}_0)} e_{img}. \tag{11}$$

We can now approximate $\boldsymbol{e}$ using finite differences:

$$\boldsymbol{e} \approx \frac{\partial Enc(Dec(\hat{\boldsymbol{z}}_0))}{\partial Dec(\hat{\boldsymbol{z}}_0)} e_{img} \approx \frac{Enc(Dec(\hat{\boldsymbol{z}}_0) + \zeta e_{img}) - Enc(Dec(\hat{\boldsymbol{z}}_0))}{\zeta} \tag{12}$$

which completely erases the need to perform memory-heavy backpropagation through the decoder model.

A step-by-step description of our joint sampling method can be found in Algorithm 2. We use 50 DDIM steps for our experiments, bicubic upsampling/downsampling for $\boldsymbol{A}$, $\delta = \zeta = 0.005$, $K = 1$, $\lambda = 0.5$. Upon observing noticeable discontinuities along the borders of the high-resolution patches, we apply a simple post-processing step by adding noise and

denoising the patches between, similar to [17]. We provide some results of the joint sampling, visualized in Figures 6,7 for the histopathology and satellite domains.

Algorithm 1 The algorithm for linear inverse problem solving proposed in [16].

Input: Diffusion model $\hat{\boldsymbol{z}}_0(\boldsymbol{z}_t)$, Enc, Dec, schedule $T_{0,\dots,M}$, subsampling operator $\boldsymbol{A}$, measurement $\boldsymbol{y}$, step size δ, # iterations K, learning rate λ
$\boldsymbol{z}_T \sim N(\mathbf{0}, \boldsymbol{I})$
for $t \in \{T_0, T_1, \dots, T_M\}$ **do**
  for $i \in \{1, 2, \dots, K\}$ **do**
    $\boldsymbol{e} = \nabla_{\boldsymbol{z}_0} ||\boldsymbol{A}Dec(\hat{\boldsymbol{z}}_0(\boldsymbol{z}_t)) - \boldsymbol{y}||_2^2$
    $\boldsymbol{g} = [\hat{\boldsymbol{z}}_0(\boldsymbol{z}_t + \delta \boldsymbol{e}) - \hat{\boldsymbol{z}}_0(\boldsymbol{z}_t)]/\delta$
    $\boldsymbol{z}_t = \boldsymbol{z}_t + \lambda \boldsymbol{g}$
  end for
  $\boldsymbol{z}_t = \text{DDIM}(\boldsymbol{z}_t, \hat{\boldsymbol{x}}_0, s)$
end for
Return: $\boldsymbol{x}_0$

Algorithm 2 The proposed modification to Algorithm 1.

Input: Diffusion model $\hat{\boldsymbol{z}}_0(\boldsymbol{z}_t)$, Enc, Dec, schedule $T_{0,\dots,M}$, subsampling operator $\boldsymbol{A}$, detail scale s, context scale s_L, step sizes δ, ζ, # iterations K, learning rate λ
$\boldsymbol{z}_T \sim N(\mathbf{0}, \boldsymbol{I})$
$\boldsymbol{z}_T^L \sim N(\mathbf{0}, \boldsymbol{I})$
for $t \in \{T_0, T_1, \dots, T_M\}$ **do**
  for $i \in \{1, 2, \dots, K\}$ **do**
    $\boldsymbol{e}_{img} = \boldsymbol{A}^T(\boldsymbol{A}Dec(\hat{\boldsymbol{z}}_0(\boldsymbol{z}_t)) - Dec(\hat{\boldsymbol{z}}_0^L))$
    $\boldsymbol{e} = [Enc(Dec(\hat{\boldsymbol{z}}_0) + \zeta \boldsymbol{e}_{img}) - Enc(Dec(\hat{\boldsymbol{z}}_0))]/\zeta$
    $\boldsymbol{g} = [\hat{\boldsymbol{z}}_0(\boldsymbol{z}_t + \delta \boldsymbol{e}) - \hat{\boldsymbol{z}}_0(\boldsymbol{z}_t)]/\delta$
    $\boldsymbol{z}_t = \boldsymbol{z}_t + \lambda \boldsymbol{g}$
  end for
  $\boldsymbol{z}_t = \text{DDIM}(\boldsymbol{z}_t, \hat{\boldsymbol{x}}_0, s)$
  $\boldsymbol{z}_t^L = \text{DDIM}(\boldsymbol{z}_t^L, \hat{\boldsymbol{x}}_0, s_L)$
end for
Return: $\boldsymbol{x}_0$

8.3. Image Inversion

In this section, we present our image inversion algorithm, which is crucial for performing the super-resolution task described in the main text. The conditioning we provide to the model is a set of SSL embeddings extracted at the highest resolution available. For instance, in histopathology, the SSL conditions are extracted at 20×. Thus, when we are given a single image at any magnification that we want to super-resolve we do not have access to this conditioning and are limited to using the model in an unconditional manner. The unconditional model is available since we randomly drop the conditioning during training, to implement classifier-free guidance [20] during sampling. However, recent works have argued that when using the diffusion model to sample with linear constraints, like super-resolution, conditioning helps in achieving better-fidelity results [10].

Inspired by those findings, we propose a simple algorithm to first *invert* the model and get conditioning for a single image, before super-resolving it. The algorithm is an adaptation of the textual inversion technique of Gal et al. [15], which has seen wide success in text-to-image diffusion models. An overview of the approach is provided in Figure 8.

Given an image $\boldsymbol{I}$ at scale s, we have access to a pre-trained latent denoiser model $\boldsymbol{\epsilon}_\theta(\boldsymbol{z}_t, t, f(\boldsymbol{e}, s))$ where $\boldsymbol{z} = Enc(\boldsymbol{I})$, g is the summarizer model and $\boldsymbol{e}$ are the SSL embeddings that describe the image. We want to draw a sample $\boldsymbol{e}$, that when provided as conditioning to the diffusion model will generate images similar to $\boldsymbol{I}$. From the latent variable perspective of diffusion models, described by Ho et al. [21], we obtain the following lower bound for the log probability of $\boldsymbol{z}$ given a condition $\boldsymbol{e}$

$$\log p(\boldsymbol{z} \mid \boldsymbol{e}) \geq -\sum_{t=1}^{T} w_t(\alpha) \mathbb{E}_{\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I})} \left[||\boldsymbol{\epsilon}_{\boldsymbol{\theta}}(\boldsymbol{z}_t, t, g(\boldsymbol{e}, s)) - \boldsymbol{\epsilon}||_2^2 \right], \quad \boldsymbol{z}_t = \sqrt{\alpha_t}\boldsymbol{z} + \sqrt{1-\alpha}\boldsymbol{\epsilon}. \tag{13}$$

We then employ variational inference to fit an approximate posterior $q(\boldsymbol{e})$ to $p(\boldsymbol{e} \mid \boldsymbol{z})$ from which we want to sample conditions given an input image. We start by defining a lower bound for $\log p(\boldsymbol{z})$

$$\begin{aligned} \log p(\boldsymbol{z}) &= \log \int_{\boldsymbol{e}} p(\boldsymbol{z}, \boldsymbol{e}) d\boldsymbol{e} = \log \int_{\boldsymbol{e}} q(\boldsymbol{e}) \frac{p(\boldsymbol{z}, \boldsymbol{e})}{q(\boldsymbol{e})} d\boldsymbol{e} \\ &= \log \mathbb{E}_{q(\boldsymbol{e})} \left[\frac{p(\boldsymbol{z}, \boldsymbol{e})}{q(\boldsymbol{e})} \right] \geq \mathbb{E}_{q(\boldsymbol{e})} \left[\log \frac{p(\boldsymbol{z}, \boldsymbol{e})}{q(\boldsymbol{e})} \right] \\ &= \mathbb{E}_{q(\boldsymbol{e})} \left[\log \frac{p(\boldsymbol{z} \mid \boldsymbol{e}) p(\boldsymbol{e})}{q(\boldsymbol{e})} \right] = L. \end{aligned} \tag{14}$$

By maximizing the bound L w.r.t. the parameters of q we minimize the KL-Divergence between the approximate posterior $q(\boldsymbol{e})$ and the real $p(\boldsymbol{e} \mid \boldsymbol{z})$. We choose a simple Dirac delta $q(\boldsymbol{e}) = \delta(\boldsymbol{e} - \boldsymbol{u})$ as our approximation, which allows us to use the bound from Eq. 13 to simplify the objective

$$
\begin{aligned}
L &= \mathbb{E}_{q(\boldsymbol{e})}\left[\log p(\boldsymbol{z} \mid \boldsymbol{e}) + \log p(\boldsymbol{e}) - \log q(\boldsymbol{e})\right] = \log p(\boldsymbol{z} \mid \boldsymbol{e} = \boldsymbol{u}) + \log p(\boldsymbol{e} = \boldsymbol{u}) \\
&= -\sum_{t=1}^{T} w_t(\alpha) \mathbb{E}_{\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I})}\left[||\boldsymbol{\epsilon}_{\boldsymbol{\theta}}(\boldsymbol{z}_t, t, g(\boldsymbol{u}, s)) - \boldsymbol{\epsilon}||_2^2\right] + \log p(\boldsymbol{e} = \boldsymbol{u}). \qquad (15)
\end{aligned}
$$

Therefore, to draw a sample from the posterior $p(\boldsymbol{e} \mid \boldsymbol{z})$ we optimize Eq. 15 w.r.t. $\boldsymbol{u}$. The result is a single point $\boldsymbol{u}$ that seeks a local mode of $p(\boldsymbol{e} \mid \boldsymbol{z})$.

For the prior term $\log p(\boldsymbol{e})$, we use a simple heuristic, implementing a penalty that maximizes the similarity between the different vectors in the SSL embeddings $\boldsymbol{e}$. This heuristic encourages the model to find embeddings that generate similar patches when used independently. For the denoising terms, we must add random Gaussian noise to the image latent $\boldsymbol{z}$ and denoise at multiple timesteps t. Instead of evaluating multiple timesteps simultaneously, we utilize an annealing schedule that starts from $t = 950$ and linearly decreases to $t = 50$ over the $n = 200$ optimization steps we perform. Overall, the proposed algorithm is similar to textual inversion [15], which utilizes the denoising loss to optimize text tokens $\boldsymbol{t}$.

In Figure 9, we provide qualitative results for our inversion approach. We present two cases, inferring the condition for $5\times$ and $2.5\times$ images. We observe that for $5\times$, which is also the scale used in our super-resolution experiments, our approach can provide conditions that faithfully reconstruct both the $5\times$ image and also give us plausible $20\times$ patches. As we increase

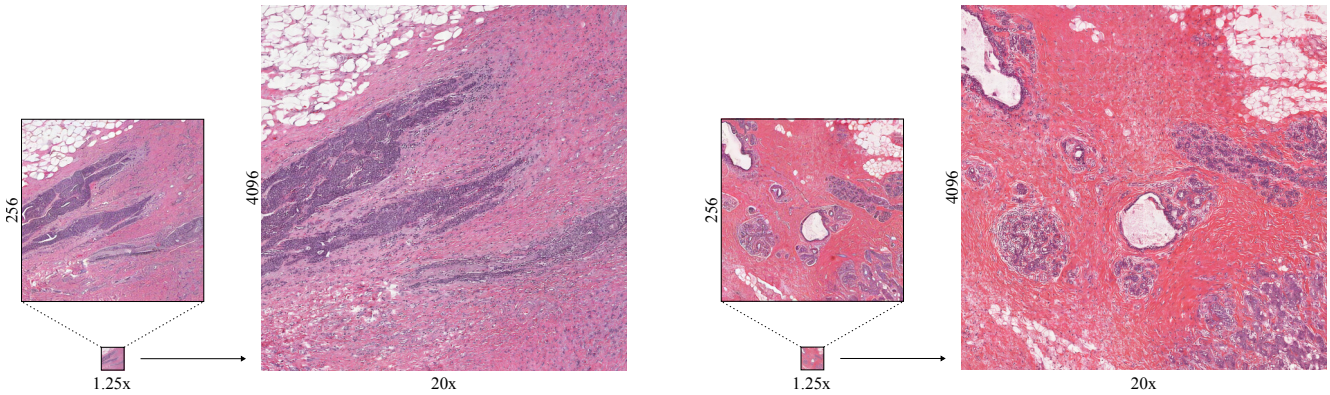

Figure 6. Joint sampling process across two different magnifications for the TCGA-BRCA ZoomLDM model. We jointly generate a 256×256 image at $1.25\times$ and a 4096×4096 image at $20\times$. The $1.25\times$ generation guides the structure of the $20\times$ image by providing the necessary global context that each $20\times$ patch is unaware of. The generated large $20\times$ image has a realistic global arrangement of cells and tissue. Best viewed zoomed-in.

Figure 7. Joint sampling process across two different resolutions for the Satellite ZoomLDM model. We jointly generate a 256×256 image at $8m$ resolution and a 2048×2048 image at $1m$. The $8m$ generation guides the structure of the $1m$ image by providing global coherence, which, otherwise, each $1m$ would be unaware of. The generated large $1m$ image has realistic global structures, with roads and forests neatly arranged across the 2048×2048 canvas. Best viewed zoomed-in.

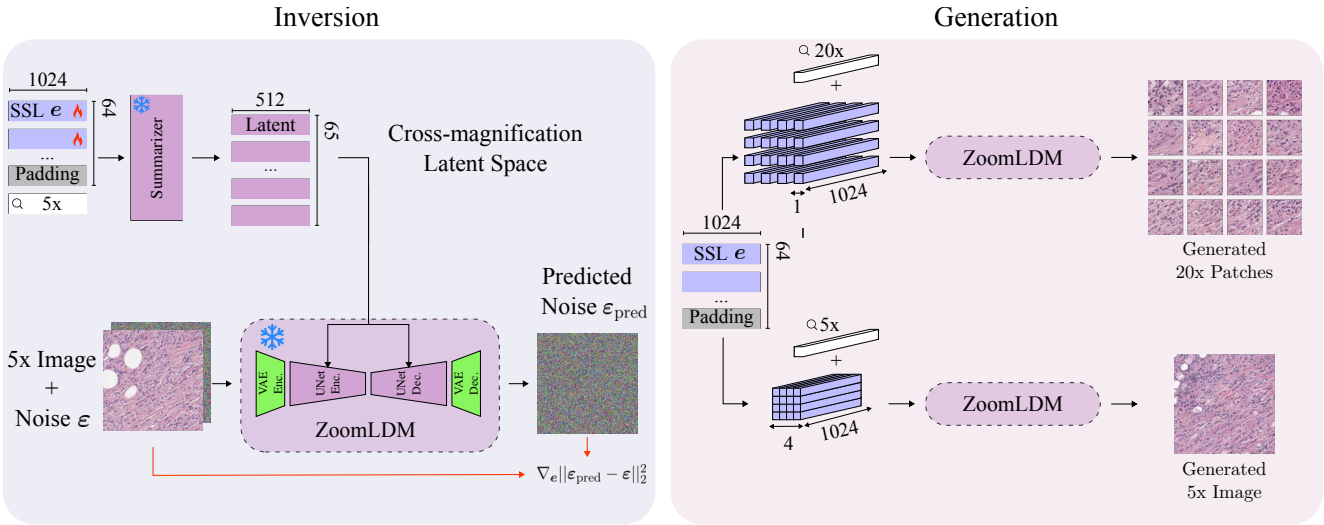

Figure 8. Figure illustrating our pipeline for the image inversion used in the super-resolution task. For a given image we first use the denoising loss to optimize the input, conditioning embeddings. We can then generate variations of the given image and high-resolution patches from it. We use those per-patch embeddings to perform super-resolution, obtaining better results than unconditional super-resolution.

the number of conditions to infer, the $2.5\times$ result remains convincing at the lower scale but struggles to provide reasonable $20\times$ patches. Future work focusing on this inversion approach could provide useful insights into the SSL embeddings used as conditioning, helping understand what they encode and the topology of the latent space created by the SSL encoder.

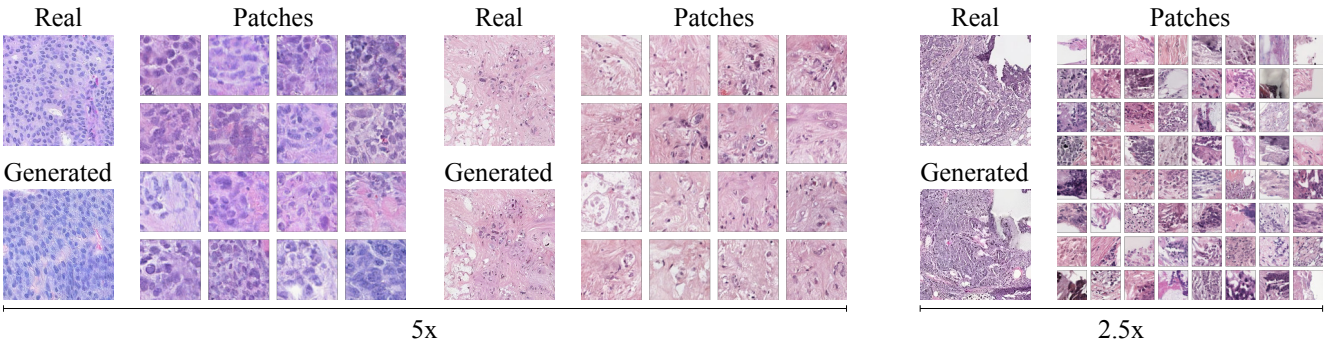

Figure 9. Examples of the image inversion algorithm. Given a real image at any magnification, we infer the SSL embeddings that generated it. We then generate a new, similar-looking image at the same magnification using those embeddings as conditioning. Using the inferred embeddings to generate single patches from the given image yields convincing results at magnifications $> 5\times$.

9. Additional results

9.1. More super-resolution baselines

In Tables 8 and 9 we provide additional baselines for the super-resolution task. We use ResShift [50, 51] and StableSR [46] to super-resolve pathology images and compare them to the zero-shot performance of ZoomLDM. Using ZoomLDM in a training-free manner (with condition inference 8.3) remains the best approach for histopathology image super-resolution.

9.2. Data efficiency and memorization

One of the arguments for training a single model for all scales is that we can learn to generate novel images even at scales with too few samples to learn from. To further demonstrate this, we use our histopathology diffusion model and sample conditions from the Conditioning Diffusion Model (CDM) to generate novel images at $0.15625\times$ magnification. At this scale, both our models have only seen ~ 2500 images and we would expect them to either generate low-quality samples or to

Table 8. Super-resolution results on TCGA-BRCA

Method	SSIM ↑	PSNR ↑	LPIPS↓	CONCH ↑	UNI ↑
ResShift v2 (15 steps) [51]	0.415	19.716	0.431	0.847	0.299
ResShift v3 (4 steps) [50]	0.525	21.528	0.314	0.866	0.311
StableSR no tiling [46]	0.515	21.644	0.315	0.862	0.390
StableSR w/ tiling [46]	0.514	21.618	0.316	0.863	0.388
ZoomLDM (Uncond)	0.591	23.217	0.260	0.936	0.680
ZoomLDM (GT Emb)	0.599	23.273	0.250	0.946	0.672
ZoomLDM (Infer Emb)	0.609	23.407	**0.229**	**0.957**	**0.719**

Table 9. Super-resolution results on BACH

Method	SSIM ↑	PSNR ↑	LPIPS↓	CONCH ↑	UNI ↑
ResShift v2 (15 steps) [51]	0.584	23.256	0.421	0.898	0.621
ResShift v3 (4 steps) [50]	0.751	26.283	0.257	0.898	0.623
StableSR no tiling [46]	0.729	26.203	0.291	0.846	0.547
StableSR w/ tiling [46]	0.729	26.200	0.293	0.845	0.538
ZoomLDM (Uncond)	0.739	29.822	0.235	0.965	0.741
ZoomLDM (GT Emb)	0.732	29.236	0.245	0.974	0.753
ZoomLDM (Infer Emb)	0.779	30.443	**0.173**	**0.974**	0.808

have memorized the training data when using a 400M parameter model in training. Contrary to that, in Figure 10, we show that the generated images are realistic and different from the ones found in the training set. For each generated image, we identify its nearest neighbor in the training data using the patch-level UNI embeddings [9], and show that they differ in shape and content. ZoomLDM can produce high-quality and unique samples for data-scarce magnifications, essentially avoiding memorization, by learning to synthesize images at all scales.

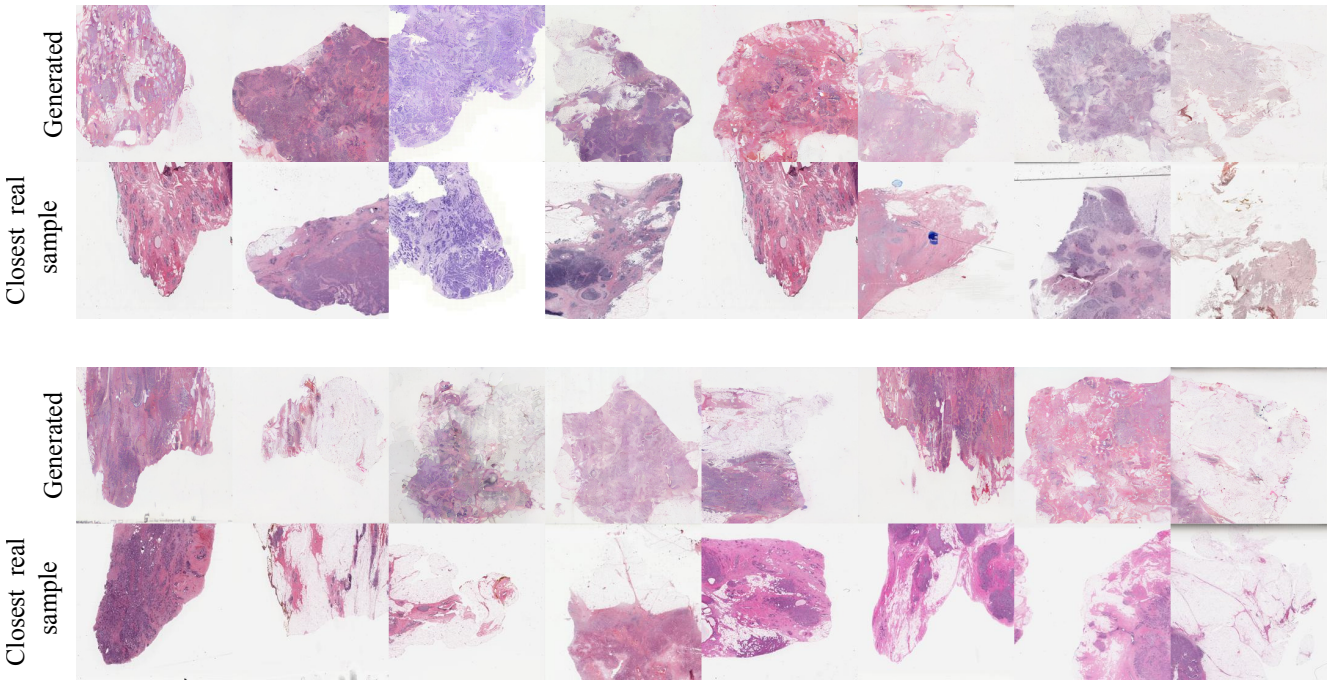

Figure 10. We present $0.15625\times$ images generated from our model and their nearest neighbors in the training dataset. Although only trained on ~ 2500 images, our 400M parameter model did not memorize the training samples and successfully synthesized novel images at that magnification.

9.3. Patches from all scales

In Figures 11 and 13, we showcase synthetic samples from ZoomLDM and the real images used to extract embeddings in histopathology and satellite. Samples from our model are realistic and preserve semantic features found in the reference patches. In data-scarce scenarios, such as $0.15625x$ magnification, achieving comparable image quality would be infeasible for a standalone model trained solely on that magnification (as indicated by the FIDs in Table 1 of the main text).

Interestingly, for magnifications below $5\times$ we find that the model can almost perfectly replicate the source image since the SSL embeddings used as conditioning contain enough information to reconstruct the patch at that scale perfectly. Although this may seem like a memorization issue, our experiments with the CDM in 9.2 show that our model has not just memorized the SSL embedding and image pairs. We believe that for these domains, this faithfulness to the conditions is advantageous as it can limit the hallucinations of the model, which are mostly unwanted in domains such as medical images.

9.4. Large images

In Figures 14,15 we present 4096×4096 px images generated from our histopathology and satellite ZoomLDM model. Readers can find more examples on histodiffusion.github.io/docs/projects/zoomldm.

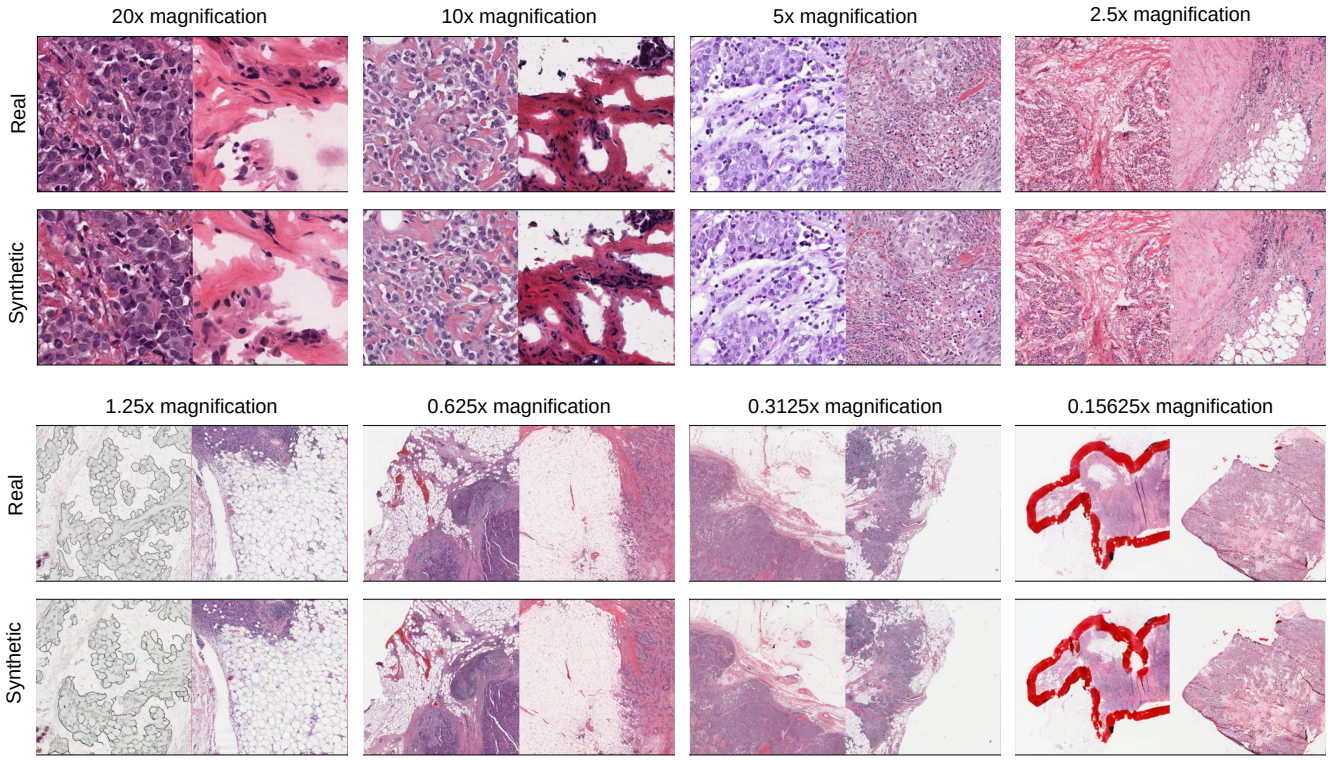

Figure 11. Synthetic patches (256×256 pixel) generated by ZoomLDM juxtaposed with the corresponding real images from TCGA-BRCA. Across all magnifications, ZoomLDM preserves the semantic features of the reference patches.

9.5. Comparison to previous works

In Figure 16, we compare our method and previous works on a single example image. We extract SSL embeddings from the 4k to replicate this image as closely as possible. We highlight our two main differences with previous methods. The method of $\infty-$Brush [26] retains some global structures but fails to produce any high-resolution details in the image. On the other hand, the patch-based model of [17] produces high-quality details but fails to capture large-scale structures that span more than a single patch. Our method solves both issues at the same time while maintaining a reasonable inference time, as discussed in the main text. We provide further comparisons to $\infty - Brush$ in Figure 17. Our generated images contain noticeably better detail.

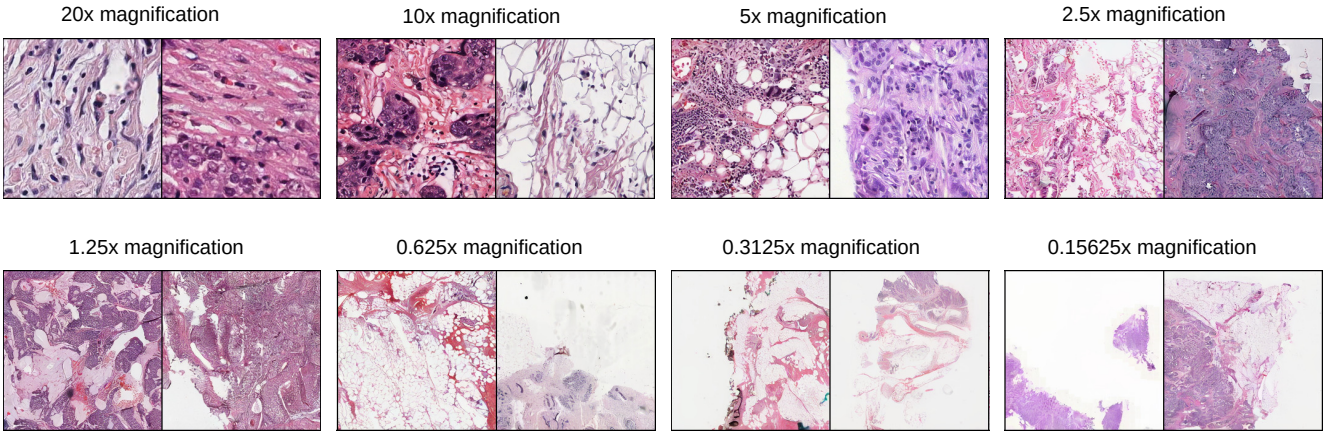

Figure 12. Images synthesized by ZoomLDM using conditions sampled from our Conditioning Diffusion model (CDM).

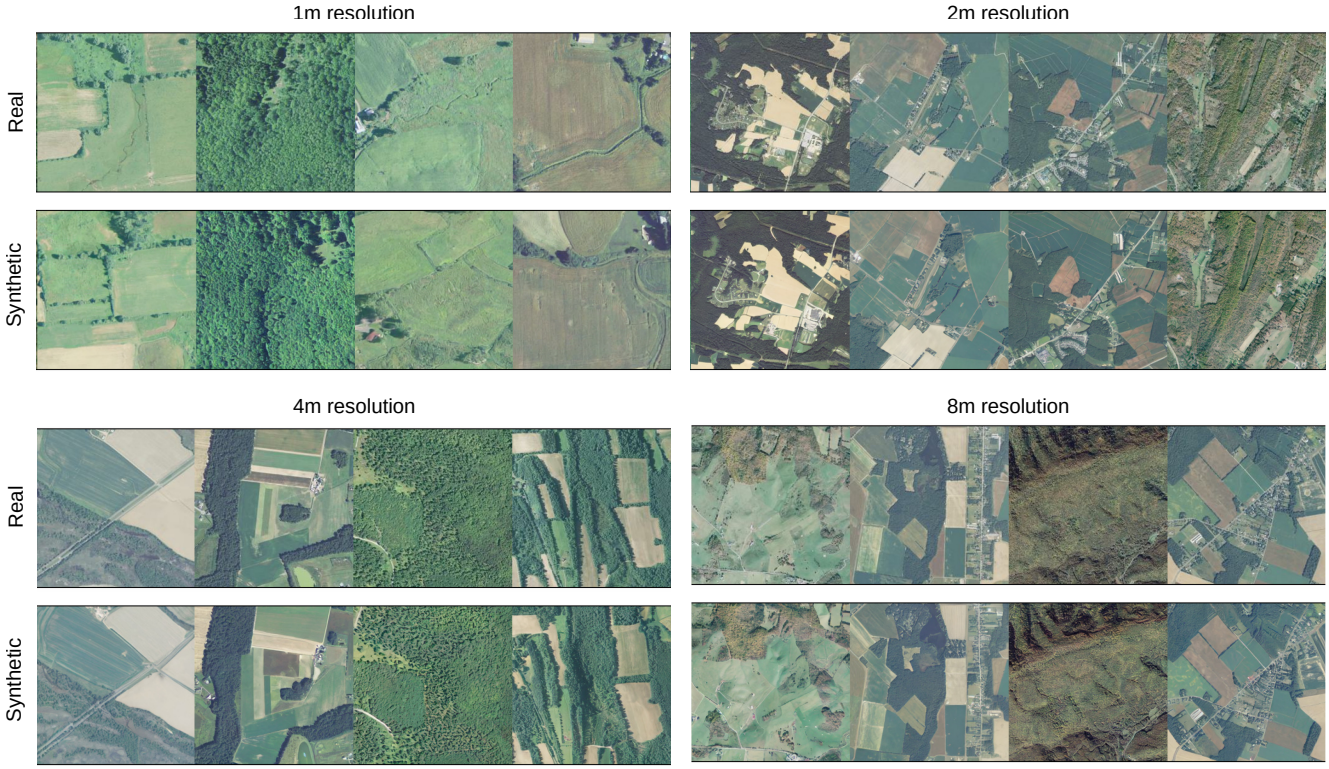

Figure 13. Synthetic patches (256×256 pixel) generated by ZoomLDM juxtaposed with the corresponding real images from NAIP

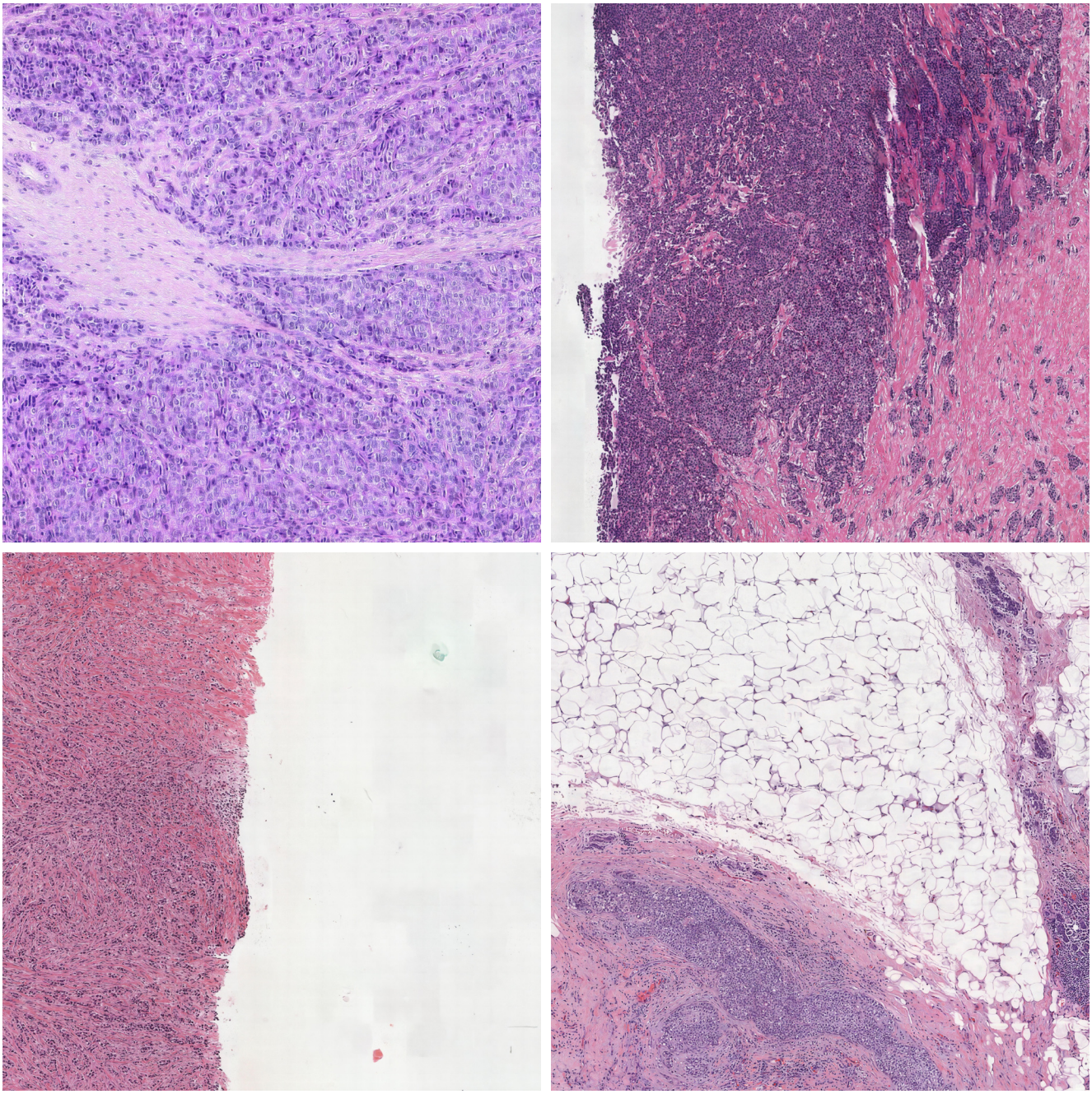

Figure 14. We present 4096×4096 images generated from our histopathology model. Our results exhibit correct global structures in terms of the arrangement of cells and tissue while also maintaining high-resolution details. We point out two weaknesses: The local model fails to maintain coherency for structures where the lower-scale image does not provide guidance, such as the thin structures in the bottom-right image. In addition, for large uniform areas, such as the background in the bottom left image, the 'stitching' of the generated $20\times$ patches is visible with noticeable discontinuities along their edges.

Figure 15. We present 4096×4096 images generated from our satellite model. The results demonstrate images with reasonable global structures that also maintain high-resolution features. A similar weakness to the pathology images is visible, with slight discontinuities among the high-resolution patch borders.

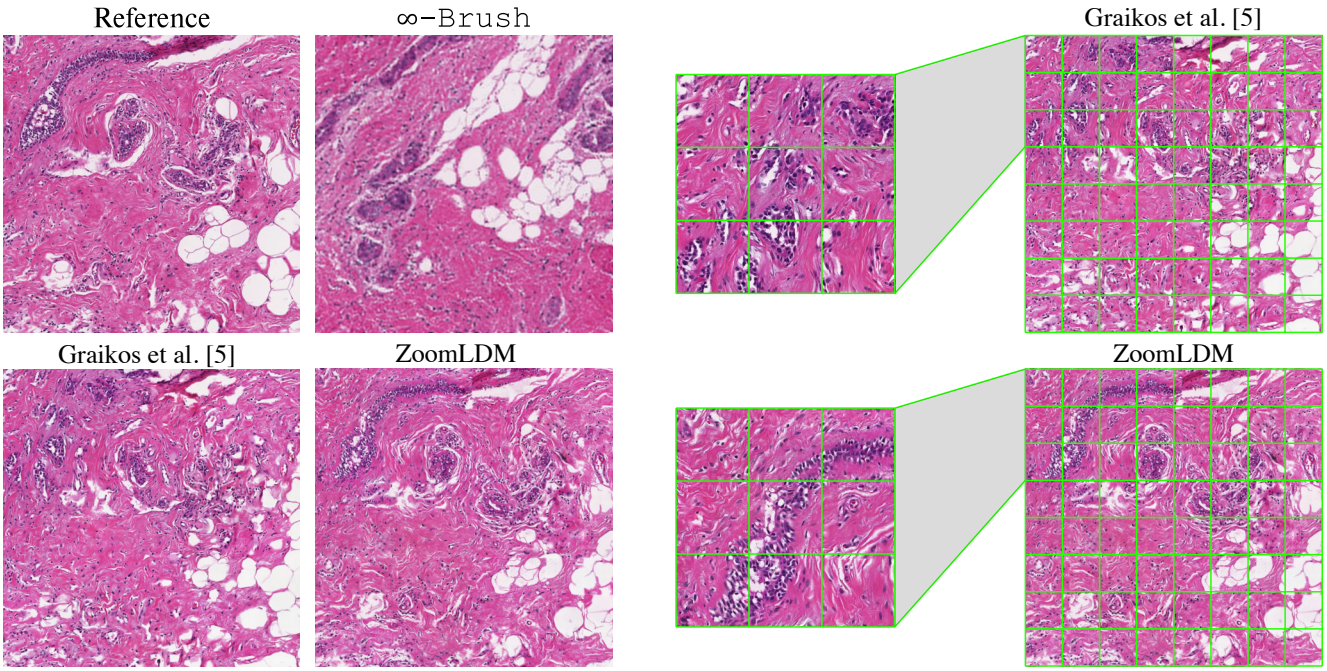

Figure 16. We compare with two recent previous methods that also generated large histopathology images. In this example, we compare a 2048×2048 image from $\infty -$ Brush and [17] to the same image generated from our model. We exceed both previous methods, with $\infty -$ Brush producing realistic global context but blurry details and [17] completely failing to capture larger scale structures.

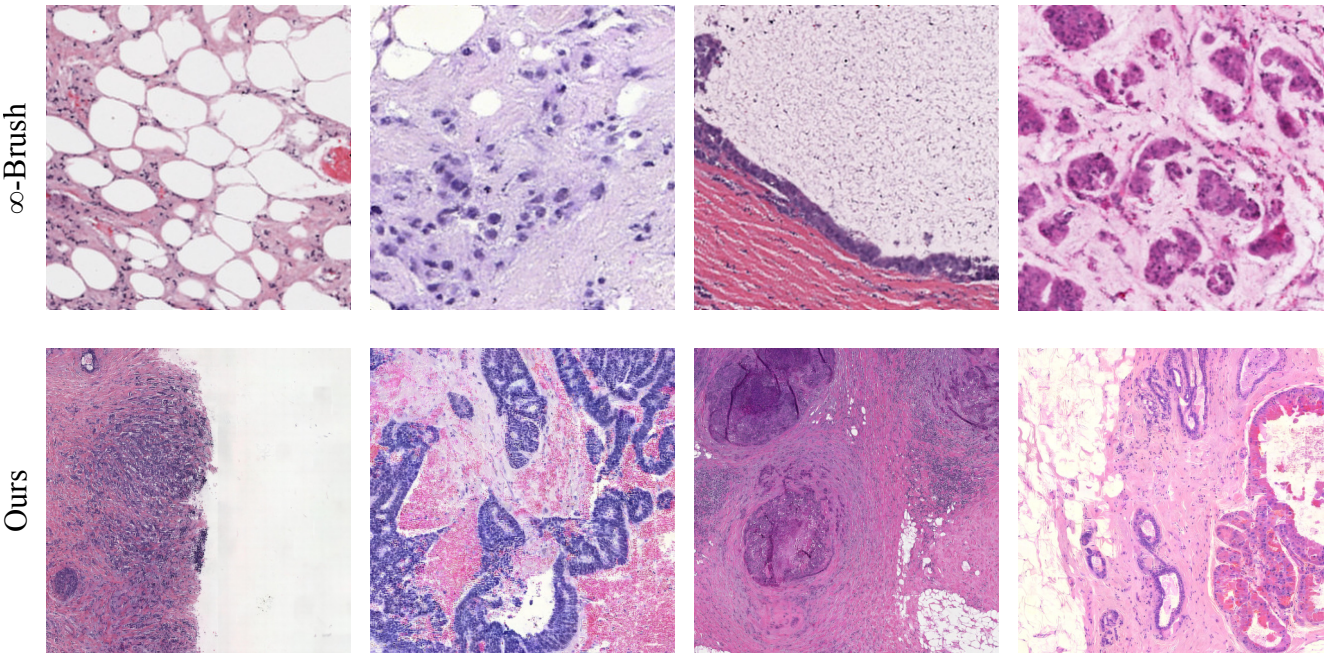

Figure 17. Comparison between $\infty -$ Brush [26] and our method.