ZoomLDM: Latent Diffusion Model for multi-scale conditional histopathology image generation

Srikar Yellapragada* Alexandros Graikos* Prateek Prasanna Rajarsi Gupta
Joel Saltz Dimitris Samaras
Stony Brook University

Abstract

Diffusion models have revolutionized image generation, yet several challenges restrict their application in digital pathology. Existing approaches struggle with precise control over the appearance of image patches, are typically confined to generating images at a single magnification level, and face limitations due to the lack of large annotated datasets which have been essential to the success of diffusion models in the natural image domain. Self-supervised encoders have emerged as a viable workaround for the annotation deficit, yet their effective application across varying magnifications presents its own set of challenges. We present ZoomLDM, a latent diffusion model tailored for generating histopathology images across a spectrum of magnifications. Central to our approach is a novel magnification-aware conditioning mechanism, that leverages a patch summarization CNN, jointly trained with the diffusion model. This CNN processes self-supervised embeddings, retaining vital information and enabling controllable image generation. ZoomLDM achieves state-of-the-art image generation quality, especially in lower magnifications where there is limited data availability. Furthermore, we introduce a second model to generate the embeddings consumed by the conditioning mechanism, eliminating the dependency on recycling existing embeddings and facilitating the generation of novel images.

1. Introduction

Diffusion models have achieved remarkable success in diverse generative tasks, from photorealistic image synthesis to audio generation [1, 13]. The availability of vast multimodal datasets [2, 20] and sophisticated conditioning techniques [9, 18] has undoubtedly fueled this progress. Latent Diffusion models (LDMs) [19] have further advanced high-resolution image generation by introducing a two-step process of encoding the image and diffusion within the latent space. In the natural image domain, foundation models like Stable Diffusion XL [18] demonstrate the potential of diffusion models for use in downstream tasks like image segmentation and classification [11, 22].

Histopathology is an active area of diffusion model development, with previous works ranging from pixel-level diffusion model for gliomas [14], to class conditional and text-conditioned LDM for various cancers [15, 24]. Despite progress, granular control over patch appearance remains challenging due to the difficulty of obtaining patch-level annotations. Annotating the entire TCGA-BRCA dataset with descriptive captions would take over 40,000 hours of expert pathologist time [6].

Additionally, diffusion models for histopathology images often restrict themselves to a single magnification level. Addressing the complete magnification spectrum is technically challenging, as the nature and scale of diagnostically relevant features vary drastically across magnifications. For instance, low magnification highlights invasive tumor edges, medium power focuses on tumor cell arrangement, and higher magnification reveals the texture and size of nuclei. In this regard, obtaining annotations becomes even more challenging because one needs patch-level annotations not just at one but for all magnifications. While [7] proposed a pixel-level diffusion model accommodating multiple magnifications, it lacks conditioning, which is necessary for better image quality and also vital in performing downstream tasks [4, 16, 24]. Self-supervised learning (SSL) emerges as a viable alternative, offering rich semantic and visual encodings suitable for downstream tasks [3, 5]. Graikos *et al*. [6] condition LDMs on SSL embeddings, reducing the need for fine-grained human labels.

While SSL holds promise, its application in this context is nontrivial, as SSL encoders are often trained on patches from a single magnification (ex HIPT [3] on $20\times$ or CTransPath [23] on $5\times$). Furthermore, the decreasing number of patches available at lower magnifications limits the possibility of training separate models for each scale. A naïve approach using an SSL encoder trained on $20\times$ patches on the full-resolution regions faces diffi-

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

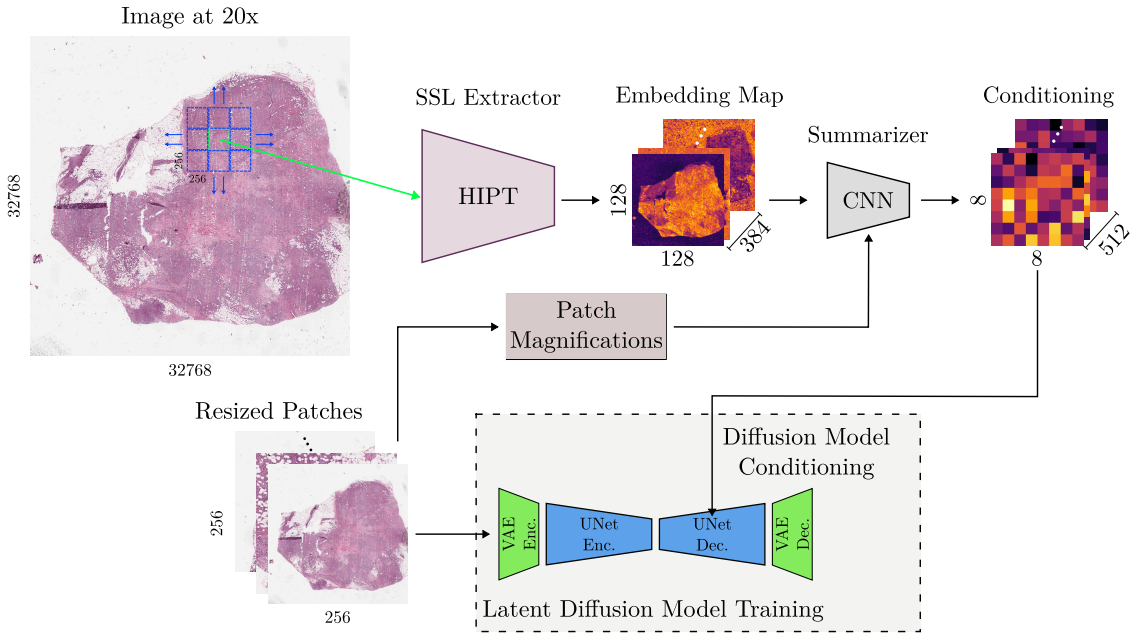

Figure 1. We extract 256×256 patches and the corresponding $20\times$ regions across all magnifications. Following this, we apply HIPT [3] on the $20\times$ region in a patch-wise manner to obtain an embedding matrix. Our CNN processes this matrix, producing an output vector of size $8 \times 8 \times 512$, which then conditions the diffusion model. The CNN is designed to be magnification-aware, enabling it to identify the appropriate level of detail required at different scales.

culties, due to the exponential growth in embedding size. For instance, while a single 256×256 patch at $20\times$ magnification yields a straightforward N-dimensional embedding, a same-sized patch at $10\times$ magnification translates into a $2 \times 2 \times N$ embedding matrix, as it corresponds to a larger 512×512 region at $20\times$. Following such progressive growth, a 256×256 patch at the lowest magnification of $0.15625\times$ results in a vast $128\times128\times N$ embedding matrix. These large embeddings are computationally prohibitive for conditioning and likely contain redundant information.

To address the above challenges, we introduce **ZoomLDM**, the first conditional latent diffusion model capable of producing image patches across the entire magnification spectrum. Our model incorporates a novel magnification-aware conditioning mechanism, consisting of a summarization CNN jointly trained with the diffusion model to efficiently process large embedding matrices. The CNN effectively compresses these matrices by up to 45 times yet retains vital information necessary for generating high-quality images.

We train ZoomLDM on patches from eight different magnifications ($20\times$, $10\times$, ..., $0.15625\times$) within the TCGA-BRCA dataset. Our image generation is comparable to the current state-of-the-art (SoTA) at higher magnifications (e.g., $20\times$ - $10\times$) [6, 24] while achieving impressive generation results at the data-scarce lower magnifications.

Since SSL embeddings are currently the only option for conditioning, our model initially lacks the flexibility to devise new conditions for synthesizing novel images, as one would for text-based generation. To address this, we also introduce an embedding diffusion model (EDM) that is trained to sample conditions at all magnifications. This approach enables high-quality image generation without the need for a database of SSL embeddings. Moreover, we argue that this added step model can be potentially guided by auxiliary information, such as slide-level pathology reports and RNA sequences for more versatile image generation.

Our contributions are the following:

- We present **ZoomLDM**, the first multi-scale conditional latent diffusion model for histopathology .
- Our novel magnification-aware conditioning mechanism leverages SSL embeddings to extract crucial information and guide generation.
- We achieve state-of-the-art image quality in histopathology image generation, particularly in settings with limited data (lower magnifications).
- We introduce a versatile embedding diffusion model that enables synthesizing novel histopathology images at different magnifications.

2. Method

We train ZoomLDM on image patches from all magnifications. Furthermore, we introduce a conditioning mechanism that processes the large SSL embedding matrix and outputs

a low-dimensional, compact representation. Figure 1 provides an overview of our method.

2.1. Unified conditional training

We begin by extracting image patches from all magnifications in a WSI, aiming to build a unified diffusion model capable of handling images across various scales. Given the labor-intensive nature of annotating histopathological images, especially across different magnifications, we leverage Self-Supervised Learning (SSL) encoders such as HIPT [3] to provide the necessary conditioning information for learning to synthesize these patches.

As previously explained, applying the HIPT encoder directly to patches from various magnifications introduces significant challenges due to the exponential growth in embedding matrix sizes. Utilizing such large embeddings directly for conditioning in image generation models poses computational difficulties, primarily due to the quadratic complexity of cross-attention mechanisms. Moreover, generating a 256×256 patch at a $0.15625\times$ magnification from a full $128 \times 128 \times N$ embedding matrix introduces redundancy. To overcome these obstacles, we propose a magnification-aware conditioning approach. This method involves training jointly with the diffusion model a summarization convolutional neural network (CNN) tailored to handle these extensive embedding matrices.

We structure the CNN architecture into four residual blocks, with each block comprising a convolutional layer (Conv), batch normalization (Batch Norm), and max pooling (Max pool) (see Supplementary for details) . The CNN is designed to compress inputs of size $128 \times 128 \times 384$ into $8 \times 8 \times 512$. To incorporate the magnification factor of the current patch into the CNN, we adopt learnable Feature-wise Linear Modulation (FiLM) [17] parameters—scale and shift—based on the magnification embedding. This design choice enables the CNN to be magnification-aware, allowing it to adapt to the appropriate level of detail required at different scales.

2.2. Embedding generation

Our image synthesis pipeline requires providing conditioning at the desired magnification. This entails extracting the necessary conditioning from a set of reference patches, which can be impractical when there is no direct access to the training data. To that end, we train a second diffusion model that learns the distribution of the conditions we provide to the Latent Diffusion Model.

The summarization CNN network transforms a map of SSL embeddings, extracted from a $20\times$ patch, into a latent representation to be consumed by the diffusion model. We choose to directly learn the distribution of the CNN output, instead of the spatial distribution of SSL embeddings, as it is less complex assuming that the summarization CNN can compress the SSL information efficiently. We also choose to guide the second model on the patch magnification to further simplify the training process.

3. Experiments

In this section, we examine the experiments conducted to validate the effectiveness of our method. We train the unified latent diffusion model, ZoomLDM, on patches from eight different magnifications and evaluate the quality of synthetic samples using both real and embedding diffusion model-sampled conditions.

3.1. Implementation details

We train all our LDM models on 3 NVIDIA H100 GPUs, with a batch size 200 per GPU. We use the training code and checkpoints provided by [19]. Our model configuration includes the LDM with a VQ-f4 autoencoder, starting with a U-Net model pre-trained on ImageNet as suggested by [24]. We set the learning rate at 10^{-4} with a warmup of 1000 steps. For the EDM, we train a Transformer decoder on the outputs of the summarization CNN that also takes in the magnification and timestep as input tokens. We learn to directly predict the final conditions from their noisy versions using the standard *linear* DDPM schedule [10]. We utilize DDIM sampling [21] with 50 steps for both models and apply classifier-free guidance [9] sampling with a scale of 1.75 to create synthetic images.

3.2. Dataset

We select 1,136 whole slide images from the TCGA-BRCA dataset. Using the code from DSMIL[12], we extract 256×256 pixel 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 HIPT encoder [3] 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 1024×1024 pixel $20\times$ region results in an embedding matrix sized $4 \times 4 \times 384$. Given that our CNN requires a constant input size, we apply nearest neighbor interpolation to upscale the embedding matrices to a uniform resolution of $128 \times 128 \times 384$.

3.3. Image quality

In Figure 2, we showcase synthetic patches alongside real images from which the embeddings were extracted. We generate 10,000 256×256 pixel patches using ZoomLDM for each magnification level and evaluate their quality using the Fréchet Inception Distance (FID) [8]. For higher magnifications like $20\times$, we conduct comparisons against the

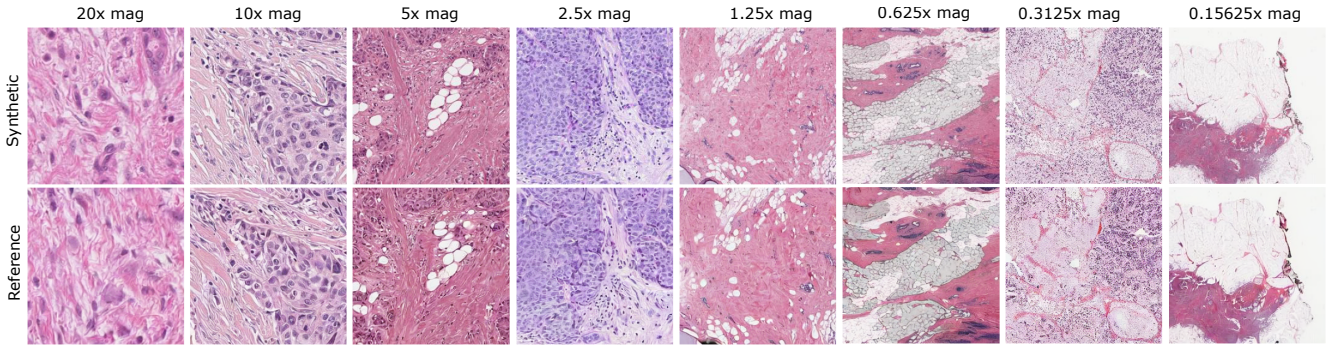

Figure 2. Synthetic patches (256×256 pixel) from ZoomLDM, juxtaposed with the reference (real) images used to generate them. Across all magnifications, ZoomLDM demonstrates consistent preservation semantic features of the reference patches.

Magnification	$20\times$	$10\times$	$5\times$	$2.5\times$	$1.25\times$	$0.625\times$	$0.3125\times$	$0.15625\times$
# Training patches	12 Mil	3 Mil	750k	186k	57k	20k	7k	2.5k
ZoomLDM (Oracle Cond.)	9.29	9.71	**9.16**	11.67	18.17	**21.08**	**23.92**	**22.5**
SoTA	**6.98** [6]	**7.64** [24]	9.74 [6]	-	-	58.98	66.28	106.14
ZoomLDM (Generated Cond.)	14.96	14.93	17.08	20.09	24.89	34.72	30.49	21.10

Table 1. FID of patches generated from ZoomLDM across different magnifications, compared with single magnification models. Our model achieves comparable performance to SoTA at higher magnifications and starts to outperform at magnifications lower than $5\times$. By Oracle Cond. we denote generating images using ground truth conditions extracted from the training set. Generated Cond. refers to synthesizing images from conditions sampled from our EDM.

state-of-the-art (SoTA) methodologies in [6, 24]. 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's performance at higher magnifications, such as $20\times$, is slightly inferior yet competitive with the SoTA. We attribute this slight discrepancy to the abundance of patches available at these magnifications, sufficient for fully training a diffusion model. Our model exhibits superior performance to [6] starting at $5\times$ magnification, where the number of patches is 16 times less than at $20\times$. At lower magnifications, ZoomLDM significantly surpasses the standalone models, highlighting the advantages of our unified architecture and conditioning approach. This unified training approach is especially advantageous when the data is insufficient to train a large diffusion model on its own fully. By leveraging data across magnifications, lower data density scenarios can harness the insights obtained from the entire dataset, improving model performance and efficiency.

3.4. Novel image synthesis

We evaluate the EDM by hierarchically synthesizing 10,000 patches at all magnifications; we first sample an embedding from the trained model and then we synthesize the corresponding image. As shown in Table 1, the generated embedding-conditioned model FID (Generated Cond.) drops when compared to the ZoomLDM that has access to the training set (Oracle Cond.). We showcase novel images in the supplementary material.

4. Conclusion

We developed ZoomLDM to address the challenge of histopathology image generation across a range of magnifications. Featuring an innovative conditioning mechanism, we achieve remarkable results in image synthesis. Our embedding diffusion model has the potential to be conditioned on diverse input sources such as text or RNA sequences. Such models could serve as a robust tool for data augmentation and exploration, laying the groundwork for developing generative foundation histopathology models. ZoomLDM holds the potential to shed light on the heterogeneity of tumor cells and various cancer gradings, and enrich our understanding of cancer's various manifestations.

Acknowledgements This research was partially supported by NCI awards 5U24CA215109, 1R21CA258493-01A1, UH3CA225021, NSF grants IIS-2123920, IIS-2212046 and Stony Brook Profund 2022 seed funding.

References

[1] 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

[2] 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

[3] 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. 1, 2, 3

[4] Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. *Advances in neural information processing systems*, 34:8780–8794, 2021. 1

[5] 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. 1

[6] 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. *arXiv preprint arXiv:2312.07330*, 2023. 1, 2, 4

[7] 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. 1

[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. 3

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

[10] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020. 3

[11] Alexander C. Li, Mihir Prabhudesai, Shivam Duggal, Ellis Brown, and Deepak Pathak. Your diffusion model is secretly a zero-shot classifier. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 2206–2217, 2023. 1

[12] 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. 3

[13] Haohe Liu, Zehua Chen, Yi Yuan, Xinhao Mei, Xubo Liu, Danilo Mandic, Wenwu Wang, and Mark D Plumbley. Audioldm: Text-to-audio generation with latent diffusion models. *arXiv preprint arXiv:2301.12503*, 2023. 1

[14] 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. 1

[15] 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. 1

[16] Alexander Quinn Nichol and Prafulla Dhariwal. Improved denoising diffusion probabilistic models. In *International Conference on Machine Learning*, pages 8162–8171. PMLR, 2021. 1

[17] Ethan Perez, Florian Strub, Harm De Vries, Vincent Dumoulin, and Aaron Courville. Film: Visual reasoning with a general conditioning layer. In *Proceedings of the AAAI conference on artificial intelligence*, 2018. 3

[18] 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

[19] 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

[20] 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

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

[22] Junjiao Tian, Lavisha Aggarwal, Andrea Colaco, Zsolt Kira, and Mar Gonzalez-Franco. Diffuse, attend, and segment: Unsupervised zero-shot segmentation using stable diffusion. *arXiv preprint arXiv:2308.12469*, 2023. 1

[23] 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. 1

[24] 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. 1, 2, 3, 4