ZoomLDM: Latent Diffusion Model for multi-scale image generation

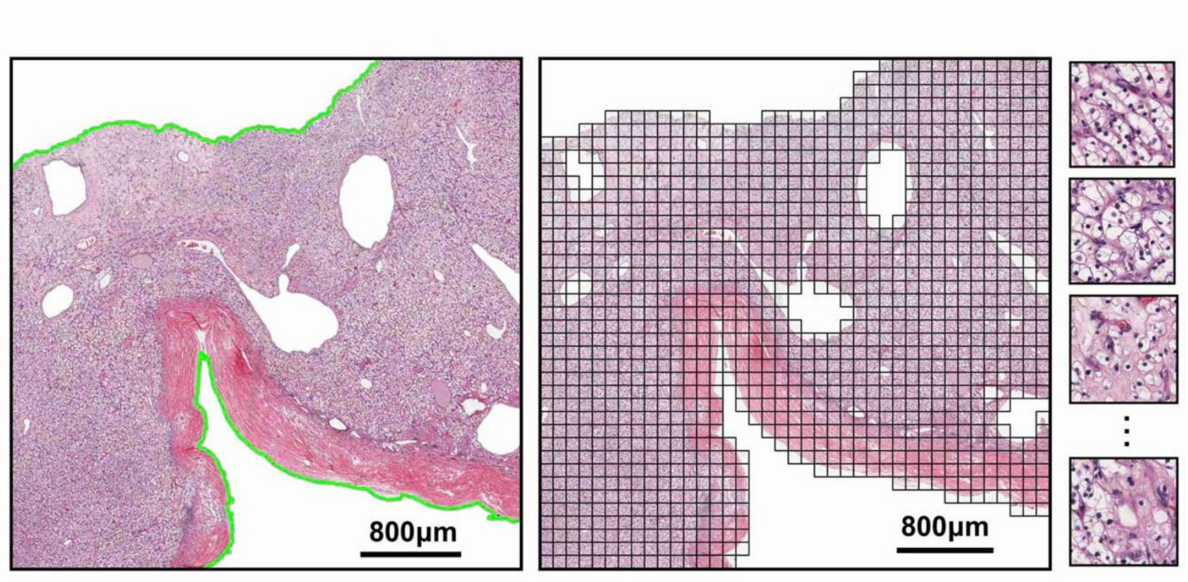
Srikar Yellapragada*, Alexandros Graikos*, Kostas Triaridis,
Prateek Prasanna, Rajarsi Gupta, Joel Saltz, Dimitris Samaras

CVPR 2025

ExHall D Poster #229
Sunday June 15, 10:30 a.m. — 12:30 p.m.

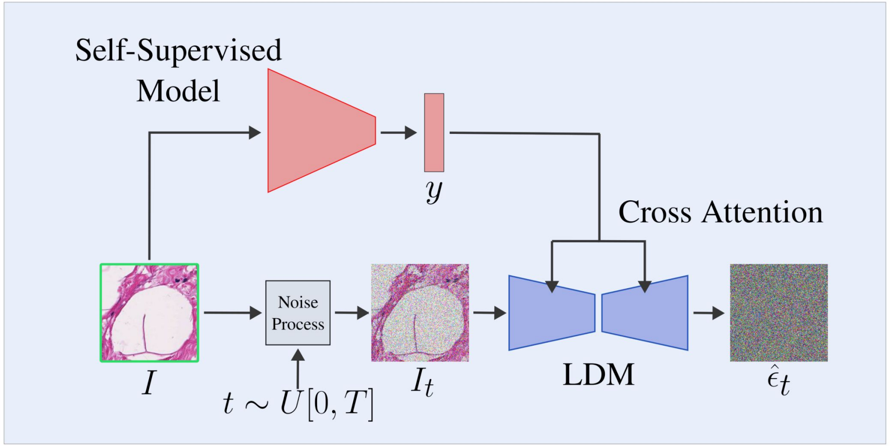
Challenges in Histopathology

- We want to generate images both at patch and WSI level
  - Labels are typically at whole slide level, eg. reports
- WSIs are inherently multi-scale



Previous work

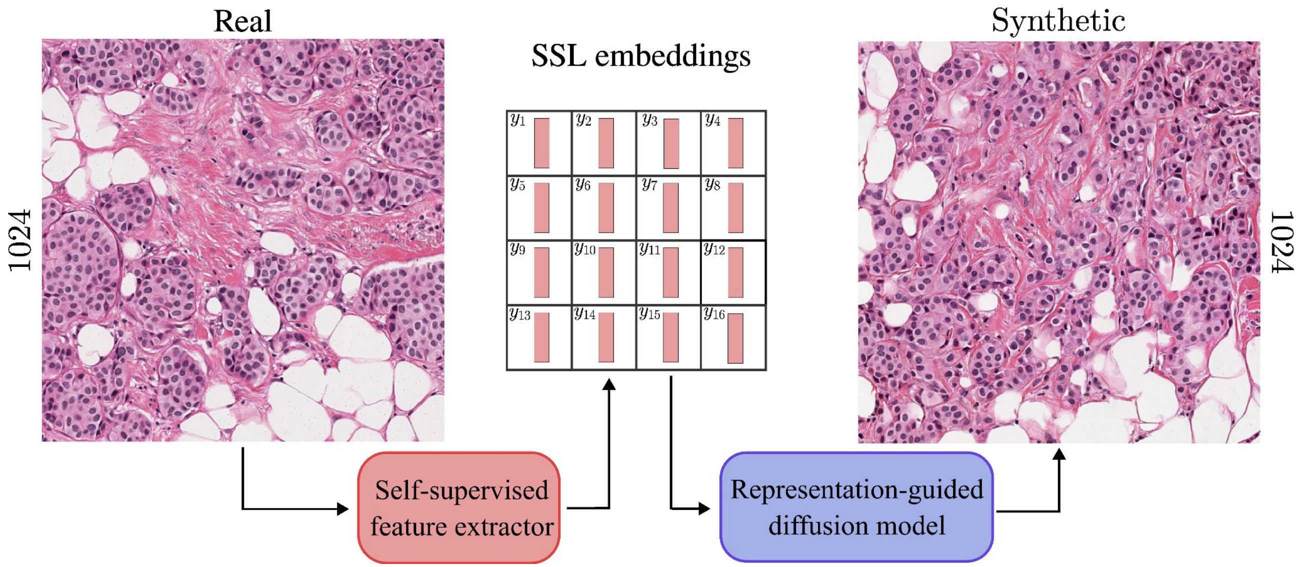
- We proposed using **representations** learned with self-supervision **in place of human annotations**



Graikos, Alexandros, et al. "Learned representation-guided diffusion models for large-image generation." CVPR 2024

Previous work

- Impractical to train directly on the entire digitized slides (32,000 x 32,000 px)
  - We introduced an algorithm to **synthesize large histopathology images** by spatially controlling the local, patch-based model

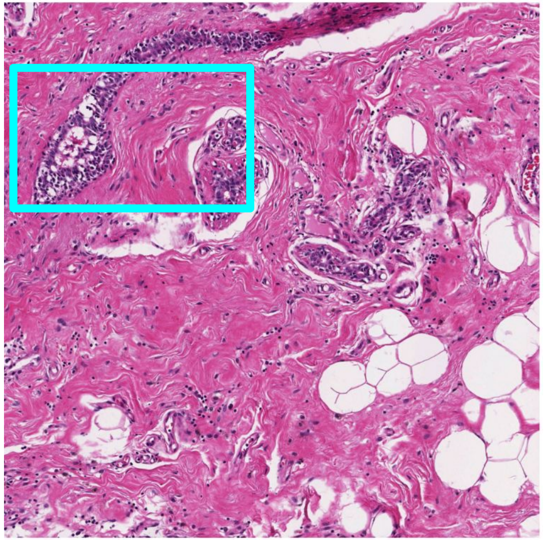


Graikos, Alexandros, et al. "Learned representation-guided diffusion models for large-image generation." CVPR 2024

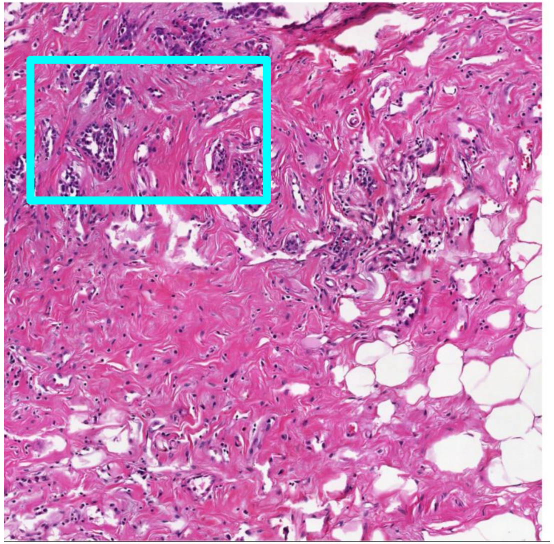
Limitations

- Models that are restricted to a single magnification/scale fail to understand global context

Reference



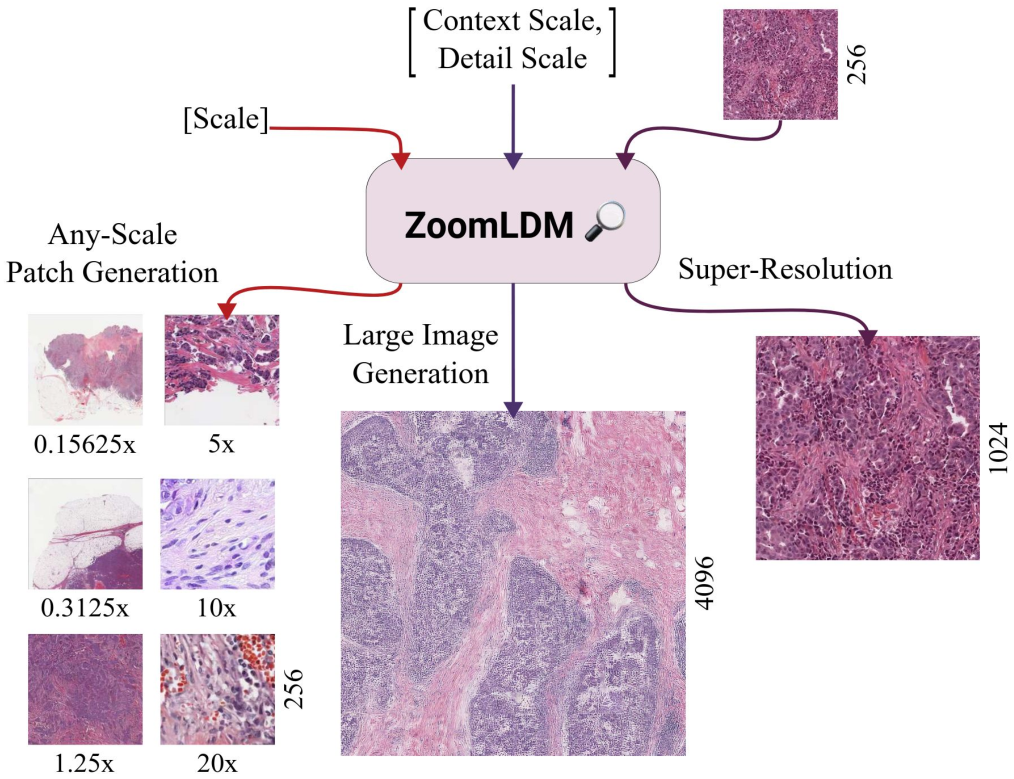
CVPR 2024



Graikos, Alexandros, et al. "Learned representation-guided diffusion models for large-image generation." CVPR 2024

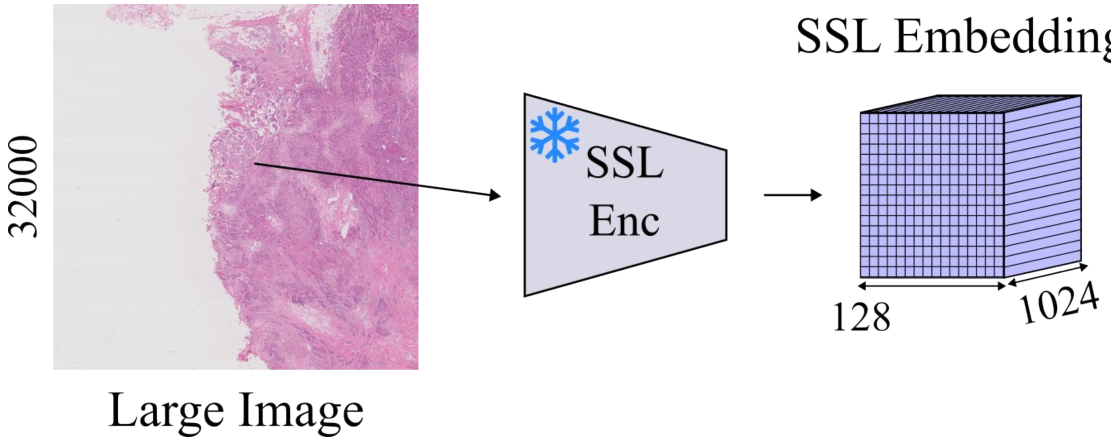
ZoomLDM - Multi-scale diffusion model

- Going multi-scale allows us to overcome this limitation.



Multi-scale data

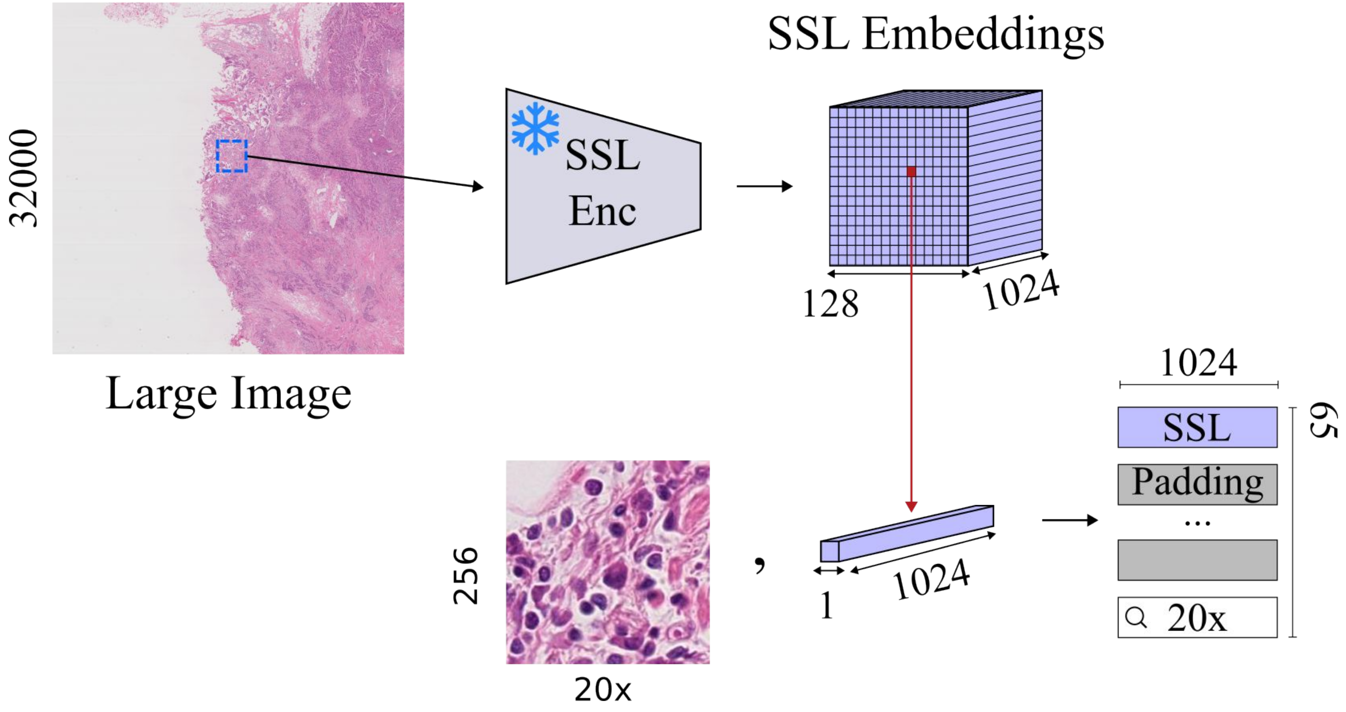
- We feed the WSI to a pre-trained SSL encoder (UNI [2]) to obtain a grid of SSL embeddings.



[2] Chen, Richard J., et al. "Towards a general-purpose foundation model for computational pathology." Nature Medicine (2024)

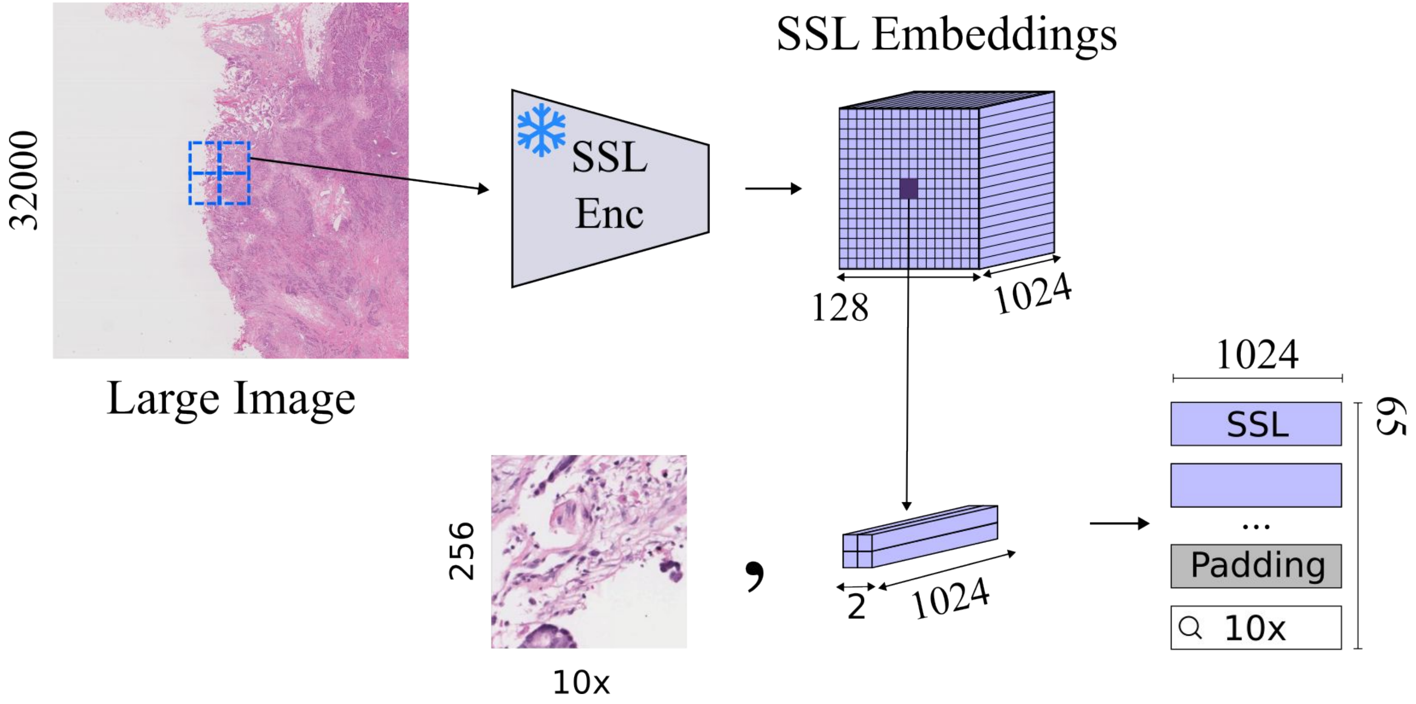
Multi-scale data : 20x

- We create (image, embedding) pairs at all magnifications
  - 20x ⇒ single SSL embedding



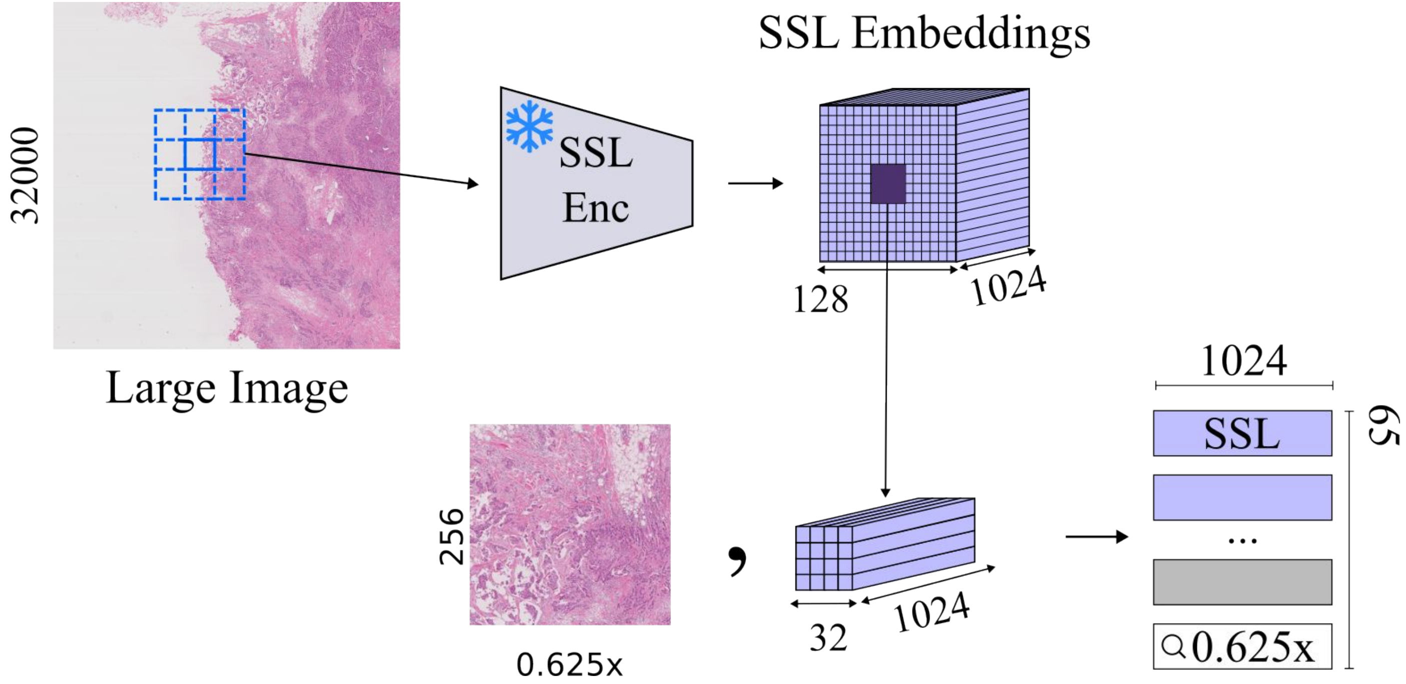
Multi-scale data : 10x

- 10x ⇒ 4 SSL embeddings



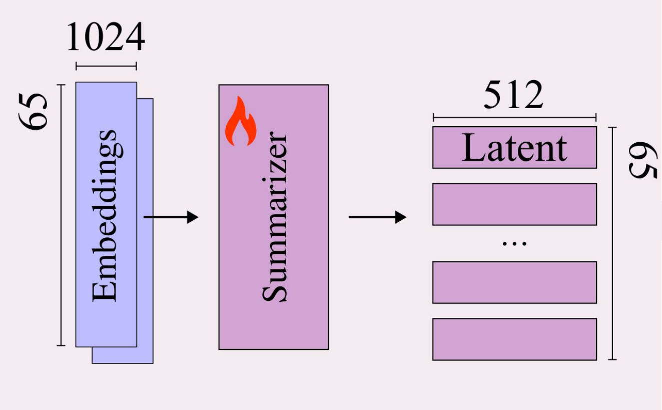
Multiscale data : Lower magnifications

- 0.625x ⇒ 32x32 SSL embeddings, average pooled to 8x8



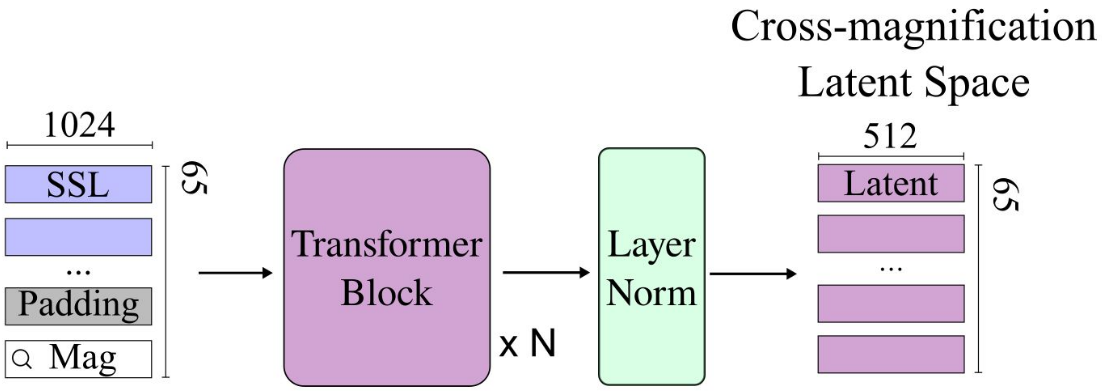
Summarizer

- We process embeddings with a Summarizer, projecting them to a shared latent space.



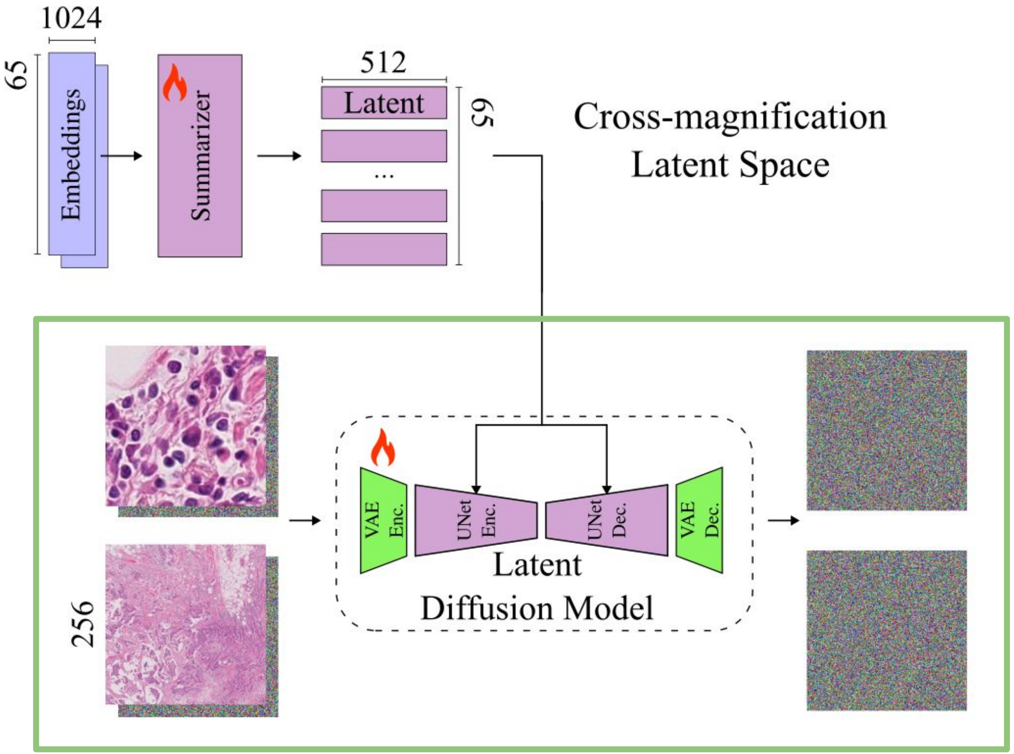
Summarizer

- Implemented with a 12-layer Transformer
- Trained jointly with the LDM



ZoomLDM - Training

- Train LDM on 256x256 patches conditioned on the summarizer's outputs



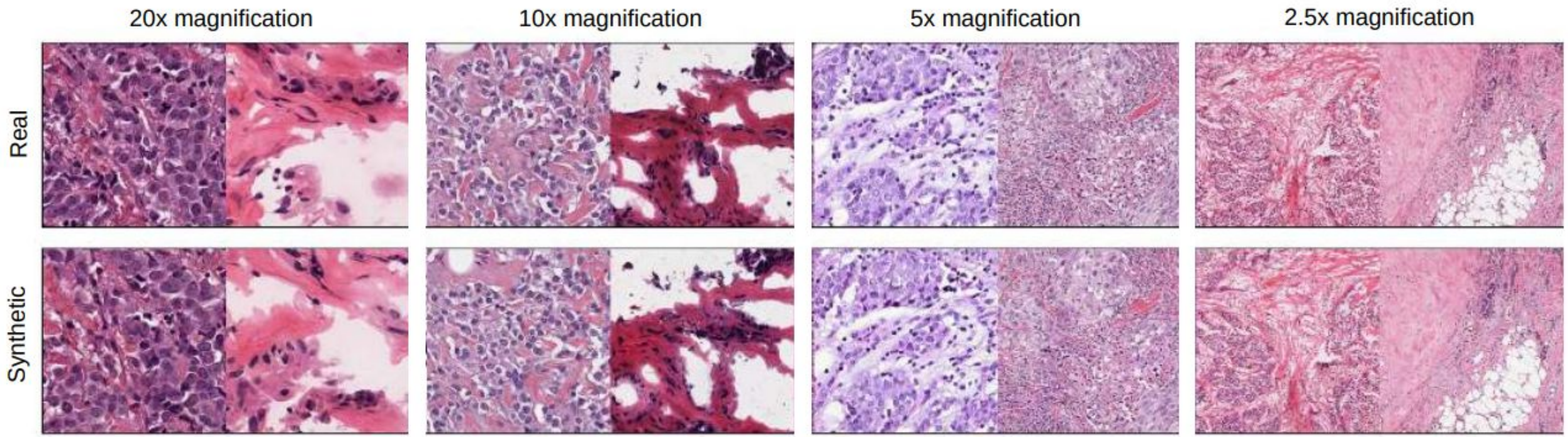
Results - Patch generation

- SoTA across all magnifications, due to parameter sharing

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

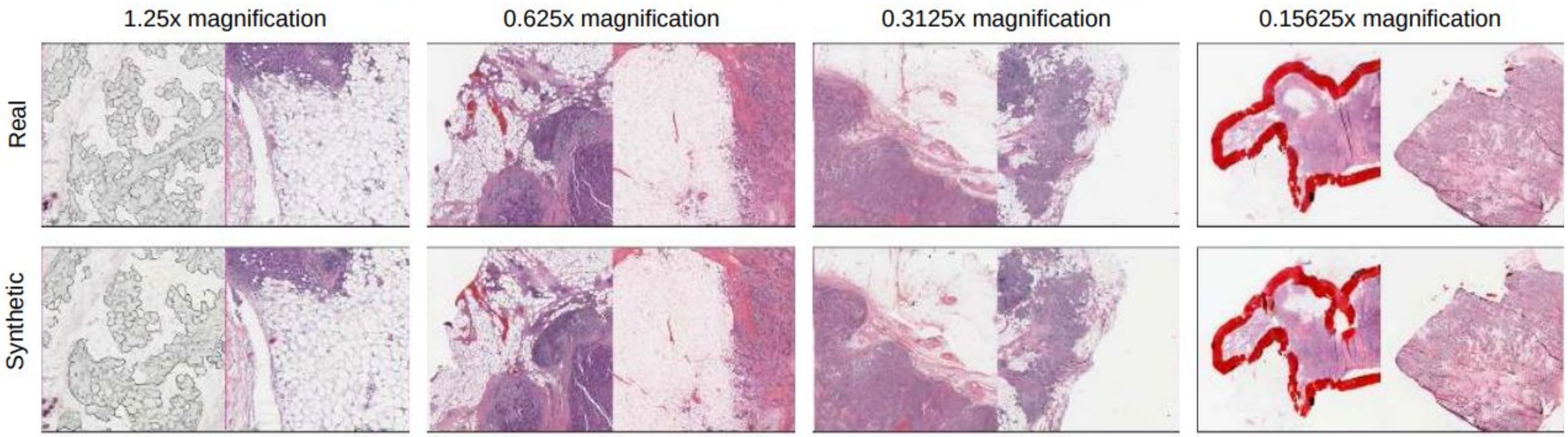
Results - Patch generation

- ZoomLDM preserves semantic features of the reference patch



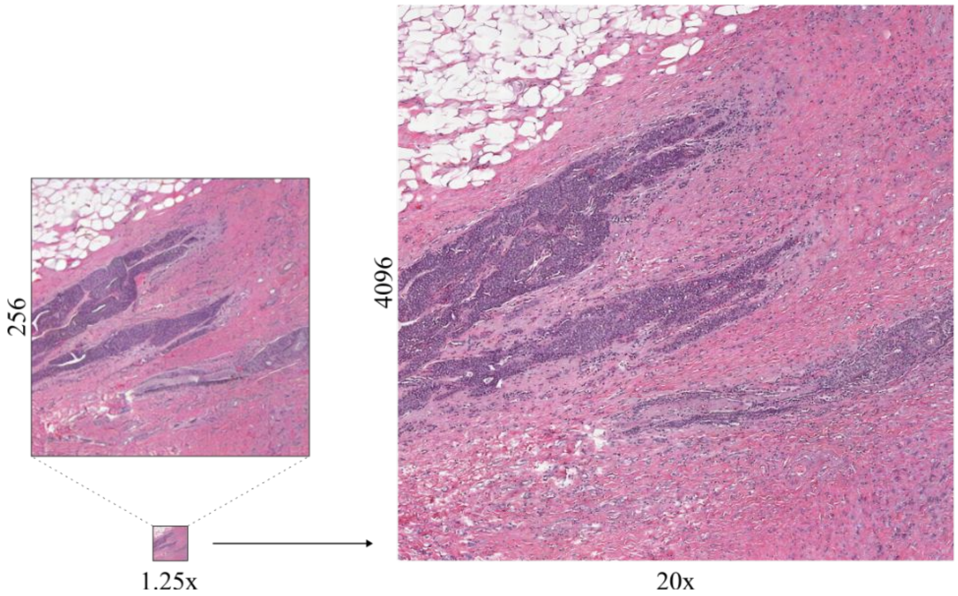
Results - Patch generation

- ZoomLDM preserves semantic features of the reference patch



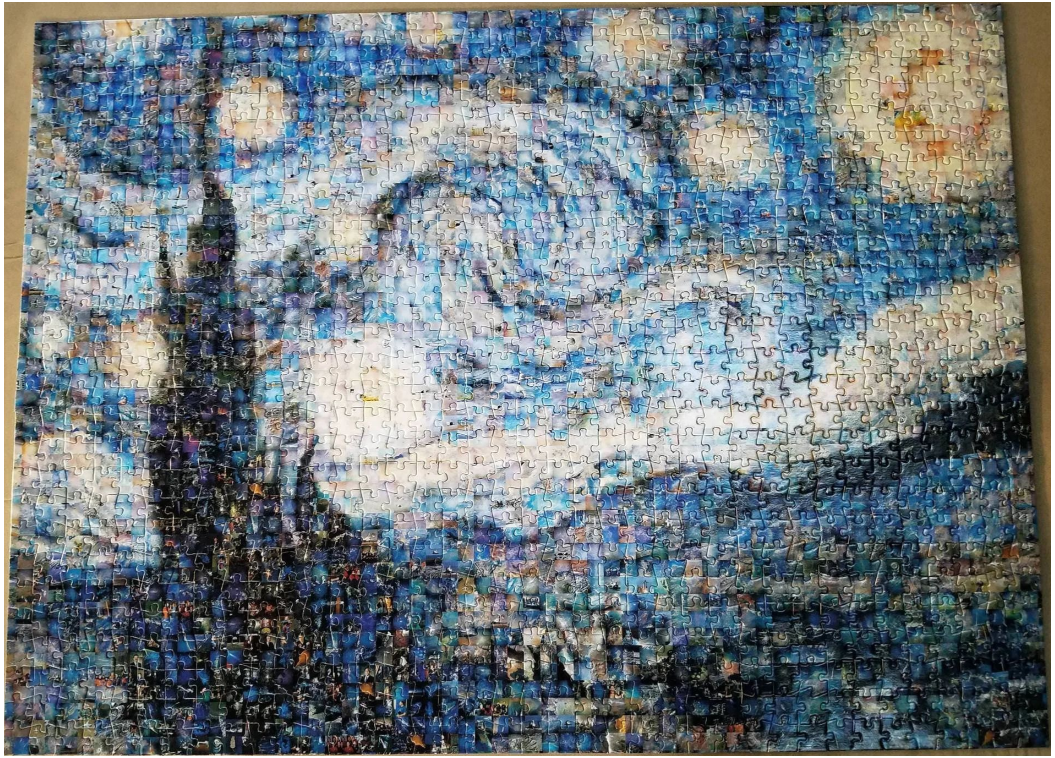
Joint multi-scale sampling

- We can generate multiple scales at the same time
  - 1.25x image guides the structure of 20x with global context.

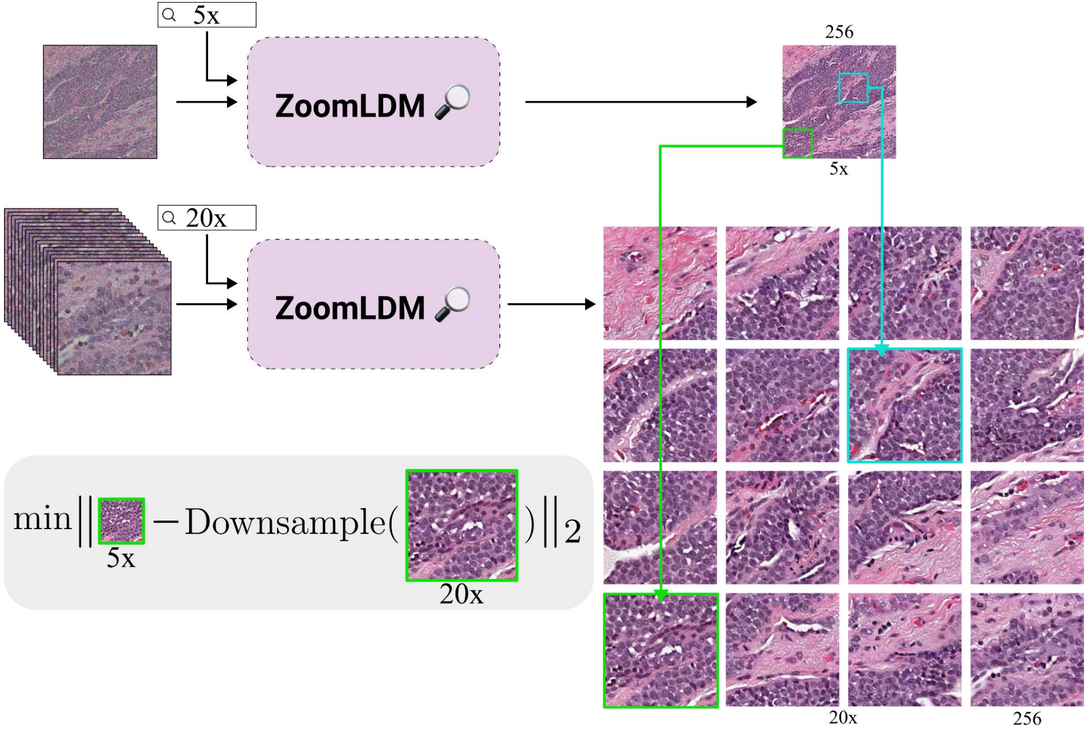


Large image generation

Similar to a photo-mosaic?



Large image generation - joint sampling



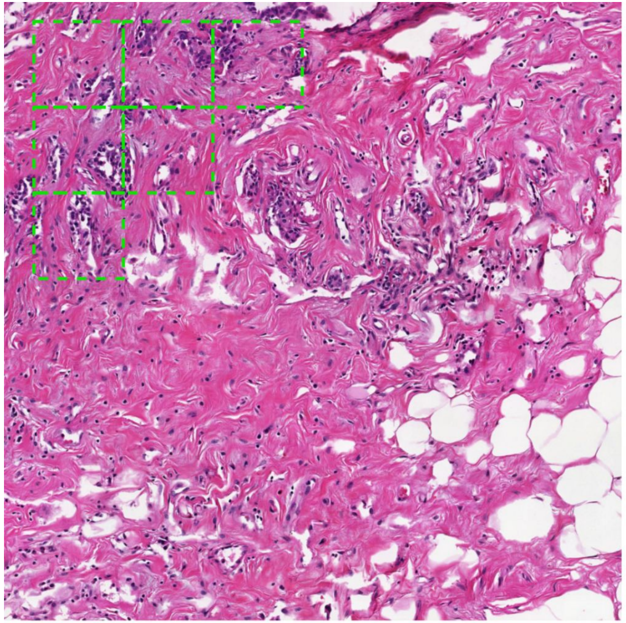
- During sampling the 5x image guides the generation of the 20x
- "Self-guidance" by minimizing the difference between the downsampled 20x patches and the 5x guide [3]
- We introduce a faster and less memory-intensive way to enforce guidance that avoids backprop

[3] Graikos, Alexandros, Nebojsa Jojic et al. "Fast constrained sampling in pre-trained diffusion models." arXiv preprint (2024)

Comparison to previous work

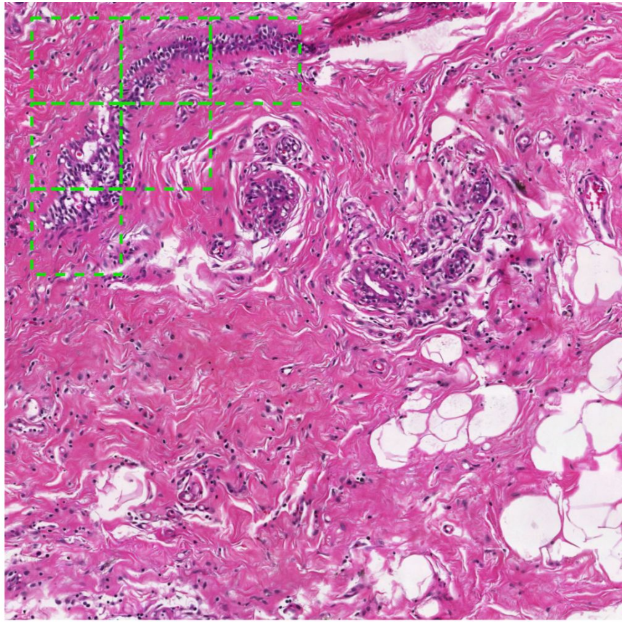
- The generated large images maintain global context without sacrificing details

CVPR 2024



1024

ZoomLDM

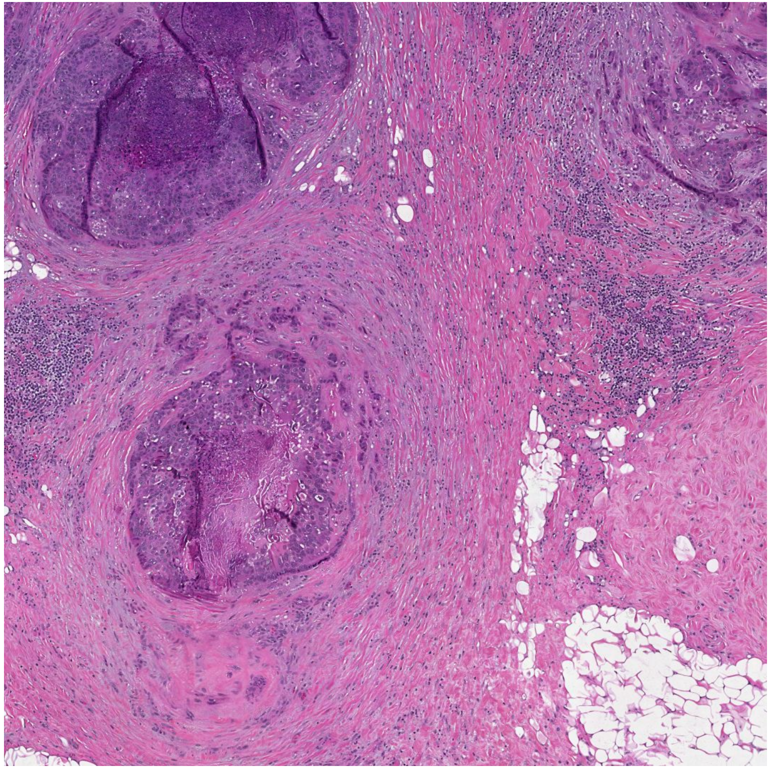


1024

Results - Large 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

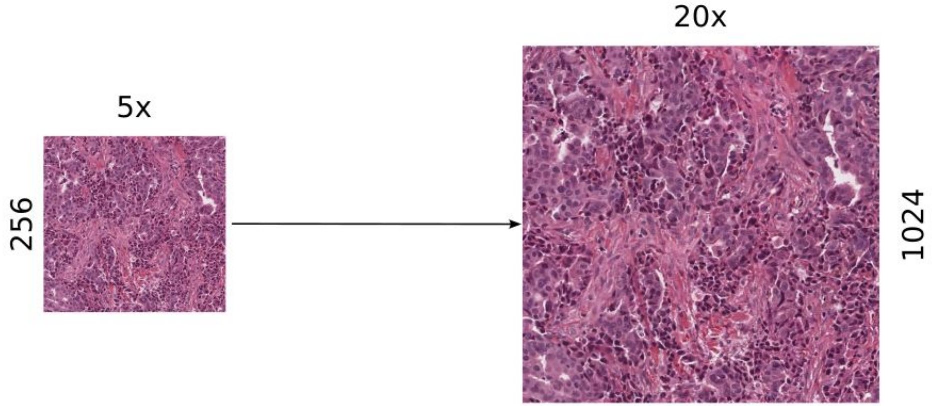
More results



Results - Satellite (NAIP)

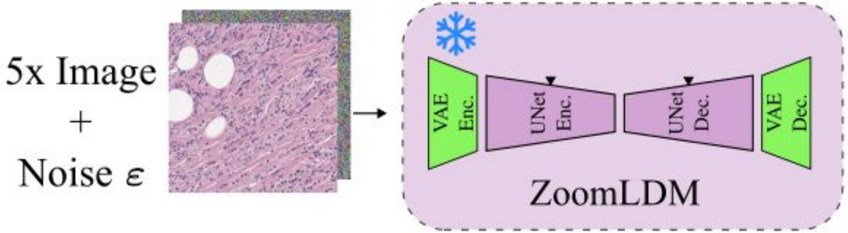
Superresolution

- 4x super-resolution ⇒ 5x to 20x magnification
- Same joint sampling algorithm
- No access to ground-truth SSL embeddings (In 20x magnification)



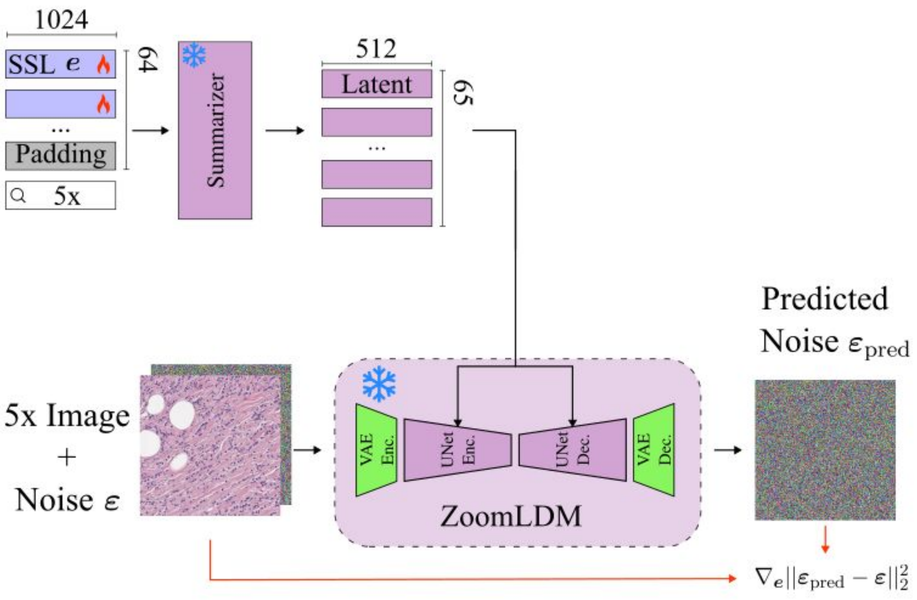
Superresolution - inversion

- We infer the 20x SSL embeddings from the 5x low-res image.



Superresolution - inversion

- We infer the 20x SSL embeddings from the 5x low-res image
  - Use the denoising loss to find the SSL embeddings that best denoise the 5x image
  - Similar to textual inversion



Results - Superresolution

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**	<u>0.180</u>	0.969	**0.810**
CompVis [39]	LR image	0.563	21.926	<u>0.247</u>	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	<u>0.680</u>	0.739	29.822	0.235	0.965	0.741
	GT emb	0.599	23.273	0.250	<u>0.946</u>	0.672	0.732	29.236	0.245	<u>0.974</u>	0.753
	Infer emb	<u>0.609</u>	<u>23.407</u>	**0.229**	**0.957**	**0.719**	<u>0.779</u>	<u>30.443</u>	**0.173**	**0.974**	<u>0.808</u>

Multiple instance learning

- ZoomLDM as feature extractor for MIL
- Can utilize features at multiple scales

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

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

Conclusion

- **ZoomLDM** is the first multi-scale LDM for large image domains
  - Shared weights across scales achieve SoTA in patch generation

- Multi-scale generation and our efficient joint sampling algorithm enable:
  - Large image generation (**4096 x 4096 pixels**)
  - Superresolution

- Multi-scale features outperform SSL in MIL tasks