

HeteroTissue-Diffuse

Semantic and Visual Crop-Guided Diffusion Models for
Heterogeneous Tissue Image Synthesis



Saghir Alfasly



Wataru Uegami



Enamul Hoq



Ghazal Alabtah



Hamid Tizhoosh

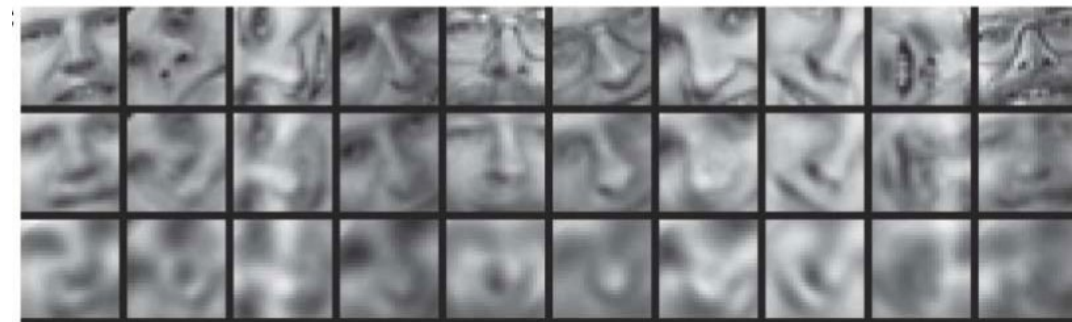
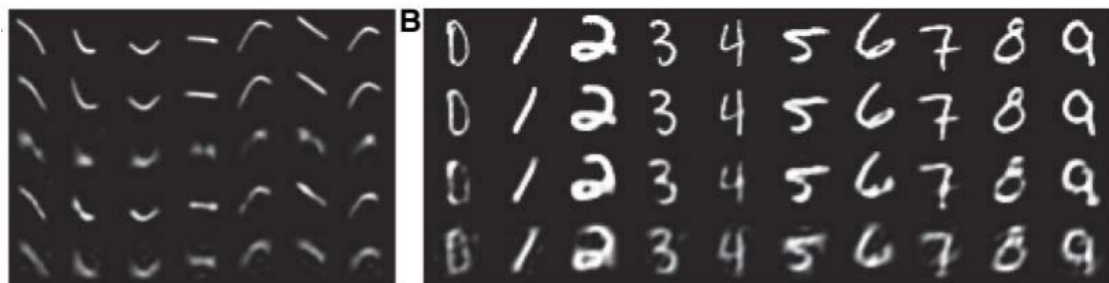
KIMIA Lab, Department of Artificial Intelligence and Informatics, Mayo Clinic, Rochester, MN, USA

December 4th, 2025

VISUAL GENERATIVE MODELS IN GENERAL DOMAIN

Visual generative model advancement

2013



2025



Sora v2



Veo v3

VISUAL GENERATIVE MODELS IN GENERAL DOMAIN

- Autoencoders and Denoising Autoencoders became popular for unsupervised feature learning.
- Variational Autoencoders (VAEs)
- GAN
- Hybrid VAE–GAN
- Autoregressive Models e.g., PixelRNN
- Diffusion Models Revolution
 - DDPM (Denoising Diffusion Probabilistic Models)
 - Stable Diffusion / Imagen / DALL·E 2
 - Video & Multimodal Diffusion (e.g., SORA, Runway Gen-3, Veo)

CHALLENGES

Histopathology image analysis faces critical challenges:

- Data scarcity
- Data privacy
- Annotation cost

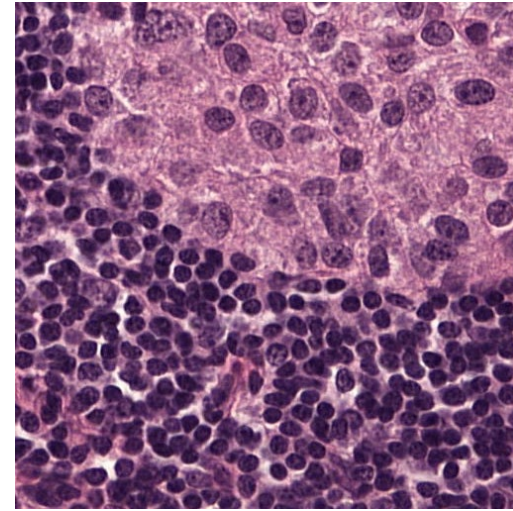
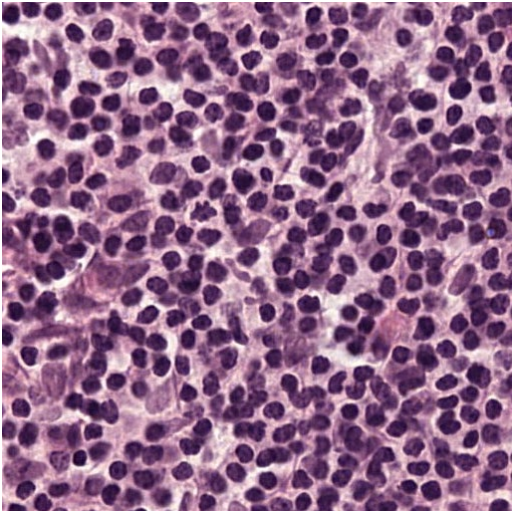
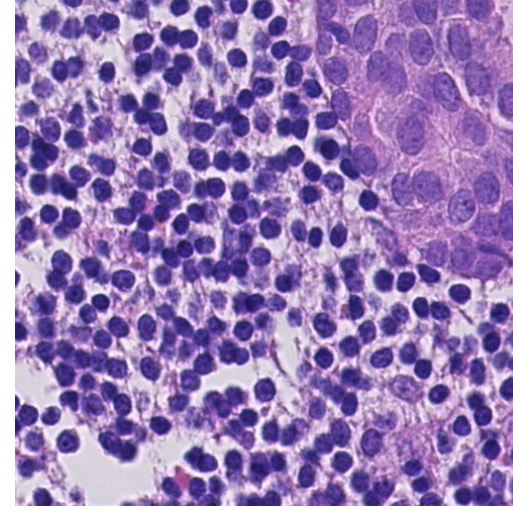
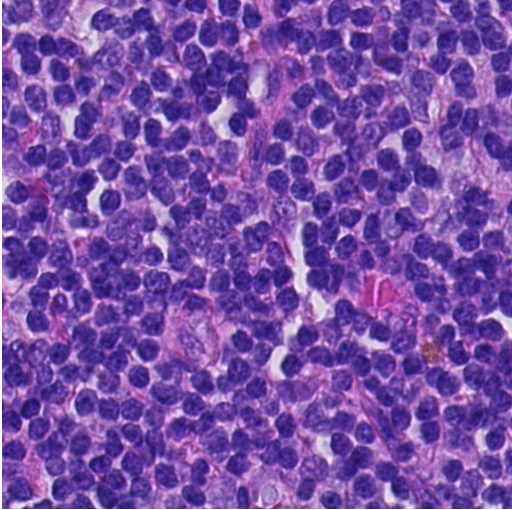
CHALLENGES

Histopathology image generation faces critical challenges:

- Limited ability to synthesize heterogeneous tissue samples
- Difficulty preserving diagnostically relevant features
- Difficult to maintain consistency across magnification levels
- Text-based prompting introduces interobserver variability
- Embedding-based approaches lose critical morphological details

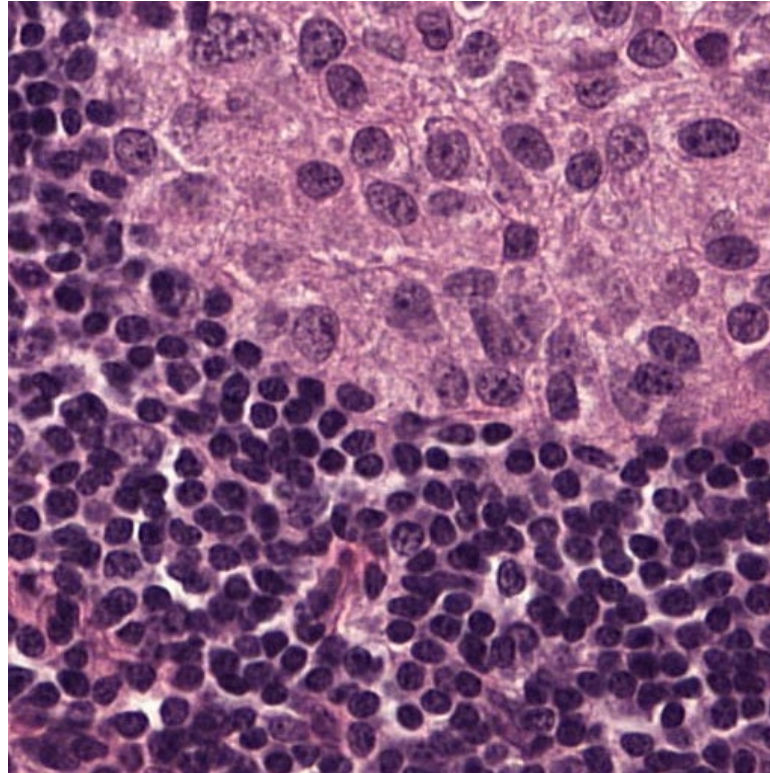
CHALLENGES

- Limited ability to synthesize **heterogeneous** tissue samples



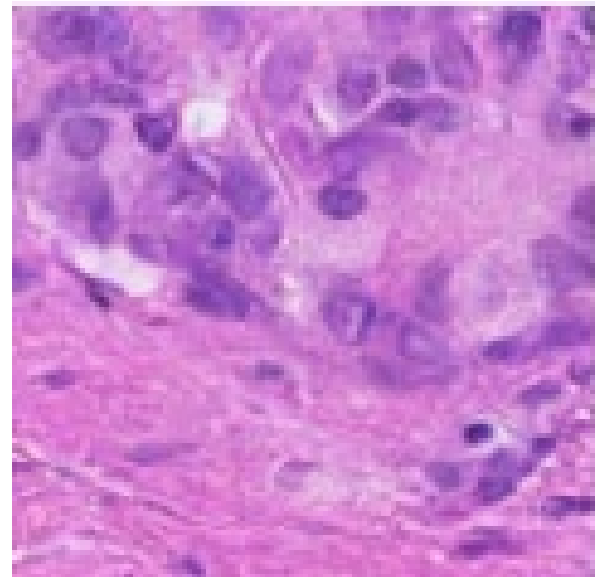
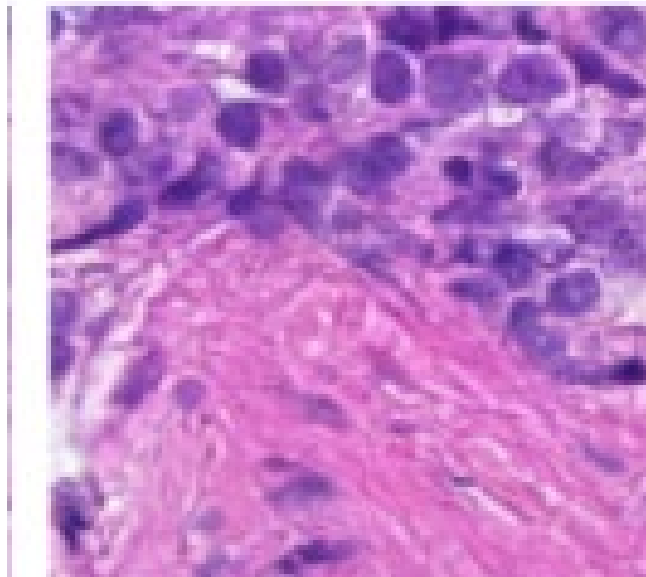
CHALLENGES

- Difficulty preserving diagnostically **relevant features**



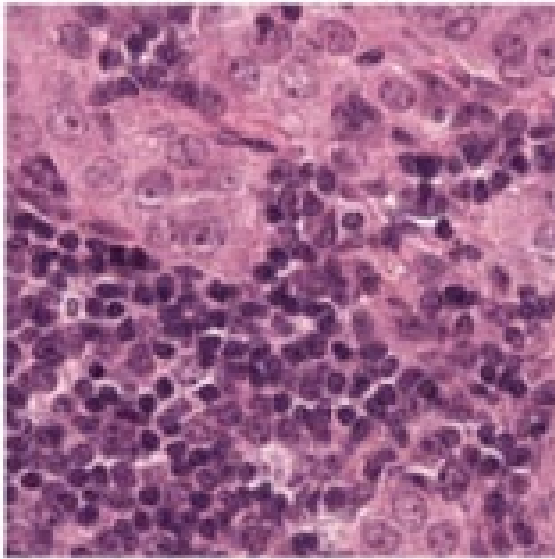
CHALLENGES

- Text-based prompting introduces **interobserver variability** and limits the **tissue nature diversity**

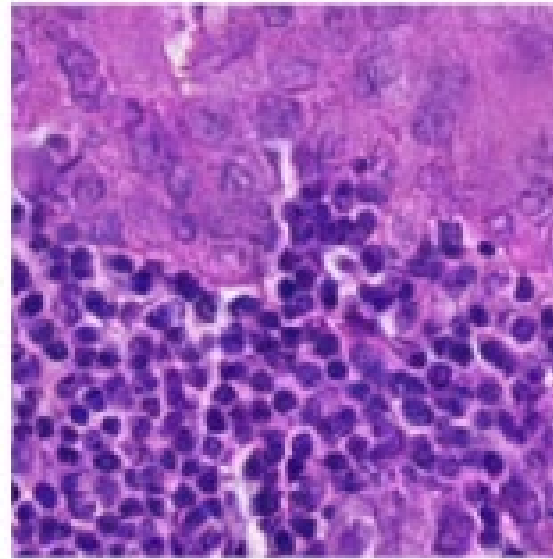


CHALLENGES

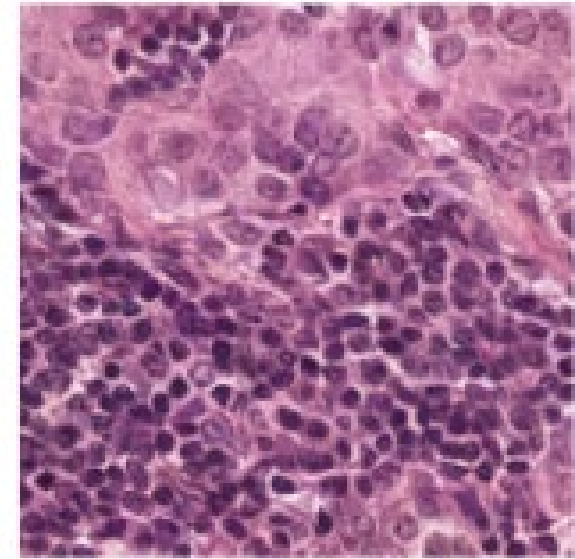
- Embedding-based approaches lose critical **morphological details**



Reference



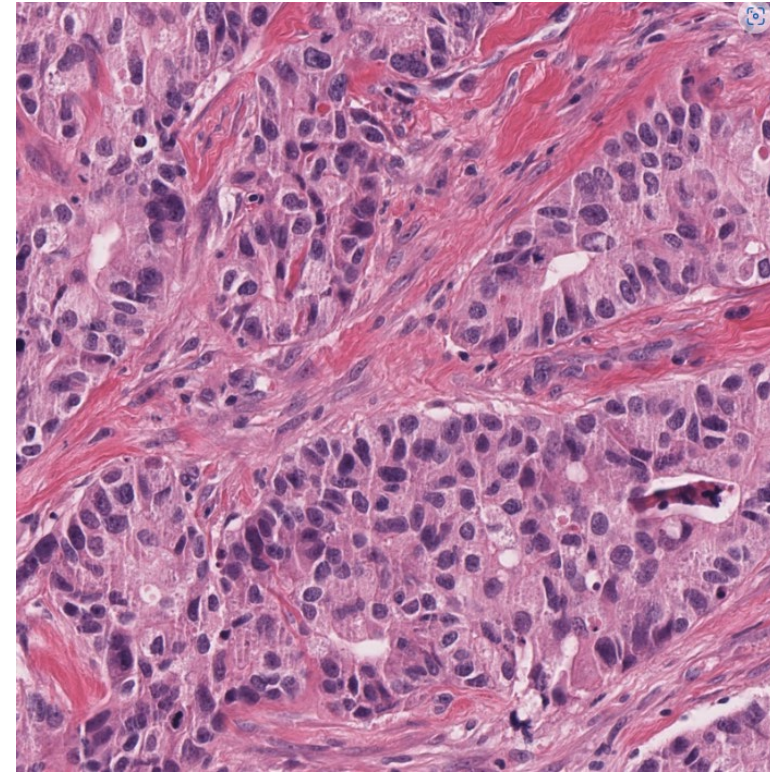
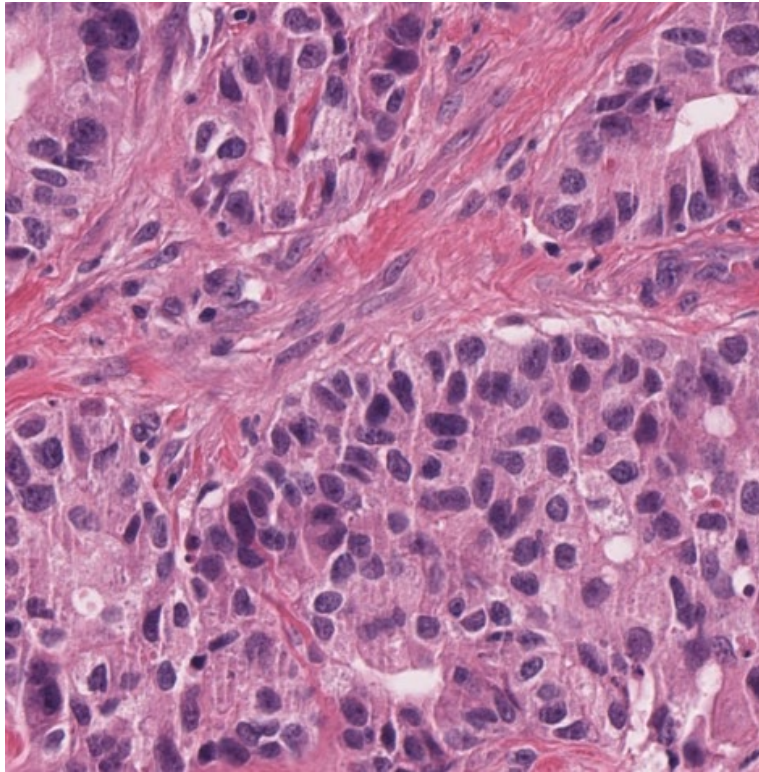
Embedding Prompt



target

CHALLENGES

- Difficult to maintain consistency across **magnification** levels



OUR SOLUTION (OVERVIEW)

A new dual-conditioning latent diffusion model that:

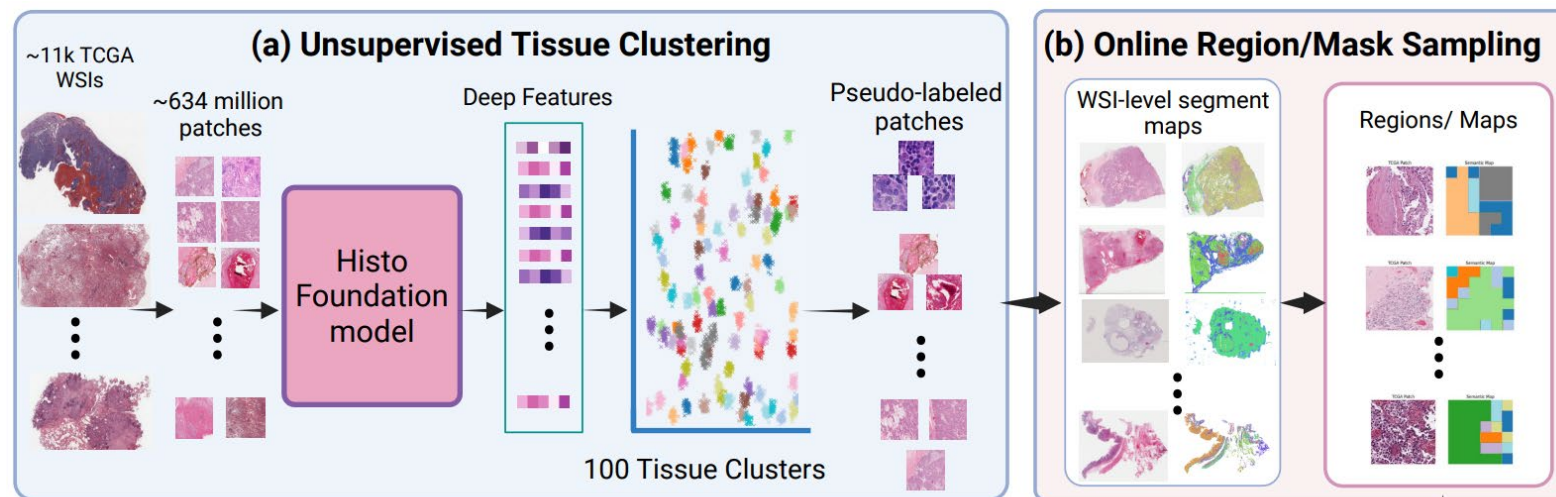
1. Combines **semantic maps** with tissue-specific **visual crops**
2. **Preserves morphological fidelity** and diagnostic features
3. Enables precise **region-specific** generation control
4. Scales to **unannotated datasets** through self-supervision

KEY CONTRIBUTIONS

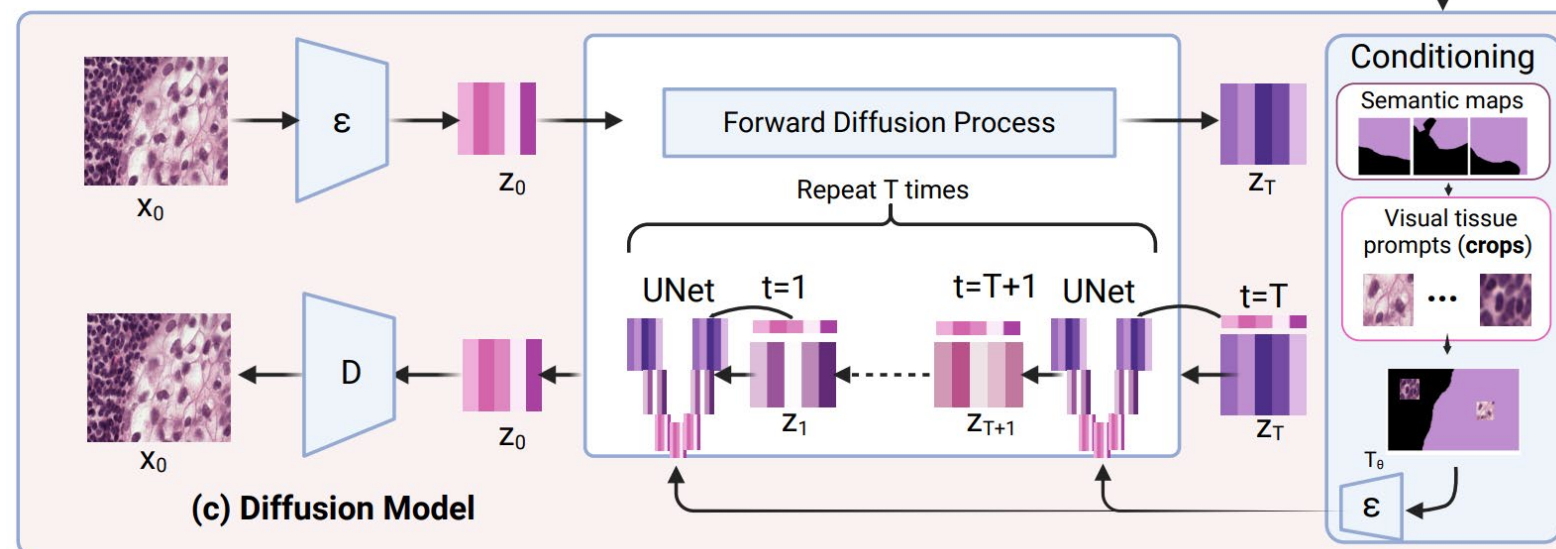
1. **Dual-conditioning architecture:** Semantic maps + raw tissue crops
2. **Self-supervised framework:** For unannotated WSIs (TCGA dataset)
3. **Comprehensive validation:** Quantitative metrics + pathologist evaluation

HETEROTISSUE-DIFFUSE FRAMEWORK

Data sampling for training

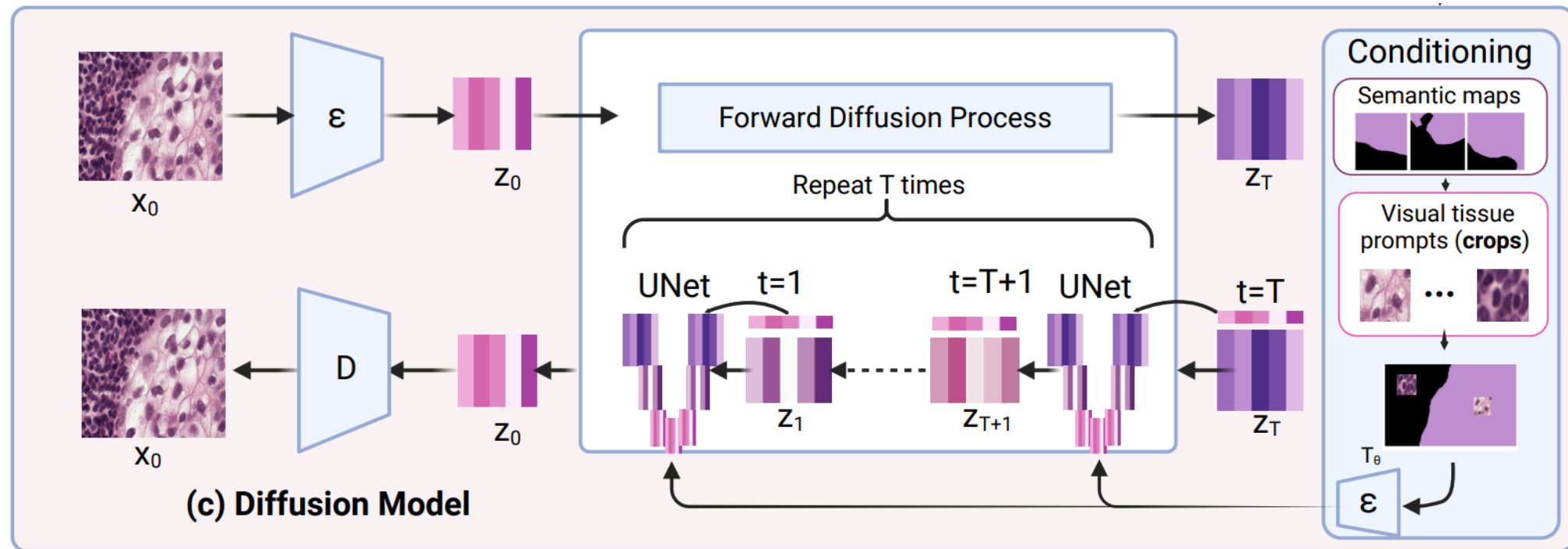


Model training



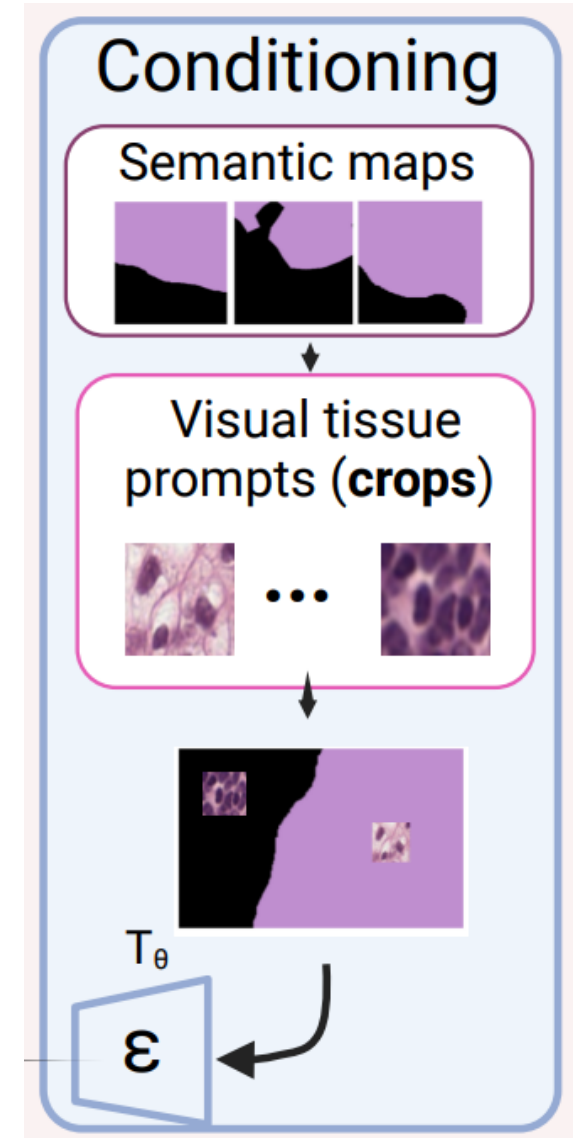
HETEROTISSUE-DIFFUSE FRAMEWORK

HeteoTissue-Diffuse training framework



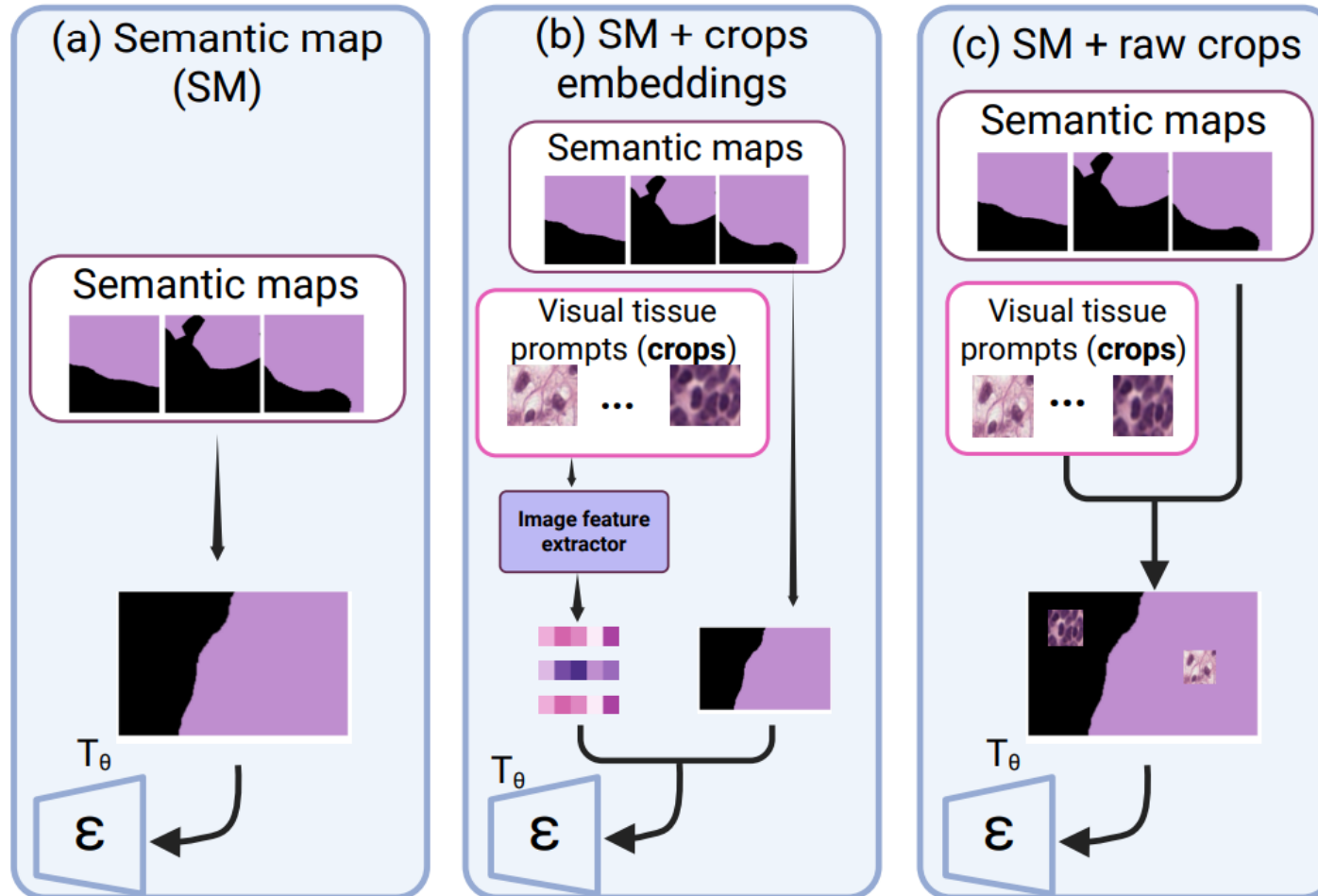
DUAL-CONDITIONING MECHANISM

- Combines spatial precision with morphological authenticity:
 - Semantic maps provide **region boundaries** and tissue **type locations**
 - Visual crops preserve **cellular texture, staining, morphology**
 - Conditioning signal: **concatenate** semantic map and crops



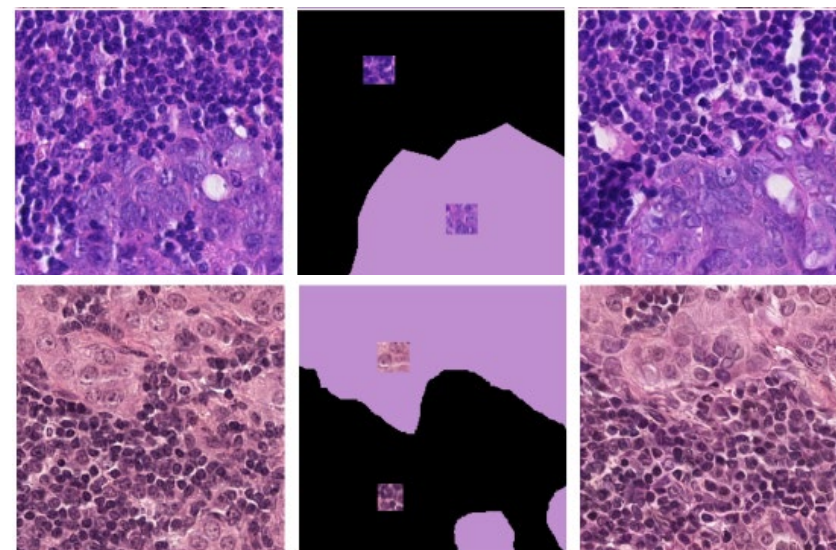
DUAL-CONDITIONING MECHANISM

Different variations of conditions (prompts)

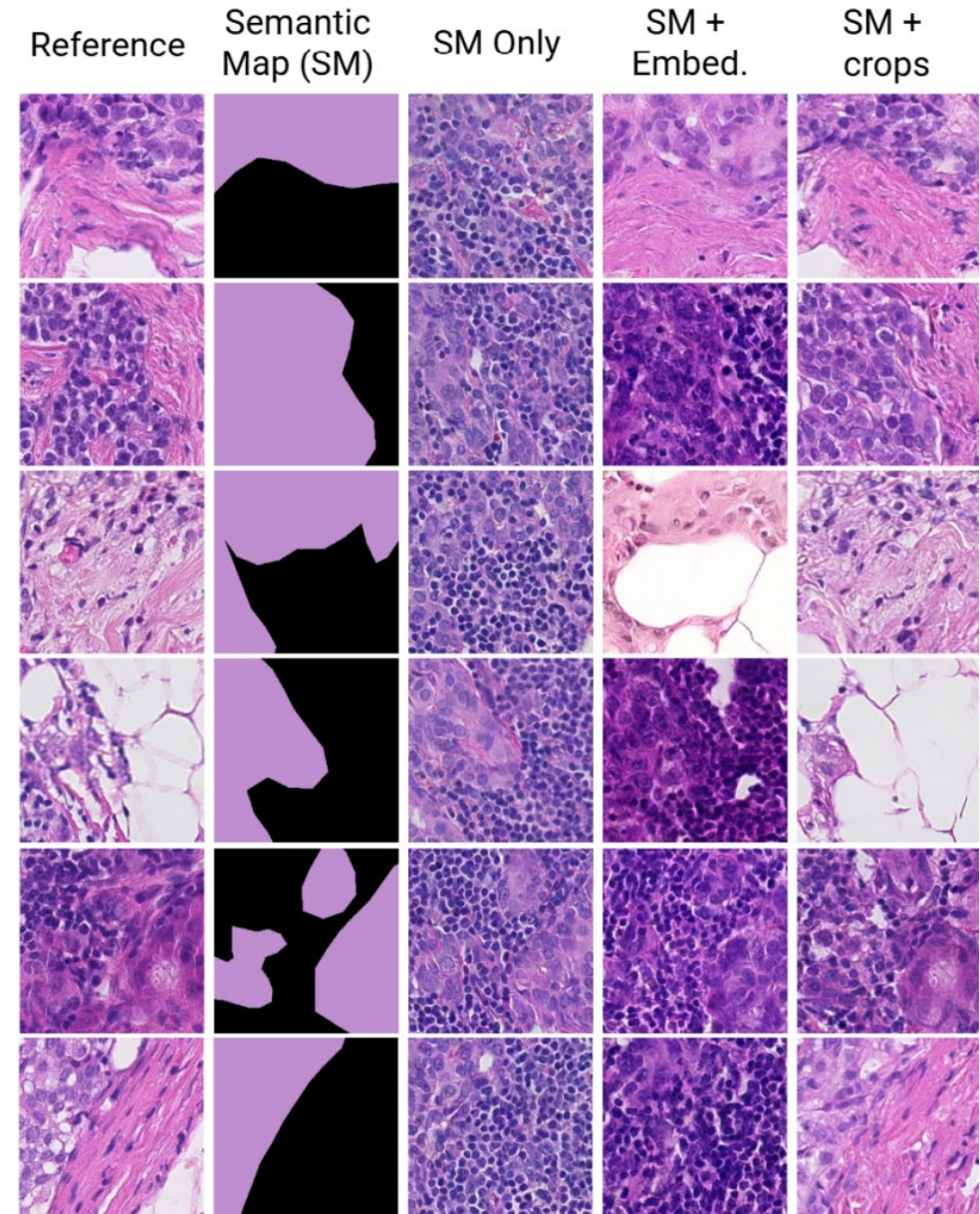
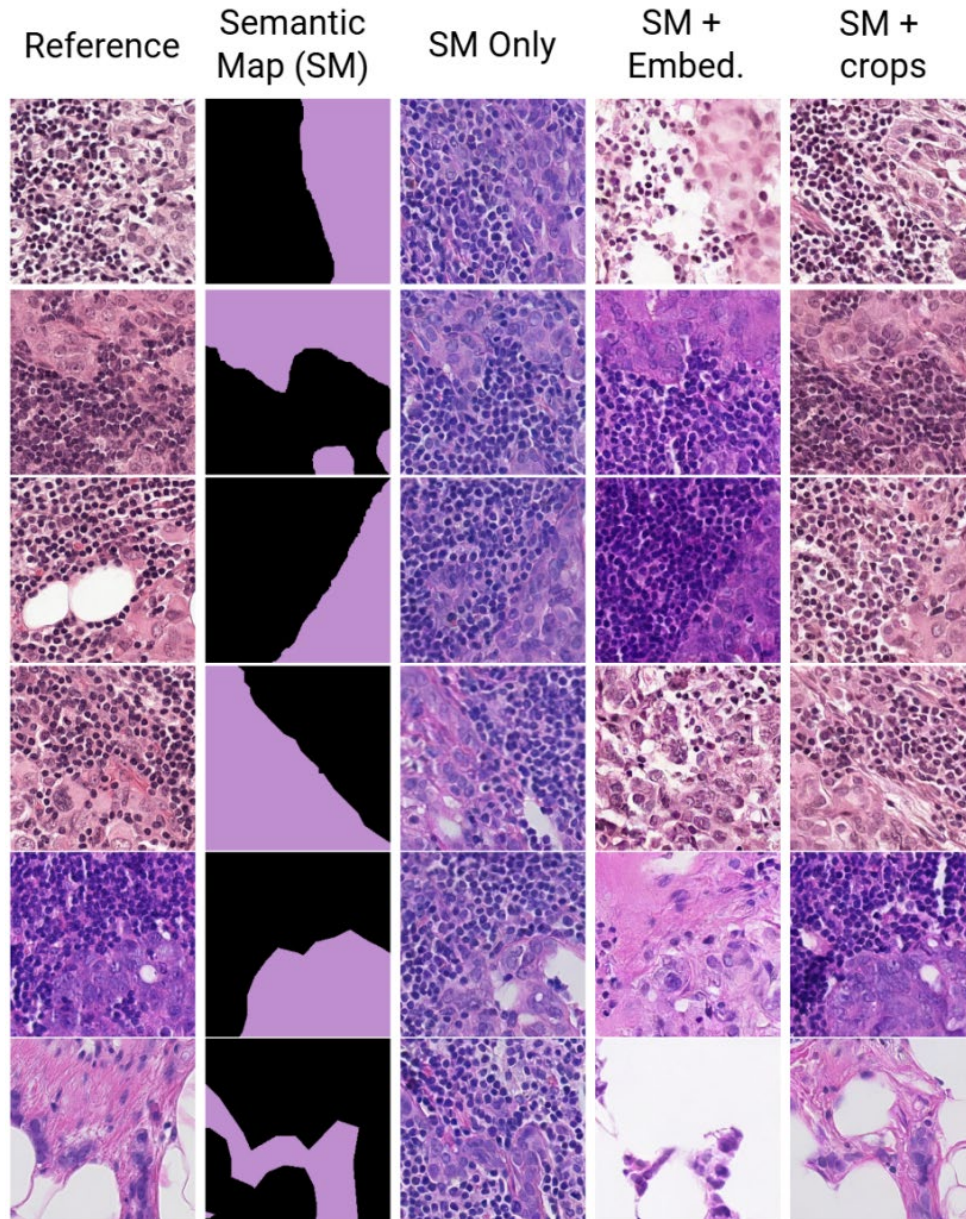


ALGORITHM: HETEROGENEOUS PATCH SAMPLING

- Sample patches where tissue ratios (**tumor/normal**) in $[0.2, 0.8]$ range
- For each class present in segmentation mask:
 - Extract **square crop** from region of that class
 - Apply optional augmentations (**rotation/flipping**)
 - **Place crop** at valid location within semantic region
- Results in training sample with both **image** and **conditional inputs**



RESULTS Generated Samples



RESULTS Quantitative Results: Frechet Distance

- Key findings: **6× reduction** in FD on Camelyon16 with prompts (430.1 → 72.0)
- **2-3× lower** FD across **Panda** and **TCGA** datasets
- **Visual prompts** consistently outperform **embedding-based** conditioning
- Different foundation model encoders show **varying sensitivity** to conditioning
- RN50-BT and DINOv2 ViT-L show greatest improvement with prompts

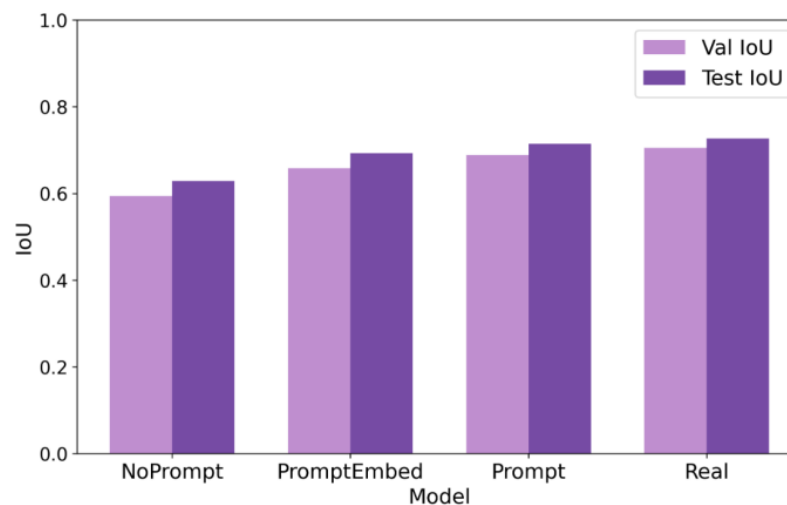
Table 1: FD Results for Prompt, Nonprompt, and crop embedding prompt conditions across CAMELYON16 [5], PANDA [7], and TCGA [38] datasets

Dataset	Cond.	Lunit-8 [18]	GigaPath [39]	H-Optimus [31]	PathDino [3]	RN50-BT [18]	DINOv2 [27]	UNI2 [10]	UNI [10]
CAM16	NP	1360.9	714.0	713.9	7540.6	430.1	122.0	139.8	70.0
	Emb. Prompt	991.3	606.6	664.7	4331.1	183.0	289.6	141.6	841.1
	Prompt	629.1	353.0	425.2	2591.5	72.0	52.7	85.2	481.4
PANDA	NP	877.8	347.3	422.2	5124.7	150.0	352.4	113.6	650.5
	Prompt	512.2	139.7	227.1	3230.9	22.8	61.4	52.4	299.9
TCGA	NP	855.1	360.4	476.0	4306.7	157.7	117.5	119.6	563.6
	Prompt	821.9	346.1	521.4	3876.7	142.9	142.1	135.1	527.9

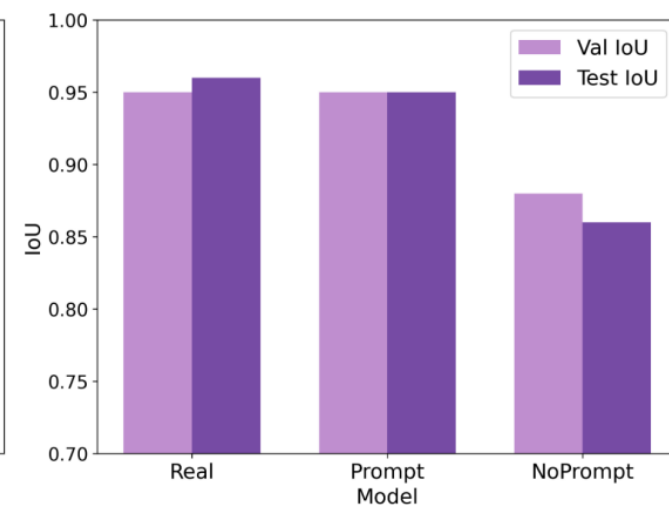
RESULTS Downstream Evaluation: Segmentation

- DeepLabv3+ models trained on synthetic data achieve:
 - 0.71 IoU on Camelyon16 (real data: 0.72)
 - 0.95 IoU on Panda (real data: 0.96)
 - Only 1-2% gap from real data training
- Models trained solely on our synthetic data **approach real-data performance**
- NoPrompt (i.e., semantic maps only) synthetic data shows **lower performance** (0.63, 0.86)
- Visual prompts critical for downstream task effectiveness

Data	Cam16	Panda
NoPrompt	0.63	0.86
PromptEmbed	0.69	0.88
Prompt	0.71	0.95
Real	0.72	0.96



(a) DeepLabv3+ performance on real/synthetic (Camelyon16)



(b) DeepLabv3+ performance on real/synthetic (Panda)

RESULTS Pathologist Evaluation

A certified pathologist with seven years of clinical experience in surgical pathology

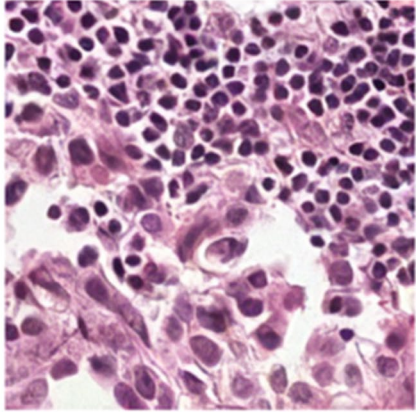
Quantitative metrics

- Overall Quality
- Structural Detail
- Nuclear Detail

Two binary assessments

- Prediction of hallucination
- Image authenticity (real versus synthetic classification)

Synth Image Evaluation by Pathologists



Overall Quality

1

2

3

4

5

Structural Details

1

2

3

4

5

Nuclear Details

1

2

3

4

5

Hallucinations?

Yes

No

Real vs Synthetic?

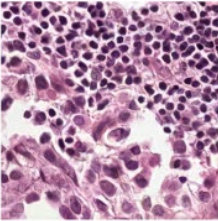
Real

Synth

Submit

RESULTS Pathologist Evaluation

Synth Image Evaluation by Pathologists



Overall Quality ☐ 1 ☐ 2 ☐ 3 ☒ 4 ☐ 5

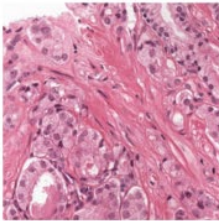
Structural Details ☐ 1 ☐ 2 ☐ 3 ☒ 4 ☐ 5

Nuclear Details ☐ 1 ☐ 2 ☐ 3 ☒ 4 ☐ 5

Hallucinations? ☐ Yes ☒ No

Real vs Synthetic? ☒ Real ☐ Synth

Synth Image Evaluation by Pathologists



Overall Quality ☐ 1 ☐ 2 ☐ 3 ☒ 4 ☐ 5

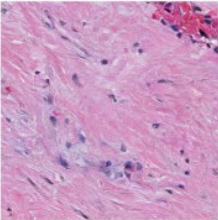
Structural Details ☐ 1 ☐ 2 ☐ 3 ☒ 4 ☐ 5

Nuclear Details ☐ 1 ☐ 2 ☐ 3 ☒ 4 ☐ 5

Hallucinations? ☐ Yes ☒ No

Real vs Synthetic? ☒ Real ☐ Synth

Synth Image Evaluation by Pathologists



Overall Quality ☐ 1 ☐ 2 ☐ 3 ☒ 4 ☐ 5

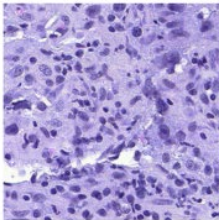
Structural Details ☐ 1 ☐ 2 ☐ 3 ☒ 4 ☐ 5

Nuclear Details ☐ 1 ☐ 2 ☐ 3 ☒ 4 ☐ 5

Hallucinations? ☐ Yes ☒ No

Real vs Synthetic? ☒ Real ☐ Synth

Synth Image Evaluation by Pathologists



Overall Quality ☐ 1 ☐ 2 ☐ 3 ☒ 4 ☐ 5

Structural Details ☐ 1 ☐ 2 ☐ 3 ☒ 4 ☐ 5

Nuclear Details ☐ 1 ☐ 2 ☒ 3 ☐ 4 ☐ 5

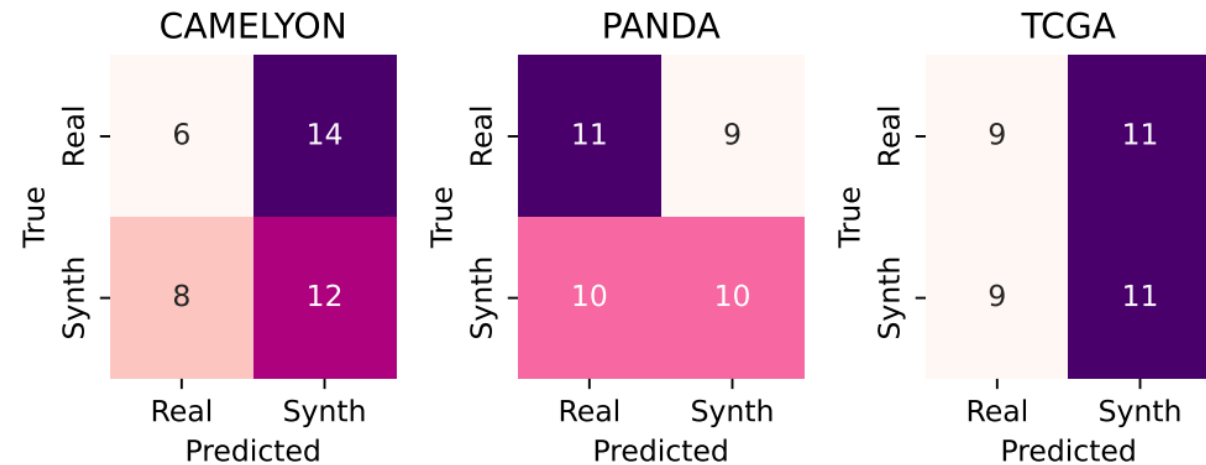
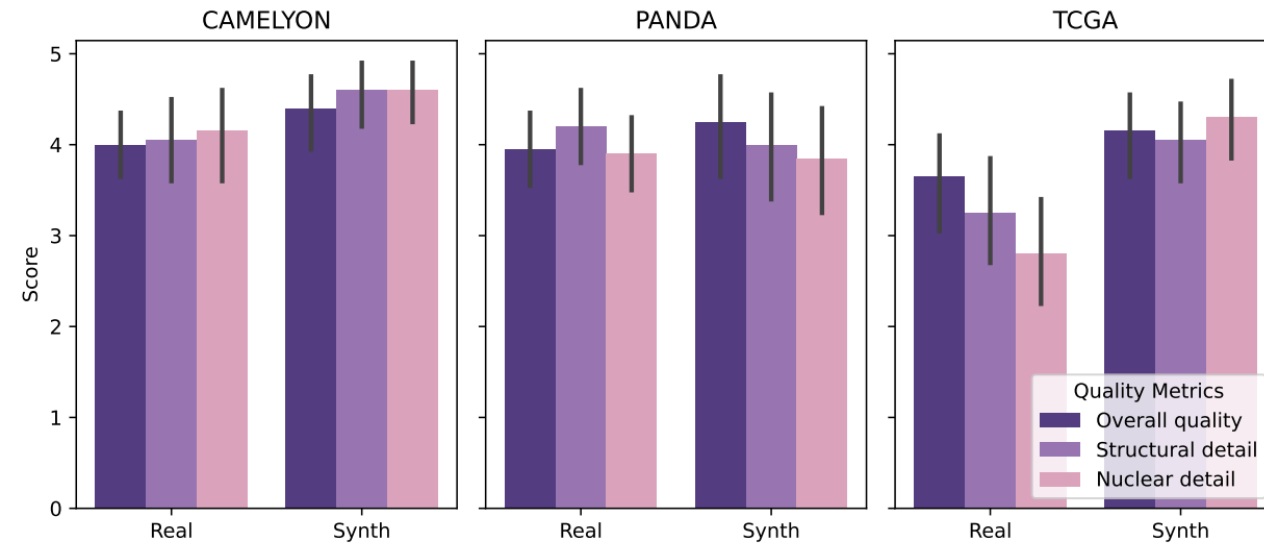
Hallucinations? ☐ Yes ☒ No

Real vs Synthetic? ☒ Real ☐ Synth

RESULTS Pathologist Evaluation

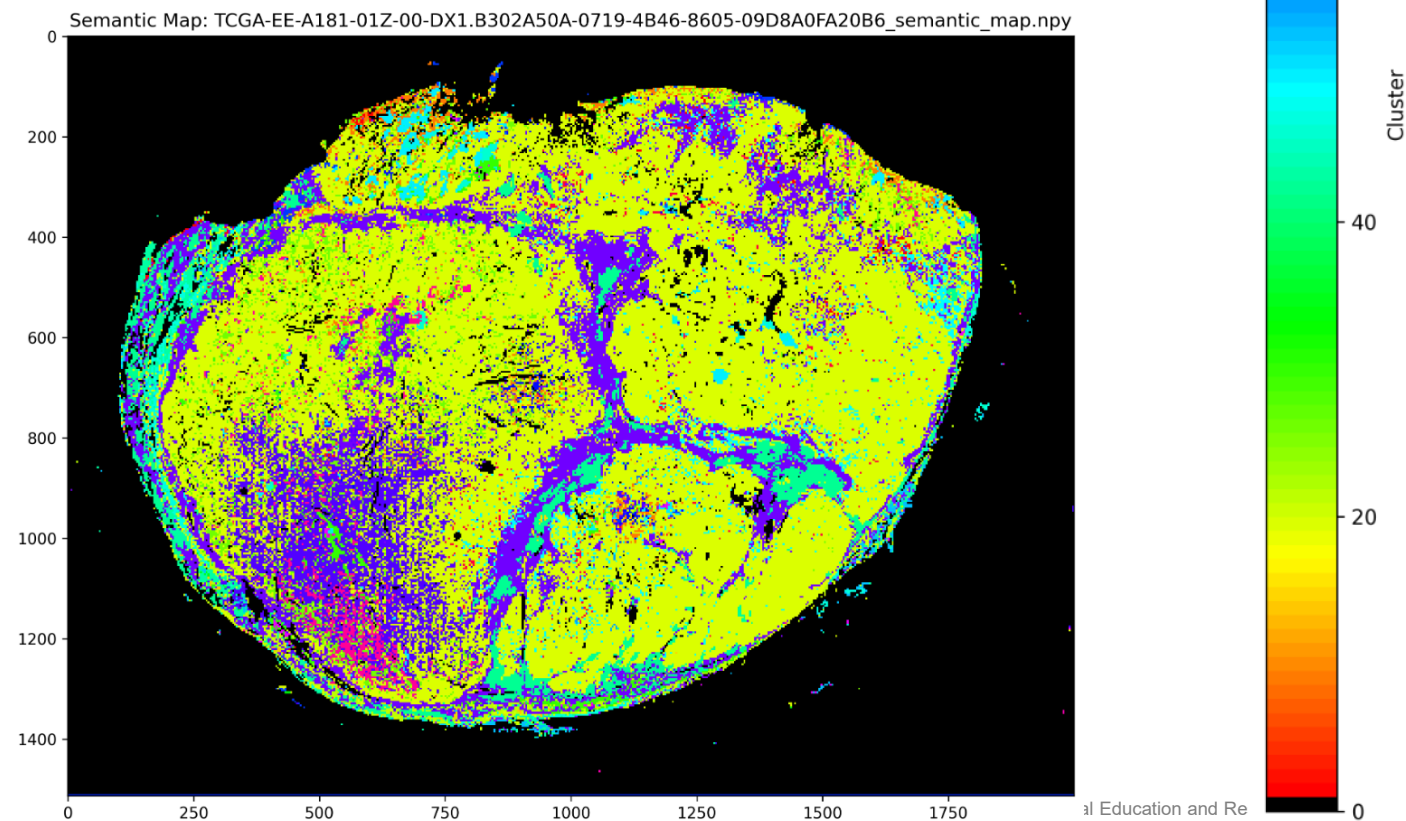
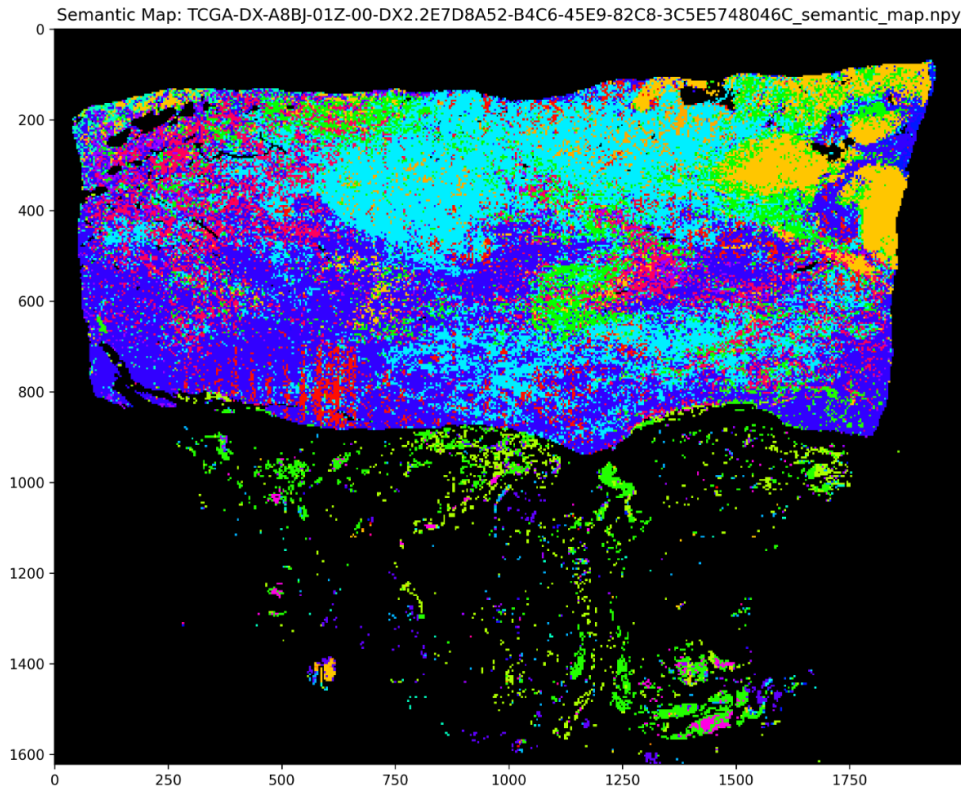
Comparable scores for real and synthetic images across:

- Overall image quality
- Histological structural detail
- Nuclear morphology accuracy



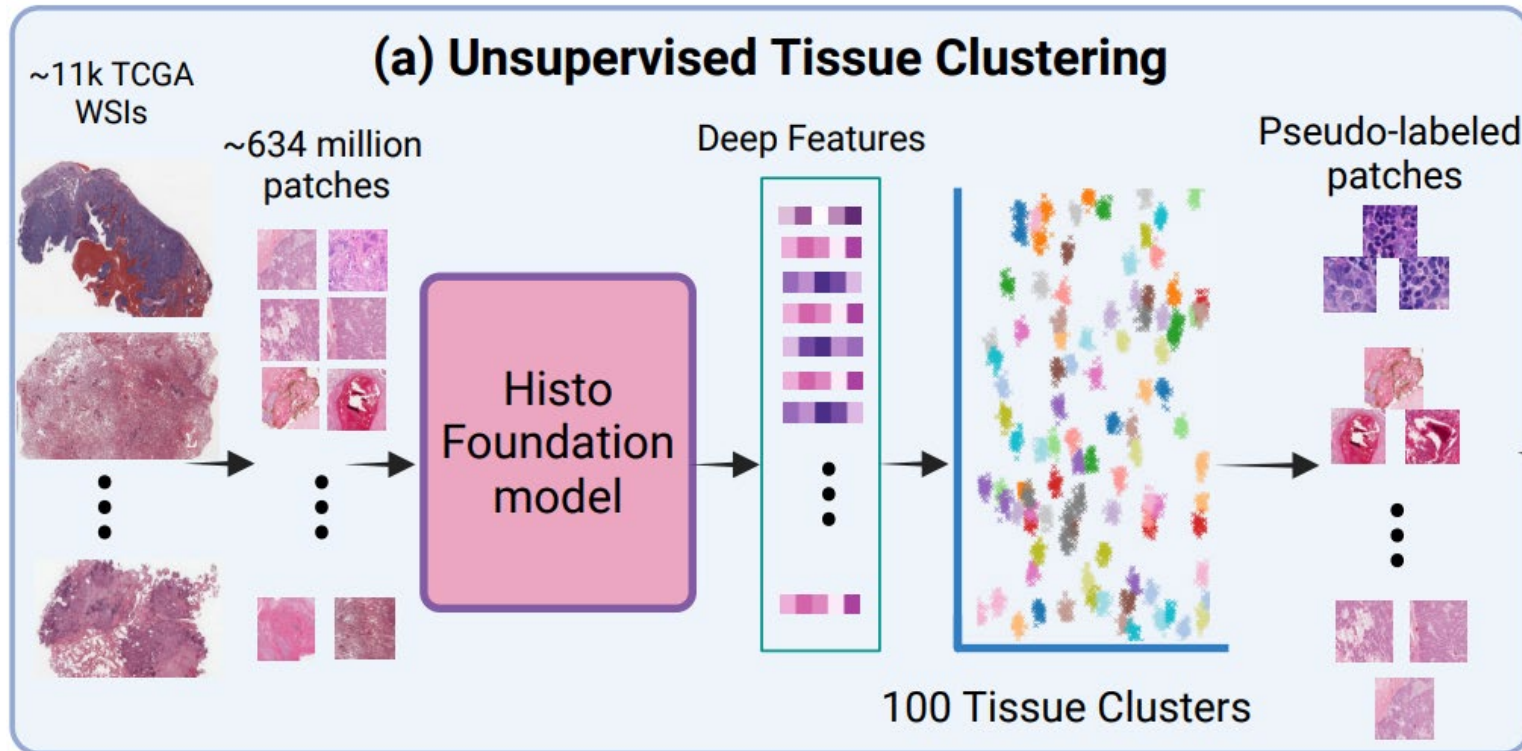
SELF-SUPERVISED EXTENSION (TCGA)

- Use 11,765 whole-slide images without manual annotation
- Three-phase approach:
 - Feature extraction with foundation models
 - Clustering (100 tissue types)
 - Multi-scale crops/semantic map generation



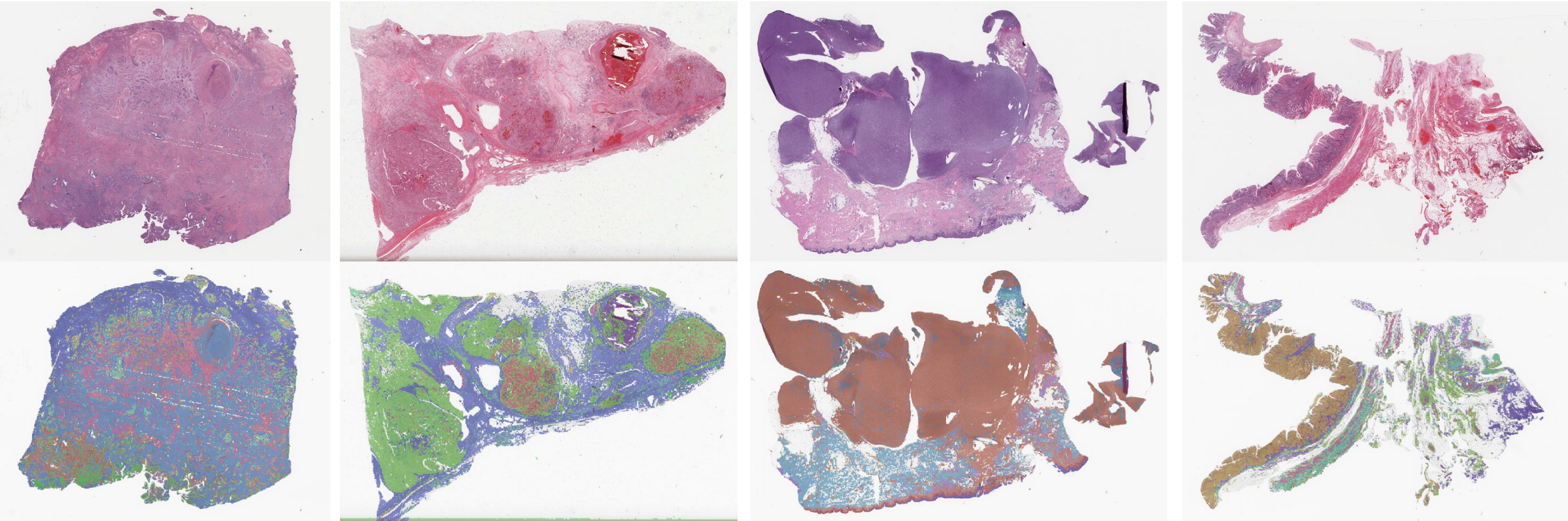
SELF-SUPERVISED EXTENSION (TCGA)

- Processed **634 million patches** at highest magnification
- Foundation model (i.e., **UNI**) embeddings for feature extraction
- Diversity-aware sampling for clustering - 1000 patches per WSI
- Total 1,174,907 patch for clustering
- **k-means clustering (GPU)** to identify 100 tissue types



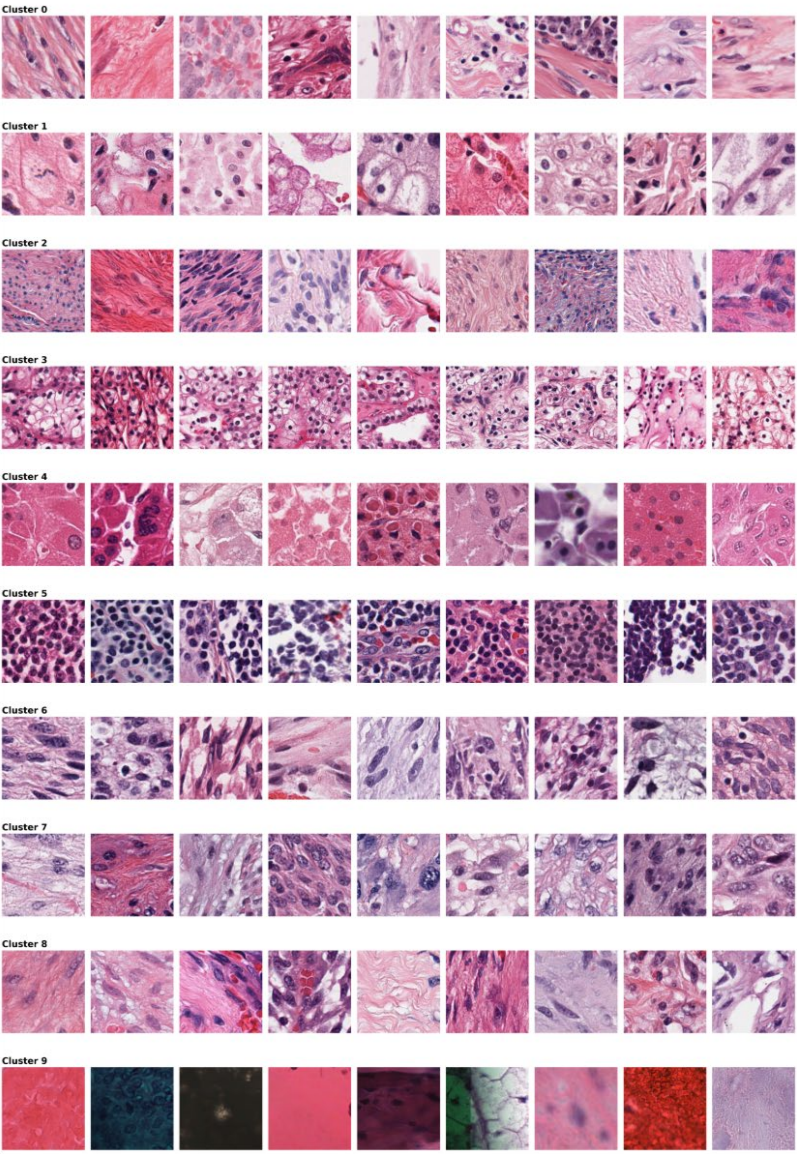
SELF-SUPERVISED EXTENSION (TCGA)

k-means clustering to identify 100 tissue types



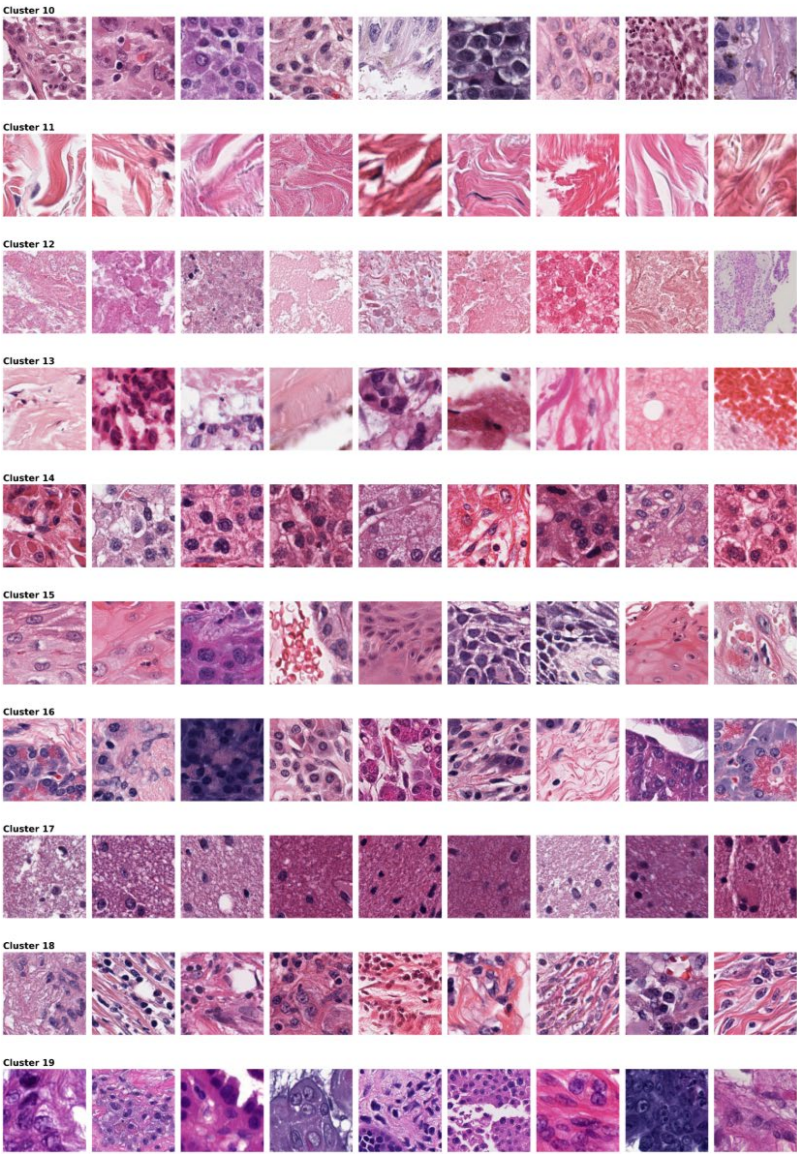
SELF-SUPERVISED EXTENSION (TCGA)

Clusters 0-9



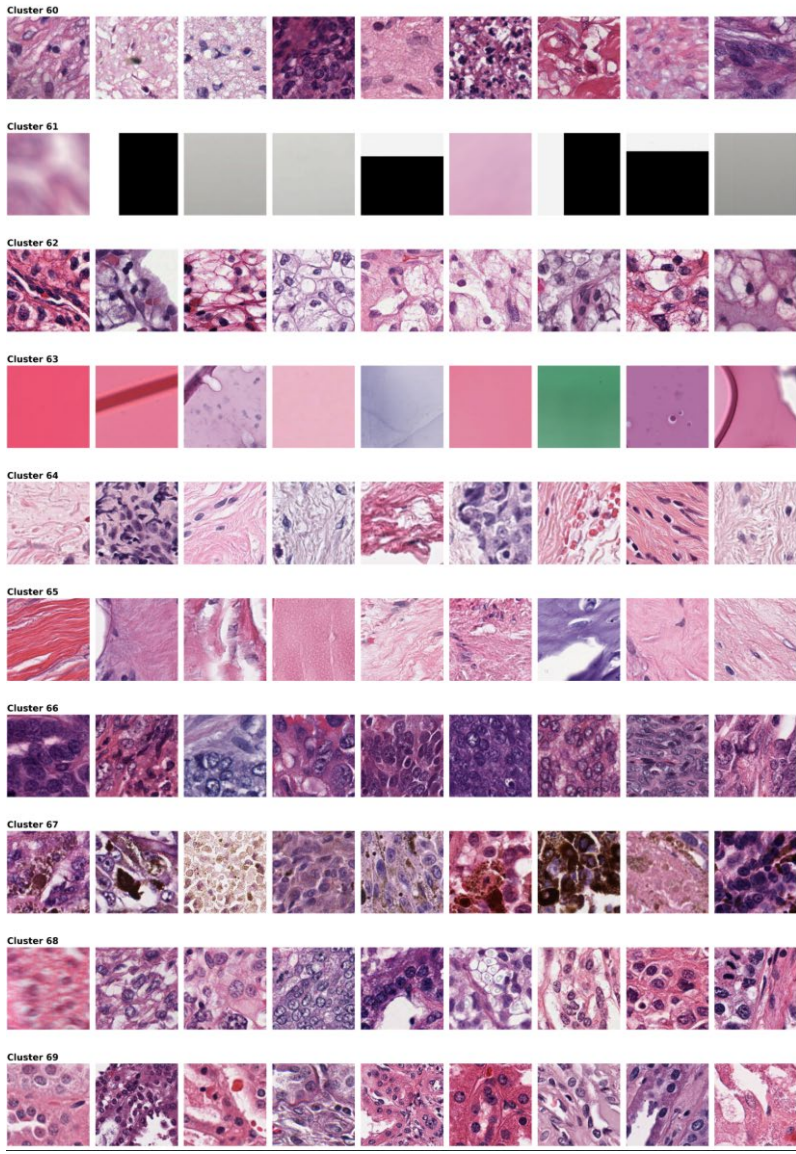
Clusters: 1-10

Clusters 10-19



Clusters: 11-20

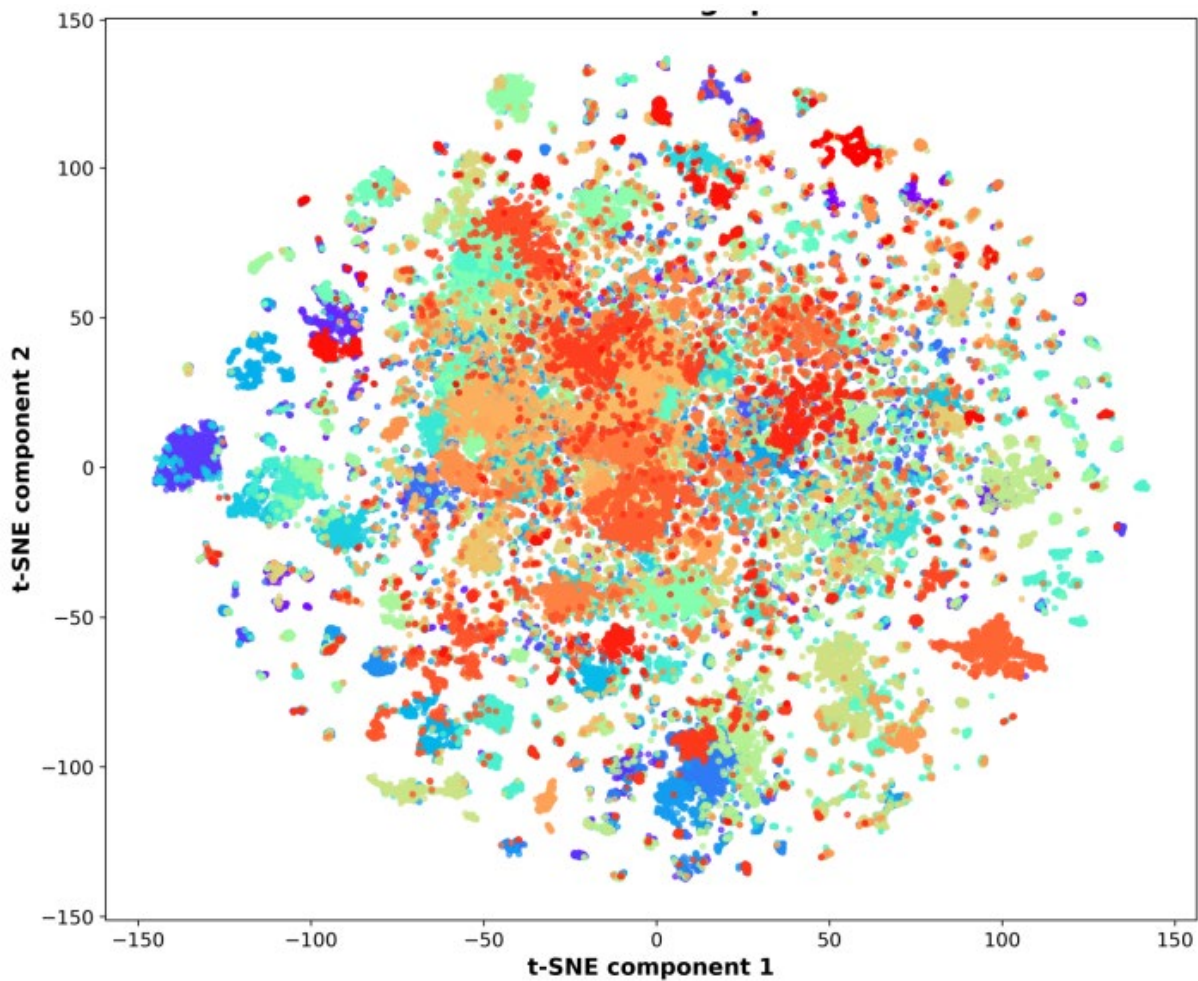
Clusters 60-69



Clusters: 61-70

SELF-SUPERVISED EXTENSION (TCGA)

k-means deep features clustering to identify 100 tissue types

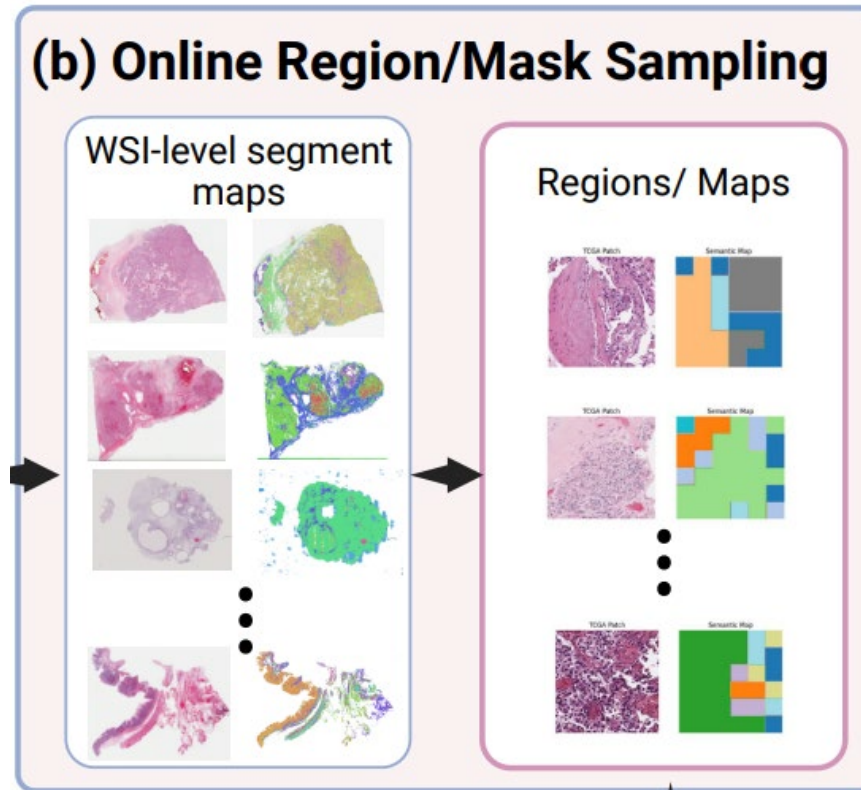


Cluster IDs					
Cluster 0	Cluster 17	Cluster 34	Cluster 51	Cluster 68	Cluster 84
Cluster 1	Cluster 18	Cluster 35	Cluster 52	Cluster 69	Cluster 85
Cluster 2	Cluster 19	Cluster 36	Cluster 53	Cluster 70	Cluster 86
Cluster 3	Cluster 20	Cluster 37	Cluster 54	Cluster 71	Cluster 87
Cluster 4	Cluster 21	Cluster 38	Cluster 55	Cluster 72	Cluster 88
Cluster 5	Cluster 22	Cluster 39	Cluster 56	Cluster 73	Cluster 89
Cluster 6	Cluster 23	Cluster 40	Cluster 57	Cluster 74	Cluster 90
Cluster 7	Cluster 24	Cluster 41	Cluster 58	Cluster 75	Cluster 91
Cluster 8	Cluster 25	Cluster 42	Cluster 59	Cluster 76	Cluster 92
Cluster 9	Cluster 26	Cluster 43	Cluster 60	Cluster 77	Cluster 93
Cluster 10	Cluster 27	Cluster 44	Cluster 61	Cluster 78	Cluster 94
Cluster 11	Cluster 28	Cluster 45	Cluster 62	Cluster 79	Cluster 95
Cluster 12	Cluster 29	Cluster 46	Cluster 63	Cluster 80	Cluster 96
Cluster 13	Cluster 30	Cluster 47	Cluster 64	Cluster 81	Cluster 97
Cluster 14	Cluster 31	Cluster 48	Cluster 65	Cluster 82	Cluster 98
Cluster 15	Cluster 32	Cluster 49	Cluster 66	Cluster 83	Cluster 99
Cluster 16	Cluster 33	Cluster 50	Cluster 67		

Figure 4: t-SNE visualization of 99,792 randomly sampled TCGA patches colored by cluster assignment using UNI foundation model features [2]. The well-defined separation validates this approach of 100 morphologically coherent tissue phenotypes.

ADAPTIVE HETEROGENEOUS REGION SAMPLING

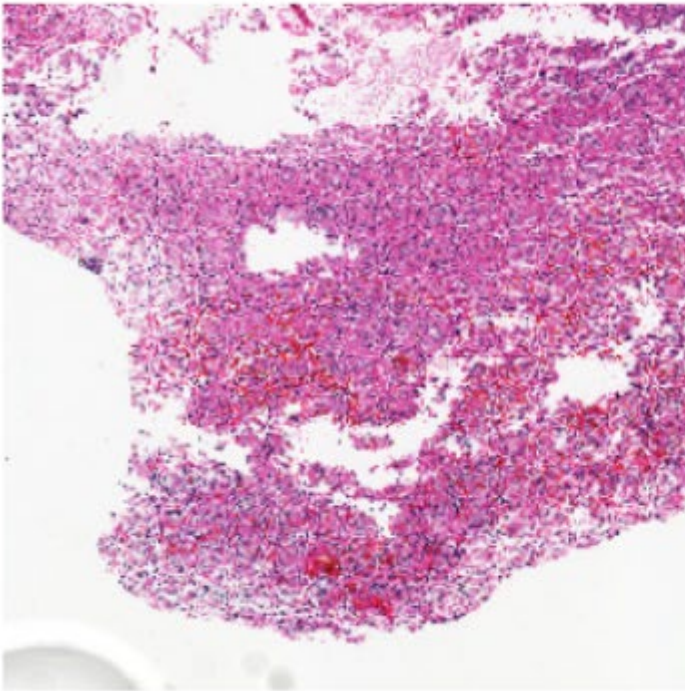
- Compute heterogeneity scores using entropy
- Identify regions with rich tissue interfaces
- Multi-scale visual crops adapted to tissue complexity
- Tissue-aware augmentations: stain variations, controlled rotations



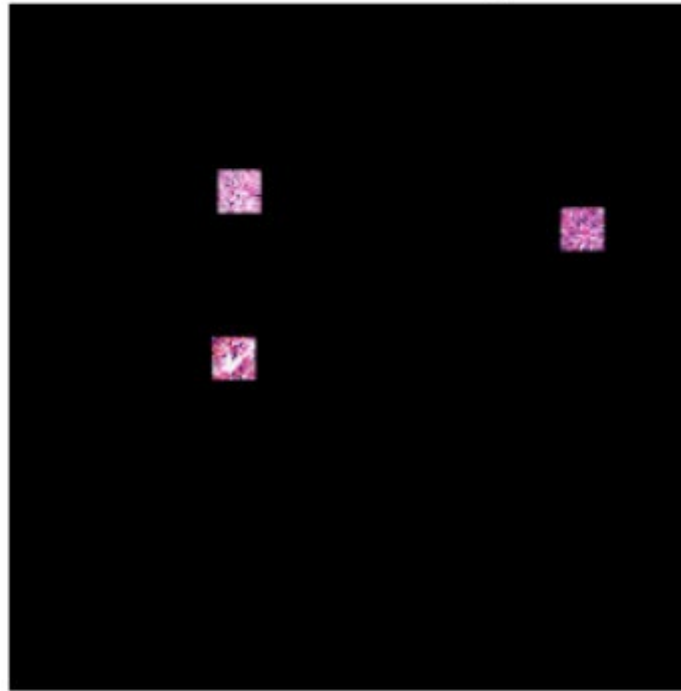
ADAPTIVE HETEROGENEOUS REGION SAMPLING

TCGA

original Image



Generated Image



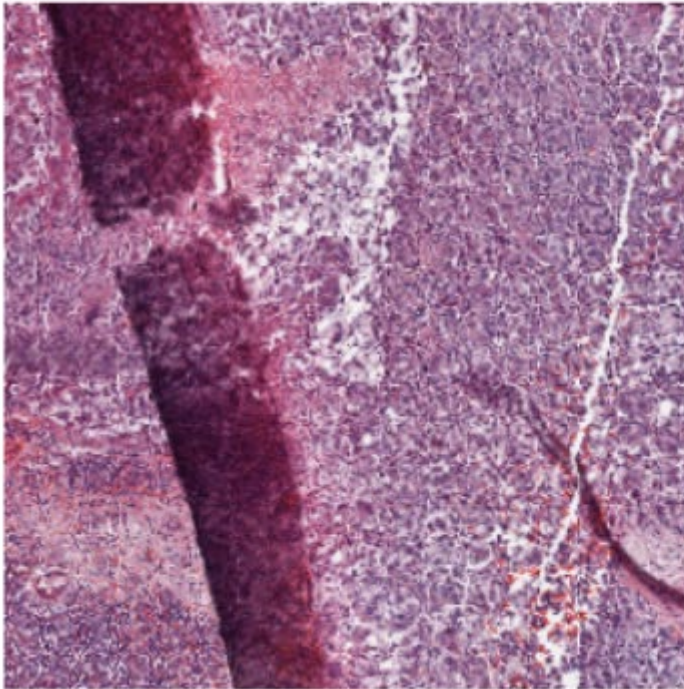
Semantic Map



ADAPTIVE HETEROGENEOUS REGION SAMPLING

TCGA

original Image



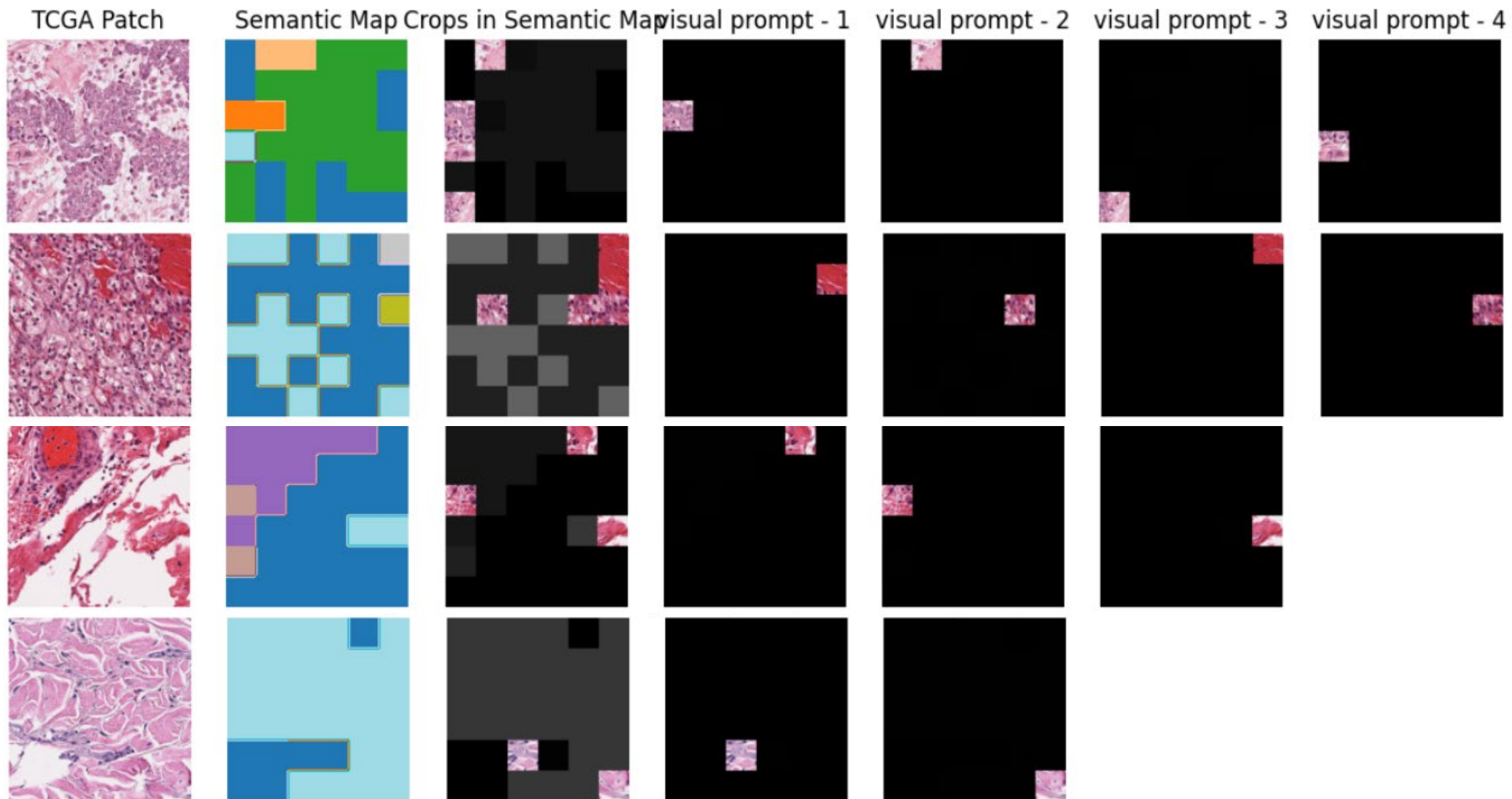
Generated Image



Semantic Map

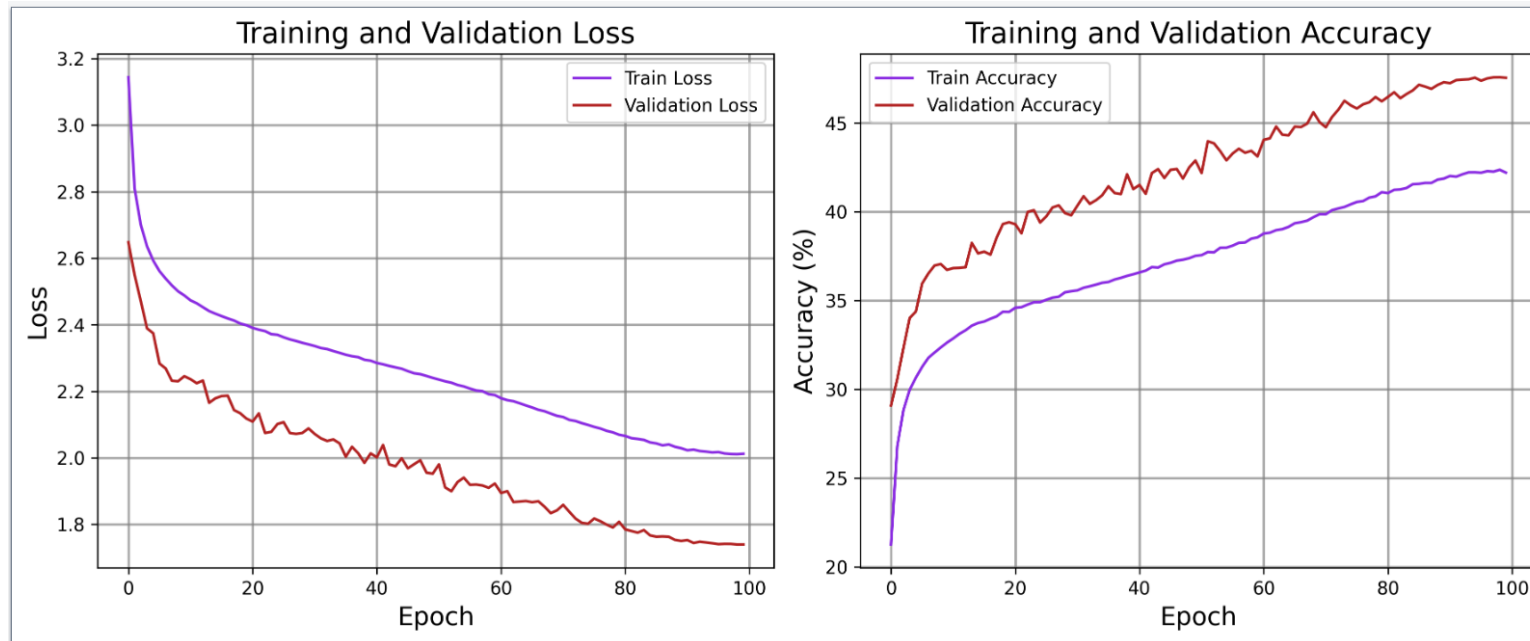


ADAPTIVE HETEROGENEOUS REGION SAMPLING



TISSUE CLASSIFIER FOR INFERENCE

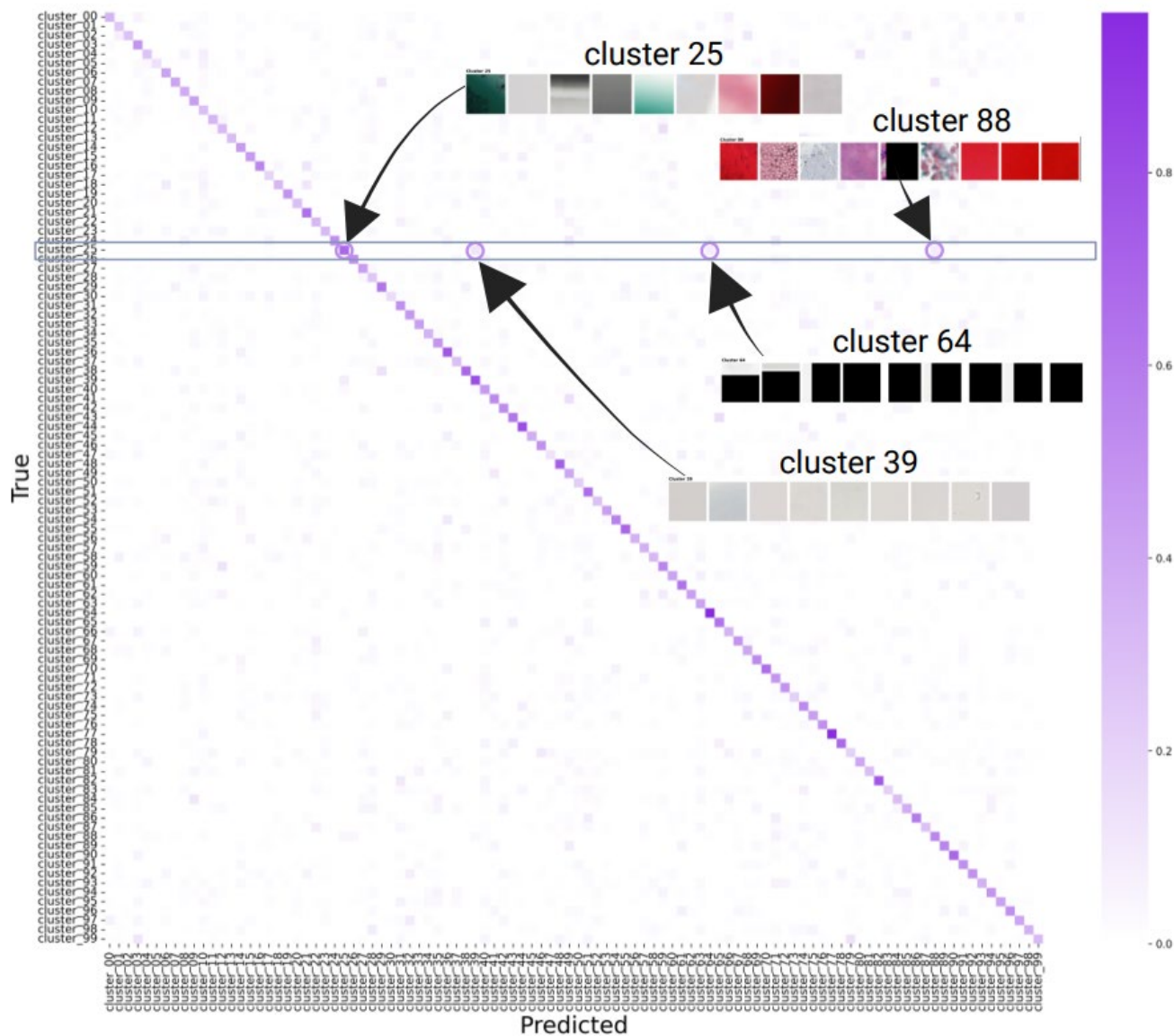
- Lightweight ViT-small architecture for efficient deployment
- Trained on ~500k patches from 11,763 TCGA WSIs
- 85% reduction in computational requirements



TISSUE CLASSIFIER

Cluster classifier performance

100 cluster



ADVANTAGES OVER EXISTING METHODS

- **Compared to text-guided approaches:** Avoids interobserver variability
- **Compared to embedding-based methods:** Preserves critical diagnostic features
- **Compared to semantic-only methods:** Maintains tissue-specific attributes
- **Compared to unconditioned models:** Precise control over tissue composition

LIMITATIONS

- **Computational requirements** for processing gigapixel WSIs
- Current focus on **H&E** (hematoxylin and eosin) stained images only
- Predefined clustering **may miss extremely rare** pathologies



Questions?