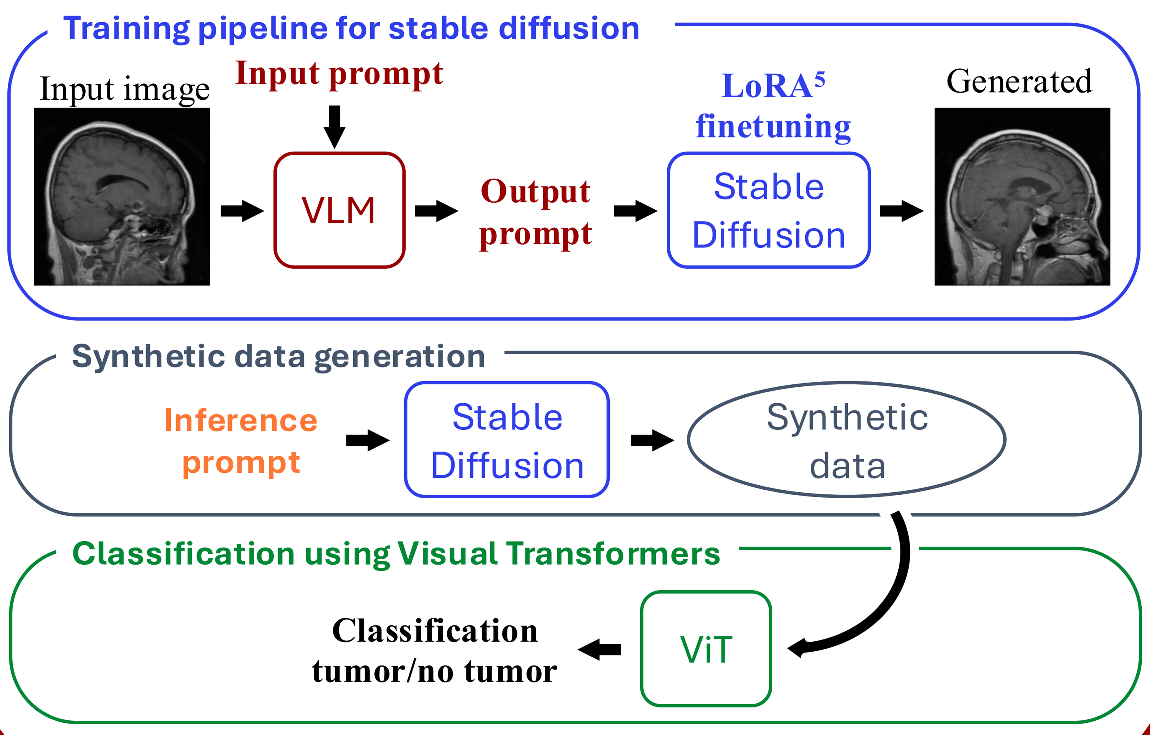
DTU Compute
Department of Applied Mathematics and Computer Science

Generative Modeling of High Fidelity Brain Tumor MRI Images Using Vision Language & Stable Diffusion Models

Jone Steinhoff (s243867), Lukas Rasocha (s233498), Mads Prip (s240577) & Petr Boska Nylander (s240466)

Introduction

The project aims to generate synthetic MRI tumor images with controllable tumor locations via Stable Diffusion prompts, and evaluate quality using FID, KID, and a classification task.



Data Preprocessing

Original training data[6]

Tumor types

pituitary	meningioma	glioma	No tumor
1457	1339	1321	1595

→ **Preprocessed training data**

Tumor	No tumor
4417	1595

Original train/test split

Training	Test
6012	1311

→ **Preprocessed train/val/test split**

Training	Val	Test
6012	262	1049

Vision Language Models

Llava[1] 8B (medical)

Tumor and location / Tumor size and location

The brain MRI shows a tumor in the pituitary region. The pituitary gland is a small, pea-sized gland located at the base of the brain, just below the hypothalamus.

Gemma[2] 12B

Tumor	Tumor location	Tumor size	Tumor shape	Tumor intensity	Brain orientation	Free parameter
tumor: yes	location: pituitary	size: medium	shape: round	intensity: bright	orientation: coronal	Brain MRI shows normal structures, with a medium-sized …

Ovis[3] 34B

Tumor	Tumor location	Tumor size	Tumor shape	Tumor intensity	Brain orientation	Free parameter
tumor: yes	location: sella turcica	size: small	shape: irregular	intensity: hyperintense	orientation: axial	general description: brain MRI shows a small irreg…

Gemini[4] Flash 2.0

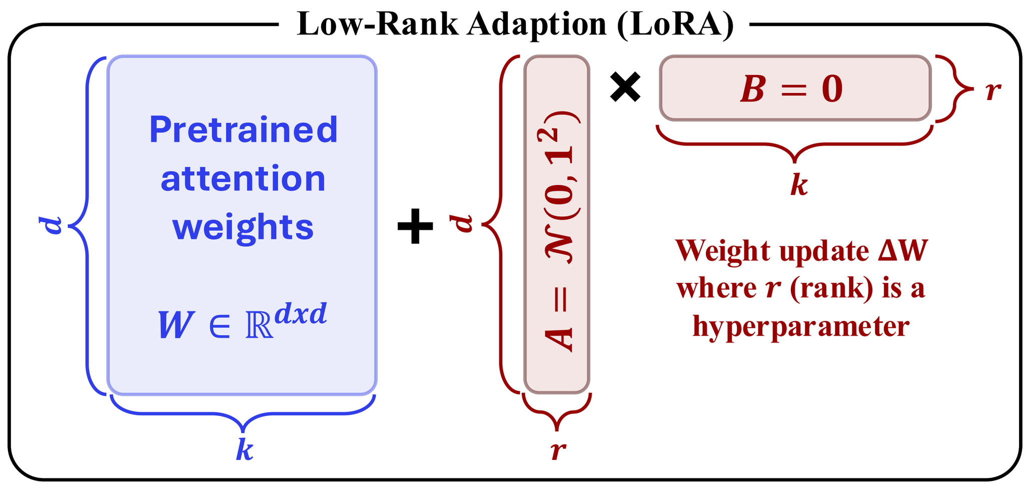
Tumor	Tumor location	Tumor size	Tumor shape	Tumor intensity	Brain orientation	Free parameter
tumor: yes	location: brain parenchyma	size: medium	shape: irregular	intensity: heterogeneous	orientation: coronal	general description: Coronal brain…

Models

Stable diffusion XL hyperparameters

- Rank **128/248**
- Resolution **512**
- Batch Size **2**
- Accumulation Steps **8**
- Learning Rate **1e-4**
- Gradient Checkpointing **True**
- Learning rate Scheduler **Cosine**
- SNR Gamma **5**
- Adam Weight Decay **0.01**
- Learning Rate Warmup Steps **1000**

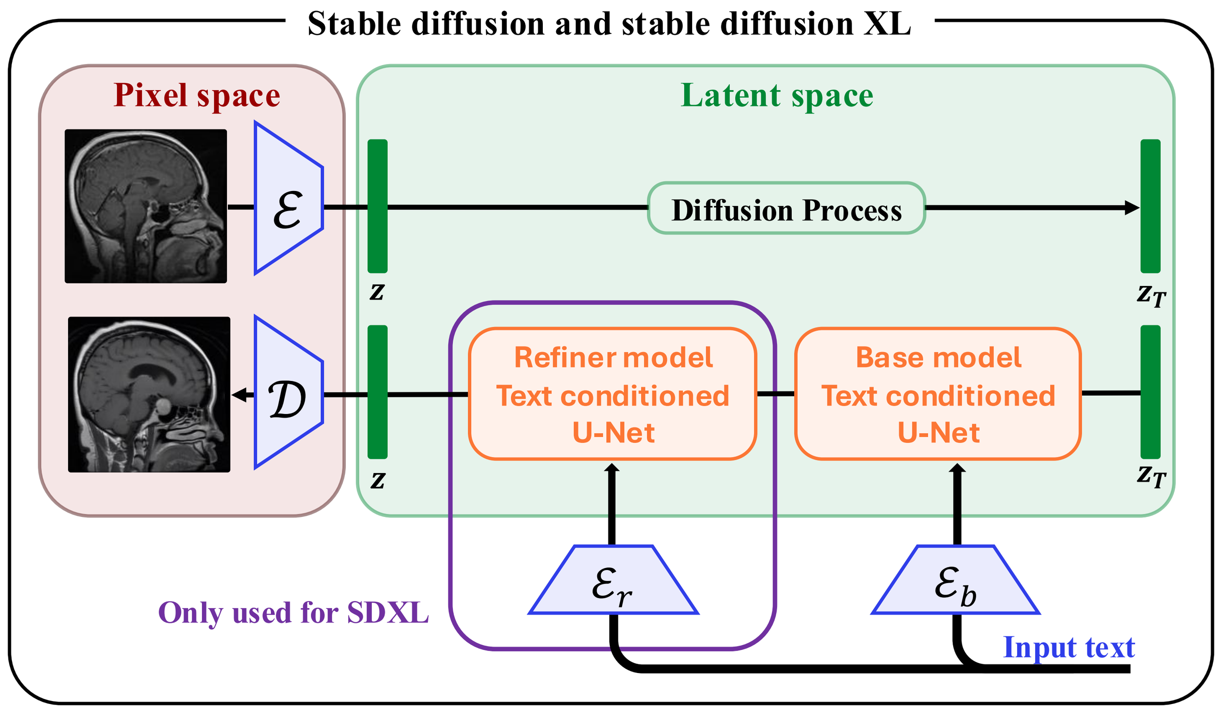
Low-Rank Adaption (LoRA)



Vision Transformers hyperparameters

- Batch Size **16**
- Image Size **512**
- Patch Size **16**
- Channels **3**
- Embedding Dimension **128**
- Number of Heads **4**
- Number of Layers **1**
- Number of Classes **2**
- Positional Encoding **Learnable**
- Pooling Method **CLS**
- Dropout Rate **0.3**
- Fully Connected Dimension **512**
- Number of Epochs **40**
- Learning Rate **1e-4**
- Warmup Steps **625**
- Weight Decay **1e-3**
- Gradient Clipping **1.0**

Stable diffusion and stable diffusion XL



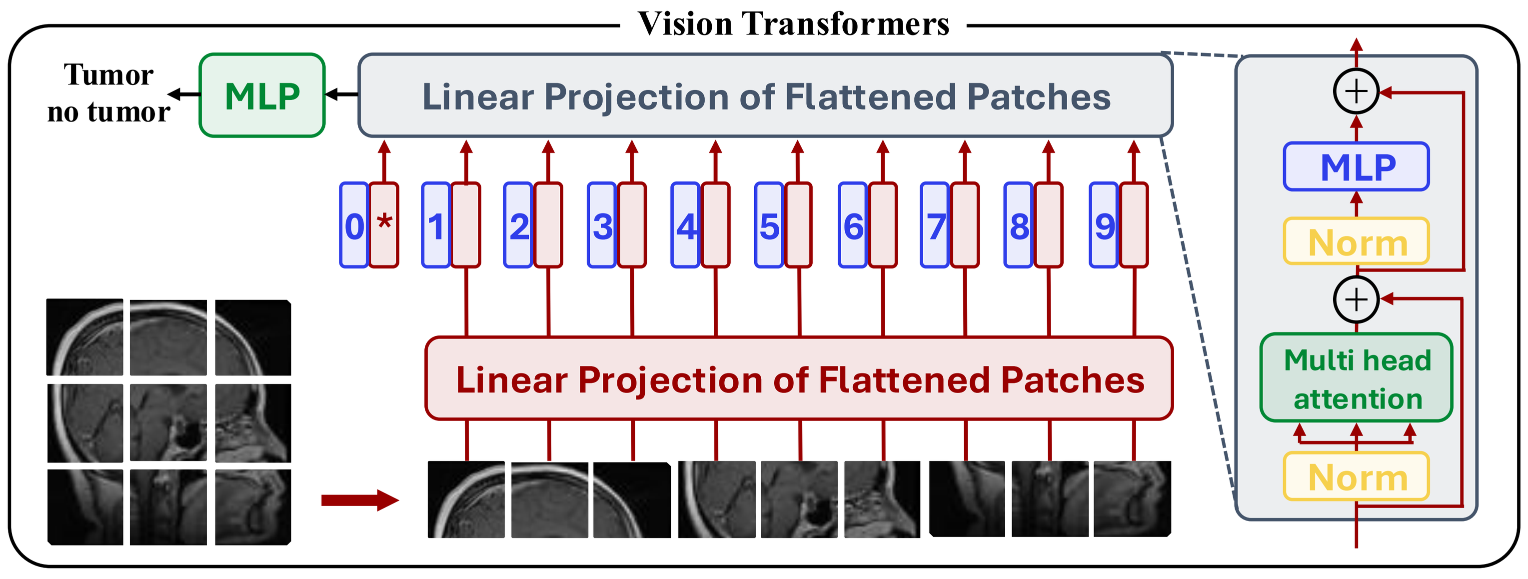
Stable diffusion hyperparameters

- Rank **128/248**
- Resolution **512**
- Batch Size **2**
- Accumulation Steps **8**
- Learning Rate **1e-4**
- Gradient Checkpointing **True**
- Learning Rate Scheduler **Cosine**
- SNR Gamma **5**
- Adam Weight Decay **0.01**
- Learning Rate Warmup Steps **500**

Stable diffusion inference

- Number of Denoising Steps **50**

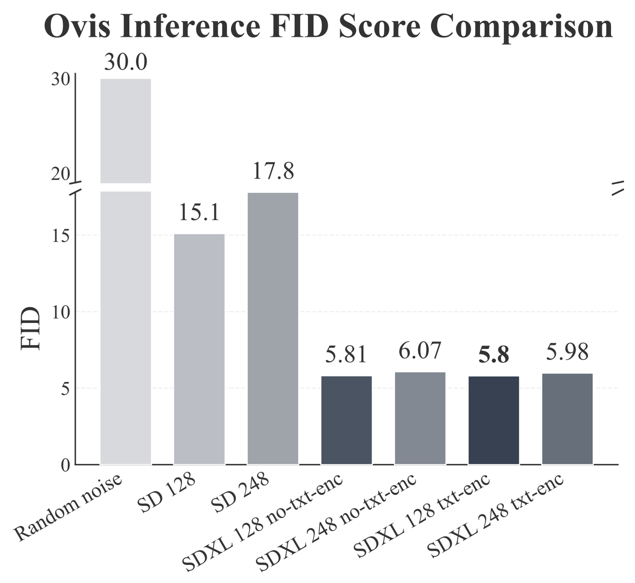
Vision Transformers

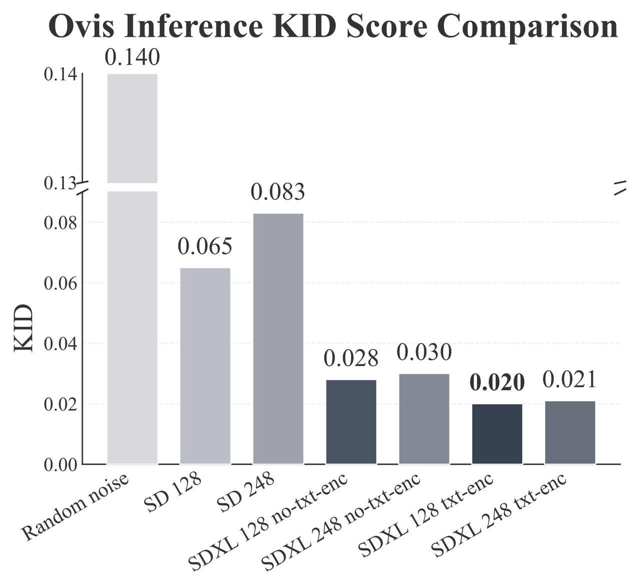


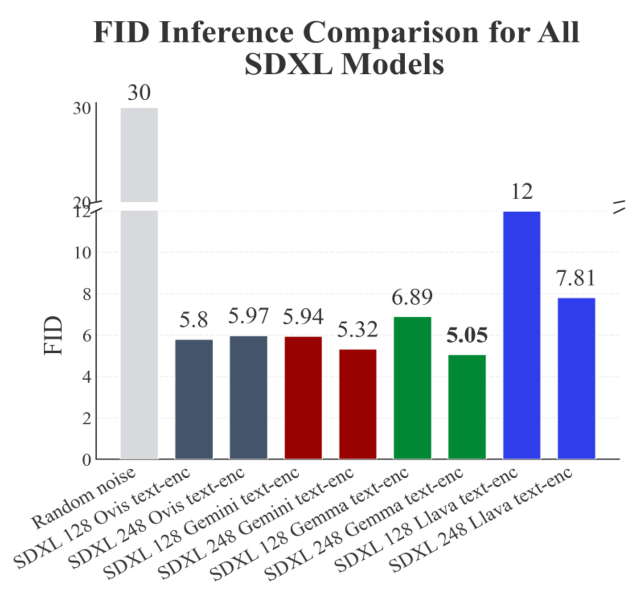
Results

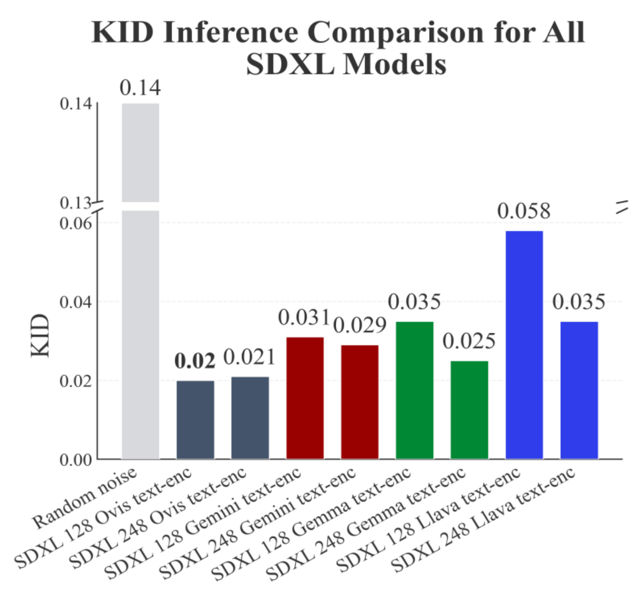
KID and FID scores show, that SDXL fine-tuned on Ovis with a trainable text encoder performs the best. Llava, Gemma and Gemini are used in the same setup to enable comparison.

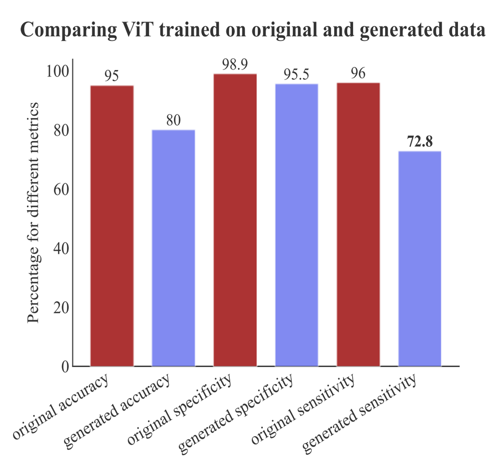
The best-performing configurations are based on Ovis (for KID) and Gemma (for FID), both incorporating a trainable text encoder and LoRA ranks of 128 and 248, respectively. Synthetic data generated with Ovis is used to classify tumor vs. non-tumor cases, in comparison to the original dataset. Accuracy, specificity and sensitivity are being compared.



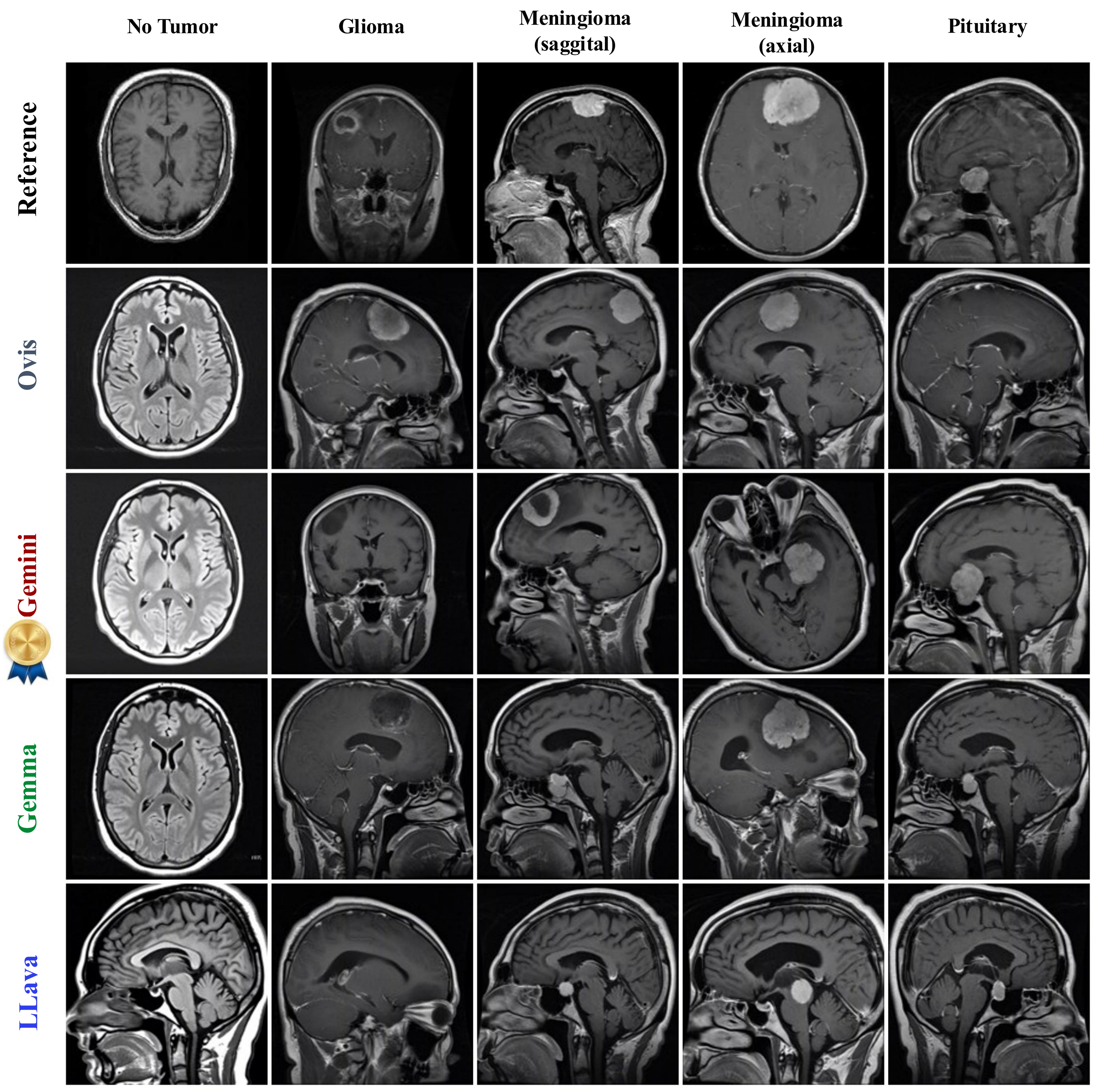








Model Acc.	Tumor	Loc.	Size	Shape	Intensity	Orient.
Ovis	1	0.66	0.65	0.5	**0.8**	0.35
Llava	1	0.45	0.05	0	0.05	0
Gemini	1	**0.82**	0.65	**0.7**	**0.8**	1
Gemma	1	0.47	**0.66**	0.5	0.64	0.8



Conclusion

- Combining automatic data labeling with VLMs with fine-tuning large diffusion models shows potential in domains with limited data availability.
- The choice of VLM impacts the quality of generated images. The best VLMs in our set-up were **Ovis 34B (KID)** and **Gemma 13B (FID).**
- SDXL consistently showed higher generative quality compared to SD-v1-5.
- Using detailed and structured medical prompts to control the generation of MRI scans shows potential, where **Gemini** prompts were followed the best.

References

1. https://github.com/microsoft/LLaVA-Med
2. https://huggingface.co/google/gemma-3-12b-it
3. https://huggingface.co/AIDC-AI/Ovis2-34B
4. https://ai.google.dev/gemini-api/docs/models
5. https://doi.org/10.48550/arXiv.2106.09685
6. https://doi.org/10.34740/kaggle/dsv/2645886